Universitat de Girona

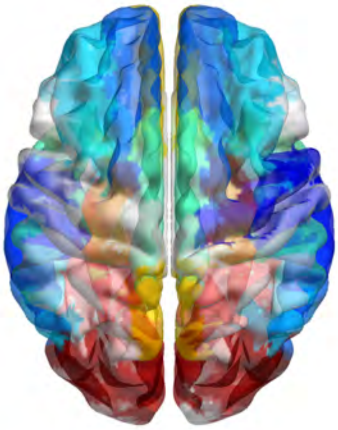
STUDY OF BRAIN COMPLEXITY USING INFORMATION THEORY TOOLS

Ester Bonmatí Coll

Universitat de Girona

DOCTORAL THESIS

Study of brain complexity using information theory tools



Ester BONMATÍ COLL

2016

Universitat de Girona

DOCTORAL THESIS

Study of brain complexity using information theory tools

Author:
Ester BONMATÍ COLL

2015

Doctoral Programme in Technology

Advisors:
Dr. Imma BOADA OLIVERAS
Dr. Anton BARDERA REIG

This manuscript has been presented to opt for the doctoral degree from the University of Girona

Science is a way of life.
Science is a perspective.
Science is the process that takes us from confusion
to understanding in a manner that’s precise,
predictive and reliable - a transformation,
for those lucky enough to experience it,
that is empowering and emotional.

Brian Greene.

Agraïments

Era 2006 quan l'Imma es va oferir per dirigir el meu primer projecte final de carrera. Ella és qui em va descobrir el món de la imatge mèdica, un món fascinant a on ha sabut transmetre i encomanar les ganes de fer investigació. Primer, amb el projecte de visualització i clustering de DTI. Més tard, amb el tumor glial, Starviewer, la màster tesis, que va ser l'inici de l'estudi de complexitat del cervell, i finalment, gairebé 10 anys més tard, amb aquesta tesis. Imma, moltes gràcies per totes les hores de treball compartides, els caps de setmana esgotadors, i pels més de 1100 emails intercanviats durant tots aquests anys. Moltes gràcies.

Aquest treball no hagués estat possible sense l'Anton, també director d'aquesta tesi. Ell és el meu referent en teoria de la informació i la màxima inspiració en l'elaboració dels mètodes. L'Imma i l'Anton, sempre amb munt munt d'idees, han esdevingut una combinació excel·lent per aquesta tesis. Moltes gràcies a tots dos. Espero que la finalització d'aquesta tesis no sigui només un tancament d'etapa i sinó que puguem seguir col.laborant durant molt de temps.

Vull agraïr també a en Miquel Feixas, per la seva contribució als articles, que han ajudat a millorar la qualitat de les publicacions. A la gent del despatx: Roger, Ferran, Màrius, Xavi, Marc R, als companys d'Starviewer, al grup de DTI, als companys de les estones de cafè: Yago, David, Pau, Nacho, sobretot a en Miquel, a en Joan i a en Suy per enviar-me a London sense saber-ho. Agraïr també a tota la resta de gent del Departament d'Informàtica i Matemàtica Aplicada i Estadística i del PIV amb què he coincidit un moment o altre i que han ajudat a tirar endavant aquest treball.

I would like to thank Stefan for his proof reading and all the discussions that have been a great inspiration for this thesis and the related publications. To Barbara, for all the support and the best company ever, the tiring weekends and, of course, the food! On the other hand, I thank to Dean, my current supervisor, who has given me the flexibility to combine the thesis with my current research and has been encouraging me to finish this work. I also thank to my mates at UCL, specially to: Rachael, Yipeng, Rene, Maria, Matt, Martin, Yi, John and obviously the Nespresso coffee machine.

Per acabar, agraïr als de casa i a la família, per la paciència que han tingut i tot el suport i facilitats que m'han donat durant tots aquests anys. A tots els amics, en especial a la Laia, per donar-me sempre ànims i per ser-hi sempre. Finalment, agraïr a tots aquells que no menciono però que d'una forma o altra han contribuït en la realització d'aquesta tesis. A tots ells, gràcies. Thank you all.

Acknowledgements

The work in this thesis has been supported by the BR grant of the University of Girona, by grant number TIN2013-47276-C6-1-R of the Spanish Government and by grant number 2014-SGR-1232 of the Catalan Government.

Data were provided, in part, by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Brain structural data is the courtesy of Prof. Jean-Philippe Thiran, Dr. Alessandro Daducci and Dr. Alessandra Griffa, Swiss Federal Institute of Technology Lausanne, Signal Processing Laboratory, Switzerland.

List of publications

Publications that support the contents of this thesis:

- Ester Bonmati, Anton Bardera, Imma Boada, Miquel Feixas and Mateu Sbert. *Hierarchical clustering based on the information bottleneck method using a control process*. Pattern Analysis and Applications, vol. 18, no.3, pages 619–637, 10.1007/s10044-015-0467-1, 2015.

- Ester Bonmati, Anton Bardera, Miquel Feixas, Imma Boada. *Novel brain complexity measures based on information theory*. Medical Image Analysis. Submitted.

- Ester Bonmati, Anton Bardera, Imma Boada. *Brain parcellation based on information theory*. Computer Methods and Programs in Biomedicine. Submitted.

List of Figures

List of Tables

Contents

Study of brain complexity using information theory tools

The human brain is a complex network that shares and processes information by using the structural paths between areas in order to perform a function. Magnetic resonance imaging techniques allow the in vivo reconstruction of the structural paths by using diffusion MRI and the mapping of the active areas by using functional MRI. The connectome models the brain as a graph where nodes correspond to brain regions and edges to structural or functional connections. In this thesis, we investigate and provide new methods to study the brain complexity and improve the understanding of the brain functioning by using information theory.

Firstly, we focus on brain parcellation, which is a key step to perform brain studies since determines the regions to be analyzed. We interpret a brain function as a stochastic process where neural impulses are modeled as a random walk by using the connectivity matrix. Using this interpretation, we first present a new hierarchical clustering method based on the information bottleneck. We describe two versions of the method, the agglomerative approach that merges elements with a minimum loss of mutual information, and the divisive approach that divides the elements with a higher gain of mutual information. The agglomerative version of the method is employed and deeply evaluated to parcellate the brain. We show that the clustered networks preserve the structure and properties of the original network but having higher mutual information.

Secondly, we focus on the definition of measures to characterize the complexity of the brain networks. We propose new global and local measures. Global measures provide quantitative values for the whole-brain network and include the entropy, the mutual information, and the erasure mutual information, which is a new measure defined by extending the mutual information. Local measures are based on different decompositions of the global measures and include the entropic surprise, the mutual surprise, the mutual predictability and the erasure surprise. These measures show local properties of the brain regions, such as the uncertainty associated to the node, or the uniqueness of the path that the node belongs to.

Finally, the consistency of the results across healthy subjects using functional or structural connectivity data, demonstrates the flexibility and robustness of the proposed methods.

Estudi de la complexitat cerebral utilitzant eines de teoria de la informació

El cervell humà és una xarxa complexa que comparteix i processa la informació mitjançant els camins estructurals per tal de realitzar una funció. La ressonància magnètica és una tècnica no invavisa que permet obtenir informació en viu de l'estructura i la composició del cos en forma d'imatges. La reconstrucció en viu dels camins estructurals es pot obtenir mitjançant l'ús de la ressonància magnètica de difusió, i el mapeig de les àrees funcionalment actives, es pot obtenir a partir de l'ús de la ressonància magnètica funcional. El connectoma és una representació del cervell en forma de graf, on els nodes corresponen a regions del cervell i les arestes a connexions estructurals o funcionals. En aquesta tesi, s'investiga i es proporcionen nous mètodes per estudiar la complexitat del cervell i millorar la comprensió del seu funcionament mitjançant l'ús de la teoria de la informació.

En primer lloc, ens centrem en mètodes de parcel.lació del cervell, el qual és un pas clau per realitzar estudis de complexitat ja que determina les regions a analitzar. En aquest treball, interpretem la xarxa cerebral com un procés estocàstic, on els impulsos neuronals es modelen com un camí aleatori. Fent servir aquesta interpretació, primer presentem un nou mètode d'agrupació jeràrquica basada en l'algorisme del coll d'ampolla (bottleneck). Proposem les dues versions del mètode: l'aglomeratiu, que agrupa els elements amb una pèrdua mínima d'informació mútua, i el divisiu, que divideix els elements de tal forma que es proporciona un major guany d'informació mútua. El mètode aglomeratiu és utilitzat i profundament avaluat com a nou mètode de parcel.lació del cervell. Es demostra que les xarxes obtingudes conserven l'estructura i les propietats de la xarxa original però amb una major informació mútua.

En segon lloc, ens centrem en la definició de mesures per a caracteritzar la complexitat de les xarxes cerebrals. Proposem noves mesures globals i locals. Les mesures globals proporcionen valors quantitatius per a la xarxa de tot el cervell i inclouen l'entropia (*entropy*), la informació mútua (*mutual information*), i la informació mútua d'esborrat (*erasure mutual information*), que és una nova mesura definida mitjançant l'extensió de la informació mútua. Les mesures locals es basen en diferents descomposicions de les mesures globals i inclouen la sorpresa entròpica (*entropic surprise*), la sorpresa mútua (*mutual surprise*), la predictabiliat mútua (*mutual predictability*) i la sorpresa d'esborrat (*erasure surprise*). Aquestes mesures mostren propietats locals de les regions del cervell, com ara la incertesa associada al node, o la singularitat del camí al qual pertany el node.

Finalment, la consistència dels resultats entre els subjectes sans a partir de dades de connectivitat funcional o estructural, demostra la flexibilitat i robustesa dels mètodes proposats.

Estudio de la complejidad cerebral utilizando herramientas de teoría de la información

El cerebro humano es una compleja red neuronal que comparte y procesa la información mediante el uso de los caminos estructurales con el fin de realizar una función. La resonancia magnética es una técnica no invasiva que permite obtener información en vivo de la estructura y composición del cuerpo en forma de imágenes. La reconstrucción en vivo de los caminos estructurales se puede obtener mediante el uso de la resonancia magnética de difusión, y el mapeo de las áreas funcionalmente activas se puede obtener mediante el uso de la resonancia magnética funcional. La idea de conectoma consiste en modelar el cerebro como un grafo donde los nodos corresponden a regiones y las aristas a conexiones estructurales o funcionales. En esta tesis, se investiga y se proporcionan nuevos métodos para estudiar la complejidad del cerebro y mejorar la comprensión de su funcionamiento mediante el uso de la teoría de la información.

En primer lugar, nos centramos en métodos de parcelación del cerebro, paso clave para realizar estudios de complejidad ya que determina las regiones a analizar. Interpretamos la red cerebral como un proceso estocástico, donde los impulsos neuronales se modelan como un paseo aleatorio. Usando esta interpretación, primero presentamos un nuevo método de agrupación jerárquica basada en el método del bottleneck. Se describen dos versiones del método: el aglomerativo, que agrupa los elementos con una pérdida mínima de información mutua, y el divisivo, que divide los elementos de tal forma que se proporciona una mayor ganancia de información mutua. El método aglomerativo es usado y profundamente evaluado como nuevo método de parcelación del cerebro. Demostramos que las redes resultantes conservan la estructura y las propiedades de la red original, pero con una mayor información mutua.

En segundo lugar, nos centramos en la definición de medidas para caracterizar la complejidad de las redes cerebrales. Proponemos nuevas medidas globales y locales. Las medidas globales proporcionan valores cuantitativos para la red de todo el cerebro e incluyen la entropía (*entropy*), la información mutua (*mutual information*), y la información mutua de borrado (*erasure mutual information*), que es una nueva medida definida mediante la extensión de la información mutua. Las medidas locales se basan en diferentes descomposiciones de las medidas globales e incluyen la sorpresa entrópica (*entropic surprise*), la sorpresa mutua (*mutual surprise*), la previsibilidad mutua (*mutual predictability*) y la sorpresa de borrado (*erasure surprise*). Estas medidas muestran propiedades locales de las regiones del cerebro, tales como la incertidumbre asociada al nodo, o la singularidad del camino al cual pertenece el nodo.

Por último, la consistencia de los resultados entre los sujetos sanos a partir de datos de conectividad funcional o estructural, demuestra la flexibilidad y robustez de los métodos propuestos.

CHAPTER 1

Introduction

Contents

1.1 Motivation

The human brain contains an extraordinary network of roughly one hundred billion of neurons capable to communicate and process information. These neurons form an organized and coordinated network where information is shared and transported using structural paths in order to perform specific functions. Current medical imaging techniques such as diffusion magnetic resonance imaging (dMRI) or functional magnetic resonance imaging (fMRI) are able to capture non-invasively the brain in vivo information required to reconstruct the structural paths as well as map the active areas of the brain. The brain connectome is the most popular approach to model the brain network by using a graph representation. In this graph, nodes correspond to brain regions and edges to structural or functional connections.

Several methods and measures have been proposed to characterize and describe properties of the brain network, however, the exact functioning of the system is still not fully understood. There are several issues that need further investigation. On the one hand, the first step to construct the brain network consists on defining the regions of interest. This procedure is usually done by a parcellation technique. Several parcellation methods have been proposed, which mainly differ in the number of regions or scale. The parcellation approach has to be chosen carefully according to the aim of the analysis, as the use of an inappropriate method may lead to erroneous conclusions. Although a large amount of unsupervised parcellation methods can be found in the literature, there are still some limitations to improve, such as the high computational cost or the dependency on a fixed number of regions. On the other hand, network measures are used to characterize brain information. Current brain network measures are able to associate disruptions with different diseases, but unfortunately, it is unknown which measures provide the best description of the brain network. Thus, novel measures are needed in order to better understand the brain functioning.

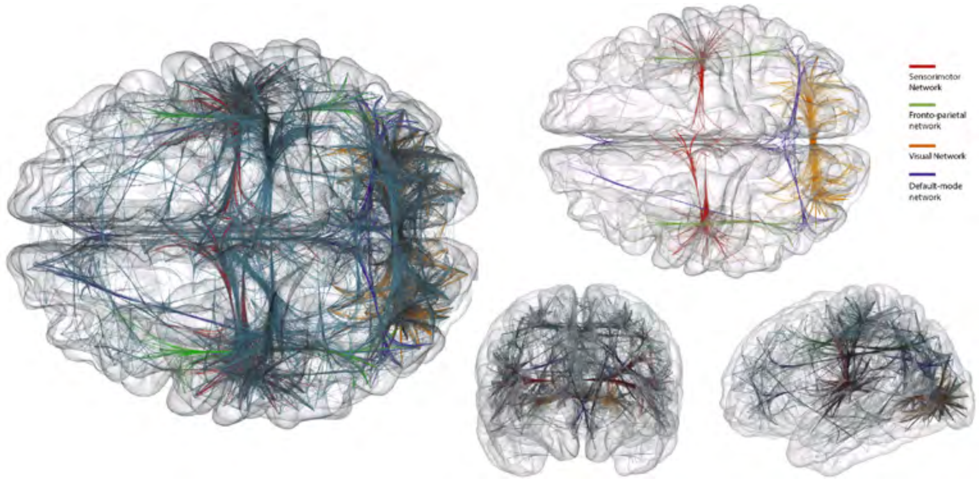


Figure 1.1: Representation of the human functional connectome [Böttger 2014]

In this thesis, motivated by the limitations of current brain parcellation techniques and the need of new brain network measures, brain complexity is investigated by applying information theory to structural and functional connectivity data. Information theory provides mathematical methods capable to quantify the uncertainty of the information in a system. Measures such as mutual information describe the dependence of the individual components, and the complexity of the system can be obtained by calculating the average of mutual information between the components and the rest of the system. Since the brain is a system of individual segregated components (areas), that integrates and shares the information, information theory measures can be used to characterize the complexity of the brain network.

1.2 Objectives

The aim of this thesis is to investigate and provide new methods to improve the understanding of the human brain complexity by using information theory. To achieve this aim two main focuses of research have been considered:

- **Brain parcellation based on clustering techniques**

  Clustering techniques organize elements into groups (or clusters) whose members are similar and dissimilar to elements belonging to other clusters. These techniques can be used to create a brain parcellation in order to define the brain regions to be studied. Since current techniques present some limitations, our purpose is to

  - Introduce a new clustering method based on information theory that overcomes some of the current limitations
  - Evaluate the proposed clustering method to parcellate functional and structural brain connectivity

  - Investigate the properties of the obtained parcellations and the robustness of the method across different subjects

- **Complex brain network measures**

  Information theory has been used previously to describe properties of complex systems successfully. However, some measures have never been applied to describe brain complexity. Because new models and measures are needed to better understand the brain functioning, our purpose is to

  - Evaluate the applicability of existing information theory measures to characterize the brain network complexity by using model networks at different scales and densities
  - Introduce new complexity measures to describe the brain network both at global and local level
  - Evaluate the proposed measures to model networks at different scales and densities.
  - Evaluate the consistency of the proposed measures across subjects by considering structural and functional networks from different patients at different scales
  - Compare the measures with well known standard measures to show new network properties that may help to improve the actual description of the human brain network

1.3 Thesis outline

This thesis is organized in six chapters. Following this introduction, the next five chapters are:

- Chapter 2: **Background and Previous Work**

  This chapter provides relevant background knowledge. First, a brief summary of the brain anatomy along with a description of the more important non-invasive magnetic resonance modalities is provided. In the second place, the main steps needed to create a brain network and measures that can be find in the literature are explained. Finally, a description of the main concepts and measures of information theory are introduced.

