ABSTRACT

Title of dissertation: CONSISTENCY OF SPECTRAL CLUSTERING WITH FUNCTIONAL MAGNETIC RESONANCE IMAGING DATA

Jessie E. Moon, Doctor of Philosophy, 2018

Dissertation directed by: Professor Eric Slud
Department of Mathematics

Functional magnetic resonance imaging (fMRI) is a non-invasive technique for studying brain activity. It uses the amount of blood flowing through a brain, referred to as the blood oxygenation level dependent (BOLD) signal. However analyzing the fMRI signals is challenging because of its complicated spatio-temporal correlation structure and its massive amount of data.

There are several brain atlases available but researchers observe that fMRI signals are not coherent even within the same area in a brain atlas. Therefore providing parcellation of a brain, especially based on its functional connectivity, is necessary to understand brain activities.

One of the techniques that are used for a brain parcellation is spectral clustering. It is a well-used technique in many areas of studies, such as physics and engineering. However, its asymptotic behavior, whether spectral clustering will produce consistent clustering as samples grow large, is not fully

clarified. In addition, there has previously been no available mathematical justification of the large-sample properties of spectral clustering when the data are dependent.

Von Luxburg et al. (2008) showed the consistency of eigenfunctions of spectral clustering under the assumption that data are independent and identically distributed. Because fMRI signals are spatially dependent, applying her results to fMRI data analysis is not appropriate. In this thesis, we extend von Luxburg's work to 3-dimensional spatially dependent data satisfying strong mixing conditions, which will be the case for fMRI data.

We applied the spectral clustering algorithm to simulated data to see how the algorithm can be affected by perturbation in a similarity matrix. There are two simulated data experiments. The first type of simulated data is similar to the stochastic block model, and the second is sampled independently from a Gaussian random field distribution with correlation.

We applied spectral clustering to various regions of interest (ROIs) both for a single subject and for multiple subjects. We also provided methods to analyze data from multiple subjects using spectral clustering and compared these methods using several criteria.

The consistency of spectral clustering with fMRI data

by

Jessie Eunyoung Moon

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Eric Slud, Chair/Advisor
Professor Shuo Chen, Co-Advisor
Professor Paul Smith
Professor Ben Kedem
Professor Xin He

Acknowledgments

First, I would like to give glory to God almighty for providing me enough patience, zeal, and strength that enabled me to complete this work successfully. I would like to thank almighty for his grace and help in all my endeavors and for bringing me this far in my academic career.

Many people have contributed in helping me complete this dissertation. First and foremost I'd like to thank my advisor, Professor Eric Slud for giving me invaluable advice and guidance. He has always made himself available for help and great advice. It has been a great honor to work with and learn from such an extraordinary individual.

I would also like to thank my co-advisor, Professor Shuo Chen. He was always patient with my progress when I struggled with work, study, and life balance. He gave great creative advice whenever I asked him for help.

I deeply appreciate my family who been always great supports. My four children, Katelyn, Jennifer, Amy, and Samuel, and my parents-in-law. They were amazingly patient and gave emotional support when I couldn't join family events as a mother due to study. I owe my deepest thanks to my children who suffered through 2017 summer with no vacation.

I would also like to acknowledge help and support from the staff members. Haydee was always encouraging with her warm heart and big hugs.

Table of Contents

List of Tables

List of Figures

Chapter 1: fMRI Data and Preprocessing

1.1 Background

Functional magnetic resonance imaging (fMRI) is a well-used technique in biomedical research since its discovery in 1991. One of its uses was for studying brain activity. When neurons become active in some area of the brain, the amount of blood flowing through that area is increased, yielding a relative surplus in local blood oxygen. Thus, fMRI scanner uses this fact to produce a fMRI image. The signal measured in fMRI depends on this change in oxygenation, and it is called the blood oxygenation level dependent (BOLD) signal.

Data from fMRI is huge because there are millions of observations for each time of the scan. Usually, the scan is repeatedly taken between 100 and 200 time-intervals, therefore the size of data is over 100 million. Not only the amount of data is large, but also data has dependency structures in time and space. Therefore the analysis of fMRI data is exceedingly complex, requiring sophisticated techniques from signal and image processing and statistics.

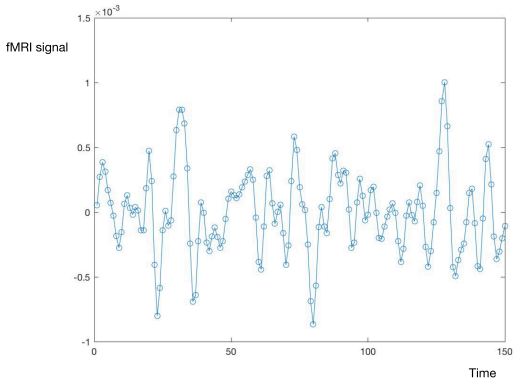


Figure 1.1: This is an example of an fMRI signal from one location in the brain over the time.

Research using fMRI has been continuously growing, as can be seen by plotting the number of papers that mention the fMRI technique from the PubMed database of biomedical literature. The following figure (Figure 1.2) is quoted from an internet published article from https://www.frontiersin.org/articles/10.3389/fnhum.2014.00462/full#h12. As shown in the graph, the number of publications increased rapidly after 2002.

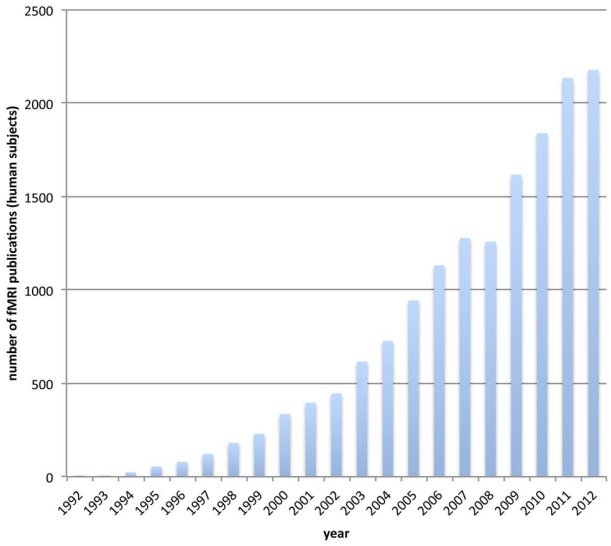


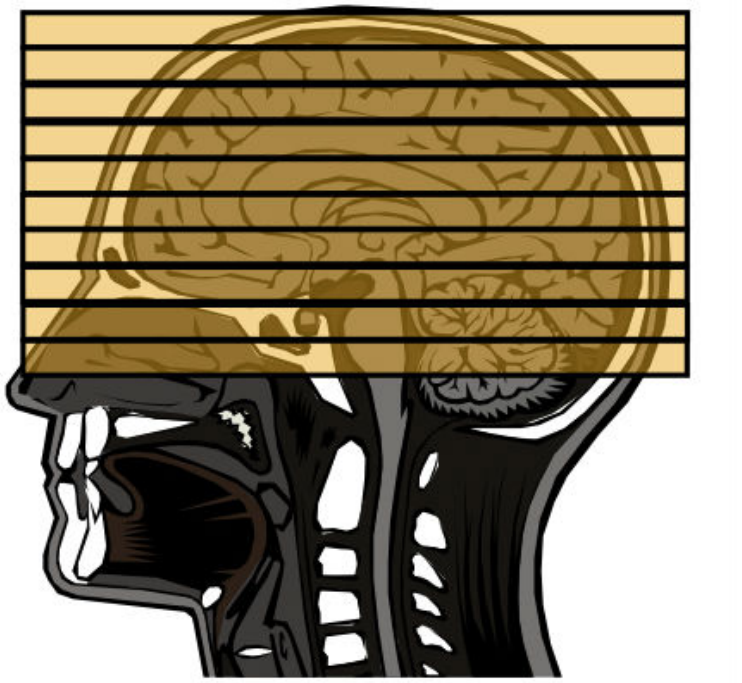
Figure 1.2: "The usage of fMRI gets increasingly popular. We depict the number of publications for each year that incorporate fMRI on human subjects. The data is based on a pubmed.org search string."

Source: https://www.frontiersin.org/articles/10.3389/fnhum.2014.00462/full#h12

1.2 Data Acquisition

When a brain is scanned by a scanner, it can be done in one long scan which might take up to 30 minutes to finish or it can be split into two or three shorter scans. In the case of several scans, brief rests might be provided. Depending on the strength of the magnetic field of a scanner, we can have a different resolution of an image. Tesla (T) is a unit measuring the strength of magnetic field of a scanner. A higher number of Tesla means that the scanner can produce a higher resolution of an image. Most standard MRI scanners produce 1.5 or 3T.

Most standard fMRI scanners produce an image with voxels of about 1 cubic millimeter in size which summarizes the activity of around 100,000 neurons. This is called the spatial resolution and one cube of the grids assigned by the scanner is called a voxel. Since a human brain is around $1200cm^3$, the total number of voxels in a high-resolution brain scan would be around 1 million. There is no standard voxel for a brain, which means that there is no set size. Voxel dimensions can range from 0.1mm x 0.1mm x 0.1mm to 2mm x 2mm x 2mm or larger, depending on the scanner's resolution and the parameters of the scan. However, the term "voxel" can be understood as a 3-dimensional extension of a pixel of MRI scanning. Therefore, each pixel in an MRI image actually corresponds to a three-dimensional (3D) voxel within the brain. The scanner typically produces these images around every 1-3 seconds, and this can be controlled by a researcher.



Among many parameters that a researcher can set with a scanner, there are two most important ones. One is to decide how often the image of the whole brain will be scanned, and the other is the spatial resolution. The time between successive whole brain scan is called Repetition time (TR). Most standard scanners have TR ranging from 2 to 3 seconds, and higher resolution scanners have TR of 1 second. The lower resolution of scanner produces larger size voxels, and a typical voxel size is about $2mm^3$. Once the scan is completed, a fMRI signal will go through several processes before it is ready for analysis. This is called preprocessing, and it will be described in the next section.

1.3 Preprocessing

Preprocessing is the necessary sequence of steps, in order to control other variations in brain scans that are not needed for our research goal, or to correct for incomplete measurements due to features of the hardware. Each researcher can decide what is preprocessing steps to include and what the order of these steps. There are many available programs to do preprocessing and most of them are free. Here I want to provide brief descriptions of the most commonly used steps of preprocessing. We are following the presentations of Poldrack (2012) and Ashby (2011) in describing each preprocessing step.

1.3.1 Slice Timing Correction

Suppose TR is T seconds, then for any slice, the time between successive complete acquisitions is T seconds. Therefore, slices consisting a whole brain image are not taken at the same. Hence if the slice-timing differences are not corrected during preprocessing, then they should be accounted for when the statistical analysis is performed. The most common preprocessing approach for correcting differences in the timing of slice acquisition is to use *interpolation* with respect to time. The idea is to take the values of observations we have, and to make a guess about how they might change over short time intervals, and then use this guess to estimate what the BOLD response was at the beginning of the TR. The most popular forms of interpolation are *linear, spline*, and

sinc.

$$sinc(t) = \frac{\sin(\pi t)}{\pi t}, \quad \text{for all } t, -\infty < t < \infty$$

1.3.2 Head Motion Correction

During the course of the complete scan, a subject can move his head even though a researcher tried to prevent the movement at the time of the scan. This must be corrected to combine the slices to create a 3-dimensional whole brain image.

We apply mathematical methods to correct head movement under the assumption that the brain does not change its shape or size when a subject moves his or her head. This is actually not true because head movement can change head shaping slightly such as flattening its shape, but since it is a minor change, we can ignore this. If the brain does not change its size or shape, it can be treated as a rigid body. Therefore head movement correction is a problem of *rigid body registration*.

Suppose that a person lies inside a scanner. Then this subject can be considered as a rigid body. We can describe any movement of a rigid body by six parameters at each time. For the center of any voxel in his or her head, we can identify it with a point in 3-dimensional space. We denote its coordinate values as (x, y, z). By convention, the z axis runs from the feet through the top of the head of a subject. The x axis runs through the subject's ears (i.e., from left to right), and the y axis runs through the back of the head and exits

the forehead.

Using this coordinate system (x, y, z), we can translate head movement as rigid body movements by combining following rigid body motions.

- Translation along the x axis,
- Translation along the y axis,
- Translation along the z axis,
- Rotation about the x axis,
- Rotation about the y axis,
- Rotation about z axis.

Each translation is parameterized by the distance moved along that axis, and each rotation is a single rotation and parameterized by the angle of rotation.

One of the standard ways of correcting head motions using rigid body movement is as follows. First, take the data from the first TR as a standard reference image. Then take the other TR data and perform rigid body movements on the data from the other TR until BOLD responses from the other TR's datasets agree as closely as possible with the data from the standard TR at each (x, y, z) coordinate point. This process is called rigid body registration and the most common standard method of head motion correction in preprocessing.

Suppose that a subject moved b_x along the x axis, b_y along the y axis, b_z along the z axis, rotated θ_x about the x axis, θ_y about the y axis, θ_z about z axis in the other TR's data from the standard one. Then the equation for the the rigid body registration is.

$$\begin{pmatrix} U_x \\ U_y \\ U_z \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_x) & \sin(\theta_x) \\ 0 & -\sin(\theta_x) & \cos(\theta_x) \end{pmatrix} \begin{pmatrix} \cos(\theta_y) & 0 & -\sin(\theta_y) \\ 0 & 1 & 0 \\ \sin(\theta_y) & 0 & \cos(\theta_y) \end{pmatrix} \times \begin{pmatrix} \cos(\theta_z) & \sin(\theta_z) & 0 \\ -\sin(\theta_z) & \cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_x \\ w_y \\ w_z \end{pmatrix} + \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix}$$

The way to find the values of the six parameters of rigid body registration equation (1.3.2) is to find the minimizer of a certain loss function. Usually, a loss function is the sum of the squared difference in BOLD responses from two datasets because we want to align the two datasets as closely as possible. Finding this minimization is usually involved with some sort of optimization algorithm about which we do not give further details.

1.3.3 Coregistering the Functional and Structural Data

When an fMRI scanner takes both functional and structural data, we call these resulting datasets the functional and structural run. Since the spatial resolution of the functional data is poor, coregistering of the functional and structural data is necessary to find out where each voxel belongs in the brain

map.

This is a matter of speed-accuracy trade-off. During a functional run, the whole brain usually gets scanned every 2-3 seconds. However, a single structural scan takes 8 or 10 minutes for the whole brain. A voxel size of functional data is usually $(2-3mm)^3$ but for the structural images, the voxel size is much smaller, less than $(1mm)^3$. Therefore the structural one has the much higher resolution.

After coregistration preprocessing, high resolution of the structural data can be used to improve spatial localization of the functional data. Since the resolution of fMRI data is so poor, it might be difficult to tell the location of the activated area from functional data alone. For example, when a certain task-related activation occurs, it might be hard to tell if it is from the supplementary motor area (SMA) or the pre-SMA because they are very small adjacent areas in the brain.

Distinguishing these areas functionally can be an important issue because the pre-SMA projects primarily to the prefrontal cortex, whereas the SMA projects primarily to motor cortex and other premotor areas. However, this issue can be easily resolved by using the structural data since we can map the functional activation onto the structural image.

1.3.4 Normalization

Since there are differences in the size and shape of individual brains, it is difficult for a researcher to assign an activated area of the brain, which is a cluster of some voxels, to a specific neuroanatomical brain structure. To resolve this issue, during the preprocessing, researchers register the structural scan of each subject separately to some standard brain. After this is done we can assume that the coordinates of all major brain structures have already been identified and published in an atlas. This process, registering a structural scan to the structural scan from some standard brain, is called normalization.

There are several brain atlases, but the earliest one is still the most widely used brain atlas, called the Talairach atlas. This atlas is based on the detailed dissection of one hemisphere of the brain of a single subject, a 60-year-old French woman. However, there has been an issue in using this atlas regarding whether this atlas can represent the human brain in general because it is based on a single subject's brain. For this reason, recently, an atlas produced by the Montreal Neurological Institute (MNI) has become popular among researchers.

The MNI atlas was created from 152 different subjects and it used high-resolution structural scans and averaged the results. It used the same axes and origin for its coordinate system to match with the Talairach system. Usually, normalization is much more complicated than rigid body motions because it

involves a linear and a non-linear transformation.

1.3.5 Spatial Smoothing

In this step, we blur the images by removing high-frequency information. It seems illogical to blur the images after trying to acquire the best possible resolution. However, there is a very important benefit of spatial smoothing, which is reducing noise. The amount of noise versus signal is measured by their ratio, called signal-to-noise ratio (SNR). Therefore, by this step, we can increase the signal-to-noise ratio. The reason why spatial smoothing is needed is because of the following. fMRI data is very noisy in general and changes in signals corresponding to a certain task can be very small. Thus, this preprocessing step, increasing the signal-to-noise ratio, can greatly increase the chances of success in an fMRI experiment.

In this preprocessing, each voxel is replaced by a weighted average of the BOLD responses in neighboring voxels. The voxel being smoothed has the greatest weight and the weight decreases with distance. Thus voxels far from the smoothed voxel contribute very little, and nearby voxels contribute the most. A researcher can set the parameter for the rate of decrease of the weight as a function of distance, and this determines the amount of smoothing.

1.3.6 Temporal filtering

Spatial filtering smoothes the signals from each voxel by averaging neighborhood voxels at each TR. However, temporal filtering smoothes the signals at each voxel by neighboring TRs. Therefore, in spatial filtering, the data lies on three-dimensional spatial maps but in temporal filtering, the data are one-dimensional time series.

The purpose of both types of filtering is to reduce noise so that a researcher can easily identify the signal. However, the type of noise that the two types of preprocessing are focusing on is different. The most common temporal filtering is called high-pass filtering. Spatial smoothing mainly reduces high-frequency noise but temporal filtering cuts off frequencies below a certain threshold. This removes signal drift that is caused by the scanner and increases SNR.

1.4 Software

A wide variety of software packages are available for fMRI data analysis and many of them are free. They are frequently updated. The most widely used package is Statistical Parametric Mapping (SPM), which is a collection of MATLAB functions. SPM is written and maintained by the Wellcome Trust Center for Neuroimaging at University College London. Another widely used software is FSL (FMRIB Software Library). FSL is produced and maintained by the FMRIB Analysis Group at the University of Oxford in England. ANFI

is a software package created and maintained by neuroimaging researcher at the National Institute of Mental Health (NIMH) in Bethesda. SPM, FSL, and ANFI are free software. There is also a commercial software called Brain Voyager.

1.5 What can we assume after preprocessing?

After preprocessing mentioned previously, we can assume that there is no effect on fMRI data due to the individual's brain size and shape, head movement of subject during the scan, and difference in time of scanning. Also, through spatial and temporal smoothing, we have increased signal-to-noise ratio. So in analyzing fMRI data, which is a time series at each voxel, we only need to consider the effect from the experiment in experimental design or its natural functional connectivities in a resting state fMRI scan.

Here is the data structure of fMRI we can assume after preprocessing.

- Data is dependent across TR and space. (The fMRI data is neither independent nor identically distributed.)
- Data consisting one time of a whole brain scan is taken concurrently.
- Subject did not move his head during the scanning.
- All brain size is the same and registered to one brain atlas.
- Data we obtain has high SNR.

Since we are standardizing brain size and anatomical structure using same atlas, we can also assume independence and identical distribution across subjects.

1.6 Motivation of research

There can be multiple goals in the statistical analysis of fMRI data. The goal can be

- localizing brain areas activated by the task;
- determining networks corresponding to brain function; and
- making predictions about psychological or disease states.

The first goal is shown in the media a lot. A researcher designs a study by asking patients to perform a certain task and seeing which areas of the brain get activated. Investigations can highlight the areas of activation. Often a researcher compares two groups of patients in order to examine if there is any difference in the locations of activated brain areas.

The second goal is related to connectivity analysis. Functional connectivity refers to the functionally integrated relationship between spatially separated brain regions. Functional connectivity is typically analyzed in terms of correlation, coherence, and spatial grouping based on temporal similarities. In my research I measure similarity using correlation.

To understand the brain's activity, dividing the brain region into smaller

pieces than the anatomical region is necessary so that small regions can be used as a unit of further analysis. The reason why anatomical region is not appropriate as the unit for further analysis is that the voxel time series have different patterns even within one small anatomical region. Examples can be shown after we define what metric we use to measure similarity.

Dividing the brain's, spatial domain into a set of non-overlapping regions or modules that show some homogeneity with respect to information is called brain parcellation. One of the most popular methods of brain parcellation is spectral clustering.

Chapter 2: Spectral Clustering

After the scans of brain volume and preprocessing, time series data can be obtained from each voxel. If two voxels' time series from fMRI are highly correlated, then we say that these two voxels are functionally connected. Using the information on how voxels are functionally correlated, we want to divide a brain area into smaller regions. We call this process clustering and these smaller regions clusters.

In graph theory, clustering is the operation of partitioning the graph into groups in such a way that the edges between different groups have very low weights (which means that points in different groups are dissimilar from each other) and the edges within a group have high weights (which means that points within the same cluster are similar to each other). Since our goal is to find a partition of the set of voxels such that voxels in different clusters are functionally dissimilar from each other and voxels within the same cluster are functionally similar to each other, we can adapt graph theory to our data.

Consider voxels as points in 3-dimensional space which are connected with measurable strength. The measurable strength is the correlation between the time series generated for the pair of voxels.

2.1 Notation

Let N be a number of voxels, which can vary depending on the resolution of the MRI scanner. Let $i = 1, \ldots, N$ be the index of voxels. Suppose that the i^{th} voxel v_i has 3- dimensional coordinate $\mathbf{s}_i = (s_{i1}, s_{i2}, s_{i3}) \in \mathbb{N}^3$. (By convention, these dimensions are called X, Y, and Z. X represents the left-right dimension, Y represents the anterior-posterior dimension, and Z represents the inferior-superior dimension.) Let j be the index of subjects and J be the total number of subjects. For multiple subjects, we denote i^{th} voxel of j^{th} subject as v_i^j and its coordinate $\mathbf{s}_i^j = (s_{i1}^j, s_{i2}^j, s_{i3}^j) \in \mathbb{N}^3$. In case of single subject analysis, with $J = 1$, the j superscript can be omitted. Let t be the index of duration of brain scans and T be the total number of brain scans, $t = 1, \ldots, T$.

Then let $\{X_i^j(t)\}$ be the observed time series at the i^{th} voxel v_i at the time of t for the subject j. In vector notation, we denote the time series by the time series $\boldsymbol{X}_i^j = (X_i^j(1), \ldots X_i^j(T))' \in \mathbb{R}^T$. In a single subject analysis, we denote it by $\boldsymbol{X}_i$.

To measure how closely two time series from voxel v_i and $v_{i'}$ are related, we use sample Pearson's correlation to estimate correlation between the time series. For a subject j, Pearson's correlation (corr) between two time series $\boldsymbol{X}_i^j$ and $\boldsymbol{X}_{i'}^j$ is

$$corr(\boldsymbol{X}_i^j, \boldsymbol{X}_{i'}^j) = \frac{1}{T-1} \sum_{t=1}^{T} \frac{(X_i^j(t) - \bar{X_i^j})(X_{i'}^j(t) - \bar{X_{i'}^j})}{sd_i^j sd_{i'}^j}, \quad i, i' = 1, \ldots N \tag{2.1}$$

where

$$sd_i^j = \sqrt{\sum_{t=1}^{T}(X_i^j(t) - \bar{X_i^j})^2}, \quad sd_{i'}^j = \sqrt{\sum_{t=1}^{T}(X_{i'}^j(t) - \bar{X_{i'}^j})^2}.$$

While attempting to use correlation as a similarity measure, we found that there are voxels with no signal. (That is, there are some voxels with $X_i = \mathbf{0}$.) These may come as a result of smoothing in preprocessing or from some minimum resolution of the measuring machinery. Therefore on these voxels, correlations are undefined. To resolve this, we can choose to omit those voxels in an analysis but omitting data that we already have may not be a good decision since we are losing some information by doing so. Therefore we can use the following definition.

Let sd_x be a sample standard deviation of x. For a small $\delta \geq 0$, define a modified correlation $\widetilde{c}_\delta(x, y)$ between x and y:

$$\widetilde{c}_\delta(x, y) = cov(\frac{x}{\max(sd_x, \delta)}, \frac{y}{\max(sd_y, \delta)}). \tag{2.2}$$

We will always use this modified correlation (2.2) when we compute correlation as a similarity measure. In this definition, time-lags are not taken into account in computing correlation because the time-scale of fMRI (defined by a parameter TR, usually 2-3 seconds) is much longer than the time-scale on which neurons transfer signals. Also, through a form of preprocessing, called slice timing correction, we assume that fMRI data have no time difference in scanning.

In many analyses, researchers often ignore any correlation that is less

than 0.5. Also, we can apply a physical constraint enforcing similarities to have a fixed small value beyond a fixed inter-voxel distance. Each voxel has 26 neighbor voxels around it. If we only measure similarities between neighborhood voxels within certain distance α, we can define similarity function as follows. For example, if we want to include only 26 voxels then $\alpha = \sqrt{2}$. Therefore for certain distance apart, we will assign very small similarity value to meet the requirement that the similarity function is bounded away from 0. We will discuss further in Chapter 3.

$$w_{\delta,\alpha,\eta}(x,y) = g_{\alpha,\eta}(\widetilde{c}_\delta(x,y)) = \begin{cases} \widetilde{c}_\delta(x,y) & \text{if} \quad \widetilde{c}_\delta(x,y) \geq 0.5 \quad \text{and} \quad d(x,y) \leq \alpha \\ \eta & \text{otherwise} \end{cases} \tag{2.3}$$

where $d(x,y)$ is a Euclidean distance between x and y and η is a very small positive number.

Since the equation (2.3) consists of two continuous functions enforcing similarities to have a fixed small value beyond a fixed inter-voxel distance, the similarity function $w_{\delta,\alpha,\eta}$ is piecewise continuous and this property is required later to show that the set of functions $w(x,\cdot)$ indexed by $x \in [-M,M]^T$ is a Glivenko-Class. These will be discussed in Chapter 3. Also $w(x,y)$ is bounded above and below and symmetric.

Let W^j be the similarity matrix for the jth subject with elements $w^j_{i,i'} = w(\boldsymbol{X}_i, \boldsymbol{X}_{i'})$ defined equal to the similarity measure of time series between voxel v_i and v'_i. For now, assume single subject analysis for convenience since we

can apply the same principle when we want to calculate W^j. A later section will discuss multiple subjects analysis.

To distinguish W from the finite matrix obtained from N time series, we can denote the similarity matrix from the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ by W_N. Then let D_N be a degree matrix, which is a diagonal matrix with diagonal entries $d_i = \sum_{l=1}^{N} w_{i,l}$.

Then we can define graph Laplacian as below.

$$L = D - W$$

$$L_N = D_n - W_N$$

We can also define two versions of normalized graph Laplacians as follows.

$$\begin{aligned}
L' &= D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} = I - H' \\
L'_N &= D_N^{-1/2} L_N D_N^{-1/2} = I - D_N^{-1/2} W_N D_N^{-1/2} = I - H'_N \\
L'' &= D^{-1} L = I - D^{-1} W = I - H'' \\
L''_N &= D_N^{-1} L_N = I - D_N^{-1} W_N = I - H''_N
\end{aligned}$$

There is a close relationship between eigenvalues and eigenvectors of four matrices L'_N, L''_N, H'_N, and H''_N. Thus properties about the spectrum of one of four matrices can be reformulated for the three other matrices as well. In particular, for studying convergence properties of spectral clustering it will make no difference whether we work with the normalization L'_N or L''_N. In the following, we will call both L'_N and L''_N normalized graph Laplacian.

2.2 Properties of graph Laplacian

Here are some properties of the spectrum of normalized and unnormalized Laplacians. Here we make the assumption that w in non-negative and symmetric, as these are the standard assumptions in spectral clustering. More properties of graph Laplacian can be found in von Luxburg (2007).

Proposition 1. (Properties of L_N)

The matrix L_N satisfies the following properties:

1. For every vector $f \in \mathbb{R}^N$ we have

$$f' L_N f = \frac{1}{2} \sum_{i,i'=1}^{N} w_{i,i'} (f_i - f_{i'})^2. \tag{2.4}$$

2. L_N is symmetric and positive semi-definite.

3. The smallest eigenvalue of L_N is 0, the corresponding eigenvector is the constant one vector $\mathbb{1}$.

4. L_N has n non-negative, real-valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$.

Proof Part (1): By the definition of d_i,

$$f' L_N f = f' D_N f - f' W_N f = \sum_{i=1}^{N} d_i {f_i}^2 - \sum_{i,i'=1}^{N} f_i f_{i'} w_{ii'}$$

$$= \frac{1}{2} \left(\sum_{i=1}^{N} d_i {f_i}^2 - 2 \sum_{i,i'=1}^{N} f_i f_{i'} w_{ii'} + \sum_{i'=1}^{N} d_{i'} {f_{i'}}^2 \right) = \frac{1}{2} \sum_{i,i'=1}^{N} w_{ii'} (f_i - f_{i'})^2.$$

Part (2): From the symmetry of W_N and D_N, L_N is also symmetric. Since $w_{ii'} \geq 0$, $f' L_N f \geq 0$ for all $f \in \mathbb{R}^N$.

Part (3): Since L_N is positive semi-definite, the smallest eigenvalue of L_N is

0, and the corresponding eigenvector is the constant one vector $\mathbb{1}$.

Part (4): (4) is a direct consequence of the parts (1)-(3).

Then similar properties can be shown for L' and L''. There is a very close relationship between the spectra of two different forms of normalized graph Laplacians: v is an eigenvector of L'' with eigenvalue λ if and only if $= D^{1/2}v$ is an eigenvector of L' with eigenvalue λ. So from a spectral point of view, the two normalized graph Laplacians are equivalent. A discussion of various other properties of graph Laplacians can be found in the literature; see, for example, Chung A.2 for the normalized and Mohar A.2 for the unnormalized case.

2.3 Spectral clustering algorithm

Two different versions of graph Laplacian yield two versions of spectral clustering, which are called "normalized" or "unnormalized" spectral clustering, respectively. The basics of algorithms can be summarized as follows. They both use k-means algorithm to assign points to k-clusters. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

Unnormalized spectral clustering

Input: Similarity matrix $W \in \mathbb{R}^{N \times N}$, number k of clusters to construct.

- Construct similarity matrix.
- Compute the unnormalized Laplacian L.
- Compute the first k eigenvectors $u_1, ..., u_k$ of L.
- Let $U \in \mathbb{R}^{Nxk}$ be the matrix containing the vectors $u_1, ..., u_k$ as columns.
- For $i = 1, ..., n$, let $y_i \in \mathbb{R}^k$ be the i-th row of U, as a column vector.
- Cluster the points $(y_i)_{i=1,...,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \ldots, C_k$.

  Output: Clusters $X_1^*, \ldots, X_k^*$ with $X_i^* = \{j | y_j \in C_i\}$.

Alternatively, using normalized Laplacian, we can also perform spectral clustering. Here is summarized algorithm of normalized spectral clustering.

Normalized spectral clustering according to Shi and Malik (2000)

Input: Similarity matrix $W \in \mathbb{R}^{N \times N}$, number k of clusters to construct.

- Construct similarity matrix.
- Compute the unnormalized Laplacian L.
- Compute the first k eigenvectors $u_1, ..., u_k$ of the generalized eigenproblem $Lu = \lambda Du$.
- Let $U \in \mathbb{R}^{N \times k}$ be the matrix containing the vectors $u_1, ..., u_k$ as columns.
- For $i = 1, \ldots, N$, let $y_i \in \mathbb{R}^k$ be the i-th row of U, as a column vector.
- Cluster the points $(y_i)_{i=1,...,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C'_1, \ldots, C'_k$.

  Output: Clusters $X^*_1, \ldots, X^*_k$ with $X^*_i = \{j | y_j \in C'_i\}$.

After $X^*_1, \ldots, X^*_k$ are computed, we can transfer the information into a different format of matrix, called adjacency matrix A : Adjacency matrix is defined to have $\{0, 1\}$ entries, with $A_{i,i'}$ equal to 1 if voxels i, i' are in the same cluster C_j, and equal to 0 otherwise.

Note that the spectral clustering algorithms presented above contain basic principles. However, the implementations used in practice can differ in

various details. We used the spectral clustering algorithm for multiple clusters that is suggested by Yu and Shi (2003).

2.4 Goal of spectral clustering in fMRI analysis

What we want to achieve is to find a partition of the brain regions such that the similarities between different clusters are very low (which means that voxels in different clusters are dissimilar from each other) and the similarities within a group are high (which means that voxels within the same cluster are similar to each other).

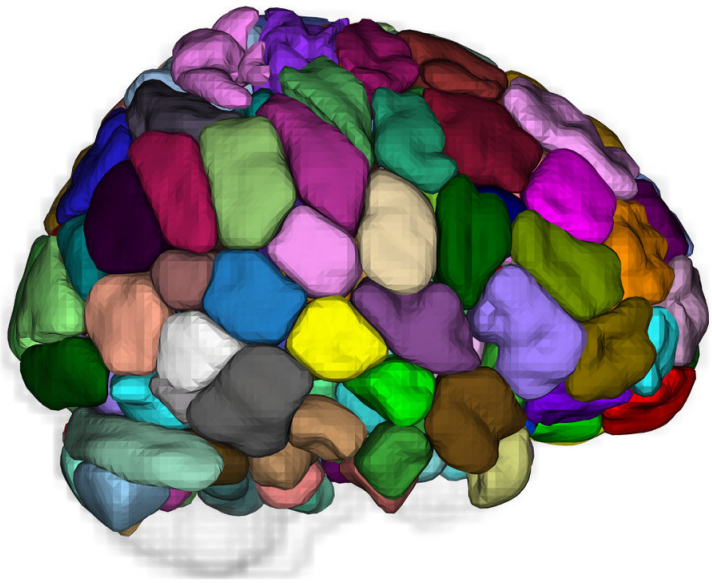


Figure 2.1: The 236-region functional parcellation used to define network nodes in the ADHD-200 dataset. From A neuromarker of sustained attention from whole-brain functional connectivity, M. Rosenberg et al., Nature Neuroscience 19, 165171 (2016)

For example, the superior temporal gyrus is one of three (sometimes two) gyri in the temporal lobe of the human brain, which is located laterally to the head, situated somewhat above the external ear. The superior temporal gyrus contains the primary auditory cortex, which is responsible for processing sounds.

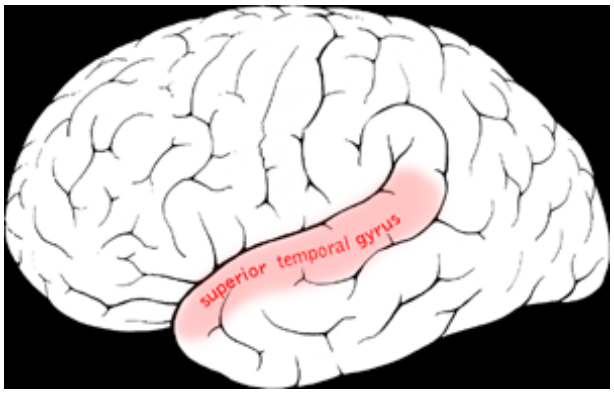


Figure 2.2: Superior temporal gyrus, from Wikipedia

When plotting the correlations *corr* as defined in the earlier section, we observed clusters of correlations even within one brain region, the superior temporal gyrus. This implies that there is a group of voxels in the same brain region that is functionally similar to voxels in the same group but dissimilar with the voxels in another group. Therefore, using the spectral clustering techinque, we want to divide brain regions into much smaller regions than the regions in brain map but larger than individual voxel. This is called brain parcellation.

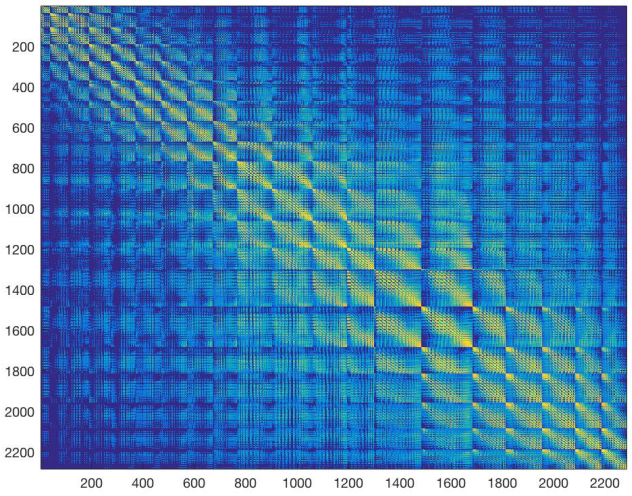


Figure 2.3: Plots of correlation in one region of interest (ROI), superior temporal gyrus. The number of voxels in this ROI is 2278. There are several clusters of correlations within one ROI. This suggests the need of brain parcellation.

2.5 Success of parcellation

Then how can we determine the success of parcellation when we deal with real brain data? One way to measure the success is to compare correlations between clusters and within clusters. We can use the idea of Fisher's discriminant to quantify this measure. Below we define Fisher's discriminant (F) in the context of spectral clustering in fMRI analysis.

Let F be the Fisher's discriminant. Suppose that in the kth cluster, there are n_k. Then these yield $n_k(n_k+1)/2$ correlations. Let $\{r_{ks}\}_{s=1}^{n_k(n_k+1)/2}$ be the correlation within the cluster k. Let μ_k be a mean correlation within

cluster and the overall mean of correlations is $\bar{\mu}$. Then we can define Fisher's discriminant (F) as

$$
\begin{aligned}
F &= \frac{S_w}{S_b}, \qquad \text{where} \\
S_b &= \frac{1}{N}\sum_{k=1}^{K} n_k(\mu_k - \bar{\mu})^2, \qquad N = \sum_{k=1}^{K} n_k, \\
\mu_k &= \frac{1}{n_k}\sum_{s=1}^{n_k(n_k+1)/2} r_{ks}, \qquad \bar{\mu} = \frac{1}{N}\sum_{k=1}^{K} n_k\mu_k, \\
S_w &= \frac{1}{N}\sum_{k=1}^{K}\sum_{s=1}^{n_k(n_k+1)/2} n_k(r_{ks} - \mu_k)^2.
\end{aligned}
$$

Then we can expect that a better clustering will yield a smaller F.

To examine compactness of clusters, we can measure diameters of clusters and compare them. However, we are going to apply the physical constraint when we apply spectral clustering to fMRI data, thus this would not be an appropriate measure for a success of clustering.