- Chapter 3: **Hierarchical Clustering Based on the Information Bottleneck**

  In this chapter, a new hierarchical clustering method based on the information bottleneck method is proposed. Two versions of the algorithm are presented, the agglomerative, that merges clusters, and the divisive, that subdivides the clusters. The method is tested by quantifying synthetic, photographic and medical images by grouping intensity bins of the image histograms. A comparison with the ground truth and the main clustering methods is provided.

- Chapter 4: **Brain Parcellation Based on Information Theory**

  In this chapter, the agglomerative clustering method presented in Chapter 3 is considered to parcellate the anatomical and functional areas of the brain at different scales. The description of the method adapted to the brain model is provided. The results based on synthetic model networks as well as human structural and functional connectivity data are presented.

- Chapter 5: **Complexity Measures Based on Information Theory**

  This chapter introduces novel brain complexity measures based on information theory. A brain function is interpreted as a stochastic process where neural impulses are modeled as a random walk. This new interpretation provides a solid theoretical framework from which we derive global and local measures. Global measures are used to quantify properties for the whole-brain network, while local measures quantify properties for the individual nodes. Experiments with synthetic model networks and human brain networks are presented in order to evaluate the performance of the measures.

- Chapter 6: **Conclusions**

  This chapter presents the conclusions of this work including a summary of the related publications and the main contributions. Finally, the future directions of the work are described.

CHAPTER 2

Background and Previous Work

Contents

The intersection of medical image computing and neuroscience is a cutting-edge emerging field and a relatively new discipline that studies the structure and the function of the nervous system by using the most advanced computer science and neuroimaging techniques. While neuroscience is focused on the study of the nervous system from the cell level to the whole structure and function, medical image computing develops computational and mathematical methods that enables to solve problems related to medical images and their clinical use. Additionally, information theory is a combination of mathematics and computer science that aims to quantify information regarding its representation. Since the brain is an organized massive network of neurons sharing information, information theory can be a suitable tool to study and characterize the brain functions.

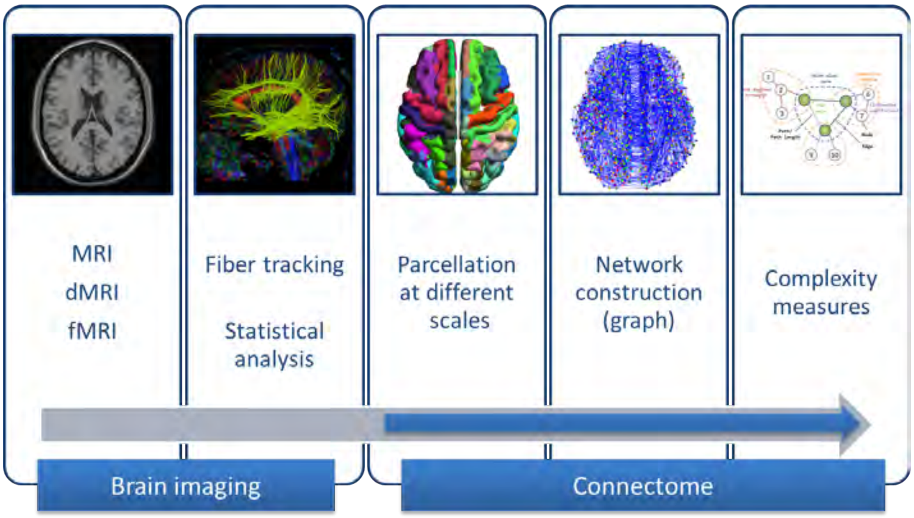


Figure 2.1: Diagram of the main aspects described in this chapter

This chapter provides relevant background knowledge required for the comprehension of this thesis. It has been structured as follows. Section 2.1 provides a brief description of the brain anatomy and the different non-invasive techniques that allow us to acquire in vivo functional and structural information. Section 2.2 provides information about the study of the brain complexity by constructing a brain network and describing the properties with complexity measures. Finally, Section 2.3 provides a description of the most relevant concepts and knowledge of information theory that are relevant for the understanding of the content in this thesis.

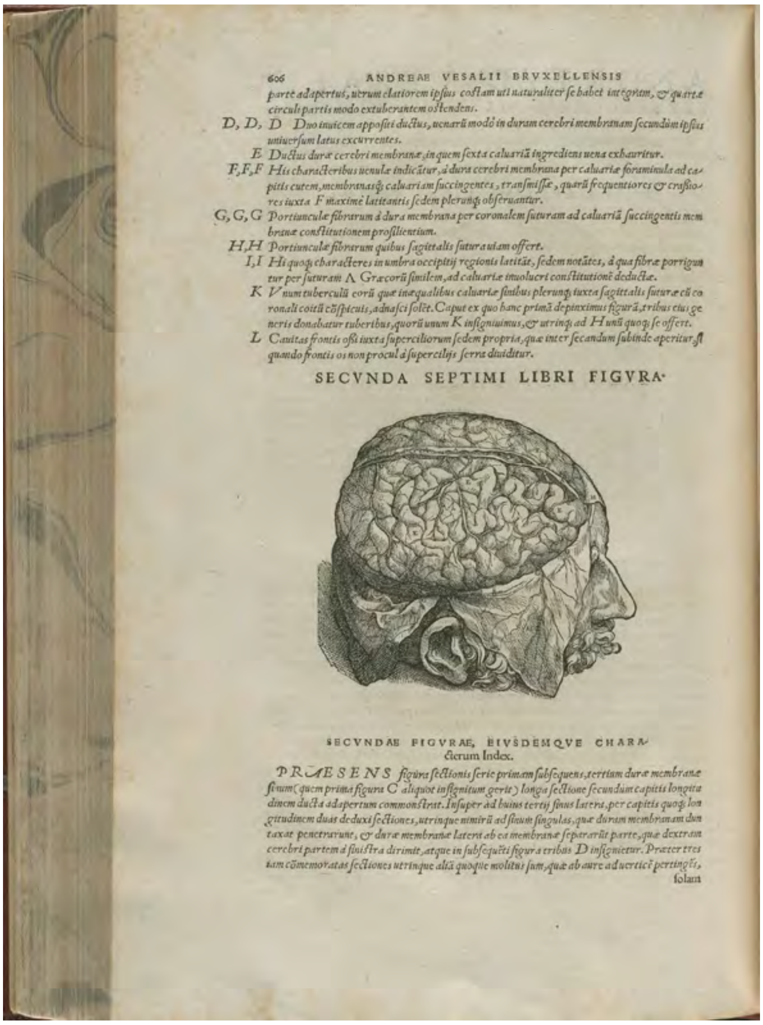
2.1 Brain imaging

Brain imaging allows the visualization of the internal structure of the brain. In this section, we review the main anatomic parts and functions of the brain as well as the different non-invasive in vivo imaging methodologies that have been considered in this thesis.

2.1.1 Brief overview of the human brain anatomy

The brain is the most complex organ in the human's body and, being part of the nervous system, is in charge of the central control over the other organs. First brain images were obtained in the Renaissance era [Papo 2014], where, at that time, the brain was represented just as a symmetrical pair of wrinkle walnut-like lobes connected to each other (see Figure 2.2).

Fortunately, thanks to science advances, more accurate and precise descriptions of brain anatomy have been provided. For instance, we know that the cerebrum is the superior and largest part of the brain, divided in two hemispheres connected by the



606 ANDREAE VESALII BRVXELLENSIS

parte adapertus, uerum elatiorem ipſius coſtam uti naturaliter ſe habet integram, & quartæ circuli partis modo extuberantem oſtendens.

D, D, D Duo inuicem appoſiti ductus, uenarũ modò in duram cerebri membranam ſecundùm ipſius uniuerſum latus excurrentes.

E Ductus duræ cerebri membranæ, in quem ſexta caluariã ingrediens uena exhauritur.

F, F, F His characteribus uenulæ indicãtur, à dura cerebri membrana per caluariæ foraminula ad capitis cutem, membranasq; caluariam ſuccingentes, tranſmiſſæ, quarũ frequentiores & craſſiores iuxta F maximè latitantis ſedem plerunq; obſeruantur.

G, G, G Portiunculæ fibrarum à dura membrana per coronalem ſuturam ad caluariã ſuccingentis membranæ conſtitutionem proſilientium.

H, H Portiunculæ fibrarum quibus ſagittalis ſutura uiam offert.

I, I Hi quoq; characteres in umbra occipitij regionis latitãt, ſedem notãtes, à qua fibræ porriguntur per ſuturam Λ Græcorũ ſimilem, ad caluariæ inuolucri conſtitutionẽ deductæ.

K Vnum tuberculũ eorũ quæ inæqualibus caluariæ ſinibus plerunq; iuxta ſagittalis ſuturæ cũ coronali coitũ cõſpicuis, adnaſci ſolẽt. Caput ex quo hanc primã depinximus figurã, tribus eius generis donabatur tuberibus, quorũ unum K inſigniuimus, & utrinq; ad H unũ quoq; ſe offert.

L Cauitas frontis oſsi iuxta ſuperciliorum ſedem propria, quæ inter ſecandum ſubinde aperitur, ſi quando frontis os non procul à ſupercilijs ſerra diuiditur.

SECVNDA SEPTIMI LIBRI FIGVRA.

SECVNDAE FIGVRAE, EIVSDEMQVE CHARActerum Index.

PRAESENS figura ſectionis ſerie primam ſubſequens, tertium duræ membranæ ſinum (quem prima figura C aliquot inſignitum gerit) longa ſectione ſecundùm capitis longitudinem ducta adapertum commonſtrat. Inſuper ad huius tertij ſinus latera, per capitis quoq; longitudinem duas deduxi ſectiones, utrinque nimirũ ad ſinum ſingulas, quæ duram membranam duntaxat penetrarunt, & duræ membranæ latera ab ea membranæ ſepararũt parte, quæ dextram cerebri partem à ſiniſtra dirimit, atque in ſubſequẽti figura tribus D inſignitur. Præter tres iam cõmemoratas ſectiones utrinque aliã quoque molitus ſum, quæ ab aure ad uerticẽ pertingẽs,

ſolam

Figure 2.2: Historical image from the Renaissance with one of the first representations of the brain [Versalius 1564] [https://www.nlm.nih.gov/exhibition/historicalanatomies/vesalius_home.html]

corpus callosum, and includes the cerebral cortex, the hippocampus and the basal ganglia. The cerebrum is in charge of the sensory processing, the olfactory system, the language, communication and the movement. The brainstem is the posterior part, underneath the cerebrum, that allows a bidirectional communication between the spinal cord and other parts of brain. Behind the brainstem, there is the cerebellum, which is in charge of the motor control.

The cerebral cortex, which is the outer thick layer of neural tissue (gray matter) of the cerebrum, can be divided in four main regions (lobes) (see Figure 2.3):

- Frontal lobe: Located at the front of the brain. It is in charge of actions such as project the consequences of current actions or distinguish between good and bad actions

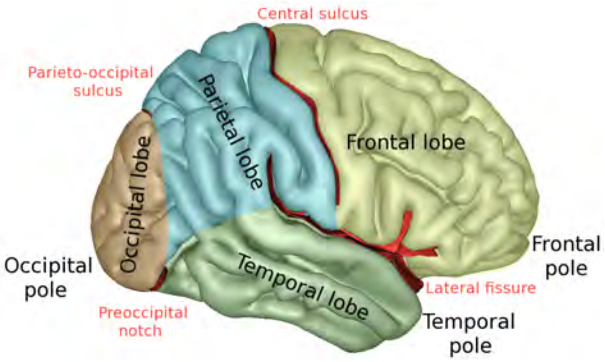


Figure 2.3: Representation of the four lobes of the cerebral cortex [https://commons.wikimedia.org/wiki/File:LobesCapts.png]

- Occipital lobe: Located at the back of the brain. It is the visual processing center
- Parietal lobe: Located between the temporal lobe and the occipital lobe. The main function is to integrate sensory information
- Temporal lobe: Located beneath the frontal and parietal lobes. It is involved in retaining visual memories, language comprehension and emotion association

An important anatomical segmentation of the brain was proposed by Korbinian Broadman, a german anatomist, in 1909 [Brodmann 1909]. The cortex was divided in 44 different areas based on the cytoarchitectural organization, which is still nowadays a popular method to segment the cortex based on the anatomy. Recently, it has been demonstrated that the original proposed anatomical partitions are also related with functional areas.

At a microscopic level, these areas contain glial cells, neurons and blood vessels. Neurons are cells able to send electrical and chemical signals (information) by using the axons (see Figure 2.4).

The brain can be studied at different scales as follows:

- *Micro-scale*: Study of the connections at the neuron level
- *Meso-scale*: Study of the connections of groups of neurons
- *Macro-scale*: Study of the connections between brain regions

In this work, we focus on anatomical and functional whole-brain connectivity at the macro-scale level, which is the appropriate level in order to study the complexity since we are interested in the connections between anatomical brain areas.

2.1.2 Magnetic resonance imaging

Magnetic Resonance Imaging (MRI) is a popular non-invasive technology to image the body. It uses magnetic fields and radio waves to produce high quality images of the

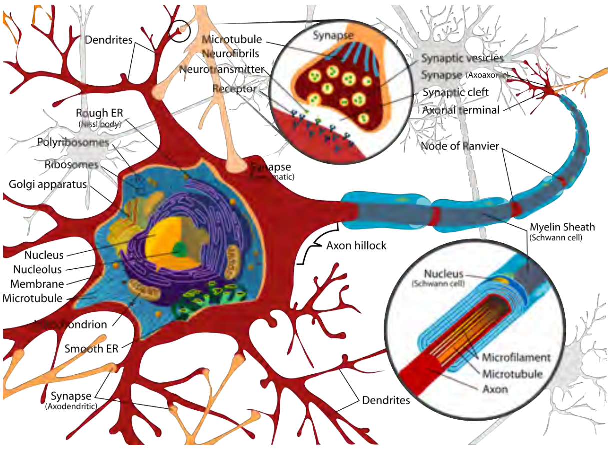
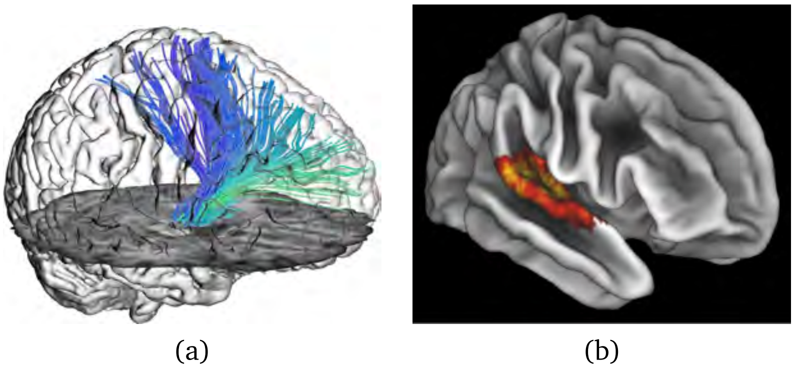


Figure 2.4: Complete neuron cell diagram [https://commons.wikimedia.org/wiki/File:Complete_neuron_cell_diagram_en.svg]


(a)

(b)

Figure 2.5: Example of fiber tracts obtained using diffusion MRI (a) [Berres 2012] and a brain activation using functional MRI (b) [Smith 2013]

body without the use of x-rays or radioactive tracers. The main advantage of using MRI is the good tissue contrast, that, depending on the mode, specific tissues can be seen better than others. For instance, *T1-weighted* images provide a good quality in visualizing normal anatomy, showing the fluid in dark color, while in *T2-weighted* images, the liquid is shown in light color and the white matter in dark color, which is useful for the diagnostic of some pathologies. There exist different MRI techniques such as diffusion MRI, angiography, spectroscopy, and functional MRI among others.

In this thesis, we focus on the *diffusion MRI* which allows the reconstruction of the structural fiber tracts (connections) across brain areas (see Figure 2.5 (a)), and the *functional MRI* which allows to find the temporal connections between activated areas (see Figure 2.5 (b)).

2.1.3 Diffusion MRI

Diffusion MRI (dMRI), also referred as diffusion tensor imaging (DTI), and the more recent variant diffusion spectrum imaging (DSI), are non-invasive magnetic resonance methods that allow the in vivo mapping of the white matter of the body. In contrast to MRI, dMRI delineates the axons within the white matter in different scanning directions to characterize the water diffusion in tissue, whereas gray matter remains equal in all directions. The different scanning directions (from six to hundreds) allow the creation of a second order symmetric positive *diffusion tensor* on a voxel level. By following the main direction of this tensor, it is possible to map the orientation of the fibers. The main challenge of this technique is the development of strategies to extract and visualize the acquired information in a comprehensive way.

The diffusion tensor D is described by a 3×3 symmetric matrix where D_{ij} is the diffusion coefficient measured in the ijth scanning direction.

$$D = \begin{pmatrix} D_{xx} & D_{xy} & D_{zx} \\ D_{yx} & D_{yy} & D_{xz} \\ D_{zx} & D_{zy} & D_{zz} \end{pmatrix} \tag{2.1}$$

The diagonalization of Equation 2.1 provides three eigenvalues, λ_1, λ_2, and λ_3 and three eigenvectors $\mathbf{e}_1, \mathbf{e}_2$ and $\mathbf{e}_3$ that define the directions of main, medium, and minimum diffusivity, respectively [Basser 1994].