2.6 Multiple subjects analysis

2.6.1 Methods

Based on a belief that there is a general parcellation of brain regions that we can apply to multiple subjects, we can perform a multi-subject analysis. In the case of multiple subjects analysis, we can suggest several ways in formulating Ws. Especially, two methods are suggested dealing with multiple subjects in Craddock (2012). Each one has benefit and drawback in achieving

clusterings.

The first method suggested in Craddock (2012) is to take an average of the simularity matrices over the subjects. Suppose that we have ${W^j}_{j=1}^J$. Then the averaged similarity matrix is $\bar{W} = \frac{1}{J}\sum_{j=1}^{J} W^j$. Next perform spectral clustering in the same way we did for a single subject analysis. This method is much faster than the other method introduced later since spectral clustering is applied only once. Then the resulting clustering can be converted to an $n \times K$ dimensional adjacency matrix A as defined in the previous section.

The second method suggested in Craddock (2012) is to apply J times of spectral clustering for each individual similarity matrices. Suppose that we have J similarity matrices W^j. Then apply clustering method for each W^j. Then the resulting clustering can be converted to J adjacency matrices A^j. Thus we have J many A^js. Then take an average of A^js to obtain $\bar{A} = \frac{1}{J}\sum_{j=1}^{J} W^j$. As a second level analysis, apply the spectral clustering to $\bar{A}$. Then we can have final adjacency matrix A. One might suggest applying a threshold to $\bar{A}$ to get the final adjacency matrix A. However, this method can yield a voxel that does not belong to any of the clusters. The rationale for the second clustering is to assign all the voxels to one of the clusters.

2.6.2 Evaluation

In this section, I want to summarize the methods that were suggested in Craddock (2012). In Craddock, the resulting clustering solutions were com-

pared using two commonly applied strategies: leave-one-out cross-validation (LOOCV) and silhouette width (SI).

To perform LOOCV, we will exclude a subject at a time and perform clustering. Next we will compare to the clustering from the single subject who is excluded. Suppose that we have an adjacency matrix generated from the clustering with the mth subject excluded. Let's denote as A_{-m}. Then compare A_{-m} to the adjacency matrix calculated by clustering the data from the mth subject only, denoted as A_m. To measure how the group and the individual level clustering solutions are different, we can use Dice's coefficient. Then the averaged value of Dice's coefficients across all possible ways to exclude one subject will give a metric to tell how similarly the multiple subject analysis is performed compared to a single subject analysis.

Dice's coefficient measures the similarity between two adjacency matrices. It is the ratio of twice the number of connections common to both matrices, divided by the total number of connections present in both matrices.

Suppose that we have two adjacency matrices A and B and they have same dimension $N \times N$. Then we define entries of $A \cap B$ as follows. For $i, j = 1, \ldots, N$,

$$(A \cap B)_{i,j} = \begin{cases} 1 \\ \\ 0, \qquad \text{otherwise} \end{cases}$$

Also, $|\cdot|$ denotes the number of non-zero entries. Then we can define Dice's

coefficient to perform LOOCV.

$$dice = \frac{2 \cdot |A_{-m} \cap A_m|}{|A_{-m}| + |A_m|}, \tag{2.5}$$

Dice's coefficient results in numbers between zero and one, where one corresponds to perfect correspondence between matrices, and zero corresponds to no similarity.

On the other hand, the silhouette width was chosen to quantify the functional homogeneity of region of interest(ROI). SI measures cluster compactness compared to cluster separation. SI has been defined in terms of both similarity and distance metrics but the similarity formulation is used here.

Let's define the average similarity, a_k, between every pair of voxels assigned to cluster C_k of clustering. Suppose that there are n_k number of voxels that is assigned to the cluster C_k. Then the average similarity, a_k is :

$$a_k = \frac{1}{n_k(n_k - 1)} \sum_{i,i' \in C_k, i \neq i'} w(v_i, v_{i'}) \tag{2.6}$$

In Craddock (2012), SI was modified from the original definition to use the average similarity between in-cluster and out-of-cluster voxels. Let's define the average similarity between in-cluster and out-of-cluster voxels b_k:

$$b_k = \frac{1}{n_k(N - n_k)} \sum_{i \in C_k} \sum_{i' \notin C_k} w(v_i, v_{i'}) \tag{2.7}$$

The silhouette width for the clustering can then be calculated from:

$$si = \frac{1}{K} \sum_{k=1}^{K} \frac{a_k - b_k}{\max\{a_k, b_k\}} \tag{2.8}$$

Negative SI values indicate an incorrect clustering and values near 1 indicate a good solution. SI was calculated for each clustering solution from

each subjects data. Therefore averaged SI across subjects will give a metric for multiple subject analysis.

Besides these two measures, in literature other methods were used to evaluate the performance of clustering. For example, Thirion (2014) A.2 used the goodness of fit to measure accuracy and also showed reproducibility of the parcellation across bootstrap samples. Many other indices, such as Rand/adjusted Rand index, Hubert index, Silhouette index, Davies-Bouldin index, Calinski-Harabasz index, Hartigan index, Weighted inter-intra index, Krzanowski-Lai index, Homogeneity, and separation index were used in Somashekara (2014) A.2. However, some of the indices require knowledge of the ground truth information which is almost never available in practice. In addition, other measures such as clustering error, the variation of information, and the Wallace Index were used to evaluate the performance of spectral clustering in Verma (2005) A.2.

In this paper, we will present LOOCV (only for multiple subjects analysis), SI and Fisher's discriminant defined in the section 2.5.

Chapter 3: Consistency of Spectral Clustering

Even though spectral clustering is frequently used in many applications, its consistency has barely been studied. Proving consistency of spectral clustering itself is not achievable. Instead, consistency of spectral clustering means actually asymptotic behavior of spectral clustering that uses eigenfunctions of Laplacian to cluster data. However since its asymptotic behavior for large N has not been fully examined, it is worthy of research. Therefore we are not trying to demonstrate consistency of spectral clustering itself. We are going to address the consistency of estimation of eigenfunctions that drive the clustering.

Von Luxburg et al. (2008) showed the consistency of spectral clustering under the setting of independent identically distributed (iid) data sample. Even though her paper's title is consistency of spectral clustering, she actually proved only that the consistency of eigenfunctions of Laplacian. Lei and Rinaldo (2015) showed the consistency under stochastic block models for nodes grouped into "communities", but this cannot be directly applied to fMRI data analysis.

As is true also for other spatial data, the time series from fMRI voxels are

not spatially independent of each other. The closer the voxels are physically, the higher the correlation between their corresponding time series tends to be. Therefore we cannot assume that fMRI data in general are independent and identically distributed. In this chapter, we will extend von Luxburg's work under assumptions of weak dependence and possibly of stationarity. Thus the consistency of spectral clustering in this thesis means the consistency of eigenfunctions of Laplacian as it is in von Luxburg.

3.1 Asymptotics for Spatial Data

When the number of voxels grows to infinity, in our notation as $N \to \infty$, we can think of two frameworks of asymptotics for spatial data. The two ways N could tend to infinity are explained in Cressie (1993). One is called infill asymptotics and the other is called increasing window asymptotics. We will summarize these in a general setting first and then describe which asymptotic situation we have for fMRI analysis when the number of voxels increases.

Suppose that we have observations $Y = (Y_{s_1}, \ldots, Y_{s_N})'$, of spatial data located at $\{s_1 \ldots, s_N, s_i = (s_{i1}, s_{i2}, s_{i3}) \in \mathbb{R}^3\}$. Let $\mathcal{U}$ be the closure of an open region in $\mathbb{R}^3$ that contains at least all location vectors $\{s_i\}$.

$$\mathcal{U} = [\min_i s_{i1}, \max_i s_{i1}] \times [\min_i s_{i2}, \max_i s_{i2}] \times [\min_i s_{i3}, \max_i s_{i3}]$$

Suppose we have a domain $\mathcal{U} \subset \mathbb{R}^3$. The first asymptotic framework is to allow more and more observations to be taken from a stationary random field, or a random field whose averages over large windows converge in the sense of

the Law of Large Numbers by increasing the domain of observation. Suppose we have $\inf\{\|s_i - s_j\| : 1 \leq i < j \leq N\} > \Delta > 0$, then $N \to \infty$ implies $|\mathcal{U}| \to \infty$, where $|\mathcal{U}|$ denotes the volume of the domain $\mathcal{U}$. Such asymptotics are called increasing window asymptotics, and they are the spatial analogue of the usual asymptotics seen in time series analysis.

On the other hand, when the spatial index ranges continuously over a fixed subset $\mathcal{U} \subset \mathbb{R}^d$, one may view $\mathcal{U}$ as a bounded domain. Then an obvious way to increase n is to take observations at locations between the existing ones. This is called infill asymptotics, where $N \to \infty$, but $0 < |\mathcal{U}| < \infty$ remains fixed.

Considering fMRI data, the domain $\mathcal{U}$ is limited to the volume of a brain. Thus, it appears to be described better by infill asymptotics at a glance. However, there is a different phenomenon from geostatistical data with infill asymptotics. In geostatistical data with infill asymptotics, as more observations are taken, the data are more closely related. In other words, their correlations are increased as $N \to \infty$. However, in fMRI data, as the number of observations is large, which are time series at voxels, we may not have the situation that the correlations between the time series continuously get increased.

As the number of voxels is huge, neighboring voxels are nearly perfectly correlated. Even though the physical brain volume is constrained and there is no actual increase of N, since the number of observation is large, we can consider this situation as $N \to \infty$. Moreover, when $N \to \infty$, correlation can die off within very small distances, yielding different parcels. Even though the

parcels are close in the brain, they can be sufficiently far apart in terms of correlation distances. In other words, even though physical distance is small between two voxels, their correlation distance is not as $N \to \infty$. Thus we can view this as increasing window asymptotics and it is still meaningful to talk about a large sample of only weakly dependent clusters. Since we have more and more statistical information as $N \to \infty$, we can assume increasing window asymptotics. In the context of observations from a large class of Gaussian random field models, it is known that Fisher's information grows infinitely as $N \to \infty$ when correlation distance is bounded below, so that we can apply increasing window asymptotics. To calculate the Fisher's information, we will assume a parametric model for covariances of fMRI data.

Mardia and Marshall (1984) showed that for a Gaussian random field, the Fisher's information goes to infinity as $N \to \infty$ under parametric assumptions. Suppose that the fMRI is a real valued Gaussian process $\{Y_t : \ t \in T \subset \mathbb{Z}^3\}$ where T is an index set. Mardia and Marshall proved a result for general index set $T \subset \mathbb{Z}^d$, but we will refer to their result for the case of $d = 3$.

Suppose that for all $t \in T$, $E\{Y_t\} = z(t)'\beta$, where $z(t) = \{z_(t), ..., z_q(t)\}'$ is a $q \times 1$ vector of nonrandom regressors and $\beta \in B$ is a parameter vector, B being an open subset of $\mathbb{R}^q$. Also let the covariance be defined by a parametric model $cov\{Y_t, Y_s\} = \sigma(t, s; \theta)$, for all $t, s \in T$, where $\theta \in \Theta$ is a $p \times 1$ parameter vector, Θ being an open subset of $\mathbb{R}^p$. We assume that $\sigma(t, s; \theta)$ is twice differentiable with respect to θ at all points on $T^2 \times \Theta$, and positive-definite in the sense that for every finite subset $T = \{t_1, \ldots, t_N\}$ of T the covariance

matrix $V_N = \{\sigma(t_i, t_j; \theta)\}$ is positive-definite. Suppose that Y_t is observed at each point to give the sample vector $\mathbf{Y}_N = \{Y_{t_1}, \ldots, Y_{t_N}\}'$. We denote the combined $(q+p) \times 1$ parameter vector by $\phi = (\beta', \theta')'$. From the formula (2.1) in Mardia and Marshall (1984), the log likelihood for ϕ is

$$L_N(\phi; \mathbf{Y}_N) = -\frac{1}{2} \log |V_N| - \frac{1}{2}(\mathbf{Y}_N - Z_N\beta)' V_N^{-1} (\mathbf{Y}_N - Z_N\beta)$$

where Z_N is an $n \times q$ regressor matrix with jth column $z_j = \{z_j(t_1), \ldots, z_j(t_N)\}'$. We assume Z_N to be rank q. The equation (2.5) in Mardia and Marshall (1984) gives the Fisher's information matrix.

$$-E(\frac{\partial^2 L_N}{\partial \phi^2}) = diag(B_\beta, B_\theta),$$

where $B_\beta = Z'VZ$, $B_\theta = tr(VV^iVV^j)$, $V^i = \frac{\partial V^{-1}}{\partial \theta_i}$, and $V^j = \frac{\partial V^{-1}}{\partial \theta_j}$.

Mardia and Marshall (1984) proved that MLE's in their parametric spatial models with increasing-window asymptotics were consistent and asymptotically normal. As part of their proof, they show that the Fisher Information (the inverse of which gives the asymptotic variance matrix for parameters) behaves in the following way.

$$\lim B_\theta^{-1} = 0 \qquad \lim B_\beta^{-1} = 0, \text{ as } N \to \infty.$$

This means that the parametric Fisher's information increases to infinity as $N \to \infty$.

3.2 Assumptions and Definitions Needed for the Proof of Consistency

Here we want to provide assumptions and definitions that are needed to show the consistency of eigenfunctions that drive spectral clustering. We will start out with general assumptions regarding the similarity function for the case of fMRI data, and later we will introduce additional assumptions and definitions that in order to demonstrate asymptotic behavior of eigenfunctions that derive spectral clustering. Our approach is similar to that of von Luxburg et al. (2008).

3.2.1 General Assumptions

Suppose we have a probability space $(\Omega, \mathcal{F}, P)$ and a real valued random field $\{X_{\mathbf{s}} : \mathbf{s} \in S \subset \mathbb{N}^3\} = \{X_{\mathbf{s}_i} : \mathbf{s}_i = (s_{i1}, s_{i2}, s_{i3}) \in S \subset \mathbb{N}^3\}$. Since fMRI data consists of time series from all voxels, we can consider it as a random field indexed by a three dimensional index set. In addition, fMRI signals are always bounded i.e. there exist a M such that $|X_{\mathbf{s}}| \leq M$. Let χ be $[-M, M]^T \subset \mathbb{R}^T$.

Let $\chi = [-M, M]^T \subset \mathbb{R}^T$ be a compact metric space and $\{X_{\mathbf{s}}\}$ be a random field, indexed by $\mathbf{s} \in S$. Here are the assumptions that we impose throughout this chapter.

A1 $w(x, y) : \chi \times \chi \to [a, b] \subset \mathbb{R}$, is symmetric, where $a > 0$.

A2 There is a partition $\{B_r\}_{r=1}^R$ such that $\cup_{r=1}^R B_r = \chi \times \chi$ and $B_r \subset \overline{int(B_r)}$.

Also, there exist Lipschitz continuous functions w_r on the compact sets $\overline{B_r}$ (denoting the closure of B_r) such that w coincides with w_r on B_r.

Note that w is bounded above since it is a piecewise Lipschitz function on a finite union of compact domains. (Thus, the finite upper bound b in A1 is a consequence of A2.)

The consistency of spectral clustering was shown under the assumption of independent identically distributed data by von Luxburg et al. (2008). Even though fMRI data are dependent spatially, if we can assume dependence dies off at some specified rate, measured through φ- or ρ'-metrics of strong mixing, we can still show similar results as von Luxburg et al. In the following sections we will summarize some definitions of mixing conditions for random fields as well as some other definitions that are needed for the proof of consistency.

3.2.2 Mixing Rates

Here are some definitions of mixing rate from Bradley (2005) and Deo (1975). Let $(\Omega, \mathcal{F}, P)$ be a probability space and consider two σ-fields $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$. Consider a random field $\{X_{\mathbf{s}}\}$ indexed by $s \in S \subset \mathbb{N}^3$.

Definition 3.2.1. (φ-mixing rate of 3-d random field from Deo (1975))
For a given r, suppose that $S = S(r)$ and $T = T(r)$ are index sets in $\mathbb{R}^3$ linked to r which satisfy the property that

$$\max_{k=1,2,3} (\inf_{\mathbf{t}\in T} t_k - \sup_{\mathbf{s}\in S} s_k) \geq r \tag{3.1}$$

for $\mathbf{s} = (s_1, s_2, s_3)$ and $\mathbf{t} = (t_1, t_2, t_3)$, then $B \in \mathcal{B}(r) = \sigma(X_{\mathbf{t}},\ \ \mathbf{t} \in T(r))$ and $A \in \mathcal{A}(r) = \sigma(X_{\mathbf{s}},\ \ \mathbf{s} \in S(r))$. Then we define

$$\varphi(r) = \sup_{S(r),T(r)\,\text{subject to (3.1)}} |P(B|A) - P(B)|. \tag{3.2}$$

Then $\varphi(0) = 1$ and clearly $\{\varphi(r)\}$ is a decreasing sequence of real numbers. If $\varphi(r) \to 0$ then we say that the random field $\{X_{\mathbf{s}}\}$ is φ-mixing. This definition will be recalled when we discuss the strict stationarity assumption.

From the definition 3.2.1, for any event $A \in \sigma(X_{\mathbf{s}}, \mathbf{s} \in S(r))$ and $B \in \sigma(X_{\mathbf{t}}, \mathbf{t} \in T(r))$ where $S(r)$ and $T(r)$ satisfy (3.1), we have

$$|P(B|A) - P(B)| \leq \varphi(r).$$

Now we want to show that the φ-mixing assumption for the random field X_s implies the same assumption for the random field $g(X_s)$ for a function g. Suppose there is a function $g : \mathbb{R} \to \mathbb{R}$, and that we define a random field $Y_{\mathbf{s}} = g(X_{\mathbf{s}})$.

Corollary 3.2.1. If $X_{\mathbf{s}}$ is φ-mixing random field with rate $\varphi(r)$ then so is $Y_{\mathbf{s}}$.

Proof. For any event $A^* \in \sigma(Y_{\mathbf{s}}, \mathbf{s} \in S(r))$ and $B^* \in \sigma(Y_{\mathbf{t}}, \mathbf{t} \in T(r))$, since $Y_{\mathbf{s}} = g(X_{\mathbf{s}})$, $\sigma(Y_{\mathbf{s}}, \mathbf{s} \in S(r)) \subseteq \mathcal{A}(r)$ and $\sigma(Y_{\mathbf{t}}, \mathbf{t} \in T(r)) \subseteq \mathcal{B}(r)$. The equality holds in these relations $\subseteq$ if g is a one-to-one function. Since $X_{\mathbf{s}}$ is a φ-mixing random field, $|P(B^*|A^*) - P(B^*)| \leq \varphi(r)$. Therefore, $Y_{\mathbf{s}}$ is a φ-mixing random field.$\square$

Definition 3.2.2. (ρ'-mixing rate from Bradley (2015))

For a given r, suppose that $S = S(r)$ and $T = T(r)$ are index sets in $\mathbb{R}^3$ linked

to r which satisfy the property that

$$\max_{k=1,2,3}(\inf_{\mathbf{t}\in T} t_k - \sup_{\mathbf{s}\in S} s_k) \geq r \tag{3.1}$$

for $\mathbf{s} = (s_1, s_2, s_3)$ and $\mathbf{t} = (t_1, t_2, t_3)$. Suppose that we have two σ-fields $\mathcal{A}(r) = \sigma(X_{\mathbf{s}},\ \mathbf{s} \in S(r))$ and $\mathcal{B}(r) = (X_{\mathbf{t}},\ \mathbf{t} \in T(r))$. Then $\forall f_1, f_2$, define the maximal coefficient of correlation

$$\rho(\mathcal{A}(r), \mathcal{B}(r)) = \sup_{f_1 \in L^2(\mathcal{A}(r)), f_2 \in L^2(\mathcal{B}(r))} |Corr(f_1, f_2)|.$$

Then we can define

$$\rho'(r) = \sup_{S(r),T(r)\,\text{subject to (3.1)}} \rho(\mathcal{A}(r), \mathcal{B}(r))$$

If $\rho'(r) \to 0$ as $r \to 0$ then we say that random field $\{X_s\}$ is ρ'-mixing. Bradley (2015) showed central limit theorem for a nonstationary random field with ρ'-mixing. We will recall this when we show the consistency of spectral clustering.