Based on λ_1, λ_2, and λ_3, different measures to quantify the diffusion tensor properties have been proposed [Basser 1996, Westin 1997, Peled 1998, Conturo 1996]. We highlight the volume ratio (VR), the relative anisotropy (RA), and the fractional anisotropy (FA), which reduce the tensor information to a simple 1D scalar. This simplification allows the representation of the tensor using grey-scale maps. In addition, different geometrical diffusion measures, such as linear anisotropy (C_l), planar anisotropy (C_p), and spherical anisotropy (C_s) have been proposed.

An important property of these measures is that are scale and rotationally invariant. The scale invariant property ensures that are not affected by the scale of the diffusion magnitude. That is, they only deal with the shape of the diffusion tensor and hence are independent of the orientation of tissue structure and image scan plane. The most popular method to analyze these maps are the grey scale and RGB color-coded maps, which are generated by mapping the measure values as intensities (see Figure 2.6 (a) and (b)). Although the visualizations represent only a little part of the tensor information, the images can be easily interpreted enabling the identification of normal and pathological brain tissue. The main limitation, due to the matrix values reduction to a scalar value, is that the directional information of the tensor is lost.

To overcome the drawbacks of scalar visualizations, glyph based techniques have been introduced. With these techniques, the D tensor is described as a 3D ellipsoid (or cuboid) where the principal axes correspond to the eigenvector directions, and the size, correspond to the eigenvalues [Westin 1994]. According to the diffusion type described, the shape of the 3D ellipsoid can be divided into three basic cases: (i) linear, when the diffusion is only along one direction (Figure 2.7 (b)), i.e., $(\lambda_1 \gg \lambda_2) \approx \lambda_3$, (ii) planar,

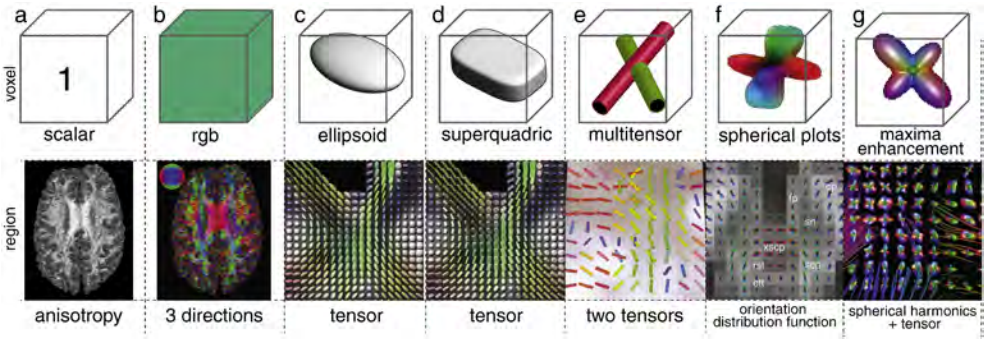
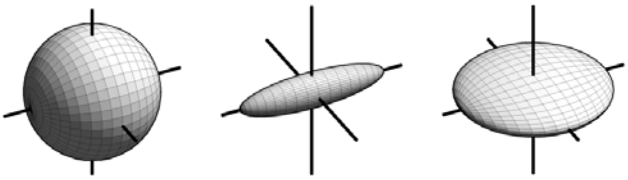


Figure 2.6: DTI tensor representations [Margulies 2013]



(a) (b) (c)

Figure 2.7: Diffusion tensor shapes: (a) spherical, (b) linear and (c) planar [Kindlmann 2004]

when the diffusion is restricted to the plane defined by the two equal eigenvalues (Figure 2.7 (c)), i.e., $(\lambda_1 \approx \lambda_2) \gg \lambda_3$, and (iii) spherical, when the diffusion is isotropic (Figure 2.7 (a)), i.e., $(\lambda_1 \approx \lambda_2) \approx \lambda_3$. Figure 2.6 (c) (d) and (e) show an example of this method applied to a dMRI image. These techniques are good in illustrating directional information at a voxel level, but fail on representing the connection between neighboring voxels.

With the introduction of the DSI, which is more sensitive in diffusion directions caused by crossing fibers, is possible to map more accurately the fiber trajectories [Wedeen 2008]. This technique has motivated the introduction of more advanced 2D visualizations methods such as spherical plots (orientation distribution function) or maxima enhancement (spherical harmonics) (see Figure 2.6 (f) and (g)).

All these representations are suitable for representing the information contained in a single slice. However, we are interested in the connections between areas, where these strategies fail. To overcome these limitations, techniques to reproduce the brain fibers have been proposed. These techniques are known as fiber tracking techniques (tractography) and are presented in the next section.

2.1.4 Tractography

Fiber tracking, or *tractography*, aims to reconstruct in vivo structural fiber tracts of the human brain (white matter pathways) from dMRI sequences. It is an emerging focus

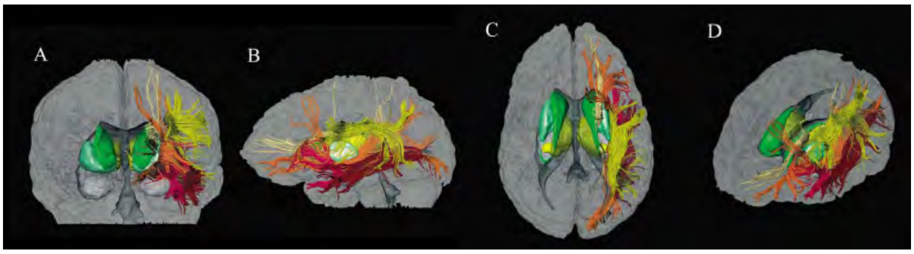


Figure 2.8: Associative fibers [Mori 2001]

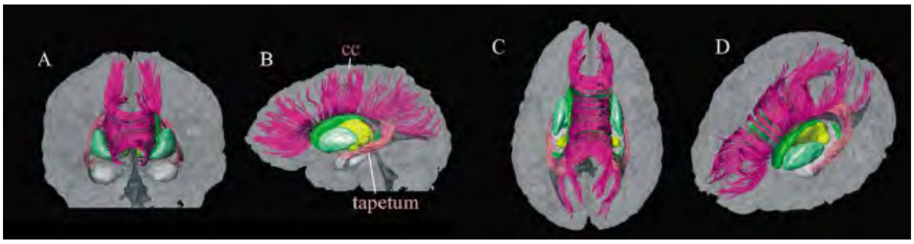


Figure 2.9: Commissural fibers [Mori 2001]

of research and the basis of current brain complexity studies, since it is the only that provides structural information that allows the definition of the *structural connectivity*. The fiber tracts do not show the actual axons, it is just a representation of the estimated trajectories.

The human brain fibers can be divided in three main groups as follows:

- *Associative fibers*: Connect brain regions from the same hemisphere and can be divided in two groups: *short association fibers* (connect adjacent areas), and *long association fibers* (connect more distant and non-adjacent areas) (see Figure 2.8)
- *Commissural fibers*: Connect the two hemispheres of the brain (see Figure 2.9)
- *Projective fibers*: Connect the cortex with lower parts of the brain or with the spinal cord (see Figure 2.10)

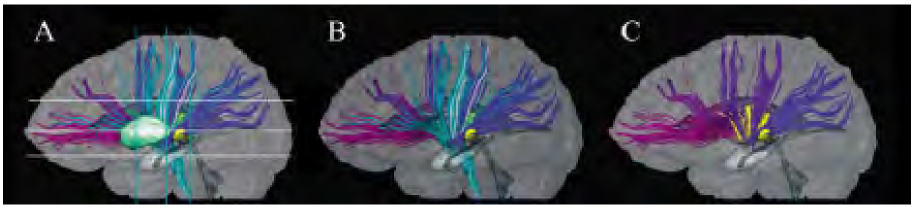


Figure 2.10: Projective fibers [Mori 2001]

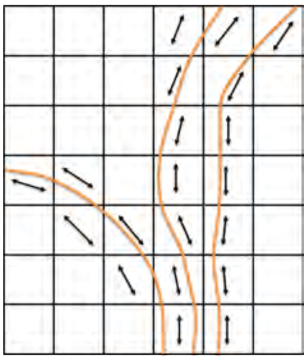


Figure 2.11: Steps on the calculation of the fibers using streamline tractography

To calculate the tracts, the directional information encoded in the diffusion tensor is used to infer patterns of continuity in the diffusion tensor field. It is assumed that the eigenvectors of the diffusion tensor give us a good estimation of the fiber orientation. Thus, the fiber path is created by following the main direction of the tensor. The main steps of a tracking algorithm include:

1. *Definition of the seeds.* The seeds are defined by a region of interest (ROI), by an atlas segmentation or by the user. Starting from each seed, and considering both ways, the tract is reconstructed

2. *Selection of an integration strategy.* A numerical integration strategy to solve the tensor function is required. Schemes such as Euler forward or second or fourth order Runge-Kutta can be used

3. *Definition of a stopping criteria*. A stopping criteria is defined in order to avoid calculations in areas where the vector field is not robustly defined, for example, areas with isotropic or planar diffusion. The stopping criteria is usually defined by the user and are commonly based on the anisotropy indices value or the curvature of the streamline

Several algorithms have been proposed to reconstruct global or local white matter fiber tracts in diffusion tensor fields either by using deterministic, probabilistic or global methods [Pujol 2015, Qi 2015]. These three main methods are described in more detail the next subsections.

2.1.4.1 Streamline tractography

Streamline tractography is the most commonly used method. It is based on a 3D streamline that starts at a seed points, and, in each step, the streamline follows the main direction of the diffusion tensor eigenvector [Conturo 1996, Mori 1999] (see Figure 2.11).

The main drawback of this approach arises with the stopping criteria, when the tensor has not a strong directional component. Image noise [Lazar 2003a], partial volume effects [Alexander 2001], and crossing, branching or merging fiber configurations [Jbabdi 2011] (see Figure 2.12) also make difficult the computation of the fiber direction. To overcome these limitations, among others, fiber-tracking algorithms based on high angular resolution acquisitions [Tuch 2002], regularization [Björnemo 2002], tensor deflection [Lazar 2003b], and stochasticity [Björnemo 2002, Hagmann 2003] have been proposed.

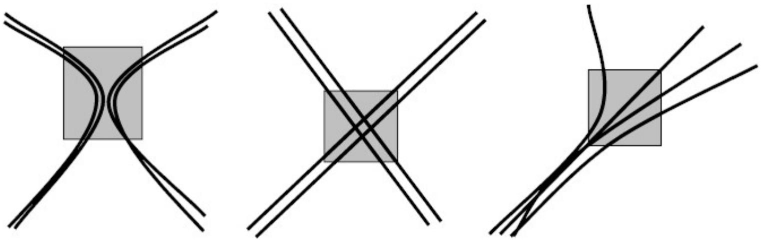


Figure 2.12: Merging, crossing, and branching fiber configurations [Berenschot 2003]

The seed points selection is also a problem of these methods. Not always is obvious where to place the seeds and, hence, important structures can be lost. On the other hand, a large number of seeds generates a large number of fibers, leading to a difficult interpretation (see Figure 2.13). To overcome these problems, techniques such as clustering that groups bundles of fibers using a similarity measure have been applied [Prados 2012].

2.1.4.2 Probabilistic tractography

As said previously, the main problem with the streamline tractography method, is the stopping criteria, that may lead to errors. Probabilistic tractography, instead of defining a stopping criteria, defines the uncertainty associated to the tracts allowing to cross areas with high uncertainty [Behrens 2007, Behrens 2014]. At every voxel, the fiber orientations are estimated to the most probable direction using an orientation density function (ODF). Then, several fibers are generated using slightly different directions. As a result, the probability of the tract is defined by measuring the percentage of fibers connecting the different areas of the brain. Figure 2.14 shows an example of the results using this method.

2.1.4.3 Global tractography

The previous methods described define the fibers orientation taking into account only local diffusion information (voxel level). The idea of global tractography is to estimate the fiber orientation by taking into account the orientation of the neighbors. This new representation leads to more accurate models, and hence, improves the confidence of the connectivity.

Figure 2.15 shows a comparison of the results that can be obtained by using the three methods described. On the left, the fibers were generated with a streamline ap-

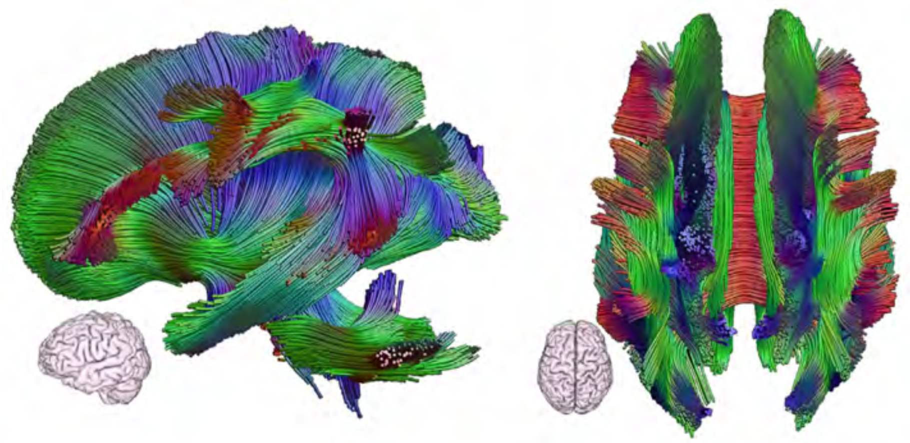


Figure 2.13: Whole-brain tractography reconstruction where colors indicate direction [Behrens 2014]

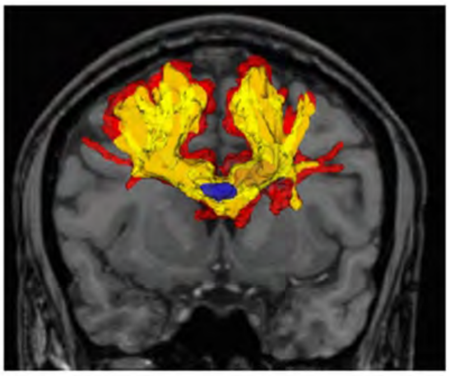


Figure 2.14: Probabilistic tractography [Campbell 2014]

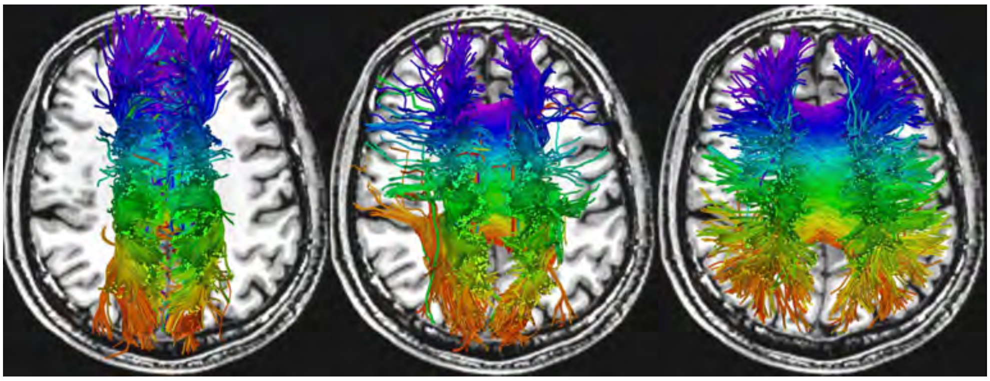


Figure 2.15: Streamline tractography (left), probabilistic tractography (center) and global tractography (right)

proach, in the middle with a probabilistic method and finally, on the right, with a global tractography method.

2.1.5 Functional MRI

Functional MRI (fMRI) is a non-invasive magnetic resonance method capable to determine the brain large-scale activity by measuring the in vivo blood flow changes [Ogawa 1990]. Functional MRI is the most used technique for mapping the activity in the brain. The procedure is similar to MRI principles, but in this case, fMRI uses the change in magnetization between oxygen-rich and oxygen-poor. Since blood flow and neuronal activity are robustly related, this technique allows to find the brain areas that are in use at a specific time.

The most popular method is the blood-oxygen-level dependent (BOLD), which can only detect differences between two states. Since the brain is always active, a new method called resting state fMRI (rfMRI) has been created to evaluate the brain activity when the subject is not doing any task. Thus, this new method allows to find the difference between a subject who is not performing any task, and the same subject performing a specific task. On the other hand, rfMRI can be used also to find anomalies in neurological or psychiatric diseases. Results are usually visualized by color-coding the activation strength for a specific time period (no more than a few seconds).

Unfortunately, the fMRI raw data is a stochastic sum of different signals including artifact noise [Kiviniemi 2003] such as respiratory fluctuations or cardiovascular cycles. Thus, is very important to filter the data before is interpreted in order to avoid wrong conclusions. Usually, the filter of the data is done by statistical procedures. In this case, the most extended method to extract the activation data is the independent component analysis (ICA) which is explained in the next section in detail, however, other methods using regions of interest, clustering and graph theory have been also presented.