From the definition 3.2.2, for any event $f_1 \in L^2(\mathcal{A}(r))$ and $f_2 \in L^2(\mathcal{B}(r))$, we have

$$|Corr(f_1, f_2)| \leq \rho'(r).$$

Now we want to show that the ρ'-mixing assumption for the random field X_s implies the same assumption for the random field $h(X_s)$ for a function h. Suppose there is a function $h : \mathbb{R} \to \mathbb{R}$ then we have a random field $Y_{\mathbf{s}} = h(X_{\mathbf{s}})$.

Corollary 3.2.2. If $X_{\mathbf{s}}$ is ρ'-mixing random field with rate $\rho'(r)$ then so is $Y_{\mathbf{s}}$.

Proof. For any event $f_1^* \in L^2(\sigma(Y_{\mathbf{s}}, \mathbf{s} \in S(r)))$ and $f_2^* \in L^2(\sigma(Y_{\mathbf{t}}, \mathbf{t} \in T(r)))$, since $Y_{\mathbf{s}} = h(X_{\mathbf{s}})$, $L^2(\sigma(Y_{\mathbf{s}}, \mathbf{s} \in S(r))) \subseteq \mathcal{A}(r)$ and $L^2(\sigma(Y_{\mathbf{t}}, \mathbf{t} \in$

$T(r))) \subseteq \mathcal{B}(r)$. The equality holds if h is a one-to-one function. Since $X_{\mathbf{s}}$ is a ρ'-mixing random field, $|Corr(f_1^*, f_2^*)| \leq \rho'(r)$. Therefore, $Y_{\mathbf{s}}$ is a ρ'-mixing random field.$\square$

3.3 Consistency Results of von Luxburg et al. (2008) under Dependent Data Structure

Here we are going to discuss some cases in which we can establish the weak law of large numbers (WLLN) that we need to follow the proof of von Luxburg et al. (2008). Theorems are similar to what von Luxburg has presented, but we cannot apply exactly the same proofs since we have different assumptions, also resulting in the different type of convergence. Some of the definitions in the empirical process have been changed. Steps in the proof needed to be re-examined because we showed the convergence in probability whereas von Luxburg showed a.s. convergence of operators. In each case, we will list out the assumptions and these will be used together with general assumptions from A1-A2 to have WLLN.

Definition 3.3.1. (Weak Law of Large Number)

Suppose that $X_{\mathbf{s}}$ is a random field indexed by multi-index $\mathbf{s} \in S \subset \mathbb{R}^3$, $\mathbf{1} \leq \mathbf{s} \leq \mathbf{n}$ and g is a real-valued measurable function on χ such that $\sup_{\mathbf{s}\in S} Eg(X_{\mathbf{s}})^2 < \infty$. If

$$\frac{1}{|\mathbf{n}|} \sum_{\mathbf{1}\leq\mathbf{s}\leq\mathbf{n}} g(X_{\mathbf{s}}) \xrightarrow{p} \mu_g \tag{3.3}$$

for all such $g : \chi \to \mathbb{R}$, where μ_g is a finite constant depending on g and the probability law of the X_s random field, then we say that we have the Weak Law of Large Numbers (for $\{X_\mathbf{s}\}$ with respect to P).

For a stationary random field, we have $\mu_g = E(g(X_\mathbf{1}))$.

3.3.1 Strict Stationarity Case

Here we want to provide assumptions under which the weak law of large numbers (WLLN) holds in the stationary case. Suppose $\{X_\mathbf{s}\}$ is a strictly stationary random field satisfying the additional conditions.

A3 $\{X_\mathbf{s}\}$ is a φ-mixing strictly stationary random field with mixing rate $\sum_{r=1}^{\infty} r^2 \varphi(r)^{\frac{1}{2}} < \infty$

A4 $EX_\mathbf{1} = \mu$ and $E(X_1^2) < \infty$

The property of φ-mixing at a rapid rate is highly plausible in the fMRI setting. The 3-dimensional index is associated with location in 3-dimensional grid and can be used to calculate distance between voxels. As the distance between two voxels gets bigger, the dependence between two fMRI signals from two locations quickly dies off as shown in the figures in Chapter 4. Therefore we can apply the lemma and theorem from Deo (1975) with a specific value of $q = 3$.

Lemma 1. Suppose that we have assumption A1-A4 hold, then

$$|\mathbf{n}|^{-1} E(S_\mathbf{n}^2) \to \sum_{\mathbf{i} \in Z^3} r(\mathbf{i}) = \sigma^2 \qquad \text{as} \quad n \to \infty \tag{3.4}$$

holds where $r(\mathbf{i}) = E(X_{\mathbf{1}} X_{\mathbf{s}_i})$ and $S_{\mathbf{n}} = \sum_{\mathbf{1} \le \mathbf{i} \le \mathbf{n}} X_{\mathbf{s}_i}$.

The proof can be found in Deo (1975). This theorem tells in particular that variance of the partial sum of X_s is finite of order $|\mathbf{n}|$, which is $o(|\mathbf{n}|^2)$. Therefore we have the weak law of large numbers by Chebyshev's inequality.

Proposition 2. Suppose $\{X_{\mathbf{s}}\}$ is a strictly stationary random field satisfying A1-A4 and $\sup_{\mathbf{s}} Eg(X_{\mathbf{s}})^2 < \infty$. Then we have the weak law of large numbers.

$$\frac{1}{|\mathbf{n}|} \sum_{\mathbf{1} \le \mathbf{s} \le \mathbf{n}} g(X_{\mathbf{s}}) \xrightarrow{p} \mu_g$$

for all $g : \chi \to \mathbb{R}$.

Proof. By Chebyshev's inequality, we have

$$P\left(\left| \frac{1}{|\mathbf{n}|} \sum_{\mathbf{1} \le \mathbf{s} \le \mathbf{n}} g(X_{\mathbf{s}}) - \mu_g \right| \ge \epsilon \right) \le \frac{Var(\sum_{\mathbf{1} \le \mathbf{s} \le \mathbf{n}} g(X_{\mathbf{s}}))}{|\mathbf{n}|^2 \epsilon^2}.$$

By the lemma 1, the right-hand side of the last inequality tends to 0 as $|\mathbf{n}| \to \infty$.

3.3.2 Non-stationary Case

Here we want to provide assumptions ensuring the weak law of large number (WLLN) for the non-stationary case. Suppose that $\{X_{\mathbf{s}}\}$ is a random field, not necessarily stationary. Suppose that we have following assumptions.

B3 $\{X_{\mathbf{s}}\}$ is a ρ'-mixing random field

B4 $\rho'(j) < 1$ for some $j \in \mathbb{N}$.

B5 $\frac{1}{N}\sum_{\mathbf{s}_i} EX_{\mathbf{s}_i} \to \mu$ and $\sup_{\mathbf{s}} E(X_{\mathbf{s}}^2) < \infty$

Then we can apply the theorem in Bradley which is the consequence of Theorem 28.9 and Theorem 28.10(I) with a specific value of $d = 3$ in Bradley (2007).

Theorem 1. Suppose that we have A1-A2— and B3-B5. Suppose $\mathbf{s} \in S \subset \mathbb{N}^3$, and $\{X_{\mathbf{s}}\}$ is a (not necessarily strictly stationary) random field such that for each $\mathbf{s} \in S$, the random variable $X_{\mathbf{s}}$ has mean zero and finite second moments. Suppose $\rho'(j) < 1$ for some $j \in \mathbb{N}$. Then for any nonempty finite set $Q \subseteq S$,

$$E|\sum_{\mathbf{s}\in Q} X_{\mathbf{s}}|^2 \leq C \sum_{\mathbf{s}\in Q} E(X_{\mathbf{s}})^2$$

where $C := j^3(1+\rho'(j))^3/(1-\rho'(j))^3$.

The proof can be found in Bradley (2007, 2015). Since the variance of the partial sum of $X_{\mathbf{s}}$ is of order $|\mathbf{n}|$ which is $o(|\mathbf{n}|^2)$, we have the weak law of large numbers by Chebyshev's inequality.

Both in strict stationary and in nonstationary cases, under assumptions A1-A4 and assumptions A1-A2, B3-B5 respectively, we have WLLN.

3.4 Glivenko-Cantelli Theorem

Here we will discuss Glivenko-Cantelli (GC) function classes and related definitions such as bracketing under our general assumptions A1-A2. Recall the definition of the weak law of large numbers in definition 3.3.1.

Definition 3.4.1. (Glivenko-Cantelli Class)

Let $\mathcal{G}$ be a class of real-valued measurable functions on $\mathbb{R}^d$ for which each of the WLLN limits (3.3) exist, with convergence in probability to a constant μ_g. If

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{|\mathbf{n}|}\sum_{\mathbf{1}\leq\mathbf{s}\leq\mathbf{n}} g(X_{\mathbf{s}}) - \mu_g\right| \xrightarrow{p} 0, \tag{3.5}$$

then we say that $\mathcal{G}$ is a Glivenko-Cantelli Class

Consider a function $w(x,\cdot) : \chi \to \mathbb{R}$, especially the particular case $w(x,\cdot) : \chi \to [a,b]$, $a > 0$, $b < \infty$ from A2. Let $S = \{1,\ldots,n_1\}\times\{1,\ldots,n_2\}\times\{1,\ldots,n_3\}$ be the locations of X_s, where $\prod_i^3 n_i = |\mathbf{n}|$. If X_s is a strictly stationary random field satisfying the assumptions A3 and A4 or if X_s is a non stationary random field satisfying the assumptions B3 to B5 , then by the law of large numbers of definition 3.3.1, for each $x \in \chi$ we have

$$\left|\frac{1}{|\mathbf{n}|}\sum_{\mathbf{s}\in S} w(x, X_{\mathbf{s}}) - E(w(x, X_{\mathbf{s}}))\right| \xrightarrow{p} 0.$$

For $\mathcal{W}$ to be a Glivenko-Cantelli class, there must hold for all $x \in \chi$

$$\sup_{w\in\mathcal{W}}\left|\frac{1}{|\mathbf{n}|}\sum_{\mathbf{s}\in S} w(x, X_{\mathbf{s}}) - E(w(x, X_{\mathbf{s}}))\right| \xrightarrow{p} 0. \tag{3.6}$$

Now we will introduce the definition of ϵ-bracket and bracketing number to prove (3.6).

Definition 3.4.2. (Bracketing number)

Suppose F is a class of measurable functions. Given two functions l and u, the bracket $[l,u]$ is the set of all functions f with $l \leq f \leq u$. An ϵ-bracket is

a bracket $l \leq f \leq u$ with $\|u - l\| < \epsilon$. The bracketing number $N_{[\,]}(\epsilon, F, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover F. The upper and lower bounds u and l of the brackets need not belong to F themselves but are assumed to have finite norms.

The definition 3.4.2 can be applied to any norm but we will be using L^2 norm when we refer to this definition.

Now we define a class of function $\mathcal{W}$ and we want to show that $\mathcal{W}$ is a Glivenko-Cantelli class, a proposition that is analogous to the proposition 11 in Luxburg (2008).

Proposition 3. Let $w : \mathcal{X} \times \mathcal{X} \to [a, b], a > 0, b < \infty$ be a similarity function satisfying A1-A2, and $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the corresponding normalized similarity function defined as $h(x, y) = w(x, y)/\sqrt{d(x)d(y)}$, where $d(x) = \int w(x, y)dP(y)$, and $g \in C(\mathcal{X})$ an arbitrary function. Then we define the following:

$$\mathcal{W} := \{w(x, \cdot); x \in \mathcal{X}\}, \qquad \mathcal{H} := \{h(x, \cdot); x \in \mathcal{X}\},$$

$$g \cdot \mathcal{H} := \{g(\cdot)h(x, \cdot); x \in \mathcal{X}\}, \qquad \mathcal{H} \cdot \mathcal{H} := \{h(x, \cdot)h(y, \cdot); x, y \in \mathcal{X}\}.$$

Under the general assumptions A1-A2, and stationary assumption A3-A4 (B3-B5 for non-stationary), the classes $\mathcal{W}$, $\mathcal{H}$ and $g \cdot \mathcal{H}$ are Glivenko-Cantelli classes.

Proof. Since $w(x, \cdot)$ is a piecewise Lipschitz function with finitely many pieces, it has piecewise bounded variation. Piecewise bounded variation on finitely many compact domains implies that it has bounded variation. Therefore, the class of these functions indexed over all $x \in \mathcal{X}$ has finitely many

ϵ-brackets by the example 19.11 on page 273 of van der Vaart (1998). Therefore taking a supremum of (3.5) over χ will also converge to zero in probability:

$$\sup_{w \in \mathcal{W}} \left| \frac{1}{|\mathbf{n}|} \sum_{\mathbf{s} \in S} w(x, X_{\mathbf{s}}) - E(w(x, X_{\mathbf{s}})) \right| \xrightarrow{p} 0.$$

Therefore $\mathcal{W}$ is Glivenko-Cantelli class. By similar argument, $\mathcal{H}$ and $g \cdot \mathcal{H}$ are Glivenko-Cantelli classes. □

Now consider empirical probability P_N. It is a linear operator mapping functions f to random variables expressed as normalized sum of the random variables $f(X_{\mathbf{s}_i})$, over $i = 1, \ldots, N$.

Definition 3.4.3. Let f be a function. The we can define P_N as follows.

$$P_N f = \frac{1}{|\mathbf{n}|} \sum_{\mathbf{1} \leq \mathbf{s} \leq \mathbf{n}} f(X_{\mathbf{s}}) = \frac{1}{|N|} \sum_{1 \leq i \leq N} f(X_{\mathbf{s}_i}).$$

Now we have established the WLLN and GC theorems that we need to show the result analogous to that of von Luxberg (2008) did under our assumptions A1-A4 or B3-B5 together with A1-A2. Let us state the conclusion similar to Theorem 15 in von Luxburg (2008). Since our similarity function is a piecewise continuous function on compact domain, we can define a space of piecewise continuous functions $C_p(\chi)$.

Definition 3.4.4. Let $\{B\}_{k=1}^{K}$ be a partition of χ such that

1. $B_k \cap B_l = \emptyset$ for $k \neq l$
2. $\cup_{k=1}^{K} B_k = \chi$

Then $C_p(\{B_1, \ldots, B_k\})$ is the space of piecewise continuous functions $f : \chi \to \mathbb{R}$ all with the same partition, defined as follows. We will denote

$C_p(\{B_1, \ldots, B_k\}) = C_P(\chi)$ for convenience. For any function $f \in C_p(\chi)$, if we restrict a domain of f to B_j, $f_{|B_j}$, then f is a continuous on $\overline{B_r}$.

$$\begin{aligned} C_p(\chi) &= C_p(\{B_1, \ldots, B_k\}) \\ &= \{f : \chi \to \mathbb{R} | \quad \forall j = 1, \ldots, K \quad \exists g_j \in C(\overline{B_r}) : g_j = f \text{ on } B_j\} \end{aligned}$$

where $\overline{B_r}$ is the closure of B_k.

Definition 3.4.5. We define norm $\|f\| = \max_{j=1,\ldots,K} \|f_{|B_j}\|_\infty$ on the space $C_p(\mathcal{X})$.

Remark: $\|f_{|B_j}\|_\infty = \|g_j\|_\infty$ where $f = g_j$ on B_j as in the definition of $C_p(\chi)$ functions f.

3.5 Construction of the Operators on $C_p(\chi) = C_p(\{B_1, \ldots, B_k\})$

As we explained earlier, demonstrating the consistency of spectral clustering is not yet achievable because it depends on properties of clustering algorithms operating on general classes of discretized spectral-operator eigenfunctions that have not yet been established by any authors. However, since spectral clustering is done by Laplacian matrix, consistency of Laplacian is the consistency of spectral clustering that von Luxburg presented. To study the convergence of normalized or unnormalized spectral clustering, we have to investigate whether the eigenvectors of the normalized or unnormalized Laplacians constructed on N sample points converge to eigenvectors defined on the underlying data space as $N \to \infty$. Since the size of the Laplacian matrix (both

normalized and unnormalized) is $N \times N$, it grows if N increases. Similarly, the dimension fo the space of eigenvectors gets larger and larger. By constructing operators on functions of a single domain, with data location points filling out that domain, we can convert the problem of spectra into the problem of convergence of operators on functions of a fixed domain. The problem is now to define convergence of operators, defined on the same space, which we do through the following construction.

First, we will introduce several linear operators on $C_p(\chi)$ corresponding to the matrices, such as Laplacian and degree matrices. For a random vector $(v_1, ..., v_N) \in \mathbb{R}^N$, we consider a random function $f \in C_p(\chi)$ such that $f(X_{\mathbf{s}_i}) = v_{\mathbf{s}_i}$ for fixed $X_{\mathbf{s}_i}$, $\quad i = 1, \ldots, N$, and extend linear operators on $\mathbb{R}^N$ to deal with such functions rather than vectors.

Let us start with the unnormalized Laplacian. Recall that L_N is defined as $D_N - W_N$ where D_N is the random diagonal matrix containing the degrees $d_i = \sum_{i'=1}^{N} w(X_{\mathbf{s}_i}, X_{\mathbf{s}_{i'}})$ as diagonal elements and W_N is the random similarity matrix. First we want to relate the random degree vector $(d_1, ..., d_N)$ to some functions in $C_p(\chi)$. We want to find an operator acting on $C_p(\chi)$ which behaves similarly to the random diagonal matrix D_N on $\mathbb{R}^N$.

We can define the true and the empirical degree functions

$$d_N(x) := \int w(x, y) dP_N(y) \in C_p(\chi), \tag{3.7}$$

$$d(x) := \int w(x, y) dP(y) \in C_p(\chi). \tag{3.8}$$

Let us analyze how the matrix D_N operates on a random vector $f = (f_1, ..., f_N)' \in \mathbb{R}^N$. For each i we have $(D_N f)_i = d_i f_i$, that is the value of the vector f at coordinate i is multiplied by the value of d_i. If we now identify $\frac{1}{N} d_i$ with $d_N(X_{\mathbf{s}_i})$ and f_i with $f(X_{\mathbf{s}_i})$, then $\frac{1}{N} D_N$ can be interpreted as a random multiplication operator. The linear operator on $C_p(\chi)$ corresponding to the random matrix $\frac{1}{N} D_N$ will be the empirical operator and true multiplication operator is defined as follows.

$$M_{d_N} : C_p(\chi) \to C_p(\chi), \qquad M_{d_N} f(x) := d_N(x) f(x), \tag{3.9}$$

$$M_d : C_p(\chi) \to C_p(\chi), \qquad M_d f(x) := d(x) f(x) \tag{3.10}$$

Next we can take look at the similarity matrix W_N. Applying it to a random vector $f \in \mathbb{R}^N$ yields $(W_N f)_i = \sum_{i'=1}^N w(X_{\mathbf{s_i}}, X_{\mathbf{s}_{i'}}) f_{i'}$. This will be represented by the empirical random and true integral operator

$$S_N : C_p(\chi) \to C_p(\chi), \qquad S_N f(x) := \int w(x, y) f(y) dP_N(y), \tag{3.11}$$

$$S : C_p(\chi) \to C_p(\chi), \qquad S f(x) := \int w(x, y) f(y) dP(y). \tag{3.12}$$

With these definitions, the operator corresponding to the unnormalized graph Laplacian $\frac{1}{N} L_N$ is the difference between two random operators, the empirical and multiplication operators:

$$U_N \quad : \quad C_p(\chi) \to C_p(\chi), \tag{3.13}$$

$$U_N f(x) \quad := \quad M_{d_N} f(x) - S_N f(x) = \int w(x, y)(f(x) - f(y)) dP_N(y) \tag{3.14}$$

$$U \quad : \quad C_p(\chi) \to C_p(\chi) \tag{3.15}$$

$$U f(x) \quad := \quad M_d f(x) - S f(x) = \int w(x, y)(f(x) - f(y)) dP(y). \tag{3.16}$$

For the case of the normalized Laplacian, we only need to work with the symmetric normalization L'_N because from a spectral point of view, the two normalized graph Laplacians are equivalent. We define the random operator $H'_N = D_N^{-1/2} W_N D_N^{-1/2}$ that operates on some vector $f = (f_1, ..., f_N)'$ by $(H'_N f)_i = \sum_{i'} \frac{w(X_{\mathbf{s_i}}, X_{\mathbf{s_{i'}}})}{\sqrt{d_i d_{i'}}} f_{i'}$. Then the eigenvalues and eigenvectors of L'_N can be computed from those of H'_N. That is, v is an eigenvector of L'_N with eigenvalue λ if and only if v is eigenvector of H'_N with eigenvalue $1 - \lambda$. Therefore, no harm will be done by studying the convergence of the eigenvalues and eigenvectors of H'_N instead of L'_N.

We can see that this is very similar to the behavior of the unnormalized similarity matrix W_N, the difference being that $w(X_{\mathbf{s_i}}, X_{\mathbf{s_{i'}}})$ is replaced by $w(X_{\mathbf{s_i}}, X_{\mathbf{s_{i'}}})/\sqrt{d_i d_{i'}}$. So we will define the following normalized empirical random and true similarity functions

$$h_N(x,y) : C_p(\chi) \to C_p(\chi), \quad h_N(x,y) := \quad w(x,y)/\sqrt{d_N(x) d_N(y)}, \tag{3.17}$$

$$h(x,y) : C_p(\chi) \to C_p(\chi), \quad h(x,y) := \quad w(x,y)/\sqrt{d(x) d(y)}, \tag{3.18}$$

and introduce the following two random empirical and one true operators:

$$T_N : C_p(\chi) \to C_p(\chi), \qquad T_N f(x) = \int h(x,y) f(y) dP_N(y), \tag{3.19}$$

$$T'_N : C_p(\chi) \to C_p(\chi), \qquad T'_N f(x) = \int h_N(x,y) f(y) dP_N(y), \tag{3.20}$$

$$T : C_p(\chi) \to C_p(\chi), \qquad T f(x) = \int h(x,y) f(y) dP(y). \tag{3.21}$$

The differences are

$$U'_N = I - T'_N, \tag{3.22}$$

$$U' = I - T. \tag{3.23}$$

Note that in the definition of these operators, the scaling factors $\frac{1}{N}$ which are hidden in P_N and d_N cancel each other. (In other words, the matrix H'_N already contains a $\frac{1}{N}$ scaling factor, contrary to the case of the matrix W_N in the unnormalized case.) Therefore, contrary to the unnormalized case we do not have to scale the matrices H'_N and H_N with a factor $\frac{1}{N}$. So the operator T'_N corresponds directly to the matrix H'_N, while the operator T_N corresponds to the matrix $H_N := (h(X_{\mathbf{s_i}}, X_{\mathbf{s_{i'}}}))_{i.i'=1,\ldots,N}$. The reason why we introduce T_N and H_N is technical. It will be easier to prove that T'_N converges to T in two steps using the operator T_N in between. We will show that T_N and T'_N get close and that T_N converges to T.

3.5.1 Convergence of Operators

Now we want to prove that the sequence of random operators T'_N converges compactly to T in probability. First we will prove pointwise convergence. Then we will prove collectively compact convergence. Combining these two, we will conclude compact covergence. Since unnormalized Laplacian and normalized Laplacian are closely related, we will present the proof for normalized case in this section.

All operators here act on $(C_p(\mathcal{X}), \|\cdot\|_\infty)$. Let $L(C_p(\mathcal{X}))$ be the set of

closed operators and we can define several types of convergence for operators. The Following definition is adapted from Chatelin (1983, Chapter 3).

Definition 3.5.1. Let T_N be a sequence of operators in $L(C_p(\chi))$ converging to $T \in L(C_p(\chi))$ according to one of the following definitions:

(a) T_N converges to T in the sense of pointwise, denoted by $T_N \xrightarrow{\tilde{p}} T$, iff for all f in $C_p(\chi)$, $T_N f \to Tf$ as $N \to \infty$.

(b) T_N converges to T in the sense of operator norm to T, iff $\|T_N - T\| \to 0$ as $N \to \infty$.

(c) T_N converges to T in the sense of collectively compact, denoted by $T_N \xrightarrow{cc} T$, iff

  (i) $T_N \xrightarrow{\tilde{p}} T$ and

  (ii) the following condition is satisfied:

  the set $\cup_{N=1}^{\infty}(T - T_n)B$ is relatively compact in $C_p(\chi)$, where $B = \{f \in C_p(\chi); \|f\| \leq 1\}$.

(d) T_N converges to T in the sense of compact convergence, denoted by $T_N \xrightarrow{c} T$, iff

  (i) $T_N \xrightarrow{\tilde{p}} T$ and

  (ii) the following condition is satisfied:

  for any sequence f_N in B, the sequence $(T - T_N)f_N$ is relatively compact in $C_p(\chi)$, where B is defined in the same way as in (d).

Proposition 4. (Proposition 3.3 from Chatelin (1983))

The following are equivalent:

(i) $T - T_N$ is compact for any integer N and $T_N \xrightarrow{c} T$ and

(ii) $T \xrightarrow{cc} T$.

To be able to extend these convergence properties to convergence in probability on random operators sequences, we want to define a metric on a space of operators.

Definition 3.5.2. (Metric defined on $L(C_p(\chi))$)

$$\gamma(T'_N, T) = \sum_{m=1}^{\infty} \sup_{\|x-x'\|\leq 1/m} \sup_{\|f\|\leq 1} 2^{-m} |T'_N f(x) - Tf(x)|$$

Then we can say that random linear operators T'_N converges w.r.t. each of the convergences in in probability to T if

$$\gamma(T'_N, T) \xrightarrow{P} .0$$

This is a convergence metrized by a weighted sum of seminorms.

Remark: Convergence under this metric implies all three convergences, pointwise, collectively compact, compact convergence, defined in definition 3.5.1. So saying that this type of convergence holds in probability for a sequence of operators on $C_p(\mathcal{X})$ is the same as saying that the metric measuring the distance between T'_N and T converges to 0 in probability. Thus, we can say that random linear operators T'_N in probability to T in the compact-convergence topology if $\gamma(T'_N, T) \xrightarrow{P} 0$. This will appear later in the section.

Proposition 5. T'_N converges pointwise to T.

$$T'_N \xrightarrow{\tilde{p}} T$$

For all $f \in C_p(\chi)$, $T'_N f \to Tf$ converges in probability.

Proof. For arbitrary $f \in C_p(\chi)$, we have

$$\| T'_N f - Tf \|_\infty \leq \| T'_N f - T_N f \|_\infty + \| T_N f - Tf \|_\infty .$$

Recall $h_N(x,y) = w(x,y)/\sqrt{d_N(x)d_N(y)}$ from (3.17) and $h(x,y) = w(x,y)/\sqrt{d(x)d(y)}$ from equation (3.18). Then the second term can be written as

$$\begin{aligned} \| T_N f - Tf \|_\infty &= \sup_{x\in\chi} |P_N(h(x,\cdot)f(\cdot)) - P(h(x,\cdot)f(\cdot))| \\ &= \sup_{g\in f\cdot\mathcal{H}} |P_N g - Pg| \to 0 \end{aligned}$$

by Proposition 3. Recall that w is bounded below by $a > 0$ in general assumption A1, i.e. $w(x,y) > a > 0$ and therefore $d_N(x) > a$ and $d(x) > a$ for all $x \in \chi$. The first term can be bounded by

$$\begin{aligned} \|T'_N f - T_N f\|_\infty &\leq \|f\|_\infty \|w\|_\infty \sup_{x,y\in\chi} |\frac{1}{\sqrt{d_N(x)d_N(y)}} - \frac{1}{\sqrt{d(x)d(y)}}| \\ &= \|f\|_\infty \|w\|_\infty \sup_{x,y\in\chi} \frac{|d_N(x)d_N(y) - d(x)d(y)|}{\sqrt{d_N(x)d_N(y)} + \sqrt{d(x)d(y)}} (\frac{1}{\sqrt{d_N(x)d_N(y)}\sqrt{d(x)d(y)}}) \\ &\leq \|f\|_\infty \frac{\|w\|_\infty}{a^2} \sup_{x,y\in\chi} \frac{|d_N(x)d_N(y) - d(x)d(y)|}{\sqrt{d_N(x)d_N(y)} + \sqrt{d(x)d(y)}} \\ &\leq \|f\|_\infty \frac{\|w\|_\infty}{2a^3} \sup_{x,y\in\chi} |d_N(x)d_N(y) - d(x)d(y)| \\ &\leq \|f\|_\infty \frac{\|w\|_\infty^2}{2a^3} |d_N(x) - d(x))| \\ &\leq \|f\|_\infty \frac{\|w\|_\infty^2}{a^3} \sup_{g\in\mathcal{W}} |P_N g - Pg|. \end{aligned}$$

Together with proposition 3, $T'_N \xrightarrow{\tilde{p}} T$ in probability.□

Proposition 6. $T'_N \xrightarrow{cc} T$ in probability.

Proof. We want to prove that, for some $N_0 \in \mathbb{N}$, the sequence of operators $(T'_N - T)_{N>N_0}$ is collective compact. Since T is compact operator, it is enough to show that $(T'_N)_{N>N_0}$ is relatively compact with respect to the norm defined in definition 3.4.5. This will be done by using an extended version of the Arzela-Ascoli theorem (e.g., Section I.6 of Reed and Simon (1980)). The same Arzela-Ascoli criterion for relative compactness is easily seen to hold for the space $C_p(\chi)$ under the norm $\|\cdot\|$.

First, we fix the random sequence f_N then the random operators T'_N. By Proposition 8 in von Luxburg (2008), we know that the operator norm of T'_N is bounded. That is $\|T'_N\| \leq \|w\|_\infty / a$ for all $N \in \mathbb{N}$. Recall that B is a unit ball in $C_p(\chi)$, $B = \{f \in C_p(\chi); \|f\| \leq 1\} \subset C_p(\chi)$. Hence, the functions in $\cup_N T'_N B$ are uniformly bounded by $\sup_{N\in\mathbb{N}, f\in B} \|T'_N f\|_\infty \leq \|w\|_\infty / a$. To prove that the functions in $\cup_{N>N_0} T'_N B$ are equicontinuous, we have to bound the expression $|g(x) - g(x')|$ in terms of the distance between x and x', uniformly in $g \in \cup_N T'_N B$. Since $B \subset C_p(\chi)$ is a subspace of piecewise continuous functions, we need to show that $\forall j = 1, \ldots, K$, $\{g_{|Bj} : g \in B\}$ is equicontinuous,

$$\begin{aligned}
\sup_{f\in B, N\in\mathbb{N}} |T'_N f(x) - T'_N f(x')| &= \sup_{f\in B, N\in\mathbb{N}} \left| \int (h_N(x,y) - h_N(x',y)) f(y) dP_N(y) \right| \\
&\leq \sup_{f\in B, N\in\mathbb{N}} \|f\|_\infty \left| \int (h_N(x,y) - h_N(x',y)) dP_N(y) \right| \\
&\leq \|h_N(x,\cdot) - h_N(x',\cdot)\|_\infty
\end{aligned}$$

Now we have to prove that the right-hand side gets small whenever the

distance between x and x' gets small:

$$\begin{aligned}
& \sup_y |h_N(x,y) - h_N(x',y)| \\
\leq\ & \frac{1}{a^{3/2}}(\|\sqrt{d_N}\|_\infty \|w(x,\cdot) - w(x',\cdot)\|_\infty + \|w\|_\infty |\sqrt{d_N(x)} - \sqrt{d_N(x')}|) \\
\leq\ & \frac{1}{a^{3/2}}(\|w\|_\infty^{1/2} \|w(x,\cdot) - w(x',\cdot)\|_\infty + \frac{\|w\|_\infty}{2a^{1/2}} |d_N(x) - d_N(x')|) \\
\leq\ & C_1 \|w(x,\cdot) - w(x',\cdot)\|_\infty + C_2 |d(x) - d(x')| + C_3 \|d_N - d\|_\infty .
\end{aligned}$$

As χ is a compact space, the piecewise continuous functions w (on the compact space $\chi \times \chi$) and d are in fact uniformly piecewise continuous (with finitely many pieces). Thus, the first two terms $\|w(x,\cdot) - w(x',\cdot)\|_\infty$, and $|d(x) - d(x')|$ can be made arbitrarily small for all x, x' whenever the distance between x and x' is small. For the third term $\|d_N - d\|_\infty$, which is a random term, we know by the Glivenko Cantelli properties of Proposition 3 that it converges to 0 in probability. This means that for each given $\epsilon > 0$ there exists some $N_0 \in \mathbb{N}$ such that, for all $N > N_0$, we have $\|d_N - d\|_\infty \leq \epsilon$ in probability. Together, these arguments show that $\cup_{N>N_0} T'_N B$ is equicontinuous in probability. By the extended Arzela-Ascoli theorem, we then know that $\cup_{N>N_0} T'_N B$ is relatively compact in probability, which concludes the proof.□

Proposition 7. $T'_N \xrightarrow{c} T'$ in probability.

Proof. Since collectively compact convergence implies compact convergence, $T'_N \xrightarrow{c} T$. Therefore $U'_N \xrightarrow{c} U$.

Proposition 6 and proposition 7 can be proved by using the metric defined in definition 3.5.2. What we have examined in two inequalities in proposition

6 can be expressed as follows.

$$\begin{aligned}\sup_{\|x-x'\|\le r,\|f\|\le 1,N\ge 1} |T'_N f(x) - T'_N f(x')| &\le C_1 \cdot \sup_{\|x-x'\|\le r} \|w(x,\cdot) - w(x',\cdot)\|_\infty \\ &+ C_2 \cdot \sup_{\|x-x'\|\le r} |d(x) - d(x')| \\ &+ C_3 \cdot \|d_N - d\|_\infty \qquad (3.24)\end{aligned}$$

Since equation (3.24) converges to 0 in probability, and the following sequence converges.

$$\sum_{m=1}^{\infty} \sup_{\|x-x'\|\le 1/m} \sup_{f\in C_p(\chi)} 2^{-m} |T'_N f(x) - Tf(x)| \xrightarrow{P} 0$$

□ Therefore we can say that random linear operators T'_N in probability to T in the compact-convergence topology.

Since the rest of the set of results needed for our conclusions are the same as those in von Luxburg (2008), we omit them.

3.6 Clustering from the Laplacian Matrix L

Von Luxburg et al. (2008) showed the consistency of spectral clustering in terms of operators and eigenfunctions of the Laplacians. However, how the consistency of operators is related to the consistency of clustering, especially when there are more than 2 clusters, needs to be explained.

First, consider the random operators U_N and T'_N defined in section 3.5. Suppose that $f \in C(\chi)$ is the eigenfunction of U_N with arbitrary eigenvalue λ, then the vector $v \in \mathbb{R}^N$ with $v_i = f(X_{s_i})$ is an eigenvector of the matrix

$\frac{1}{N}L_N$ with eigenvalue λ. Similarly if, $f \in C(\chi)$ is an eigenfunction of T'_N with arbitrary eigenvalue μ, then the vector $v \in \mathbb{R}^N$ with $v_i = f(X_{s_i})$ is an eigenvector of the matrix H'_N with eigenvalue μ. Therefore it is sufficient to explain how the clustering is defined in terms of eigenvectors of L_N and T'_N.

When $K = 2$, the clustering can be directly obtained by the eigenfunction of the smallest non-zero eigenvalue. Suppose that $f \in C(\chi)$ is the eigenfunction of U_N or T'_N. Then we can define a partition by following rule.

$$\text{If } sign(f(X_{s_i})) > 0 \text{ then } s_i \in C_1.$$

$$\text{If } sign(f(X_{s_i})) < 0 \text{ then } s_i \in C_2.$$

Suppose $K > 2$ and $f_1, \ldots, f_K$ are the eigenfunctions of the Laplacian. Then we have k eigenvectors $v_i^k = f_k(X_{s_i})$. Let us create a matrix with the columns $\{v^k\}_{k=1}^K$ and denote it by V. Then consider the rows y_i of V as points in $\mathbb{R}^K$ and apply the k-means algorithm to assign rows to the partition elements $C_1, \ldots, C_k$.

The k-means algorithm is a popular machine learning technique for classification. Given a set of observations $(s_1, s_2, \ldots, s_N)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the N observations into $K(\leq N)$ sets $C = \{C_1, C_2, \ldots, C_K\}$ so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance).

We will show next how we obtain clustering from the eigenvectors of the Laplacian when $K > 2$. Now consider the ideal case when there are completely separated K clusters. Then the similarity matrix W has a block diagonal form,

and the same is true for the matrix L:

$$\begin{pmatrix} L_N^{(1)} & & & \\ & L_N^{(2)} & & \\ & & \ddots & \\ & & & L_N^{(K)} \end{pmatrix} \tag{3.25}$$

Even though this similarity matrix of block diagonal is not obtainable in the real fMRI data analysis, it is still helpful to understand the relation between the eigenvalues of Laplacian and spectral clustering. In the real fMRI data analysis, we can assume there is a very small positive off diagonal entry $\eta > 0$ after re-ordering the entries of Laplacian.

In chapter 2, Proposition 1 showed that the smallest eigenvalue is always 0 and its corresponding eigenvector is $\mathbb{1}$. Furthermore, the tutorial on spectral clustering by von Luxburg (2007) showed the following proposition on page 4.

Proposition 8. (Number of connected components and the spectrum of L) Let G be an undirected graph with non-negative weights. Then the multiplicity k of the eigenvalue 0 of L equals the number of connected components $A_1, ..., A_k$ in the graph. The eigenspace of eigenvalue 0 is spanned by the indicator vectors $\mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_k}$ of those components.