2.1.6 Independent component analysis

The most common method to obtain the voxels representing active regions of the brain is the independent component analysis (ICA) [McKeown 1998]. ICA is a computational method based on statistics that separates a signal into non-overlaping spatial and time components, and allows the filtering of noise. It is assumed that the subcomponents are statistically independent non-Gaussian signals, and the number of subcomponents are be less or equal to the original number of signals (variables). This method finds the independent components by maximizing the statistical independence of the subcomponents, such that the variables are independent, or in other words, that the mutual information is 0, and the mutual information between the original variables and the independent variables is as higher as possible. This approach uses the Kullback-Leiber divergence and maximum entropy explained in Section 2.3.3. After applying ICA to the data, the voxels that are activated when performing a task are identified.

2.2 The human connectome

The roughly one hundred billion of neurons in the brain constitute a *complex network* where information is shared and transported by using the structural paths in order to perform a function. The first representation of the complex brain network was by Felleman et al. that represented the connections between areas by defining a *connectivity matrix* [Felleman 1991]. Later, it was introduced the idea of the *connectome* [Hagmann 2005, Sporns 2005] that characterizes the brain network using connectivity matrices and graph theory at different scales from dMRI, rfMRI or considering both techniques together [Hagmann 2007, Hagmann 2010, Sporns 2013]. The introduction of the brain connectome has lead to an increasingly interest in the study of the brain complexity.

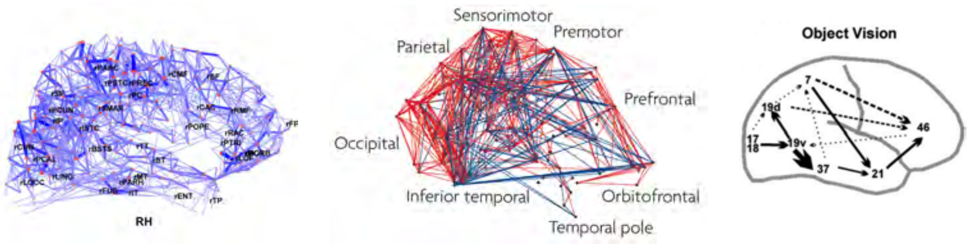
In this section, we first explain the concept of complex brain network, afterwards we provide a detailed description on the brain network construction from structural and functional connectivity data, and finally, the main graph theory measures to describe networks are provided.

2.2.1 Brain complexity

The *brain network* is a complex system that acts as a group of efficient integrated (non-connected regions sharing information) and segregated (small group of highly connected regions that perform a specific function) areas [Sporns 2005]. Thus, it can not be studied as a group of independent elements. Although the big amount of promising methods and discoveries, the mapping of the structure and the functionality of the brain network is still far to be clear.

Focusing on macro-scale studies there are three main aspects to be considered:

- Structural connectivity: Physical connections between regions by finding the white matter fibers (tracts) using dMRI (see 2.16 (a))



(a) (b) (c)

Figure 2.16: (a) Structural connectivity [Hagmann 2008] (b) functional connectivity [Achard 2006] (b) effective connectivity [McIntosh 1994]

- Functional connectivity: Temporal connections between brain regions, often by finding statistical correlation using fMRI (see 2.16 (b))

- Effective connectivity: Combination of anatomical and functional connectivity that shows the influence of a neural system over another as a network of directional effects (see 2.16 (c))

However, the strong relation between the structural and functional systems [Hagmann 2008, Damoiseaux 2009, Caspers 2013] suggests that future advances in understanding the functioning of the brain should focus on the study of both aspects together.

The first requirement in order to create the brain network is the definition of the nodes. This procedure involves a segmentation of the brain (parcellation) into non-overlaped areas. Once the nodes are defined, the next step is to estimate the connections (edges) between the nodes by using functional or structural information. Next section provides details on the definition of the nodes and Section 2.2.3 explains how to create the different brain networks from dMRI or fMRI.

2.2.2 Brain parcellation

Brain parcellation consists in subdividing the brain into different subregions, according to a predefined criteria (i.e., anatomy, structure, function...), in order to define the nodes of the brain network. This parcellation can be done at different scales.

Originally, brain parcellations were created by using ex-vivo architectonic characteristics leading to the creation of anatomical atlas (see Section 2.1.1). Broadmann's atlas [Brodmann 1909] is the most popular so far, and not taking into account any structural of functional connectivity information, is still today one of the most in use. Nowadays, there exist several methods and techniques (see Figure 2.17). They mainly differ in scale and number of regions. However there is no unanimity in which parcellation use, and the most suitable method will depend on the aim of the study. In Section 4.2 a literature review of parcellation methods is provided.

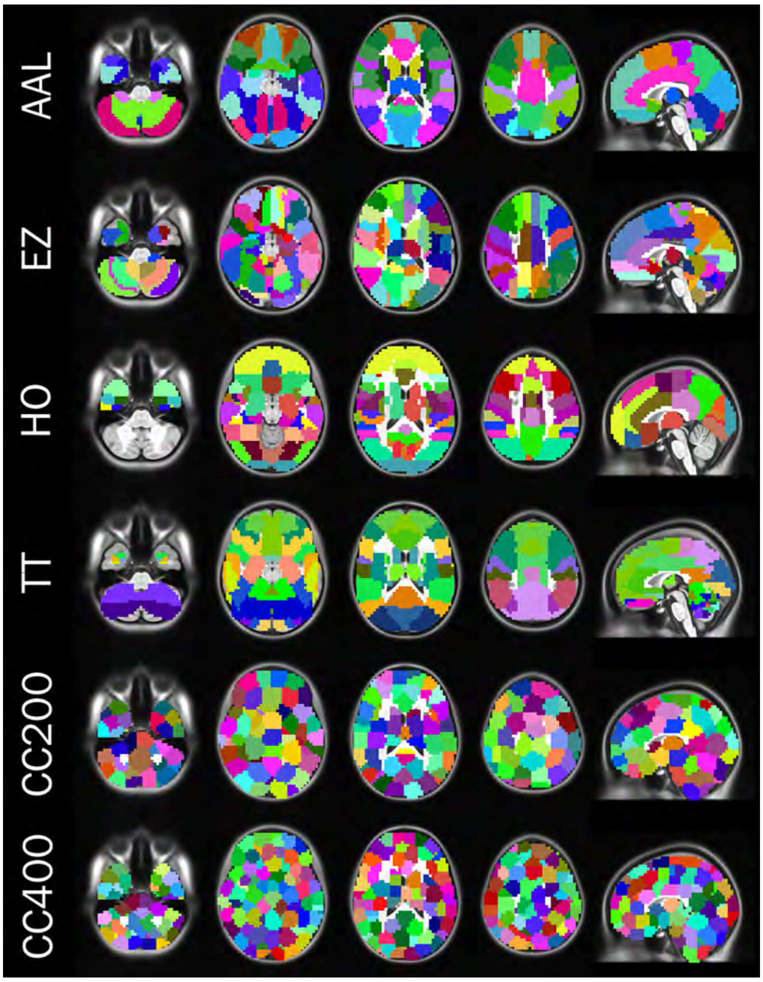


Figure 2.17: Different parcellations of the brain. First four rows correspond to anatomical parcellations, and the two bottom rows correspond to functional parcellations [Craddock 2013]

2.2.3 Brain graph

By computing the connections (functional or structural) between all possible pair of regions defined by the parcellation method (see Section 2.2.2), the connectivity matrix can be computed. It is a symmetric matrix where rows and columns represent brain regions, and values represent the number of connections. From this matrix, is straightforward to construct the graph, which simplifies and helps the understanding of the network by using graph theory. In this case, the nodes of the graph represent brain regions, and the edges the binary or weighted connections [Bullmore 2011, Sporns 2011, Wu 2013]. Thus, the comparison between subjects using the same parcellation, is straightforward since is just a matrix comparison [Nakagawa 2013, Horn 2014]. Note that, the fact of having a functional-based parcellation, does not restrict the possibility of defining the edges with structural information, or inversely.

As it has been mentioned before, the structure (from dMRI) and the functionality (from fMRI) of the brain are very close related [Damoiseaux 2009]. On the one hand, structural connections may predict functional connections [Honey 2007, Hagmann 2008]. On the other hand, if no structure exists, is not possible to be functionally connected. For this reason, a key element in the future studies of the brain network will be to understand how the functional network is using the structural network. In the next subsections we provide a brief summary of both connectivity systems.

2.2.3.1 Structural network

White matter tracts obtained from the dMRI are the representation of the paths that enable the transport of the information in the brain. This information allows the study of structural connections between brain areas predefined by a parcellation method. The connectivity matrix or graph, is created by calculating the fiber tracts that connect all possible pairs of regions. Binary graphs (undirected and unweighted edges) are the most popular [Bullmore 2011] but weighted graphs (i.e., taking the amount fibers connecting areas as a strength) and directed graphs (i.e., taking the influence of one region in another) can also be obtained. Figure 2.18 shows the whole-brain structural networks represented as a weighted undirected graph. Figure 2.19 shows a *connectogram*, a circular representation of the connectome [van Horn 2012].

The consistency of the structural network across subjects and scales allows the study of global properties of the brain. For instance, the study of the brain network using graph theory, has discovered that the brain network is a *small-world*, which means that most of the nodes are not neighbors but it is possible to go from one node to all the other ones efficiently (with a short path length). A part from the global structure, there are two main aspects to study: the segregation and the integration. *Segregation* refers to the organization of specialized groups of neurons to perform specific tasks (communities). This is shown by having groups of highly efficient and locally connected nodes with a high clustering (number of triangles), which lows the random error. On the other hand, *integration* refers to the coordination of the shared information between the different segregated groups of specialized neurons. It is demonstrated with the average

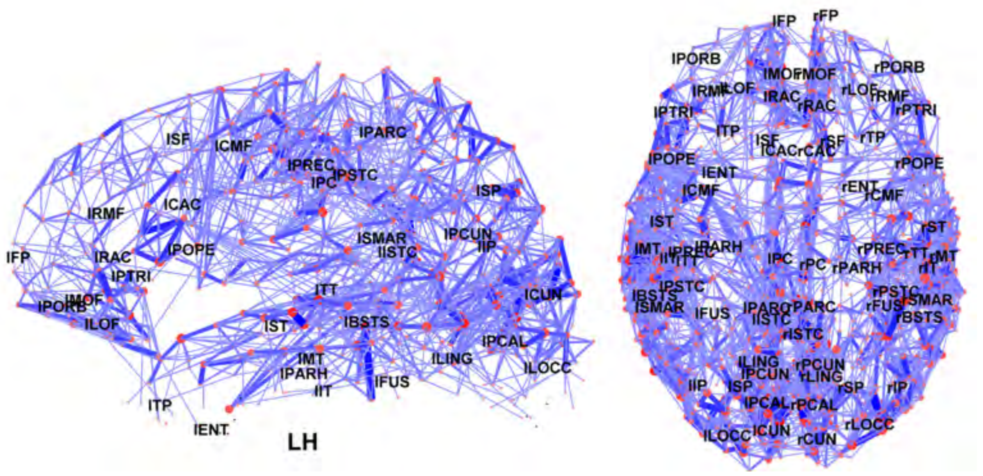


Figure 2.18: Whole-brain structural network represented as a weighted undirected graph. Nodes represent anatomical areas [Hagmann 2008]

short path length that allows a global efficient distributed processing. This integration is possible due to a main core of nodes with a high strength and betweennees that are highly connected with the modular communities. All these properties are mandatory in order to integrate and segregate all the information. Random networks are efficient but not able to process information. Figure 2.20 shows a representation of the structural modules and the main hubs for the human brain.

Synthetic model networks can be used to study specific characteristics. For example, lattice networks are suitable to study the wiring cost properties and random networks are suitable to study the efficiency. However, small-world networks are the better synthetic approximation having a balance between wiring cost and efficiency.

2.2.3.2 Functional network

Functional connectivity is an increasing field in neuroimaging that studies the temporal dependence between active brain regions. This dependence is measured statistically by finding the fluctuations using the fMRI BOLD series in resting state (see Section 2.1.5). Similar to the structural network, the matrix that defines the functional connections can be also defined as a graph, where nodes represent brain areas (defined by a parcellation method) and the edges the temporal functional connection between the regions.

In order to find the functional edges, the average time series associated to each node are used to estimate the edges between nodes, usually by finding the pairs of nodes with similar time series. Since the average time series has negative values, frequently, the matrix is thresholded in order to build the graph. The threshold can be found by applying a statistical significance test to each edge (i.e., *t-test*).

Some connectivity patterns have been found in functional networks. For instance, it has been shown that it has a hierarchical organization [Zhou 2006] with small-world properties, a mean path length similar to random networks and a much higher clus-

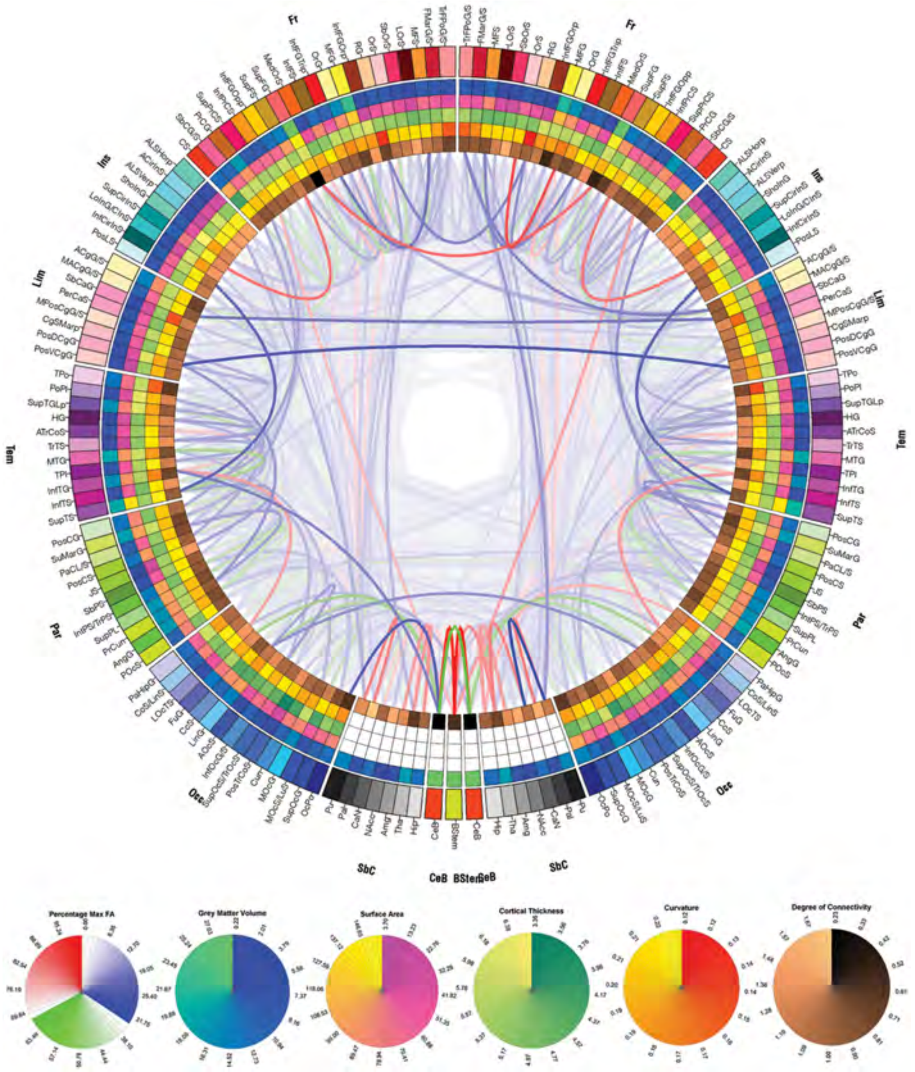


Figure 2.19: Connectogram. The connectome is represented as a ring where values of different measures and the connections between areas are displayed in the center [van Horn 2012]

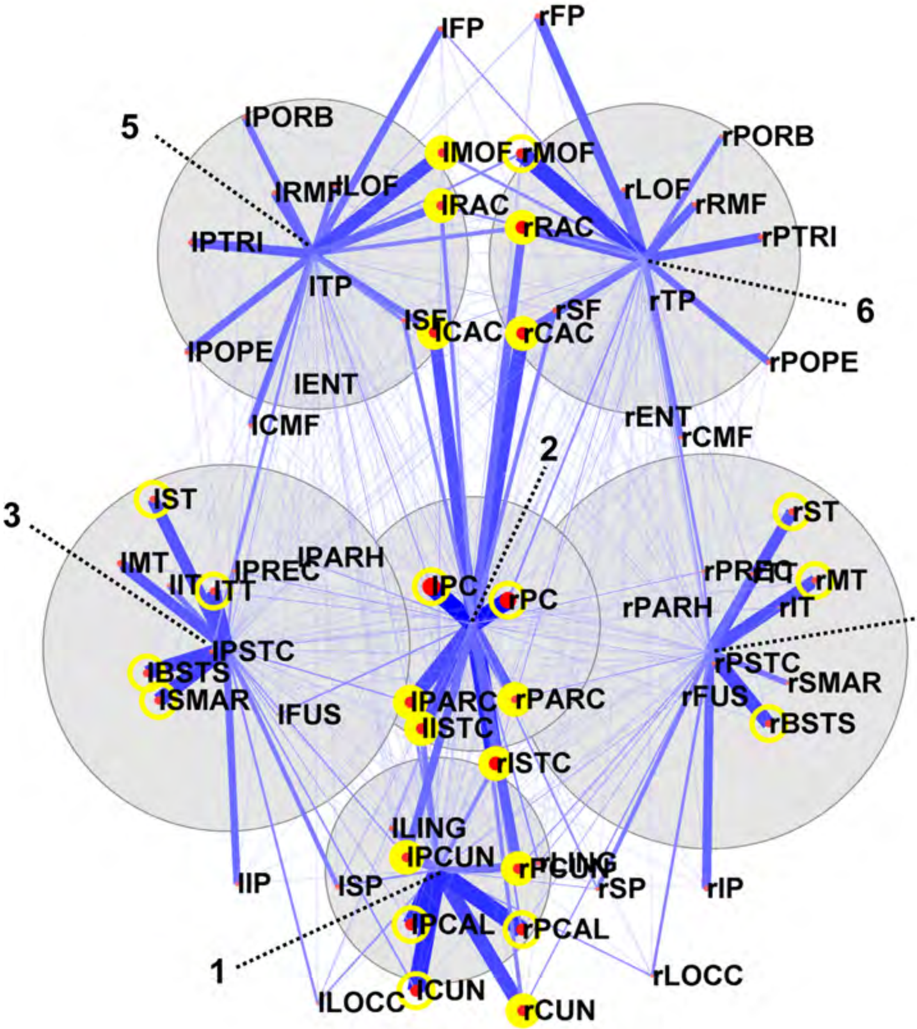


Figure 2.20: Modules (gray circles) and hubs (yellow nodes) of the whole-brain [Hagmann 2008]

tering coefficient [Achard 2006]. Other studies have reported a pattern of highly connected anatomical separated areas during rest, which form the resting state network and is consistent across subjects and scales [Beckmann 2005, Damoiseaux 2009].

Although most of the functional connections have an equivalent structural connection, there are functional connections with no direct path, which means that the functional connections use a third region in order to communicate and transfer information.

2.2.4 Graph theory measures

Graph theory is a key element in the study of the topology of the brain network by using complexity measures, regardless the information that has been used to construct the network (functional or structural). Graph theory measures can describe local properties (node level) or global properties (network level) and do not take into account the anatomy.

It is still unknown which are the measures that describe best the brain network. Additionally, the results obtained strongly depend on the quality of the connectivity matrices and the parcellation method, which are not final but are the best approximation [Kennedy 2013, Stephan 2013]. However, it has been shown that it is possible to associate network disruptions with different diseases using complexity measures [van den Heuvel 2010, Meskaldji 2013, Sato 2013, Sporns 2013, Crossley 2014]. Thus, novel measures showing new properties are needed in order to better understand the functioning of the brain network [Papo 2014].

In the next subsections, we describe some relevant measures. Although the most popular connectivity matrices are binary, most of the measures have the binary and undirected version and the weighted directed and undirected versions.

2.2.4.1 Local measures

Local measures describe properties of the nodes such as the pattern of connections of the node or the way the node is connected (topology) [Stam 2007, Rubinov 2010, Kaiser 2011]. Is of interest to describe the node's properties given that it helps to characterize the brain network architecture and organization.

The most well known local measures include:

- *Degree or degree centrality*: Number of connections of a node [Bullmore 2009]. It provides information about how highly connected is the node

- *Strength*: Similar to the degree but considering the node weights. It provides information about how strongly connected is the node

- *Density*: Percentage of the number of connections among all possible connections. Thus, the higher the density, the lower the variability

- *Betweenness centrality*: Percentage of short paths that include the node (the shortest path length is the minimum distance or steps between two nodes). It is related with the efficiency, since efficient networks require a short path length (nodes

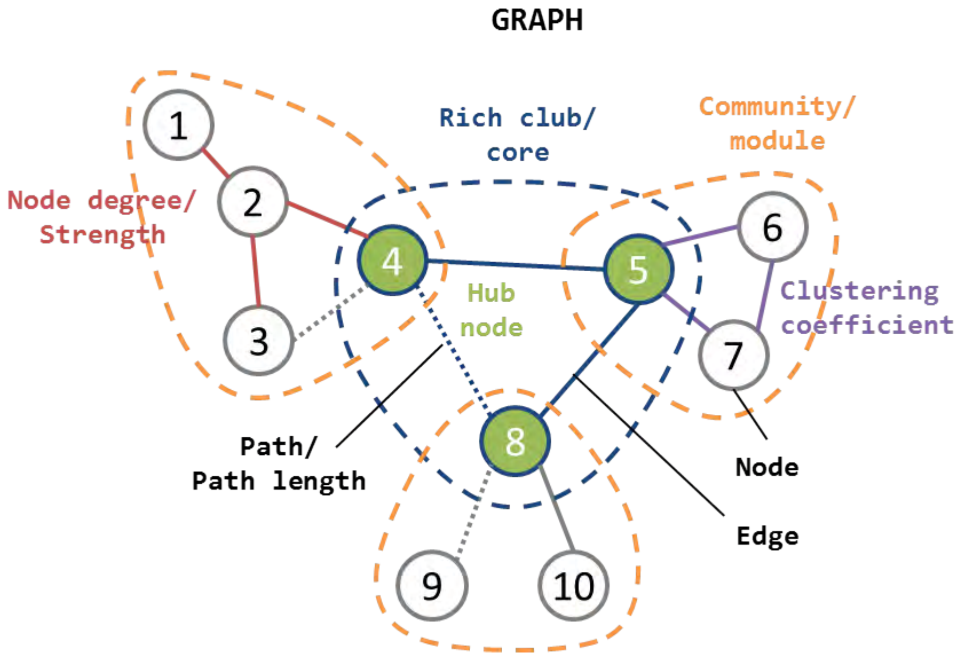


Figure 2.21: Schematic graph diagram showing the most relevant properties and measures described

with high betweenneess centrality). It is also used to find hub nodes, as the main characteristic of the hubs nodes is the short path length

The diagram in Figure 2.21 provides an schematic representation of some of these measures and properties.

Collectively, these measures can describe networks more globally, by showing specific properties, such as small-world, scale-free or hierarchical structures. A good summary, including more measures and the respectively formulas can be found in [Rubinov 2010, Papo 2014].

2.2.4.2 Global measures

The aim of global measures is to describe the overall network structure of the brain. These measures are mainly quantitative and help to find global differences otherwise not identified. Global measures can be divided in two main groups, those that describe integration, and those that describe segregation.

Measures that describe integration include:

- *Characteristic path length*: Average shortest path length [Watts 1998]. It describes how close on average a node is connected to all the other ones. Random networks have a short characteristic path length, which means that are efficient networks in terms of sharing and integrating information

- *Global efficiency*: Average inverse shortest path length [Latora 2001]. It describes how efficiently the network shares information. Unlike the characteristic path length, this measure is useful when computed on disconnected networks

On the other hand, measures that describe segregation include:

- *Clustering coefficient*: Number of neighbors of a node that are also connected to each other forming a triangle [Watts 1998]
- *Transitivity*: Normalized clustering coefficient [Newman 2003b]
- *Modularity or community*: Group of nodes highly interconnected but poorly connected to other groups of nodes [Newman 2003a]

In addition, other properties of the networks are described by the following concepts:

- *Core*: Group of highly and mutually interconnected nodes where all the possible connections exist
- *k-core*: Subnetwork containing only the nodes with a degree greater than a threshold *k*
- *Hubs*: Nodes that connect communities, usually with a high degree, short average path length and high centrality
- *Rich-club*: Set of nodes with a high degree more densely connected than the average of the network
- *Hierarchical modularity*: Multiscale structure within a community

The diagram in Figure 2.21 provides an schematic representation of some of these measures and properties.

Studies such as [Kennedy 2013] suggests that the information distribution and integration in the brain is governed by a structurally and functionally central circuit with different areas acting as a hub. These hubs are densely interconnected forming a *rich-club* [Colizza 2006, Harriger 2012, van den Heuvel 2012]. Watts et al. studied the anatomical connectivity of the nervous system of C. elegans showing an evidence of small-world properties [Watts 1998]. Later, other studies demonstrated that the human brain's network has also small-world properties [Sporns 2004]. These networks are highly clustered like lattice networks and with small path lengths like random graphs, having a balanced segregation and integration.

2.3 Information theory: the basics

In 1948, Claude Shannon published "A mathematical theory of communication" [Shannon 1948] which marks the beginning of information theory. In this paper, he defined measures such as entropy and mutual information, and introduced the fundamental laws of data compression and transmission. Information theory deals with the transmission, storage, and processing of information, and is used in fields such as physics, computer science, mathematics, statistics, economics, biology, linguistics, neurology, learning, image processing, and computer graphics.

In this section, we present some basic concepts of information theory. Good references on information theory can be found in the books written by Cover and Thomas [Cover 1991], and Yeung [Yeung 2002].

2.3.1 Entropy

Let X be a discrete random variable with alphabet $\mathscr{X}$ and probability distribution $\{p(x)\}$, where $p(x) = Pr\{X = x\}$ and $x \in \mathscr{X}$, the *Shannon entropy* $H(X)$ of a discrete random variable X with values in the set $\mathscr{X} = \{x_1, x_2, \ldots, x_n\}$ is defined by

$$H(X) = -\sum_{x\in\mathscr{X}} p(x) \log p(x), \tag{2.2}$$

where $p(x) = Pr[X = x]$, the logarithms are taken in base 2 (entropy is expressed in bits), and we use the convention that $0 \log 0 = 0$, which is justified by continuity. In this thesis, $\{p(x)\}$ will be also denoted by $p(X)$ or simply p. This notation will be extended to two or more random variables.

We can use interchangeably the notation $H(X)$ or $H(p)$ for the entropy, where p is the probability distribution $\{p_1, p_2, \ldots, p_n\}$. As $-\log p(x)$ represents the *information* associated with the result x, the entropy gives us the *average information* or *uncertainty* of a random variable. Uncertainty and information can be seen as opposite sides of the same coin. While entropy quantifies the uncertainty that we have before an event, information is a measure of the uncertainty reduction after the event.

Some other relevant properties of the entropy are [Shannon 1948]:

1. $0 \leq H(X) \leq \log n$

   - $H(X) = 0$ if and only if all the probabilities except one are zero, this one having the unit value, i.e., when we are certain of the outcome.
   - $H(X) = \log n$ when all the probabilities are equal. This is the most uncertain situation.

2. If we equalize the probabilities, entropy increases.

When $n = 2$, the *binary* entropy (Figure 2.22) is given by

$$H(X) = -p \log p - (1-p) \log(1-p), \tag{2.3}$$

where the variable X is defined by

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

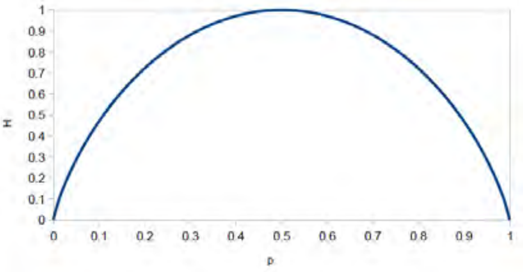


Figure 2.22: Plot of binary entropy

If we consider another random variable Y with probability distribution $p(y)$ corresponding to values in the set $\mathscr{Y} = \{y_1, y_2, \dots, y_m\}$, the *joint entropy* of X and Y is defined as

$$H(X,Y) = -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log p(x,y), \tag{2.4}$$

where $p(x,y) = Pr[X = x, Y = y]$ is the joint probability.

The *conditional entropy* $H(Y|X)$ of a random variable Y given a random variable X is defined by

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathscr{X}} p(x)H(Y|x) & (2.5)\\ &= \sum_{x\in\mathscr{X}} p(x)\left(-\sum_{y\in\mathscr{Y}} p(y|x)\log p(y|x)\right) & (2.6)\\ &= -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log p(y|x), & (2.7)\end{aligned}$$

where $p(y|x) = Pr[Y = y|X = x]$ is the conditional probability of y given x and $H(Y|x)$ is the entropy of Y given x.

The Bayes theorem expresses the relation between the different probabilities:

$$p(x,y) = p(x)p(y|x) = p(y)p(x|y). \tag{2.8}$$

If X and Y are *independent*, then $p(x,y) = p(x)p(y)$.

In other words, the conditional entropy can be described as a *channel* that the input is the random variable X and the output is the random variable Y. Then, $H(X|Y)$ corresponds to the uncertainty of the channel's input from the receiver's point of view, and vice versa for $H(Y|X)$. Note that in general $H(X|Y) \neq H(Y|X)$.

The following properties are also fulfilled:

1. $H(X,Y) \leq H(X) + H(Y)$
2. $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
3. $H(X) \geq H(X|Y) \geq 0$

2.3.2 Mutual information

The *mutual information* $I(X;Y)$ between two random variables X and Y is defined by

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) && (2.9)\\
&= H(Y) - H(Y|X) && (2.10)\\
&= -\sum_{x\in\mathscr{X}} p(x)\log p(x) + \sum_{y\in\mathscr{Y}}\sum_{x\in\mathscr{X}} p(x,y)\log p(x|y) && (2.11)\\
&= \sum_{x\in\mathscr{X}} p(x)\sum_{y\in\mathscr{Y}} p(y|x)\log\frac{p(y|x)}{p(y)} && (2.12)\\
&= \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. && (2.13)
\end{aligned}$$

Mutual information represents the amount of information that one random variable, the output of the channel, tells about a second random variable, the input of the channel, and vice versa, i.e., how much the knowledge of X decreases the uncertainty of Y and vice versa. Therefore, $I(X;Y)$ is a measure of the shared information between X and Y.

Mutual information $I(X;Y)$ has the following properties:

1. $I(X;Y) \geq 0$ with equality if, and only if, X and Y are independent.

2. $I(X;Y) = I(Y;X)$

3. $I(X;Y) = H(X) + H(Y) - H(X,Y)$

4. $I(X;Y) \leq H(X)$

The relationship of all the measures explained above can be expressed by the Venn diagram, as shown in Figure 2.23.

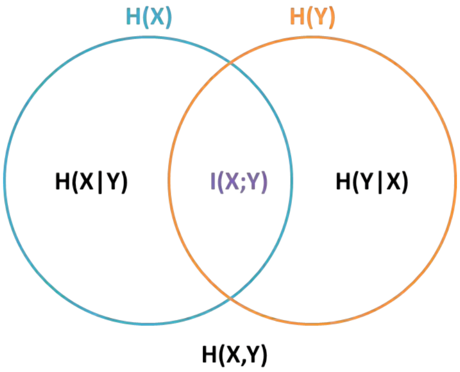


Figure 2.23: Venn diagram of a discrete channel

2.3.3 Kullback-Leibler distance

The *relative entropy* or *Kullback-Leibler distance* $D_{KL}(p,q)$ between two probability distributions p and q [Cover 1991, Yeung 2002], that are defined over the alphabet $\mathscr{X}$, is given by

$$D_{KL}(p,q) = \sum_{x\in\mathscr{X}} p(x)\log\frac{p(x)}{q(x)}, \tag{2.14}$$

where, from continuity, we use the convention that $0\log 0 = 0$, $a\log\frac{a}{0} = \infty$ if $a > 0$, and $0\log\frac{0}{0} = 0$.

The relative entropy is "a measure of the inefficiency of assuming that the distribution is q when the true distribution is p" [Cover 1991].

The relative entropy satisfies the *information inequality* $D_{KL}(p,q) \geq 0$, with equality only if $p = q$. The relative entropy is also called *discrimination* and it is not strictly a distance, since it is not symmetric and does not satisfy the triangle inequality. Moreover, we have to emphasize that the mutual information can be expressed as

$$I(X;Y) = D_{KL}(p(x,y), p(x)p(y)). \tag{2.15}$$

2.3.4 Decomposition of mutual information

Given a communication channel $X \to Y$, mutual information can be decomposed in different ways to obtain the information associated with a value (or symbol) in $\mathscr{X}$ or $\mathscr{Y}$. Next, we present different definitions of information that have been proposed in the field of neural systems to investigate the significance associated to stimuli and responses [DeWeese 1999, Butts 2003].

For random variables S and R, representing an ensemble of stimuli $\mathscr{S}$ and a set of responses $\mathscr{R}$, respectively, the mutual information (see Equations 2.10 and 2.12) is given by

$$\begin{aligned} I(S;R) &= H(R) - H(R|S) && (2.16)\\ &= H(R) - \sum_{s\in\mathscr{S}} p(s)H(R|s) && (2.17)\\ &= \sum_{s\in\mathscr{S}} p(s) \sum_{r\in\mathscr{R}} p(r|s)\log\frac{p(r|s)}{p(r)}, && (2.18) \end{aligned}$$

where $p(r|s)$ is the conditional probability of value r given a known value s, and $p(S) = \{p(s)\}$ and $p(R) = \{p(r)\}$ are the marginal probability distributions of the input and output variables of the channel, respectively.