Each of the blocks L_k is a proper graph Laplacian on its own, namely the Laplacian corresponding to the subgraph of the k-th connected component. As it is the case for all block diagonal matrices, we know that the spectrum of L is given by the union of the spectra of L_k, and the corresponding eigenvectors

of L are the eigenvectors of L_k, filled with 0 at the positions of the other blocks. As each L_k is a graph Laplacian of a connected graph, we know that every L_k has eigenvalue 0 with multiplicity 1, and the corresponding eigenvector is the constant one vector on the i-th connected component. Thus, the matrix L has as many eigenvalues 0 as there are connected components, and the corresponding eigenvectors are the indicator vectors of the connected components.

This is an extended result from the theorem in Mohar (1991, p.5).

Theorem 2. Let G be a graph and W be a graph similarity matrix with its entries non-negative. Then:

(a) Laplacian $L = D - W$ has only real eigenvalues,

(b) L is positive semidefinite, its smallest eigenvalue is $\lambda_1 = 0$ and a corresponding eigenvector is $(1, 1, ..., 1)^t$. The multiplicity of 0 as an eigenvalue of L is equal to the number of components of G .

Therefore, in the ideal case when there are completely separated K clusters, we know that the eigenvectors of L and L'' are piecewise constant. Therefore if s_i and $s_{i'}$ are in the same cluster C_j, then they are mapped to exactly the sample point y_i. Since the clustering algorithm is applied to set of the points $y_i \in \mathbb{R}^K$, it will be able to extract the correct clusters.

3.7 Approximating RatioCut and Ncut for Arbitrary k

The Laplacian contains information about how many clusters are in the given sets of nodes. However, spectral clustering algorithm uses the k smallest eigenvalues (counting multiplicity) and their corresponding eigenvectors to assign N points into k clusters. In this section, we want to explore why clustering can be achieved by the k smallest eigenvalues and their corresponding eigenvectors. Since the clustering problem can be viewed as an optimization problem of a certain objective function such as RatioCut or Ncut defined below, the solution can be achieved by the eigenvalues and eigenvectors of the Laplacian as a consequence of the RayleighRitz theorem.

First, let us explain how clustering is related to the solution of an eigenvalue problem of the Laplacian. Given a similarity matrix W, we can denote sum of similarity between two sets A and B as $W(A,B) = \sum_{i\in A, i'\in B} w_{ii'}$. Let us introduce objective functions Cut and Ratio Cut (RatioCut). Suppose there are clusters $C_1,\ldots,C_K$. Let $vol(A) = \sum_{i\in A} d_i$. Then define cut and RatioCut as follows:

$$cut(C_1,\ldots,C_k) = \frac{1}{2}\sum_{k=1}^{K} W(C_k,\overline{C}_k)$$

$$RatioCut(C_1,\ldots,C_k) = \frac{1}{2}\sum_{k=1}^{K}\frac{W(C_k,\overline{C}_k)}{|C_k|} = \frac{1}{2}\sum_{k=1}^{K}\frac{cut(C_k,\overline{C}_k)}{|C_k|}$$

$$Ncut(C_1,\ldots,C_k) = \frac{1}{2}\sum_{k=1}^{K}\frac{W(C_k,\overline{C}_k)}{vol(C_k)} = \frac{1}{2}\sum_{k=1}^{K}\frac{cut(C_k,\overline{C}_k)}{vol(C_k)}$$

Given a partition of V (the set of N nodes of the graph) into $C_1, \ldots, C_k$, we define k indicator vectors $h_k = (h_{k1}, \ldots, h_{kn})'$ by

$$h_{ki} = \begin{cases} \frac{1}{\sqrt{|C_k|}} & \text{if} \quad v_i \in C_k \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, n; k = 1, \ldots, K$$

Then we define the matrix $H \in \mathbb{R}^{n \times K}$ as the matrix containing those k indicator vectors as columns. Observe that the columns in H are orthogonal to each other, that is $H'H = I$. Then we can see that

$$h_k' L h_k = \frac{cut(C_k, \overline{C}_k)}{|C_k|} = (H'LH)_{kk}.$$

Thus

$$RatioCut(C_1, \ldots, C_k) = \sum_{k=1}^{K} h_k' L h_k = \sum_{k=1}^{K} (H'LH)_{kk} = tr(H'LH).$$

Therefore the problem of minimizing $RatioCut(C_1, \ldots, C_k)$ can be rewritten as

$$\min_{C_1, \ldots, C_k} tr(H'LH) \quad \text{subject to } H'H = \text{I}.$$

We can relax the problem by allowing the entries of the matrix H to take arbitrary real values. Then the relaxed problem becomes:

$$\min_{H \in \mathbb{R}^{n \times K}} tr(H'LH) \quad \text{subject to } H'H = \text{I}.$$

This is a trace minimization problem and by the Rayleigh-Ritz theorem in the Section 5.2.2.(6) of Lütkepohl (1997), the solution is given by choosing H as the matrix which contains the first K eigenvectors of L as columns.

Proposition 9. (Rayleigh-Ritz Theorem) If A is a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and if $(u_1, \cdots, u_n)$ is any orthonormal basis of eigenvectors of A, where u_i is a unit eigenvector associated with λ_i, then

$$\min_{x \neq 0} \frac{x^T A x}{x^T x} = \lambda_1$$

(with the minimum attained for $x = u_1$), and

$$\min_{x \neq 0, x \in \{u_i, \ldots, u_{i-1}\}^{\perp}} \frac{x^T A x}{x^T x} = \lambda_i$$

(with the minimum attained for $x = u_i$), where $2 \leq i \leq n$.

Equivalently, if $W_k = V_{k-1}^{\perp}$ is the subspace spanned by $(u_k, ..., u_n)$, (with $V_0 = (0)$) then

$$\lambda_k = \min_{x \neq 0, x \in W_k} \frac{x^T A x}{x^T x} = \min_{x \neq 0, x \in V_{k-1}^{\perp}} \frac{x^T A x}{x^T x} \qquad \text{for } k = 1, \ldots, n.$$

Using the fact that $\sum {h_k}^T L h_k = tr(H^T L H)$ and $\lambda_i > 0$ for all i, we can get following

In order to obtain a partition of the graph, we need to re-transform the real-valued eigenvectors into a discrete indicator vector. The standard way is to use the k-means algorithms on the rows of H. Consider the entries of ith row of H as points in $\mathbb{R}^K$ and apply k-means algorithm to assign these points to the partition $C_1, \ldots, C_k$.

Chapter 4: Data Analysis

4.1 Algorithm of Spectral Clustering

There are many different variations in performing spectral clustering. In this chapter, we will use the algorithm suggested by Yu and Shi (2003) which uses unnormalized Laplacian. For the computation software, we used MATLAB version R2016b.

Our goal is to parcellate a brain area into K smaller clusters. The idea of spectral clustering is to separate points in different groups, called clusters, according to their similarities. What we want to achieve is to have large similarities within the same cluster and much smaller similarities across the clusters. We have shown that the spectral clustering can be restated as the minimization problem of a certain objective function and minimizer can be obtained by solving for the eigenvectors of Laplacian, corresponding to the K smallest eigenvalues.

The algorithm suggested by Yu and Shi (2003) used the normalized version of Laplacian, $L'' = I - D^{-1}W$. Instead of finding eigenvectors of L'', it finds eigenvectors of $D^{-1}W$. Thus instead of K eigenvectors corresponding to the K smallest eigenvalues, it finds K eigenvectors corresponding to the K

largest eigenvalues.

4.1.1 Summary of algorithm

First, Steps 1 to 2 find eigenvectors of the normalized Laplacian, $D^{-1}W$. Then the algorithm normalizes the eigenvectors so that they lie on the unit hypersphere centered at the origin in Step 3. Since we have the continuous optimum, we transform to a discrete solution. Therefore, Steps 4 to 7 find a discrete solution that satisfies the binary constraints, yet is closest to the continuous optimum using K-means clustering.

4.1.2 Algorithm in Steps

Here is the algorithm detail step by step from Yu and Shi (2003). Given weight matrix W and a number of classes K:

1. Compute the degree matrix D.

2. Find the optimal eigensolution Z^* by:

$$D^{-\frac{1}{2}}WD^{-\frac{1}{2}}\bar{V}[K] = \bar{V}[K]\text{Diag}(s), \bar{V}[K]^T\bar{V}[K] = I,$$

$$Z^* = D^{-\frac{1}{2}}\bar{V}.$$

3. Normalize Z^* by: $\widetilde{X}^* = \text{Diag}(\text{diag}^{-\frac{1}{2}}(Z^*Z^{*T}))Z^*$

4. Initialize X^* by computing R^* as:

   $R_1^* = [\widetilde{X}^*(i,1), \ldots \widetilde{X}^*(i,K)]^T, \text{random } i \in [n]$

   $c = 0_{n\times 1}$

For $k = 2, \ldots, K$, do: $c = c + abs(\widetilde{X}^* R^*_{k-1})$

$R^*_k = [\widetilde{X}^*(i,1), \ldots, \widetilde{X}^*(i,K)]^T, i = \arg\min c$

5. Initialize convergence monitoring parameter $\bar{\phi}^* = 0$.

6. Find the optimal discrete solution X^* by:

   $\widetilde{X} = \widetilde{X}^* R^*$

   $X^*(i,l) = \langle l = \arg\max_{k \in [K]} \widetilde{X}(i,k) \rangle, i \in \mathbb{V}, l \in [K]$.

7. Find the optimal orthonormal matrix R^* by:

   $X^{*T} \widetilde{X}^* = U \Omega \widetilde{U}^T, \Omega = \mathrm{Diag}(\omega)$

   $\phi = \mathrm{tr}(\Omega)$

   If $|\bar{\phi} - \bar{\phi^*}|$ <machine precision, then stop and output $X^* \bar{\phi^*} = \bar{\phi}$

   $R^* = \widetilde{U} U^T$

8. Go to Step 6.

We applied the algorithm to simulated data and real fMRI data from Autism Brain Imaging Data Exchange (ABIDE) group. Several criteria defined in section 2.5 were presented to measure the quality of clusterings, such as Fisher's discriminant, silhouette (SI), and Dice's coefficient. We provided average Dice's coefficient for the simulated data, provided Fisher's discriminant and silhouette for the single subject data analysis and provided Fisher's discriminant, silhouette and Dice's coefficient from LOOCV for the multiple subject data analysis.

4.2 Simulated Data Analysis

In this section, we applied the algorithm in Yu and Shi (2003) to two simulated data. The two simulated datasets do not reflect realistic characteristics of fMRI data because the correlation matrix from fMRI data cannot have such a block-diagonal structure, rather it will have strips of non zero blocks in the similarity matrix. However, since we know the ground truth of clustering for the simulated data, we can compare the clustering result with the ground truth. As a way of comparison, after we get the adjacency matrix from clustering algorithm, we compare with the ground truth adjacency matrix.

Dice's coefficient was introduced as a measurement of multiple subject analysis in Chapter 2. Originally Dice's coefficient can be calculated as long as there are two matrices to compare. In the Chapter 2, we provided the equation 2.5 to compare how many common entries A_{-m} and A_m have. We can have slightly different definition of Dice's coefficient to compare two matrices from clustering result from simulated data and ground truth.

Notations $\cap$ and $|\cdot|$ in the definition of following Dice's coefficient are same as the ones in Chapter 2. Suppose that we have two adjacency matrices A and B and they have same dimension $N \times N$. Then we define entries of $A \cap B$ as follows. For $i, j = 1, \ldots, N$,

$$(A \cap B)_{i,j} = \begin{cases} 1 & \text{if } A_{i,j} = B_{i,j} \\ 0 & \text{otherwise.} \end{cases}$$

Also, $|\cdot|$ denotes the number of non-zero entries. Then we can define Dice's coefficient to compare two adjacency matrices from clustering result of simulated data and ground truth.

$$Dice = \frac{2 \cdot |A_{\text{ground truth}} \cap A_{\text{simulation}}|}{|A_{\text{groud truth}}| + |A_{\text{simulation}}|}. \tag{4.1}$$

The average Dice's coefficient between ground truth adjacency matrix and adjacency matrix from simulated data was provided to measure how close these two matrices are.

4.2.1 Simulation 1

Since the algorithm takes a similarity matrix as an input, we do not need to simulate the fMRI time series to see the performance of the algorithm. We only need to create a similarity matrix. We will consider one of the most simple similarity matrices with block diagonal structure. Suppose that we have 320 voxels from 10 well-separated clusters. Each cluster will have 32 voxels. Assume that the voxels within the clusters have the correlation greater than zero and the voxels between the clusters have a very small positive correlation almost close to zero. The idea is very similar to the stochastic block model. See Lei and Rinaldo (2014). That is we assign certain numbers as the similarities between voxels in the same cluster and assign zero as the similarities between voxels from different clusters.

Consider $r_1, \ldots, r_q$ to be random numbers sampled from standard normal

distribution. We sampled $q = 31 * (31 + 1)/2 + 32$ random numbers for the correlations within clusters. For the rest of the correlation we assigned a very small positive number η.

Then we added blurred areas at the borders of clusters. Let β be the size of a blurred area. For every 32th voxels, neighborhood voxels within a distance $\sqrt{2}\frac{\beta-1}{2}$ will have correlation that is not equal to η. Let $BB = \{v_{32}, v_{64}, \ldots, v_{320}\}$ be the set of 32th voxels. Then we generated random numbers from standard normal distribution for the correlations between voxels within the distance of $\sqrt{2}\frac{\beta-1}{2}$ from voxels in BB.

$$w_{i,i'} = \begin{cases} r \sim \text{N(0,1) iid}, & \text{if } v_i \text{ and } v_{i'} \text{ are from the same cluster} \\ r' \sim \text{N(0,1) iid}, & \text{if } v_i \text{ and } v_{i'} \text{ are not from the same cluster,} \\ & v_i \in BB \text{ and } |v_i - v_{i'}| \leq \sqrt{2}\frac{\beta-1}{2} \\ \eta > 0, & \text{otherwise} \end{cases} \tag{4.2}$$

The Figure 4.1 illustrates the simulated data and the ground truth we want to recover is shown Figure 4.2.

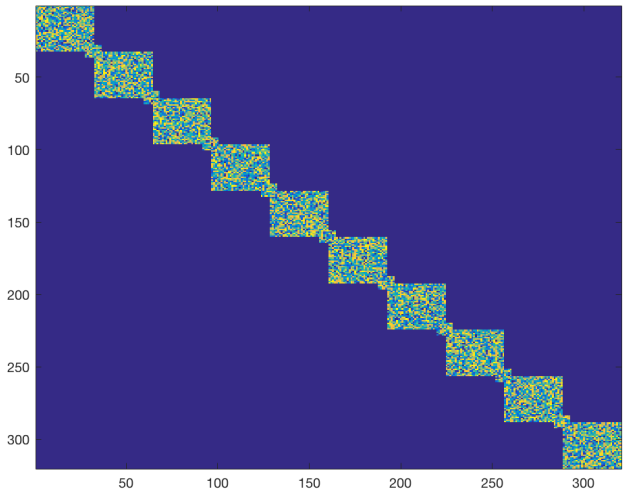


Figure 4.1: Simulated data using the equation 4.2 with 320 points with 10 clusters when $\beta = 9$

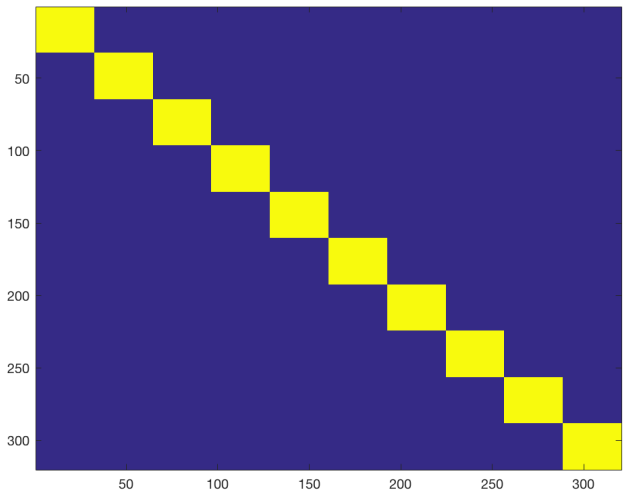


Figure 4.2: Ground truth

We repeatedly generated data 1000 times and performed clustering. The following figure shows the averaged result of clustering. (Figure 4.3).

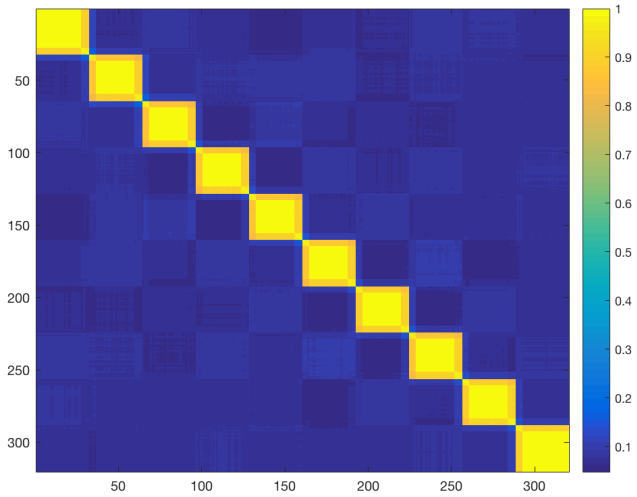


Figure 4.3: Average of resulting spectral clustering after 1000 repeats

Also we compared the adjacency matrix from the clustered results of 1000 simulated data and adjacency matrix from ground truth by Dice's coefficient using equation 4.1. Figure 4.4 is the histogram of the Dice's coefficient for 1000 simulations, and the average of Dice's coefficients is 0.726 and standard deviation is 0.057.

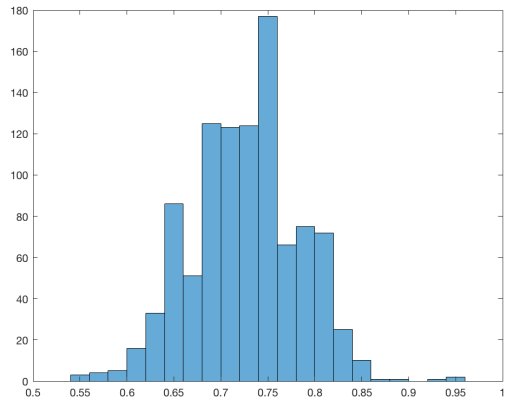


Figure 4.4: Histogram of Dice's coefficient for 1000 repeats

We are also interested is whether the algorithm will give the ground truth clusters as the blurred areas get large. As shown in the Figure 4.1 through Figure 4.4, when the blurred area gets larger, clustering results become worse.

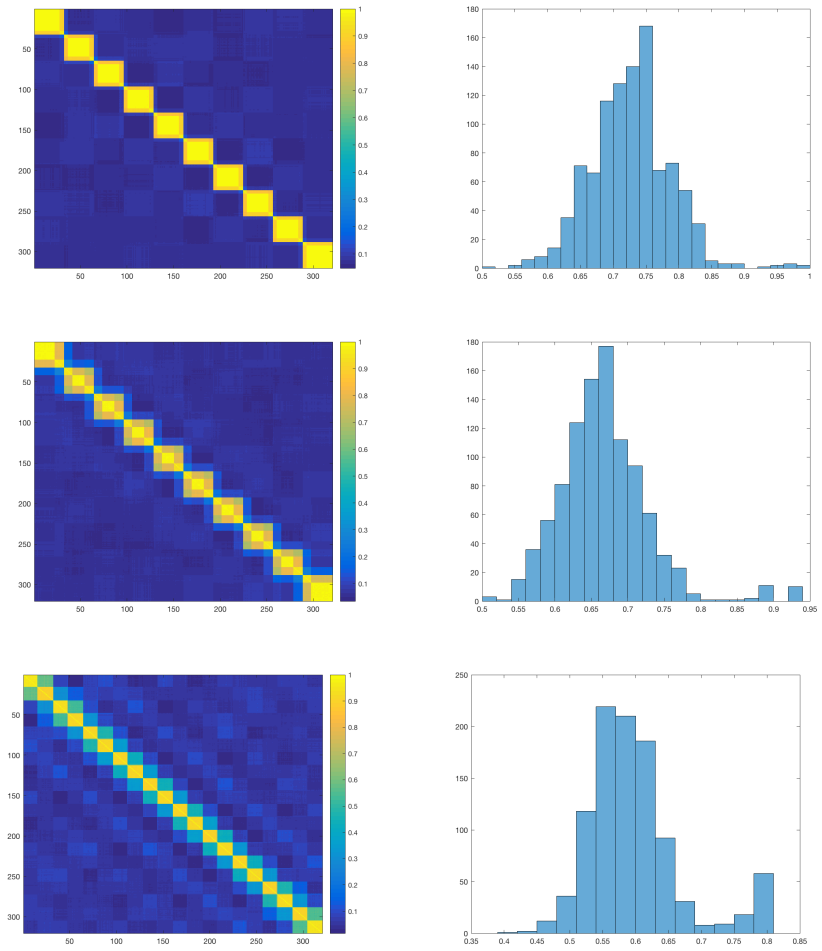


Figure 4.5: Clustering results and histograms of Dice's coefficient by β values.

Clustering results, β=9 (Left, Top row)

Dice's coefficient, β=9 (Right, Top row)

Clustering results, β=19 (Left, Middle row)

Dice's coefficient, β=19 (Right, Middle row)

Clustering results, β=33 (Left, Bottom row)

Dice's coefficient, β=33 (Right, Bottom row)

Dice's coefficient is defined in equation (4.1).

Average Dice's coefficient by different sizes of blurred area (β) and its

standard deviation are given in the table 4.1

β	Mean	SD
9	0.727	0.058
19	0.672	0.064
33	0.601	0.074

Table 4.1: Dice's coefficient in equation (4.1) by different sizes of blurred area, β.

From this simulation, we can conclude that clustering algorithm recovered ground truth adjacency matrix successfully until the size of blurred area is not greater than half of the block diagonal size. However, when the size of blurred area becomes greater than the half size of block diagonal, clustering algorithm did not recover ground truth matrix.

4.2.2 Simulation 2

Differently from simulation 1, now we generate a realization of a 3D Gaussian random field at a sequence of locations, $i = 1, \ldots, 320$, broken into 10 blocks. First we will generate the random variables, then generate iid sequence of 10 for time $t = 1, \ldots, 10$.

Then the generated data $X_i(t)$, where $i = 1, \ldots, 320,\ t = 1, \ldots, 10$ would be Gaussian mean 0 and iid across t and have a block-wise correlation structure as follows.

$$Cov(X_i(t), X_{i'}(t)) = \begin{cases} \sigma^2 & \text{if } i = i' \\ \rho^2 * \sigma^2 & \text{if } i \text{ and } i' \text{ are distinct within the same block} \\ \eta > 0 & \text{otherwise} \end{cases} \tag{4.3}$$

In the equation 4.3, η is always much less than ρ^2. If $\eta = 0.0001$, then the similarity matrix is close to a block diagonal matrix because there are very small positive entries for outside of blocks, and the average Dice's coefficient in equation (4.1) was 0.474 (SD 0.119) when $\sigma = 1, \rho = .2$.

We also examined how the clustering result is affected by different values of η. As expected as η increase, clustering result became worse in terms of average Dice's coefficient. The table 4.2 is showing the results by different η values.

η	Mean	SD
0.0001	0.474	0.119
0.001	0.419	0.096
0.01	0.282	0.042
0.05	0.246	0.014
0.1	0.241	0.013

Table 4.2: Dice's coefficient in equation (4.1) by different η values

4.3 Real Data Analysis

4.3.1 Data Description

The Preprocessed Connectomes Project (PCP) released preprocessed neuroimaging data for public use and opened for sharing. The preprocessed neuroimaging data from the Autism Brain Imaging Data Exchange (ABIDE) are now available for public use. From 2013 to 2017, about 34 publications used the data from ABIDE. Background and data description is available in online. Also, the data is available for download in http://preprocessed-connectomes-project.org/abide/index.html.

"Autism, or autism spectrum disorder (ASD), refers to a range of conditions characterized by challenges with social skills, repetitive behaviors, speech

and nonverbal communication, as well as by unique strengths and differences." (https://www.autismspeaks.org/what-autism) Previously it was considered rare, but ASD is now recognized to occur in more than 1% of children.

ABIDE is a collaboration of 16 international imaging sites that have aggregated and are openly sharing neuroimaging data from 539 individuals suffering from ASD and 573 typical controls. Data were preprocessed by five different teams using their preferred tools. For our analysis, we chose a dataset from one site and that is processed with the Configurable Pipeline for the Analysis of Connectomes (CPAC), which included options of skull stripping, template-based registration, automatic tissue segmentation, anatomical/functional coregistration, volume realignment, slice timing correction, intensity normalization, temporal filtering, nuisance signal correction, median angle correction, spatial smoothing, and motion scrubbing.

Here is a summary of demographic information of data. There are 110 subjects in total. Autism group has 55 subjects with average age of 12.7 years old. Control group has 55 subjects with average age of 14.1 years old at the time of the scan. Autism group has 46 male and 9 female, and the control group has 40 male and 15 female. In next two sections, we will perform spectral clustering for single subject data and multiple subjects data.

4.3.2 Single Subject Analysis

Here we provide a single subject fMRI data analysis using spectral clustering. This is a male subject who was 16.8 years old at the time of the scan. He is a subject from the Autism group. Due to the computational burden, we have chosen a single area for spectral clustering, and the area is labeled as superior occipital gyrus from the right hemisphere, based on AAL Single-Subject Atlas. The occipital lobe is one of the four major lobes of the cerebral cortex in the brain of mammals. It is the visual processing center of the mammalian brain containing most of the anatomical region of the visual cortex.

We use the modified correlation defined in the Chapter 2, the equation (2.2) and (2.3).

$$\widetilde{c}_\delta(x,y) = cov(\frac{x}{\max(sd_x,\delta)}, \frac{y}{\max(sd_y,\delta)}).$$

$$w_{\delta,\alpha,\eta}(x,y) = g_{\alpha,\eta}(\widetilde{c}_\delta(x,y)) = \begin{cases} \widetilde{c}_\delta(x,y) & \text{if} \quad \widetilde{c}_\delta(x,y) \geq 0.5 \quad \text{and} \quad d(x,y) \leq \alpha \\ \eta & \text{otherwise} \end{cases}$$

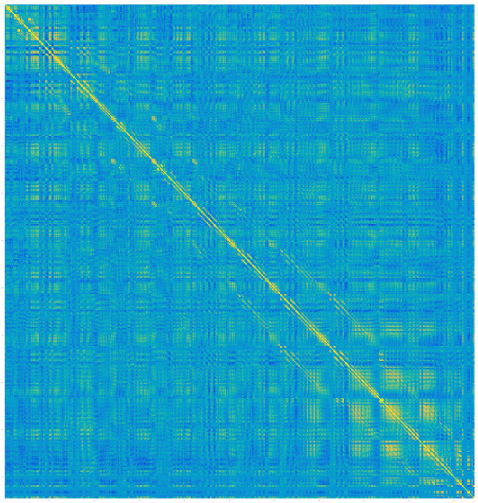


Figure 4.6: Plots of correlation equation 2.2

The δ is a positive number, required to ensure the continuity of the similarity function w. Also, η was a positive number to meet the bounded below requirement for the proof. Since we believe that neighborhood voxels have the most correlations, we only consider the correlations of voxels within a certain rage of distance that is parametrized as α. To check this, we can plot correlations with $\alpha = \infty$ first.

Figure 4.6 is the plot of the correlations defined in the equation 2.2 when $\delta = 0.001$. These are the correlations of all pairs of voxels with $\alpha = \infty$. For this particular subject, we observed high correlations between voxels within certain range.

In finding adjacency matrix, similarities between neighborhood voxels play important role. For each voxel, there are 26 neighborhood voxels around

it. If we include only these 26 neighborhood voxels then we can set $\alpha = \sqrt{2}$. In addition. if we only compute the correlations within certain distance α, we can set α to be different numbers. By setting smaller α, we can have a much simpler similarity matrix. Simpler matrix means a matrix with fewer non-zero elements. When $\alpha = 1.7$, if we choose correlation greater than equal to 0.5 as defined in the equation 2.3, we obtain the following similarity matrix.

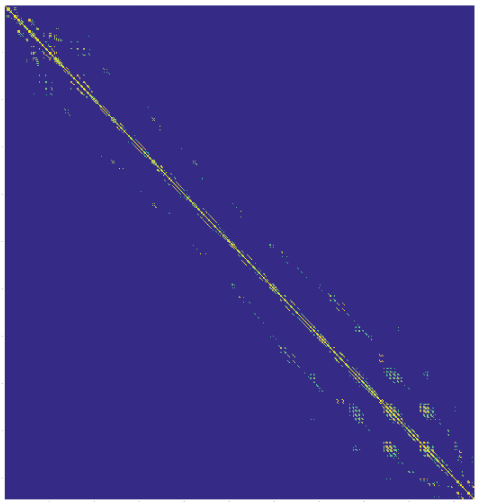


Figure 4.7: Plots of correlation defined as 2.3 with $\alpha = 1.7$. There are strips of non-zero elements.

After using the modified correlation as defined in the equation 2.3, we applied the spectral clustering algorithm by Yu and Shi (2003). Resulting clusters are shown in below when $K = 5$.

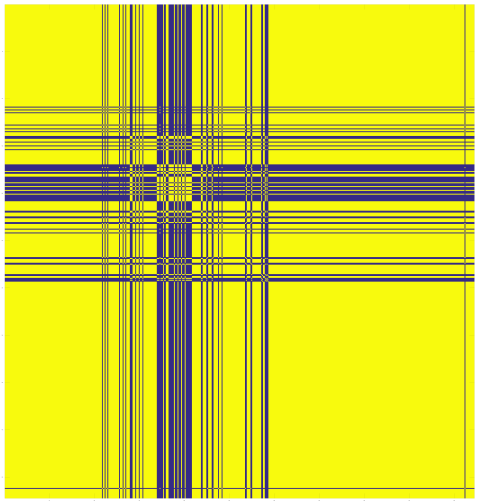


Figure 4.8: Plots of resulting clusters

The Fisher's discriminant was 0.1001 and silhouette was 1.0000 as a result of clustering. Sizes of clusters are (460 14 0 45 3). Even though we set the parameter K, the number of clusters, equal to 5, we obtained 4 cluster. Since Yu and Shi (2003) algorithm seeks the clustering differently from K-means algorithm, it does not give the same number of clusters as defined parameter. This coincide with what the Figure 4.7 showed. Many of voxels at the right bottom seem to be highly correlated and there are about 3-4 different clusters in the figure.

Following three figures (Figure 4.9 to Figure 4.11) are resulting clusters mapped with brain.

Figure 4.9: Clustering results, sagittal

Figure 4.10: Clustering results, coronal

Figure 4.11: Clustering results, axial

Here we want to discuss about the choice of α. As α gets larger, the similarity matrix includes correlations from more pairs of voxels. Therefore, it is expected to include wider band of correlations. In other words, if we increase α, we will have wider band in the correlation matrix based on the definition of the equation 2.3. Although we cannot display in details, we were actually able to see wider band when we increased α with major depressive disorder (MDD) data.

However, for this particular subject, we did not see huge difference of similarity matrices as α grows because there were high correlations between voxels within certain range and very small correlations when two voxels are outside of the certain range based on 4.6. Therefore for this subject, the correlations within occipital lobe area dies fairly quickly as the distance between

two voxels grows. Thus the impact of changes in value of α might be different by each dataset.

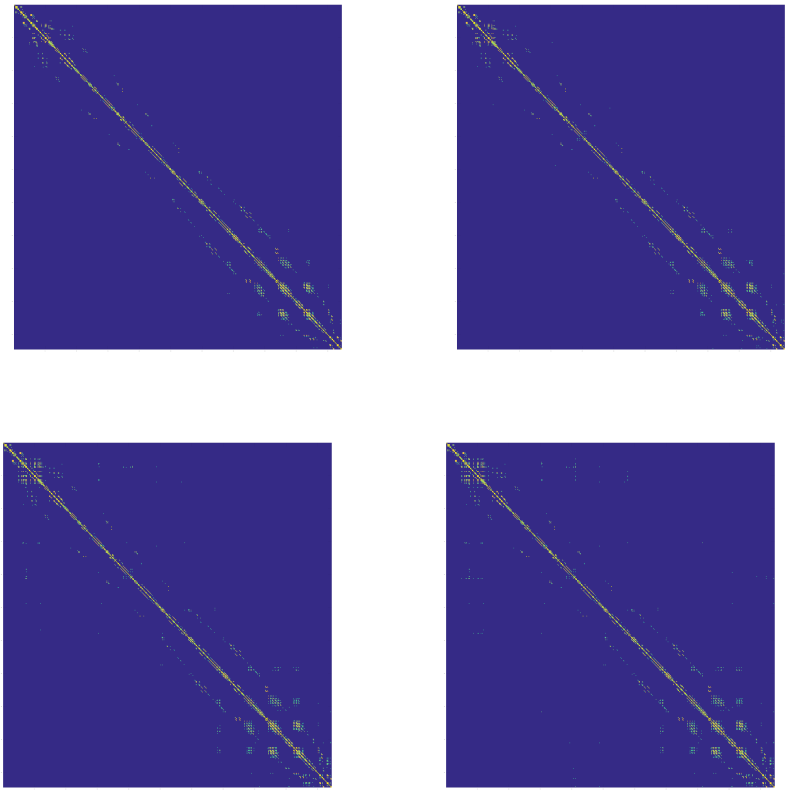


Figure 4.12: Plots of correlation by different number of clusters α, α=1.7 (Top, Left), 2.2 (Top, Right), 3 (Bottom, Left), 4.5 (Bottom, Right)

Here are the Fisher's discriminant and silhouette by different value of α. Dice's coefficient is not applicable because it is to compare two adjacency matrix. Based on Silhouette and Fisher's discriminant we can conclude that smaller α yields better clustering.

α	Fisher's discriminant	Silhouette	Dice's coefficient
1.7	0.1001	1	NA
2.2	0.0983	1	NA
3	0.1075	1	NA
4.5	0.2007	0.9896	NA

Table 4.3: Silhouette and Fisher's discriminant by α

As shown in the table, the smaller α yielded better clustering as expected. Now we will fix $\alpha = 1.7$ and see how the clustering results are changed as the number of clusters K increases.

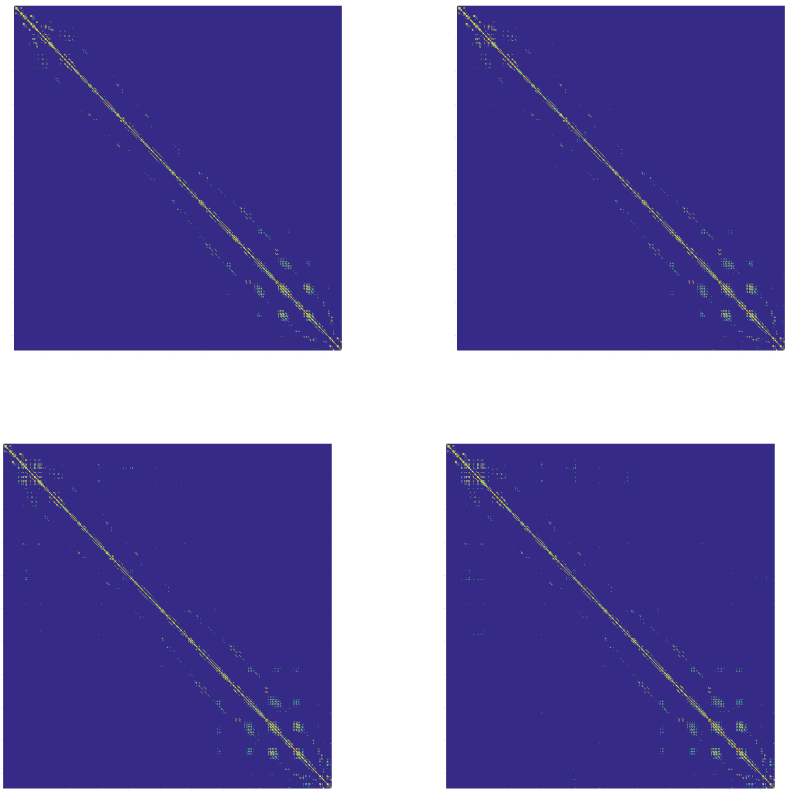


Figure 4.13: Similarity matrices by with different α values, α=1.7 (Top, Left), 2.2 (Top, Right), 3 (Bottom, Left), 4.5 (Bottom, Right)

Number of cluster (K)	Fisher's discriminant	Silhouette	Dice's coefficient
5	0.1001	1.0000	NA
10	0.0940	0.9893	NA
20	0.0989	0.9790	NA
30	0.0894	0.9702	NA

Table 4.4: Silhouette and Fisher's discriminant by K

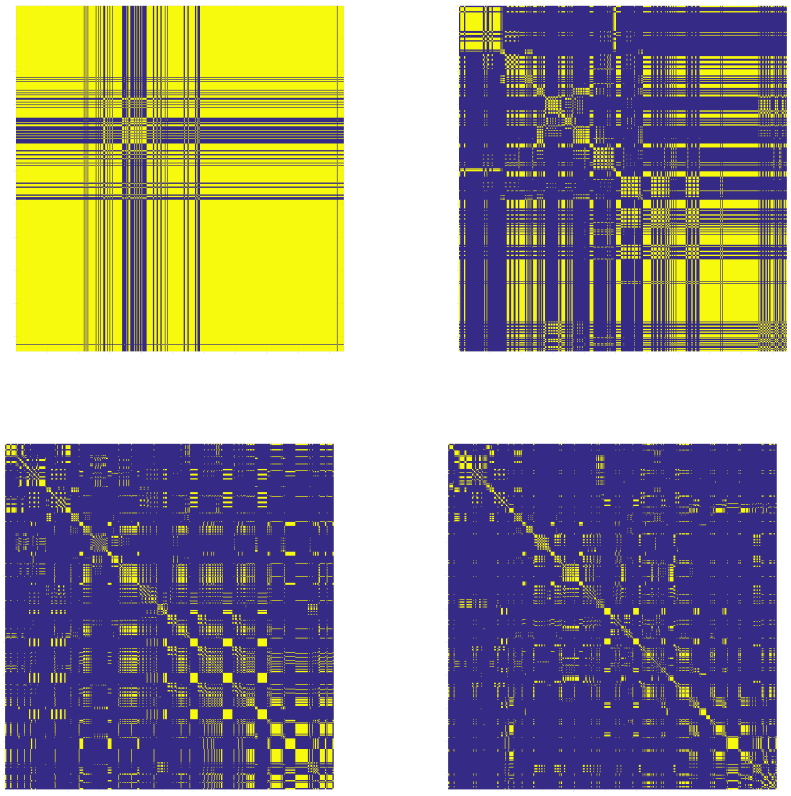


Figure 4.14: Clustering result by different number of clusters K, K=5 (Top, Left), 10 (Top, Right), 20 (Bottom, Left), 30 (Bottom, Right)

Increasing the number of clusters does not guarantee good clustering. There seems to be an optimal number of clusters for each of the brain area. For occipital robe of the right hemisphere, the optimal number of clusters K seems to be 4. However, the number of cluster K really depends on the choice of area.