To quantify the information associated to each stimulus or response, $I(S;R)$ can be decomposed as

$$\begin{aligned} I(S;R) &= \sum_{s\in\mathscr{S}} p(s)I(s;R) && (2.19)\\ &= \sum_{r\in\mathscr{R}} p(r)I(S;r), && (2.20) \end{aligned}$$

where $I(s;R)$ and $I(S;r)$ represent, respectively, the information associated to stimulus s and response r. Thus, $I(S;R)$ can be seen as a weighted average over individual contributions from particular stimuli or particular responses. The definition of the contribution $I(s;R)$ or $I(S;r)$ can be performed in multiple ways, but we present here the three most basic definitions denoted by I_1, I_2 [DeWeese 1999], and I_3 [Butts 2003].

Given a stimulus s, three specific information measures that fulfill Equation 2.19 are defined:

- The *surprise* I_1 can be directly derived from Equation 2.18, taking the contribution of a single stimulus to $I(S;R)$:

$$I_1(s;R) = \sum_{r \in \mathscr{R}} p(r|s) \log \frac{p(r|s)}{p(r)}. \tag{2.21}$$

  This measure expresses the surprise about R from observing s. It can be shown that I_1 is the only positive decomposition of $I(S;R)$ [DeWeese 1999]. This positivity can be shown by observing that $I_1(s;R)$ is the Kullback-Leibler distance [Cover 1991] between the conditional probability $p(R|s)$ and the marginal distribution $p(R)$.

- The *specific information* I_2 [DeWeese 1999] can be derived from Equation 2.17, taking the contribution of a single stimulus s to $I(S;R)$:

$$\begin{aligned} I_2(s;R) &= H(R) - H(R|s) \\ &= -\sum_{r \in \mathscr{R}} p(r) \log p(r) + \sum_{r \in \mathscr{R}} p(r|s) \log p(r|s). \end{aligned} \tag{2.22}$$

  The specific information I_2 of a particular response is defined as the reduction uncertainty in the stimulus gained by the observation of that response [Butts 2003]. Thus, this measure expresses the change in uncertainty about R when s is observed. Note that I_2 can take negative values. This means that certain observations s do increase our uncertainty about the state of the variable R.

- The *stimulus-specific information* I_3 [Butts 2003] is defined by

$$I_3(s;R) = \sum_{r \in \mathscr{R}} p(r|s) I_2(S;r) \tag{2.23}$$

  and also fulfills Equation 2.19 (for a proof, see [Butts 2003]). The most informative (or significant) stimuli are those that cause the most informative responses. Thus, a large value of $I_3(s;R)$ means that the states of R associated with s are very informative in the sense of $I_2(S;r)$ (i.e., the specific information associated with response r). That is, the most informative input values s are those that are related to the most informative output values r. Observe that $I_1(s;R)$ and $I_2(s;R)$ are obtained from both distributions $p(R)$ and $p(R|s)$, while $I_3(s;R)$ is a weighted sum of the measure $I_2(S;r)$, which is obtained from distributions $p(S)$ and $p(S|r)$.

Note that, similar to the above definitions for a stimulus s, the information associated to a response r can be defined. The properties of positivity and additivity of these measures have been studied in [DeWeese 1999, Butts 2003]. A measure is additive when the information obtained about S from two observations, $r_1 \in \mathscr{R}_1$ and $r_2 \in \mathscr{R}_2$, is equal to that obtained from r_1 plus that obtained from r_2 when r_1 is known. While I_1 is always positive and non-additive, I_2 can take negative values but is additive, and I_3 can take negative values and is non-additive. On the one hand, because of the additivity property, DeWeese and Meister [DeWeese 1999] prefer I_2 against I_1 since they consider that additivity is a fundamental property of any information measure. On the other hand, Butts [Butts 2003] proposes some examples that show how I_3 identifies the most significant stimuli.

2.3.5 Jensen's inequality

Some important properties of information measures can be deduced from the Jensen's inequality [Cover 1991].

A function $f(x)$ is *convex* over an interval (a, b) (the graph of the function lies below any chord) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2). \qquad (2.24)$$

A function is strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$. A function $f(x)$ is *concave* (the graph of the function lies above any chord) if $-f(x)$ is convex.

For instance, $x \log x$ for $x \geq 0$ is a strictly convex function, and $\log x$ for $x \geq 0$ is a strictly concave function [Cover 1991].

Jensen's inequality: If f is convex on the range of a random variable X, then

$$f(E[X]) \leq E[f(X)], \qquad (2.25)$$

where E denotes expectation. Moreover, if $f(x)$ is strictly convex, the equality implies that $X = E[X]$ with probability 1, i.e., X is a deterministic random variable with $Pr[X = x_0] = 1$ for some x_0.

One of the most important consequences of Jensen's inequality is the information inequality $D_{KL}(p\|q) \geq 0$. Other previous properties can also be derived from this inequality.

Observe that if $f(x) = x^2$ (convex function), then $E[X^2] - (E[X])^2 \geq 0$. So, the variance is invariably positive.

If f is substituted by the Shannon entropy, which is a concave function, we obtain the *Jensen-Shannon inequality* [Burbea 1982]:

$$JS(\pi_1, \pi_2, \ldots, \pi_n; p_1, p_2, \ldots, p_n) \equiv H\left(\sum_{i=1}^{n} \pi_i p_i\right) - \sum_{i=1}^{n} \pi_i H(p_i) \geq 0, \qquad (2.26)$$

where $JS(\pi_1, \pi_2, \ldots, \pi_n; p_1, p_2, \ldots, p_n)$ is the *Jensen-Shannon divergence* of probability distributions $p_1, p_2, \ldots, p_n$ with prior probabilities or weights $\pi_1, \pi_2, \ldots, \pi_n$, fulfilling $\sum_{i=1}^{n} \pi_i = 1$. The JS-divergence measures how 'far' are the probabilities p_i from their likely joint source $\sum_{i=1}^{n} \pi_i p_i$ and equals zero if and only if all the p_i are equal. It is important to note that the JS-divergence is identical to $I(X;Y)$ when $\pi_i = p(x_i)$ and $p_i = p(Y|x_i)$ for each $x_i \in \mathscr{X}$, where $p(X) = \{p(x_i)\}$ is the input distribution, $p(Y|x_i) = \{p(y_1|x_i), p(y_2|x_i), \ldots, p(y_m|x_i)\}$, $n = |\mathscr{X}|$, and $m = |\mathscr{Y}|$ [Burbea 1982, Slonim 2000b].

2.3.6 Markov process

A *Markov process* [Cover 1991], or *Markov chain*, is a discrete stochastic process defined over a set of states $\mathscr{X} = \{x_1, x_2, \ldots, x_n\}$ which is described by a *transition probability matrix* P. In each step, the process makes a transition from its current state i to a new state j with *transition probability* P_{ij}. The transition probabilities only depend on the current state. A Markov process can also be seen as a sequence of random variables $X_k, k = 0 \ldots \infty$, in which each $X_k, k \geq 1$, depends only on the previous X_{k-1} and not on the ones before. Thus, $P_{ij} = p(x_j^k | x_i^{k-1}) = Pr[X_k = x_j | X_{k-1} = x_i]$.

For a *stationary Markov process*, the probabilities of finding the particle in each state i converge to a *stationary distribution* $w = \{w_1, \ldots, w_n\}$ after a number of steps. The stationary or equilibrium probabilities w_i fulfill the relation $w_i = \sum_{j=1}^{n} w_j P_{ji}$ and also the reciprocity relation $w_i P_{ij} = w_j P_{ji}$.

In particular, a Markov process can be considered as a chain of random variables complying with

$$H(X_t|X_1, X_2, \ldots, X_{t-1}) = H(X_t|X_{t-1}) \tag{2.27}$$

An important result is the following theorem: For a stationary Markov chain, with stationary distribution w_i, the entropy rate or information content is given by

$$\begin{aligned} h &= \lim_{t\to\infty} \frac{1}{t} H(X_1, X_2, \ldots, X_t) \\ &= \lim_{t\to\infty} H(X_t|X_{t-1}) \\ &= H(X_2|X_1) = -\sum_{i=1}^{t} w_i \sum_{j=1}^{t} P_{ij} \log P_{ij} \end{aligned} \tag{2.28}$$

where w_i is the equilibrium distribution and P_{ij} is the transition probability from state i to state j. Entropy rate represents the *average information content* per output symbol [1] [Cover 1991]. It is the "uncertainty associated with a given symbol if all the preceding symbols are known" and can be viewed as "the intrinsic *unpredictability*" or "the irreducible *randomness*" associated with the chain [Feldman 1998].

Finally, the *excess entropy* or *effective measure complexity* [Crutchfield 1983, Shaw 1984, Grassberger 1986, Szépfalusy 1986] of an infinite chain is defined by

$$E = \lim_{t\to\infty} (H(X_1, X_2, \ldots, X_t) - th) \tag{2.29}$$

[1] At least, h exists for all stationary stochastic processes.

where h is the entropy rate of the chain and t is the length of this chain. The excess entropy can be interpreted as the mutual information between two semi-infinite halves of the chain. Another way of viewing this, is that excess entropy is the *cost of amnesia* – the excess entropy measures how much more random the system would become if we suddenly forgot all information about the left half of the string [Feldman 1997]. For a stationary Markov process, excess entropy coincides with mutual information, and, hence, in this context, mutual information can be seen as a measure of the system structure.

Chapter 3

Hierarchical Clustering Based on the Information Bottleneck

Contents

3.1 Introduction

Clustering techniques organize elements into groups (or clusters) whose members are similar and dissimilar to elements belonging to other clusters. The core issue in clustering is the similarity estimation. Once a similarity measure is chosen, clustering is formulated as an optimization problem [Xu 2005]. More details about distance metrics can be found in [Yu 2012]. Clustering is used in many different fields, such as engineering, computer sciences, life and medical sciences, earth sciences, social sciences [Hartigan 1975, Everitt 2001] and, more recently, in cartoon animation [Yu 2011, Yu 2013]. This technique has been extensively investigated and it is still an active area of research. Some reference papers on the topic are [Hansen 1997, Jain 1988, Jain 1999, Duda 2001, Xu 2005, Lam 2014].

Clustering techniques can be grouped into two main categories: partition-based methods that iteratively divide the dataset by minimizing a pre-defined distance function [MacQueen 1967] and hierarchy-based methods that organize data in a hierarchy of clusters [Nagpal 2013]. These last can be classified as divisive or agglomerative [Jain 1988, Kaufman 1990, Jain 1999, Everitt 2001]. Divisive methods split up the dataset into smaller clusters in a top-down process until each cluster contains only one element or a pre-defined number of clusters is reached. Agglomerative methods follow a bottom-up procedure that starts with n singleton clusters and form a hierarchy by successively merging the clusters until a desired number of clusters is obtained. Recently, new hierarchical clustering techniques have been proposed to deal with large-scale datasets in data mining and other fields [Zhao 2005, Hsu 2007, Tao 2007, Kersting 2010, Cai 2014].

The *information bottleneck* method provides a different view of the clustering problem [Tishby 1999, Slonim 2006, Xu 2014]. The method considers that each variable X that takes values from an alphabet $\mathscr{X}$ occurs together with a corresponding variable Y, which takes values from an alphabet $\mathscr{Y}$, where X represents the data and Y is a control variable that holds some correlation with X, and it requires the control variable Y to be different from the variable to be clustered X. Clustering is performed in the variable X to compress the description of the elements while preserving as much information as possible about Y. From this statement, there is no need to explicitly define a similarity measure, since the similarity of the elements is defined from the optimization principle itself.

A stochastic Markov process, $\mathbf{X}$, is an indexed sequence of random variables, $\{X_0, X_1, \ldots, X_t, X_{t+1}, \ldots\}$, which take values from an alphabet $\mathscr{X}$. Inspired by the hierarchical clustering and the information bottleneck method, we propose a new method to cluster the states of a stationary Markov process preserving the maximum mutual information between consecutive variables, X_t and X_{t+1}, which extends the applicability of the information bottleneck-based methods. Note that, in this case, X_t and X_{t+1} take values over the same alphabet $\mathscr{X}$, and, for this reason, the standard information bottleneck method cannot be applied, since the states of both variables X_t and X_{t+1} are simultaneously clustered. The agglomerative and divisive versions of the proposed hierarchical clustering technique are presented and applied to image quantization (see Figure 3.1).

This chapter has been structured as follows. Section 3.2, we present an overview of previous work on information theory, the information bottleneck method, and the Markov processes. In Section 3.3, we introduce the mathematical basis of our framework. In Section 3.4, we present the agglomerative and divisive versions of the proposed hierarchical clustering approach. In Section 3.5, we describe the application of the proposed algorithms to image quantization and, in Section 3.6, we discuss the obtained results. Finally, in Section 3.7, we present conclusions and future work.

3.2 Background

3.2.1 Information bottleneck method

The information bottleneck method, introduced by Tishby et al. [Tishby 1999], extracts a compact representation of the random variable X, denoted by $\widehat{X}$, with minimal loss of MI with respect to an additional or relevant random variable Y. In this framework,

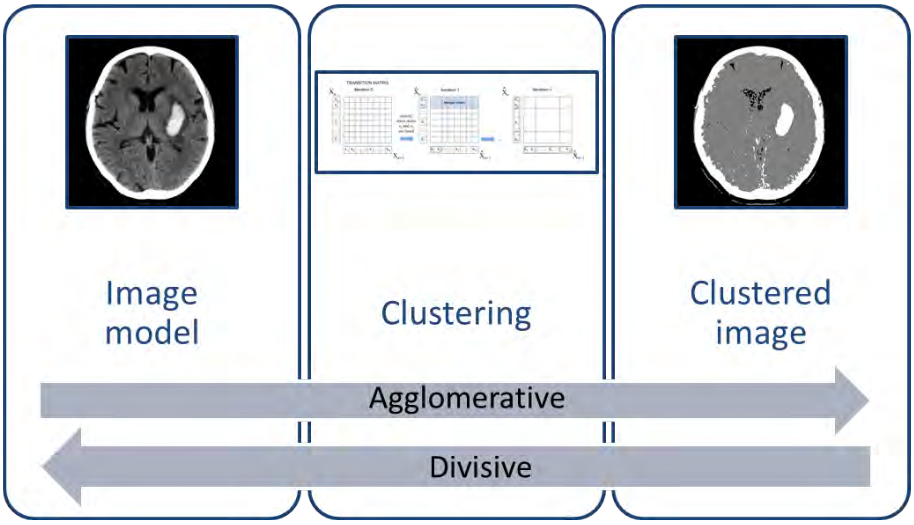


Figure 3.1: Diagram of the main aspects described in this chapter

X represents the dataset to be clustered, $x_i \in \mathcal{X}$ is the i-th element of this dataset, and $\widehat{X}$ represents the set of data clusters and $\widehat{x}_k \in \widehat{\mathcal{X}}$ is the k-th cluster. It is also assumed that Y represents a feature set that has a certain degree of correlation with respect to the dataset X, so that the clustering of X is guided by the maximal preservation of this correlation, i.e., by the maximization of the mutual information between $\widehat{X}$ and Y. In this thesis, the variable Y is also called *control variable* as it controls the clustering process of X.

Soft [Tishby 1999] and hard [Slonim 2000a] partitions of X can be adopted. In the first case, every value $x \in \mathcal{X}$ can be assigned to every cluster $\hat{x} \in \widehat{\mathcal{X}}$ with some conditional probability $p(\hat{x}|x)$ (soft clustering). In the second case, every value $x \in \mathcal{X}$ is assigned to only one cluster $\hat{x} \in \widehat{\mathcal{X}}$ (hard clustering).

We focus our attention on the *agglomerative information bottleneck* method [Slonim 2000a] which is a hard clustering technique. Given a cluster $\hat{x}$ defined by $\hat{x} = \{x_1, \ldots, x_l\}$, where l is the number of elements of the cluster and $x_k \in \mathcal{X}$ ($\forall k \in [1..l]$), and given probability distributions $p(\hat{x})$ and $p(y|\hat{x})$ defined by

$$p(\hat{x}) = \sum_{k=1}^{l} p(x_k), \tag{3.1}$$

$$p(y|\hat{x}) = \frac{1}{p(\hat{x})} \sum_{k=1}^{l} p(x_k, y) \quad \forall y \in \mathcal{Y}, \tag{3.2}$$

the following properties are fulfilled:

- The decrease in the mutual information from $I(X;Y)$ to $I(\widehat{X};Y)$ due to the clustering of $x_1, \ldots, x_l$ is given by

$$\delta I_{\hat{x}} = p(\hat{x}) JS(\pi_1, \ldots, \pi_l; p(Y|x_1), \ldots, p(Y|x_l)) \geq 0, \tag{3.3}$$

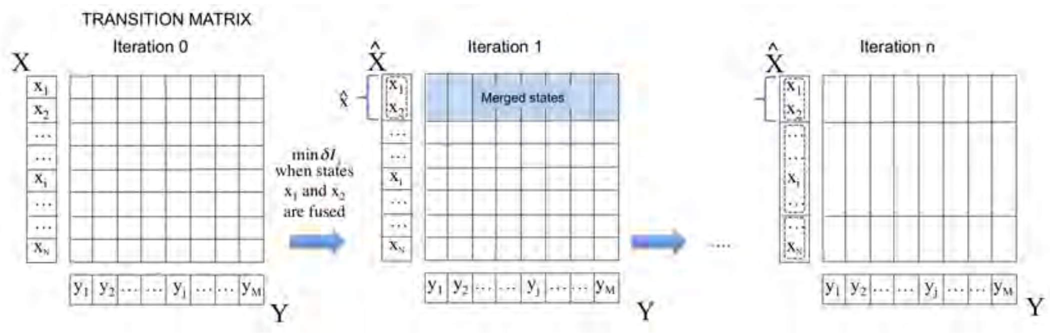


Figure 3.2: Diagram of standard agglomerative information bottleneck method

where $\pi_k = \frac{p(x_k)}{p(\hat{x})}$. An optimal clustering algorithm has to minimize $\delta I_{\hat{x}}$.