To see how the spectral clustering works, we applied Yu and Shi (2003) algorithm to several other areas with parameter $K = 10, \alpha = 1.7, \delta = 0.001$.

Name of areas from AAL atlas	Fisher's discriminant	Silhouette	Dice's coefficient
Postcentral gyrus (L)	0.1461	0.0014	NA
Hippocampus (R)	0.0560	0.9915	NA
Cerebellum Crust 1 (L)	0.1247	0.9955	NA
Thalamus (R)	0.0573	0.9818	NA

Table 4.5: Results of clustering applied to several regions,
(L): Left Hemisphere, (R): Right Hemisphere

4.3.3 Multiple Subjects Analysis

For the multiple subjects analysis, there are two ways we discussed in the Chapter 2. We will apply Method 1 and Method 2 from the Chapter 2 to the identical datasets and will see which performs better in terms of Fisher's discriminant, silhouette, and Dice's coefficient. Just for our convenience we

chose single area, Postcentral gyrus from left hemisphere.

4.3.3.1 The Method 1

The first approach of handing multiple subjects is following. Once we obtain W^j, we take the average of W^j. Then we apply the spectral clustering algorithm. Here is the summary of the resulting clustering assuming that we want to have 10 clusters.

If we choose $\delta = 0.001$, $\alpha = 1.7$ and choose correlations greater than equal to 0.5 as defined in 2.3, then we obtain following average similarity matrix.

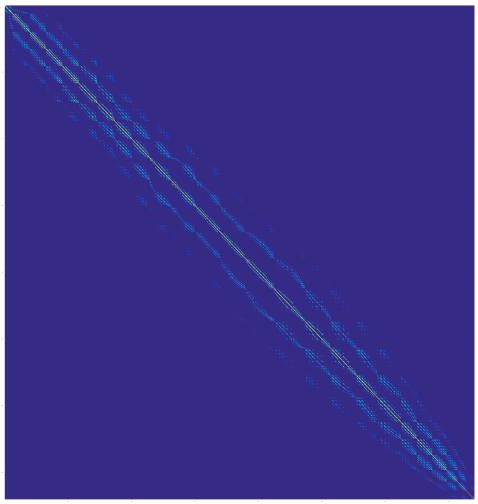


Figure 4.15: Plots of averaged similarity matrix

And we can apply the spectral clustering algorithm by Yu and Shi (2003). Then the resulting clusters are shown in the Figure 4.16.

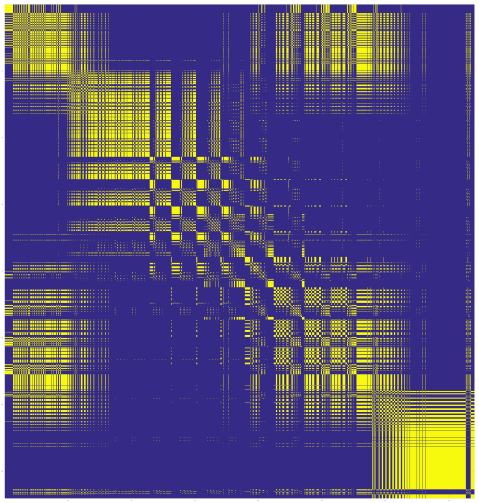


Figure 4.16: Clustering results

The Fisher's discriminant is 0.149 and silhouette is 0.8985. Also averaged Dice's coefficient is 0.601. Following three Figure 4.17 to 4.19 are resulting clusters mapped with brain. Size of clusters are following: 257 1 0 246 156 3 307 78 230 69.

Figure 4.17: Clustering results, sagittal

Figure 4.18: Clustering results, coronal

Figure 4.19: Clustering results, axial

By taking average similarity matrix of multiple subjects and perform spectral clustering, we have obtained Fisher's discriminant = 0.149 and SI = 0.8985. After the method 2 is applied we will compare two methods to see which method provides better spectral clustering results in terms of Fisher's discriminant and SI.

4.3.3.2 The Method 2

The second approach of handling multiple subjects is following. Once we obtain W^j, we applied the spectral clustering to each of W^j. Then we get the adjacency matrices A^j from each clustering results for J subjects. Then we take the average of A^j. Since the averaged adjacency matrix does not have binary entries any more, we apply the spectral clustering once more to

cluster averaged adjacency matrix. This will give a final clustering for multiple subjects analysis.

Suppose that we want to have 10 clusters. If we choose $\delta = 0.001$ and $\alpha = 1.7$ and choose correlations greater than equal to 0.5 as defined in equation 2.3, then we obtain J similarity matrices and Figure 4.20 is a plot of one of them as an example.

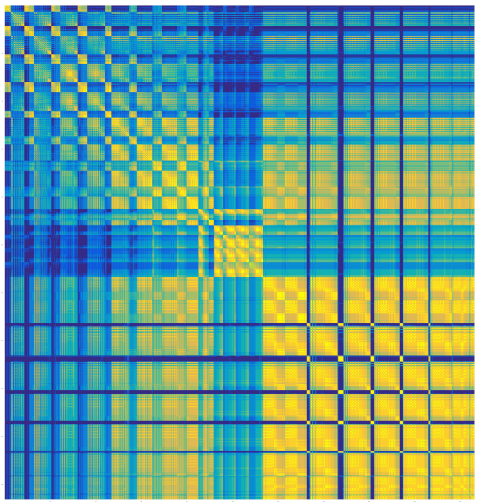


Figure 4.20: Plots of correlation defined as 2.3

We repeats this process for J times, so that we can obtain J similarity matrices. Then we apply the spectral clustering algorithm by Yu and Shi (2003) to each of J similarity matrices. Then we compute the average adjacency matrix from J adjacency matrices. The we apply second time of clustering to get the final adjacency matrix for multiple subjects analysis. Resulting clusters are shown in Figure 4.21 .

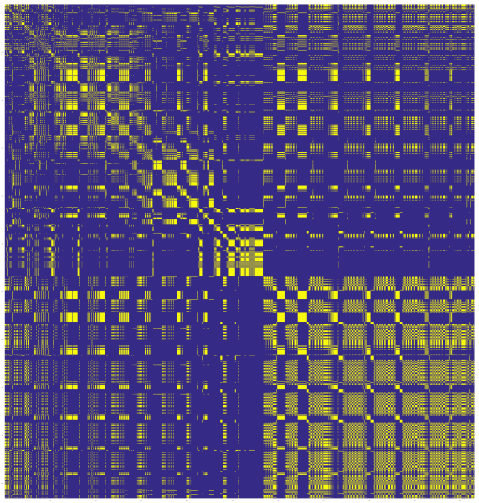


Figure 4.21: Plots of resulting clusters

Its Fisher's discriminant is 1.101 and silhouette is 0.676. Also averaged Dice's coefficient is 0.551. Size of clusters is following: 200 341 146 265 234 2 0 4 132 23. Following three Figure 4.17 to 4.19 are resulting clusters mapped with brain.

Figure 4.22: Clustering results, sagittal

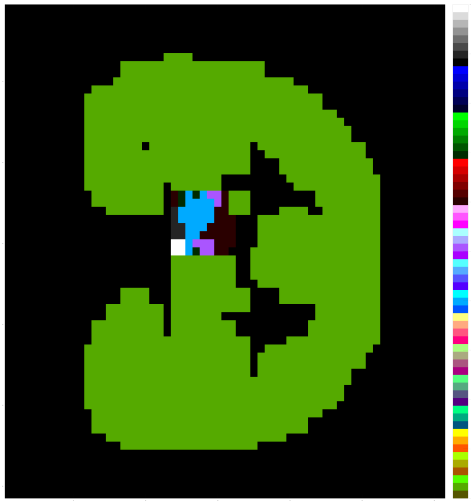


Figure 4.23: Clustering results, coronal

Figure 4.24: Clustering results, axial

For Postcentral gyrus area in the left hemisphere, method 1 performed better based on Fisher's discriminant and silhouette. However, this results may change by selecting different regions of brain and by selecting different numbers of clustering. From what we observed, there seem to be a optial choice of K, the number of clusters, for each regions of brain area although we are not displaying details in this paper. Therefore, we pefromed the same comparison using different areas of brain fixing the number of cluster to be eqal to 10, $K = 10$.

The table 4.6 is the result of clustering using Method 1 for several different regions and the table 4.7 is the one using Method 2.

Name of areas from AAL atlas	Fisher's discriminant	Silhouette	Dice's coefficient
Postcentral gyrus (L)	3.371	0.938	0.837
Hippocampus (R)	5.537	0.885	0.855
Cerebellum Crust 1 (L)	2.946	0.916	0.890
Thalamus (R)	3.507	0.877	0.834

Table 4.6: Results of clustering for several regions by Method 1

Name of areas from AAL atlas	Fisher's discriminant	Silhouette	Dice's coefficient
Postcentral gyrus (L)	0.412	0.073	0.953
Hippocampus (R)	0.914	0.162	0.852
Cerebellum Crust 1 (L)	1.614	0.042	0.863
Thalamus (R)	2.177	0.142	0.842

Table 4.7: Results of clustering for several regions by Method 2

Comparing two approaches for multiple subjects analysis, the method 1 yielded a better clustering in most case than the method 2, in terms of three criteria. However, this can be changed if we use different variations of spectral clustering algorithm.

4.4 Summary of Results

In previous section in Chapter 4, we explored two simulated datasets to see the effect of sized of blurred areas. As the β increases, the size of blurred areas grows larger. Until the $beta = 33$ that makes the size of blurred area to be the half size of clusters, spectral clustering algorithm still gave clear algorithm. Form this experiment, we observed that the spectral clustering algorithm can fail if the datasets dont have large enough size of cluster.

In the real data analysis, equations (2.2) and (2.3) define a similarity between two voxels. In equations (2.2) and (2.3), there are three parameters $delta$, α, and η. The α decides how many neighborhood voxels are included in creating a similarity matrix, δ decides what is the threshold for meaningful correlation. η is a very small positive number to ensure convergence of spectral clustering. Reason why we do not want to have a large η is that the spectral clustering can fail if η is too big. Recall that the algorithm failed in simulated datasets when there are larger blurred areas. If η is too big, then correlations between voxels can be close to or even smaller than η because fMRI signals can be very weak. The best case is to have η is zero, but with $\eta = 0$ we violate the assumptions of similarity matrix that we need to prove consistency.

4.5 Further Research

There are more to explore in data analysis. First, we can see how clustering is changed as we choose different parameters and what is the scientific meaning of different parameters. We can choose different numbers for each parameter and compare clustering results.

The second possible further research is to choose different similarity functions. We can choose many different similarity functions and compare clustering results. Since many different scientific areas are using spectral clustering, we can choose similarity functions from literatures and modify for fMRI data structure. However, to ensure the consistency of spectral clustering, we need to examine if similarly functions satisfy the assumption.

The third feasible research is to see if spectral clustering can be used as diagnosis tool or classification tool. For example, patients with mental diseases can have different connectivity between voxels and having different connectives can yield different results of clustering. Therefore, using clustering results, we may be able to determine which group a patient belongs to between disease and healthy groups.

Chapter 5: Conclusion

Through this thesis, we provided mathematical justification for consistency of spectral clustering. We used the results from von Luxburg (2008) and extended her proof to dependent data. This extension is a useful contribution because in many cases the nature of data set cannot satisfy independent assumption that von Luxburg (2008) assumed. For example, when we apply spectral clustering to a two-dimensional (2D) photo, we cannot assume that all data from each pixel are independent because characteristics of data such as brightness and shades from a pixel are affected by neighborhood data. Therefore, mathematical examination under dependent data assumptions are our contribution in spectral clustering.

We started exploratory data analysis that can lead further researches. Using different similarity functions and choosing different parameters in a similarity functions can be the next research. We can also extend research to see if spectral clustering can be used as an aid of diagnosis tool. Even though what has been done in this thesis was explanatory, the work in this thesis connected mathematical justification and real data analysis that can lead to further research.

Appendix A: Propositions from von Luxburg (2008)

Here we want to provide additional propositions from von Luxburg (2008).

A.1 Relations between the spectra of the operators

The main point about all the constructions above is that they enable us to transfer the problem of convergence of the Laplacian matrices to the problem of convergence of a sequence of operators on $C(\chi)$. Now we want to establish the connections between the spectra of operators defined in Chapter 3.

Proposition 10. (Spectrum of U'_N, Propostion 9 from von Luxburg (2008))

(1) If $f \in C(\chi)$ is an eigenfunction of U'_N with eigenvalue λ, then the vector $v = \rho_N f \in \mathbb{R}^N$ is an eigenvector of the matrix L'_N with eigenvalue λ.

(2) Let $\lambda \neq 1$ be an eigenvalue of U'_N with eigenfunction $f \in C(\chi)$, and $v_i := (v_1, \ldots, v_N) := \rho_N f \in \mathbb{R}^N$. Then f is of the form

$$f(x) = \frac{1/n \sum_{i'} w(x, X_{i'}) v_{i'}}{1 - \lambda}. \tag{A.1}$$

(3) If v is an eigenvector of the matrix L'_N with eigenvalue $\lambda \neq 1$, then f defined in by (A.1) is an eigenfunction of U'_N with eigenvalue λ.

This proposition establishes a one-to-one correspondence between the eigenvalues and eigenvectors of L'_N and U'_N , provided that they satisfy $\lambda \neq 1$. The condition $\lambda \neq 1$ needed to ensure that the denominator of equation (A.1) does not vanish. As a side remark, note that the set $\{\mathbf{1}\}$ is the essential spectrum of U'_N. Thus, the condition $\lambda \neq 1$ can also be written as $\lambda \notin \sigma_{ess}(U'_N)$, which will be analogous to the condition on the eigenvalues in the unnormalized case. This condition ensures that λ is isolated in the spectrum.

A.2 Convergence in the unnormalized case

We can apply similar approach to prove the convergence for unnormalized Laplacian L_N. The first step is relating the eigenvectors of $\frac{1}{N}L_N$ to those of U_N. Next step is to prove that U_N converges to U compactly. By considering the multiplication operator part M_{d_N} and the integral operator part S_N of U_N separately. It will turn out that the multiplication operator M_{d_N} converge to M_d in operator norm, and the integral operators S_N converge to S collectively compactly.

Proposition 11. (Spectrum of U_N)

1.The spectrum of U_N consists of the compact integral $rg(d_N)$, plus eventually some isolated eigenvalues with finite multiplicity. The same holds for U and $rg(d_N)$.

2.If $f \in C(\chi)$ is an eigenfunction of U_N with arbitrary eigenvalue λ, then the vector $v \in \mathbb{R}^n$ with $v_i = f(X_i)$ is an eigenvector of the matrix $\frac{1}{N}L_N$ with

eigenvalue λ.

3. Let $\lambda \notin rg(d_N)$ be an eigenvalue of the matrix U_N with eigenfunction $f \in C(\chi)$, and $v_j = f(X_i)$. Then f is the form

$$f(x) = \frac{\frac{1}{N}\sum_j w(x, X_j)v_j}{d_N(x) - \lambda}. \tag{A.2}$$

4.If v is an eigenvector of the matrix $\frac{1}{N}L_N$ wth eigenvalue $\lambda \notin rg(d_N)$.

This proposition establishes a one-to-one correspondence between the eigenvalues and eigenvectors of $\frac{1}{N}L_N$ and U_N, provided they satisfy $\lambda \notin rg(d_N)$. The condition $\lambda \notin rg(d_N)$ is needed to ensure that the denominator of equation (A.2) does not equal 0.

Proposition 12. $U_N \xrightarrow{c} U$.

Proof. We consider the multiplication and integral operator parts of U_N separately. Similarly to Proposition 6, we can prove that the integral operators S_N converge collectively compactly to S in probability, and, as a consequence, also $S_N \xrightarrow{p} S$. For the multiplication operators, we have operator norm convergence

$$\|M_{d_N} - M_d\| = \sup_{\|f\|_\infty \leq 1} \|d_N f - df\|_\infty \leq \|d_N - d\|_\infty \xrightarrow{p} 0.$$

by the Glivenko-Cantelli Proposition 3. As operator norm convergence implies compact convergence, we also have $M_{d_N} \xrightarrow{p} M_d$. Finally, it is easy to see that the sum of two compactly converging operators also converges compactly.□

Bibliography

[1] Ashby, F.G. (2011), *Statistical analysis of fMRI data*, The MIT Press.

[2] Chatelin, F. (1983), *Spectral Approximation of Linear Operators*, Academic Press, New York.

[3] Chung, F. (1997), *Spectral Graph Theory*, Conference Board of the Mathematical Sciences, Washington.

[4] Craddock, R et al. (2012), A whole brain fMRI atlas generated via spatially constrained spectral clustering, Hum Brain Mapp. **33**, 1914-1928.

[5] Cressie, N. (1993), *Statistics for Spatial Data*, New York, Wiley.

[6] Kang, J., Bowman, F. D., Mayberg, H., and Liu, H. (2016), A depression network of functionally connected regions discovered via multi-attribute canonical correlation graphs, NeuroImage, **141**, 431-441.

[7] Kessler, R.C. et al. (2003), The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R), Jama **289**, 3095-3105.

[8] Mohar, B (1991), The Laplacian spectrum of graphs, Graph Theory, Combinatorics, and Applications **2**, 871-898.

[9] Poldrack, R. et al. (2012), *Handbook of functional MRI data analysis*, Cambridge University Press.

[10] Reed, M. and Simon, B. (1980), *Functional Analysis I*, Academic Press, New York.

[11] Shi, J. and Malik, J. (2000), Normalized Cuts and Image Segmentation, IEEE Trans. Pattern Analysis and Machine Intelligence, **22**, 888-905.

[12] Somashekara, M.T. and Manjunatha D. (2014), Performance evaluation of spectral clustering algorithm using various clustering validity indices, International Journal of Electronics Communication and Computer Engineering **5**.

[13] Stelzer, J. et al. (2014), Deficient approaches to human neuroimaging, Front. Hum. Neurosci., **1**

[14] Thirion, B et al. (2014), Which fMRI clustering gives good brain parcellations?, Front. Neurosci.

[15] Van der Vaart (1998), *Asymptotic Statistics*, Cambridge University Press, New York.

[16] Van der Vaart, A.W. and Wellner, J.A. (1996), *Weak Convergence and Empirical Processes*, Springer, New York.

[17] Verma, D. and Meila, M. (2005), A comparison of spectral clustering algorithms, Technical Report, Department of CSE University of Washington Seattle.

[18] Von Luxburg, U. (2007), A tutorial on spectral clustering, Statistics and Computing **17**.

[19] Von Luxburg, U.G et al. (2008), Consistency of spectral clustering, The Annals of Statistics, **36**, 555586.

[20] Yu, S. and Shi, J. (2003), Multiclass Spectral Clustering, Proc. IEEE Int'l Conf., **1**, 313-319.