- A clustering of l components can be obtained by $l-1$ consecutive clustering of pairs of components.

When considering the clustering of two states x_i and x_j into a cluster $\hat{x}$, such that

$$p(\hat{x}) = p(x_i) + p(x_j) \tag{3.4}$$

and

$$p(y|\hat{x}) = \frac{p(x_i)p(y|x_i) + p(x_j)p(y|x_j)}{p(\hat{x})}, \tag{3.5}$$

Equation 3.3 is simplified in the following form:

$$\begin{aligned} \delta I_{\hat{x}} &= I(X;Y) - I(\widehat{X};Y) \\ &= p(\hat{x}) JS\left(\pi_i, \pi_j; p(Y|x_i), p(Y|x_j)\right), \end{aligned} \tag{3.6}$$

where $\pi_i = \frac{p(x_i)}{p(\hat{x})}$ and $\pi_j = \frac{p(x_j)}{p(\hat{x})}$. The JS-divergence $JS\left(\pi_i, \pi_j; p(Y|x_i), p(Y|x_j)\right)$ between two states can be interpreted as a measure of dissimilarity between them with respect to the output variable. In Figure 3.2, the evolution of the transition matrix (i.e. the matrix that each row corresponds to $p(Y|x_i)$) for the different iterations of the algorithm is shown. In the first iteration, the states 1 and 2 (those that minimize the information loss) are grouped, which results in the merging of rows 1 and 2. The loop is repeated until the desired final number of states is reached.

Dhillon et al. [Dhillon 2003] presented a co-clustering algorithm applied to word-document clustering that simultaneously clusters X and Y into disjoint or hard clusters. An optimal co-clustering algorithm has to minimize the difference $I(X,Y) - I(\widehat{X},\widehat{Y})$. In this case, the loss in mutual information is given by

$$\begin{aligned} \delta I_{\hat{x};\hat{y}} &= I(X,Y) - I(\widehat{X};\widehat{Y}) \\ &= D_{KL}(p(X,Y), q(X,Y)), \end{aligned} \tag{3.7}$$

where $q(X,Y)$ is a distribution of the form

$$q(x,y) = p(\hat{x},\hat{y})p(x|\hat{x})p(y|\hat{y}), x \in \hat{x}, y \in \hat{y}.$$

Dhillon et al. [Dhillon 2003] used this method to simultaneously cluster words, represented by the variable X, and documents, represented by the variable Y. Note that X and Y take values in different alphabets. Thus, joining two states in X does not directly imply which states are joined in Y, and vice versa.

Slonim et al. [Slonim 2006] generalized the information bottleneck to multivariate variables and introduced different variations of the information bottleneck algorithm. In particular, the authors proposed a symmetric information bottleneck framework where both input and output variables are simultaneously merged. In this approach, each random variable is clustered using different partitions, obtaining the clustering of the variable X that maximally preserves the mutual information with Y and the clustering of the variable Y that maximally preserves the mutual information with X. Recently, the information bottleneck has also been extended by multi-view features and has been applied in classification and recognition tasks [Xu 2014].

Other related works are inspired by the non-negative matrix factorization (NMF) that decomposes a non-negative matrix by approximating a product of two non-negative factor matrices. NMF was proposed by Paatero and Tapper [Paatero 1994] and it has been demonstrated that the NMF method can be applied in image clustering and segmentation [Guan 2012a, Wang 2013b]. Several methods have been proposed for NMF [Wang 2013b], although each approach has advantages and disadvantages and the method of choice is often application dependent [Pauca 2004]. Recently, some efficient solvers have been proposed to improve the NMF method, such as NeNMF [Guan 2012b] that sequentially optimizes one matrix factor with another fixed by using Nesterov's method, MahNMF [Guan 2012a] that minimizes the Manhattan distance between the data matrix and its low-rank approximation and is more robust than the traditional NMF, or using Procrustes rotations [Huang 2014].

3.3 Mathematical framework

In this section, we present the mathematical framework of our clustering approach. We assume that we have a stationary Markov process, $\mathbf{X} = \{X_0, X_1, \ldots, X_t, X_{t+1}, \ldots\}$, defined over the alphabet $\mathscr{X}$. This alphabet is given by $\{1, 2, \ldots, N\}$ and represents 1D scalar values or high-dimensional data. From now on, we will consider that the stationary distribution of $\mathbf{X}$ is represented as $p(x_k) = Pr[X_t = k]$, where $k \in \mathscr{X}$. A summary of the notation used in this section is shown in Table 3.1.

As we have previously mentioned, our clustering criterion is to obtain the minimum loss of mutual information between consecutive states of the Markov process $\mathbf{X}$.

Proposition 1. *The total MI loss when merging the states i and j of a stationary Markov*

Table 3.1: Summary of the notation used in this thesis

$\mathbf{X}$	Markov process
$\mathcal{X}$	Alphabet where the Markov process takes values
$\widehat{\mathbf{X}}$	clustered Markov process
X_t	state of the Markov process at time t
$\widehat{X}_t$	state of the clustered Markov process at time t
$\widetilde{X}_t$	variable at time t that only considers the states that participate in the merging keeping their relative probabilities
$p(X_t)$	probability density function of the Markov process at time t
$p(x_i) = Pr[X_t = i]$	probability of the state i at time t of the stationary distribution of the Markov process
$p(x_i, x_j) = Pr[X_t = i, X_{t+1} = j]$	joint probability of the state i at time t and the state j at time $t+1$ of the Markov process
$p(X_{t+1} \vert x_i^t)$	probability density function of the Markov process at time $t+1$ conditioned on X_t takes the value i
$p(x_j^{t+1} \vert x_i^t)$	probability of the state j at time $t+1$ conditioned on the state i at time t

process $\mathbf{X}$ *is given by*

$$\begin{aligned}
\delta I_{\widehat{ij}}(\mathbf{X}) &= I(X_t;X_{t+1}) - I(\widehat{X}_t;\widehat{X}_{t+1}) \\
&= p(\hat{x})JS\left(\pi_i,\pi_j;p(X_{t+1}|x_i^t),p(X_{t+1}|x_j^t)\right) \\
&\quad +p(\hat{x})JS\left(\pi_i,\pi_j;p(X_t|x_i^{t+1}),p(X_t|x_j^{t+1})\right) \\
&\quad -p(\hat{x},\hat{x})I(\widetilde{X}_t;\widetilde{X}_{t+1}). \qquad (3.8)
\end{aligned}$$

where $p(\hat{x}) = p(x_i)+p(x_j)$, $p(\hat{x},\hat{x}) = \sum_{s=i,j}\sum_{t=i,j} p(x_s,x_t)$ *(i.e., the probability to stay in the same clustered state* $\widehat{ij}$ *during two consecutive states,* X_t *and* X_{t+1}*),* $\pi_j = \frac{p(x_j)}{p(\hat{x})}$*, and* $I(\widetilde{X}_t;\widetilde{X}_{t+1})$ *represents the mutual information between the variables* X_t *and* X_{t+1} *when only the states that participate in the merging are considered, that is,*

$$I(\widetilde{X}_t;\widetilde{X}_{t+1}) = \sum_{s=i,j}\sum_{t=i,j} p(\tilde{x}_s,\tilde{x}_t)\log\frac{p(\tilde{x}_s,\tilde{x}_t)}{p(\tilde{x}_s)p(\tilde{x}_t)}, \qquad (3.9)$$

where $p(\tilde{x}_s,\tilde{x}_t) = \frac{p(x_s,x_t)}{p(\hat{x},\hat{x})}$, $p(\tilde{x}_s) = \frac{\sum_{t=i,j} p(x_s,x_t)}{p(\hat{x},\hat{x})}$, *and* $p(\tilde{x}_t) = \frac{\sum_{s=i,j} p(x_s,x_t)}{p(\hat{x},\hat{x})}$.

Proof. The MI loss when two states i and j are merged to obtain the fused state $\widehat{ij}$ can be seen as a two-step clustering. In the first step, we consider the MI loss of the variable X_t, which is given by

$$\delta I^1_{\widehat{ij}}(\widehat{X}_t;X_{t+1}) = I(X_t;X_{t+1}) - I(\widehat{X}_t;X_{t+1})$$

$$= \quad p(\hat{x})JS\left(\pi_i, \pi_j; p(X_{t+1}|x_i^t), p(X_{t+1}|x_j^t)\right), \tag{3.10}$$

where $p(\hat{x}) = p(x_i) + p(x_j)$ and $\pi_i = \frac{p(x_i)}{p(\hat{x})}$, $\pi_j = \frac{p(x_j)}{p(\hat{x})}$. This process coincides with the standard agglomerative bottleneck method. Thus, the loss of mutual information is given by Equation 3.6.

In the second step, we consider the MI loss when merging the same states i and j of the variable X_{t+1}, once the same states of the variable X_t have already been merged. This information loss is given by

$$\begin{aligned}
\delta I^2_{\widehat{ij}}(\widehat{X}_t;\widehat{X}_{t+1}) &= I(\widehat{X}_t;X_{t+1}) - I(\widehat{X}_t;\widehat{X}_{t+1}) \\
&= p(\hat{x})JS\left(\pi'_i, \pi'_j; p(\widehat{X}_t|x_i^{t+1}), p(\widehat{X}_t|x_j^{t+1})\right) \\
&= p(\hat{x})JS\left(\pi'_i, \pi'_j; p(X_t|x_i^{t+1}), p(X_t|x_j^{t+1})\right) \\
&\quad -p(\hat{x},\hat{x})I(\widetilde{X}_t;\widetilde{X}_{t+1}),
\end{aligned} \tag{3.11}$$

where $\pi'_i = \frac{p(x_i)}{p(\hat{x})}$, $\pi'_j = \frac{p(x_j)}{p(\hat{x})}$, and $I(\widetilde{X}_t;\widetilde{X}_{t+1})$ represents the mutual information between the variables X_t and X_{t+1} when only the states that participate in the merging, i and j, are considered (see Equation 3.9).

The previous result can be proved as follows:

$$\begin{aligned}
\delta I^2_{\widehat{ij}}(\widehat{X}_t;\widehat{X}_{t+1}) &= I(\widehat{X}_t;X_{t+1}) - I(\widehat{X}_t;\widehat{X}_{t+1}) \\
&= p(\hat{x})JS\left((\pi'_i, \pi'_j; p(\widehat{X}_t|x_i^{t+1}), p(\widehat{X}_t|x_j^{t+1})\right) \\
&= p(\hat{x})JS(\pi'_i, \pi'_j; p(X_t|x_i^{t+1}), p(X_t|x_j^{t+1})) \\
&\quad +p(\hat{x})(-p(\hat{x}^t|\hat{x}^{t+1})\log p(\hat{x}^t|\hat{x}^{t+1}) + p(x_i^t|\hat{x}^{t+1})\log p(x_i^t|\hat{x}^{t+1}) \\
&\quad +p(x_j^t|\hat{x}^{t+1})\log p(x_j^t|\hat{x}^{t+1})) \\
&\quad -p(\hat{x})\pi'_i((-p(\hat{x}^t|x_i^{t+1})\log p(\hat{x}^t|x_i^{t+1}) + p(x_i^t|x_i^{t+1})\log p(x_i^t|x_i^{t+1}) \\
&\quad +p(x_j^t|x_i^{t+1})\log p(x_j^t|x_i^{t+1})) \\
&\quad -p(\hat{x})\pi'_j((-p(\hat{x}^t|x_j^{t+1})\log p(\hat{x}^t|x_j^{t+1}) + p(x_i^t|x_j^{t+1})\log p(x_i^t|x_j^{t+1}) \\
&\quad +p(x_j^t|x_j^{t+1})\log p(x_j^t|x_j^{t+1})) \\
&= p(\hat{x})JS\left(\pi'_i, \pi'_j; p(X_t|x_i^{t+1}), p(X_t|x_j^{t+1})\right) \\
&\quad -p(\hat{x},\hat{x})\log p(\hat{x},\hat{x}) + \sum_{k=i,j} p(x_k,\hat{x})\log p(x_k,\hat{x}) \\
&\quad + \sum_{l=i,j} p(\hat{x},x_l)\log p(\hat{x},x_l) - \sum_{k=i,j}\sum_{l=i,j} p(x_k,x_l)\log(p(x_k,x_l)) \\
&= p(\hat{x})JS\left(\pi'_i, \pi'_j; p(X_t|x_i^{t+1}), p(X_t|x_j^{t+1})\right) \\
&\quad - \sum_{k=i,j}\sum_{l=i,j} p(x_k,x_l)\log(p(x_k,x_l)) + p(\hat{x},\hat{x})\log(p(\hat{x},\hat{x}))
\end{aligned}$$

$$\begin{aligned}
& + \sum_{k=i,j} \sum_{l=i,j} p(x_k, x_l) \log \left(p(x_k^t | \hat{x}^t, \hat{x}^{t+1}) \right) \\
& + \sum_{k=i,j} \sum_{l=i,j} p(x_k, x_l) \log \left(p(x_k^{t+1} | \hat{x}^t, \hat{x}^{t+1}) \right) \\
= \quad & p(\hat{x}) JS \left(\pi_i', \pi_j'; p(X_t | x_i^{t+1}), p(X_t | x_j^{t+1})) \right) - p(\hat{x}, \hat{x}) I(\widetilde{X}_t; \widetilde{X}_{t+1})
\end{aligned}$$

Equation 3.8 can be obtained by adding the results of Equation 3.10 and Equation 3.11. □

In conclusion, three different terms contribute to the MI loss. The two first terms are given by the Jensen-Shannon divergence between the conditional probabilities between the states to be merged in the past to future direction and future to past direction, respectively. These terms have to be minimized to obtain the minimal MI loss and represent how the states are related to the other states. Observe that these terms coincide with the MI loss in the standard agglomerative information bottleneck algorithm. The third term is given by the mutual information between the merged states. This term has to be maximized and is related to how the merged states interact between them. Note that this term does not have an equivalent term in the standard agglomerative information bottleneck method.

3.3.1 Variation of the MI loss

One of the main drawbacks of this method compared to the original agglomerative information bottleneck algorithm is that the MI loss between any pair of states has to be recomputed when two other states are clustered. Fortunately, the computation of the variation of the MI loss can be computed in constant time for each pair of states to be clustered from the following result.

Proposition 2. *The variation of the MI loss associated with the states k and l when the states i and j have been merged is given by*

$$\begin{aligned}
\Delta \delta I_{\widehat{kl}}^{\widehat{ij}}(\widehat{\mathbf{X}}) \quad &= \quad \delta I_{\widehat{kl}}(\widehat{\mathbf{X}}_{\widehat{ij}}) - \delta I_{\widehat{kl}}(\widehat{\mathbf{X}}) \\
&= \quad -p(\hat{x}_{\widehat{kl}}, \hat{x}_{\widehat{ij}}) I(\widetilde{X}_t^{\widehat{kl}}; \widetilde{X}_{t+1}^{\widehat{ij}}) - p(\hat{x}_{\widehat{ij}}, \hat{x}_{\widehat{kl}}) I(\widetilde{X}_t^{\widehat{ij}}; \widetilde{X}_{t+1}^{\widehat{kl}}),
\end{aligned} \tag{3.12}$$

where $\widehat{X}$ represent the clustered variable by merging the states k and l before the merging of the states i and j, $\widehat{X}_{\widehat{ij}}$ represent the clustered variable after the merging of i and j, $p(\hat{x}_{\widehat{kl}}, \hat{x}_{\widehat{ij}}) = \sum_{s=k,l} \sum_{t=i,j} p(s,t)$, and $I(\widetilde{X}_t^{\widehat{kl}}; \widetilde{X}_{t+1}^{\widehat{ij}})$ is defined as

$$I(\widetilde{X}_t^{\widehat{kl}}; \widetilde{X}_{t+1}^{\widehat{ij}}) = \sum_{s=k,l} \sum_{t=i,j} p(\tilde{x}_s, \tilde{x}_t) \log \frac{p(\tilde{x}_s, \tilde{x}_t)}{p(\tilde{x}_s) p(\tilde{x}_t)} \tag{3.13}$$

and represents the mutual information when only i, j for X_t and k, l for X_{t+1} are considered.

Proof. The previous result can be proved as follows:

$$\begin{aligned}
\Delta\delta I_{\widehat{kl}}^{\widehat{ij}}(\widehat{\mathbf{X}}) &= p(\hat{x}_{\widehat{kl}})(-p(\hat{x}_{\widehat{ij}}^{t+1}|\hat{x}_{\widehat{kl}}^{t})\log p(\hat{x}_{\widehat{ij}}^{t+1}|\hat{x}_{\widehat{kl}}^{t}) + p(x_i^{t+1}|\hat{x}_{\widehat{kl}}^{t})\log p(x_i^{t+1}|\hat{x}_{\widehat{kl}}^{t}) \\
&\quad +p(x_j^{t+1}|\hat{x}_{\widehat{kl}}^{t})\log p(x_j^{t+1}|\hat{x}_{\widehat{kl}}^{t})) \\
&\quad -p(x_k)(-p(\hat{x}_{\widehat{ij}}^{t+1}|x_k^t)\log p(\hat{x}_{\widehat{ij}}^{t+1}|x_k^t) + p(x_i^{t+1}|x_k^t)\log p(x_i^{t+1}|x_k^t) \\
&\quad +p(x_j^{t+1}|x_k^t)\log p(x_j^{t+1}|x_k^t)) \\
&\quad -p(x_l)(-p(\hat{x}_{\widehat{ij}}^{t+1}|x_l^t)\log p(\hat{x}_{\widehat{ij}}^{t+1}|x_l^t) + p(x_i^{t+1}|x_l^t)\log p(x_i^{t+1}|x_l^t) \\
&\quad +p(x_j^{t+1}|x_l^t)\log p(x_j^{t+1}|x_l^t)) \\
&\quad +p(\hat{x}_{\widehat{kl}})(-p(\hat{x}_{\widehat{ij}}^{t}|\hat{x}_{\widehat{kl}}^{t+1})\log p(\hat{x}_{\widehat{ij}}^{t}|\hat{x}_{\widehat{kl}}^{t+1}) + p(x_i^{t}|\hat{x}_{\widehat{kl}}^{t+1})\log p(x_i^{t}|\hat{x}_{\widehat{kl}}^{t+1}) \\
&\quad +p(x_j^{t}|\hat{x}_{\widehat{kl}}^{t+1})) \\
&\quad -p(x_k)(-p(\hat{x}_{\widehat{ij}}^{t}|x_k^{t+1})\log p(\hat{x}_{\widehat{ij}}^{t}|x_k^{t+1}) + p(x_i^{t}|x_k^{t+1})\log p(x_i^{t}|x_k^{t+1}) \\
&\quad +p(x_j^{t}|x_k^{t+1})\log p(x_j^{t}|x_k^{t+1})) \\
&\quad -p(x_l)(-p(\hat{x}_{\widehat{ij}}^{t}|x_l^{t+1})\log p(\hat{x}_{\widehat{ij}}^{t}|x_l^{t+1}) + p(x_i^{t}|x_l^{t+1})\log p(x_i^{t}|x_l^{t+1}) \\
&\quad +p(x_j^{t}|x_l^{t+1})\log p(x_j^{t}|x_l^{t+1})) \\
&= -p(\hat{x}_{\widehat{kl}},\hat{x}_{\widehat{ij}})\log p(\hat{x}_{\widehat{kl}},\hat{x}_{\widehat{ij}}) + \sum_{r=k,l} p(x_r,\hat{x}_{\widehat{ij}})\log p(x_r,\hat{x}_{\widehat{ij}}) \\
&\quad + \sum_{q=i,j} p(\hat{x}_{\widehat{kl}},x_q)\log p(\hat{x}_{\widehat{kl}},x_q) - \sum_{r=k,l}\sum_{q=i,j} p(x_r,x_q)\log\big(p(x_r,x_q)\big) \\
&\quad -p(\hat{x}_{\widehat{ij}},\hat{x}_{\widehat{kl}})\log p(\hat{x}_{\widehat{ij}},\hat{x}_{\widehat{kl}}) + \sum_{r=i,j} p(x_r,\hat{x}_{\widehat{kl}})\log p(x_r,\hat{x}_{\widehat{kl}}) \\
&\quad + \sum_{q=k,l} p(\hat{x}_{\widehat{ij}},x_q)\log p(\hat{x}_{\widehat{ij}},x_q) - \sum_{r=i,j}\sum_{q=k,l} p(x_r,x_q)\log\big(p(x_r,x_q)\big) \\
&= -p(\hat{x}_{\widehat{kl}},\hat{x}_{\widehat{ij}})I(\widetilde{X}_t^{\widehat{kl}};\widetilde{X}_{t+1}^{\widehat{ij}}) - p(\hat{x}_{\widehat{ij}},\hat{x}_{\widehat{kl}})I(\widetilde{X}_t^{\widehat{ij}};\widetilde{X}_{t+1}^{\widehat{kl}}).
\end{aligned}$$

□

3.3.2 Reversible Markov process

A stationary Markov process is said to be reversible if $p(x_i,x_j) = p(x_j,x_i), \forall i,j \in \mathscr{X}$. In this situation, the two latter propositions can be simplified as follows. The MI loss when merging the states i and j is given by

$$\begin{aligned}
\delta I_{\widehat{ij}}(\mathbf{X}) &= I(X_t;X_{t+1}) - I(\widehat{X}_t;\widehat{X}_{t+1}) \\
&= 2p(\hat{x})JS\left(\pi_i,\pi_j;p(X_{t+1}|x_i^t),p(X_{t+1}|x_j^t)\right) \\
&\quad -p(\hat{x},\hat{x})I(\widetilde{X}_t;\widetilde{X}_{t+1}), \qquad (3.14)
\end{aligned}$$

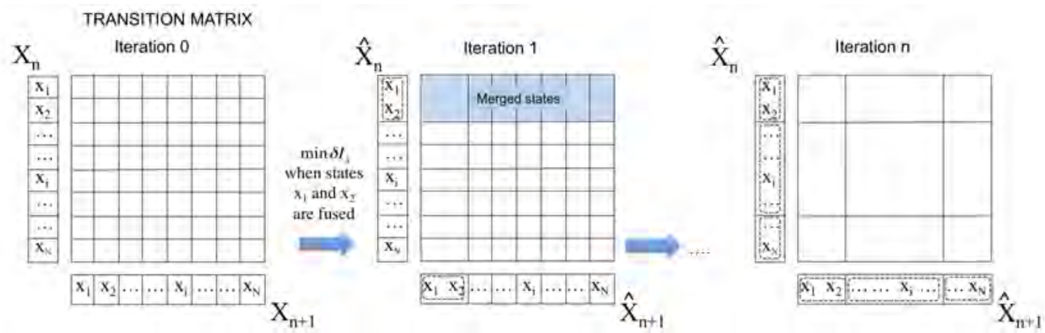


Figure 3.3: Diagram of the proposed agglomerative information bottleneck method

where $I(\widetilde{X}_t;\widetilde{X}_{t+1})$ has been defined in Equation 3.9. The variation of MI loss between the states k and l when the states i and j are merged is given by

$$\Delta\delta I^{\widehat{ij}}_{\widehat{kl}}(\widehat{\mathbf{X}}) \quad = \quad -2p(\hat{x}_{\widehat{kl}},\hat{x}_{\widehat{ij}})I(\widetilde{X}^{\widehat{kl}}_t;\widetilde{X}^{\widehat{ij}}_{t+1}), \qquad (3.15)$$

where $I(\widetilde{X}^{\widehat{kl}}_t;\widetilde{X}^{\widehat{ij}}_{t+1})$ has been defined in Equation 3.13.

3.4 Hierarchical clustering algorithms

In this section, we describe two different greedy algorithms based on the theoretical framework presented in the previous section. The first method is based on an agglomerative bottom-up strategy, while the second one is based on a divisive top-down strategy.

3.4.1 Agglomerative algorithm

Inspired by the agglomerative information bottleneck method applied to a single variable [Slonim 2000a], we propose a new clustering algorithm, the key idea of which is shown in Figure 3.3. The procedure requires as input the joint probability distribution between the two consecutive states of the Markov process, X_t and X_{t+1}, and the number of final clusters k. Initially, this algorithm assigns each state of the alphabet to a different cluster and, then, it greedily merges the pair of clusters that minimizes the loss of mutual information between these consecutive states. This last step is repeated until a pseudo-optimal partition in k clusters is obtained. Observe that this partition is not the global optimal partition, since a sequence of locally optimal merges does not guarantee a globally optimal result.

The proposed algorithm is described in Figure 3.4. First, the MI loss $\delta I_{\widehat{ij}}(\mathbf{X})$ due to the merging of each pair of states is computed and stored in a matrix. Then, the main loop begins by searching the minimum value of this matrix and, thus, finding the states m and n such that their merging causes the minimum loss of mutual information. Once s and t are identified, the MI loss associated to the states i and j, such that neither i nor j correspond to s or t, are recomputed according to Equation 3.12. Subsequently, the

```
Input
    Joint probability distribution: $p(X_t, X_{t+1})$
    Number of clusters: $k \in \{1..n\}$
Output
    A partition of $X$ into $k$ clusters
Computation
    $\widehat{X} \leftarrow X$
    $\forall i,j \in \{1..n\}$.compute($\delta I_{\widehat{ij}}(\mathbf{X})$) (see Eq. (3.8))
    while $|\widehat{X}| > k$ do
        $s,t \leftarrow \min_{i,j}(\delta I_{\widehat{ij}}(\widehat{\mathbf{X}}))$
        Update, using Eq. (3.12), $\delta I^{st}_{\widehat{ij}}(\widehat{\mathbf{X}})\ \forall i,j \in \{1..n\} | i,j \neq s,t$
        $\hat{x} \leftarrow \mathrm{merge}(x_s, x_t)$
        $\widehat{X} \leftarrow (\widehat{X} - \{x_s, x_t\}) \bigcup \{\hat{x}\}$
        Update $\delta I_{\widehat{\hat{x}j}}(\widehat{\mathbf{X}})\ \forall j \in \{1..n\}$
        Update $\delta I_{\widehat{j\hat{x}}}(\widehat{\mathbf{X}})\ \forall j \in \{1..n\}$
    end while
    return $\widehat{X}$
```

Figure 3.4: The agglomerative clustering algorithm

states s and t are merged and the MI loss between this new merged state and each one of the other states is computed. The main loop is repeated until the desired number of clusters is reached.

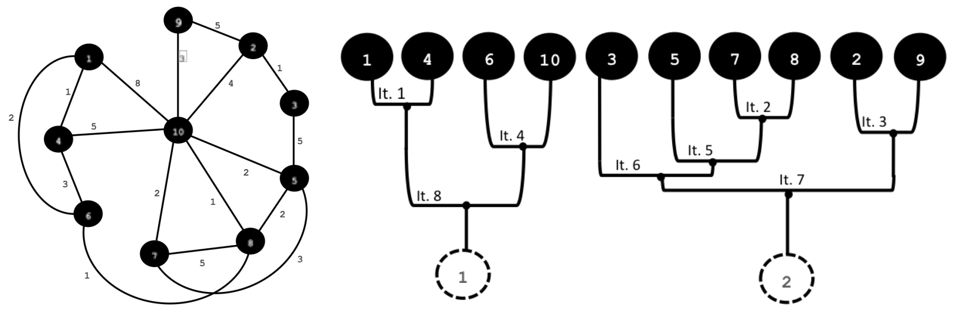
Additionally to the number of clusters, our procedure can also be stopped when a mutual information ratio (MIR) is achieved, where MIR is defined by

$$MIR(\mathbf{X}) = \frac{I(\widehat{X}_t; \widehat{X}_{t+1})}{I(X_t; X_{t+1})}. \qquad (3.16)$$

In this case, the number of clusters will depend on how fast the mutual information channel loses information throughout the agglomerative algorithm. Note that the final number of clusters will change between different Markov processes.

Observe that the main loop is repeated $n-k$ times and the computational cost of the loop is $O(n^2)$. Thus, the global computational cost is given by $O(n^3)$, where n is the number of states of the process $\mathbf{X}$ and it is considered that $n \gg k$. This cost is achieved thanks to the fact that the computational cost of Equation 3.12 is constant. Note that if the recomputation of the information loss for each pair of states was computed with the original Equation 3.8, with cost $O(n)$, the final cost would be given by $O(n^4)$. Let us note that some optimization strategies, such as the use of a heap structure instead of a distance matrix [Virmajoki 2004], would not reduce the computational complexity $O(n^3)$, since the computational complexity bottleneck is not only given by the minimum distance search (as in the pairwise nearest neighbour agglomerative method) but also by the distance computation.

To illustrate how the algorithm proceeds, this has been applied to a random walk on the weighted graph represented in Figure 3.5 (a). The nodes of this graph are labelled with a number and the edges have a positive weight W_{ij}, where i and j in-



(a) (b)

Figure 3.5: (a) Simple weighted graph and (b) the corresponding clustering hierarchy

dicate the nodes that are connected by the edge. For a random walk on a weighted graph, the probabilities of the stationary distribution are given by $p(x_i) = \frac{W_i}{2W}$, where $W_i = \sum_k W_{ik}$ is the sum of the weights that emanate from the node i, and W is the total sum of weights [Cover 1991]. The conditional probability is given by the expression $p(x_j^{t+1}|x_i^t) = \frac{W_{ij}}{W_i}$, that is, the weight of the edge between the nodes i and j divided by the sum of the edges that emanate from i. Figure 3.5 (b) shows, for each iteration, the nodes that have been clustered. In this case, the process performs eight iterations and returns two clusters: the first is composed of nodes 1, 4, 6, and 10, and the second by nodes 3, 5, 7, 8, 2, and 9.

3.4.2 Divisive algorithm

For some alphabets, each state represents the quantization of a scalar value that can be sorted in a certain order. In this case, a good strategy to hierarchically cluster this kind of variables consists in a top-down divisive clustering. To better illustrate the algorithm, we describe in Figure 3.6 how it proceeds. The method begins with an alphabet with a unique state and at each iteration a state is divided into two parts following a certain dissimilarity criterion. Since the states follow a given order, the partition can be simply given by a threshold value that splits one state into two different states. Note that, if the states represent categories, this strategy cannot be applied since the number of possible configurations at each step is a combinatorial value. In our approach, the criterion to decide the threshold value is given by the maximization of the mutual information gain.

This algorithm is described in Figure 3.7. First, the variable $\widehat{X}$ is initialized with a unique state $\hat{x}$. Then, the information gain that would be achieved if $\widehat{X}$ was divided according to the threshold i is computed. At this initial step, the partition at the threshold i divides $\widehat{X}$ into two symbols $\hat{x}_1$ and $\hat{x}_2$, where $\hat{x}_1$ and $\hat{x}_2$ are composed of the values in the partitions $\{1..i\}$ and $\{i+1..n\}$ of the original alphabet $\mathscr{X}$, respectively. Then, the main loop begins by searching the threshold that obtains a maximum gain of information and dividing the variable $\widehat{X}$ according to this threshold. Afterwards, the information gain associated with the splitting of $\hat{x}$ in the states $\hat{x}_i$ and $\hat{x}_j$ has to be computed using Equation 3.8, while the variation of the information gain associated with the other states has to be updated according to Equation 3.12. This main loop is repeated until

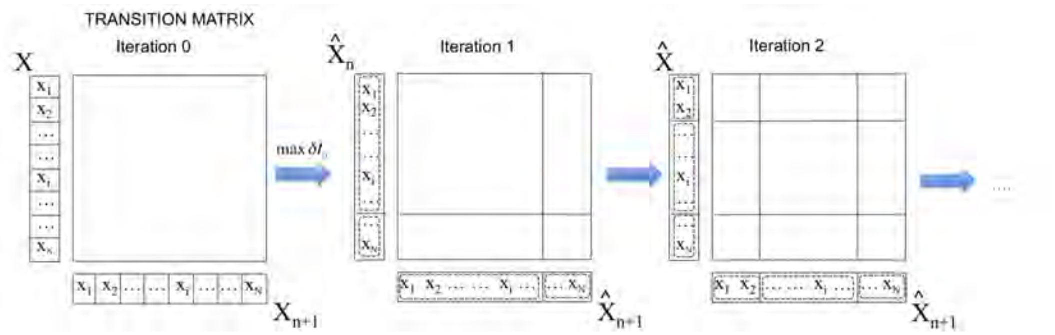


Figure 3.6: Diagram of the proposed divisive information bottleneck method

the final number of clusters, k, is reached. As in the agglomerative algorithm, the MIR (Equation 3.16) can be used as stopping criterion.

For this algorithm, the computational cost of the main loop, which is done k times, is given by $O(n^2)$. Thus, the final computational cost is given by $O(kn^2)$. Since in general $n \gg k$, this divisive algorithm will be faster than the agglomerative one. Unfortunately, the divisive algorithm can only be applied when the states of the random variable take sorted values and this reduces its application to a limited number of clustering problems. For instance, the divisive algorithm cannot be applied to the example shown in Figure 3.5.

In Figure 3.8, we illustrate how the divisive approach proceeds when it is used to segment the Lena image (Figure 3.8 (a)) using the stationary Markov process given by a random walk in the image (see Section 3.5.1 for more details). Figure 3.8 (b-e) shows the gain of mutual information according to the selected threshold value for the first four iterations of the method. In the first iteration, the algorithm computes the gain of mutual information of the channel for each threshold value (Figure 3.8 (b)). Then, the threshold that maximizes the mutual information is selected (in this case, the threshold is equal to 127) and the gains corresponding to the other thresholds are recomputed according to Equation 3.8, since, at the first iteration, all values belong to the cluster that has been split (Figure 3.8 (c)). Then, the threshold with maximum gain of mutual information is selected as the new threshold (in this case, the threshold is equal to 79). Note that the information gains of the thresholds greater than 127 (the first threshold) are updated according to Equation 3.12, since they do not belong to the cluster that has been split, while the information gains for the other thresholds are recomputed according to Equation 3.8 (Figure 3.8 (d)). This process is repeated for each iteration until the desired number of clusters is reached (Figure 3.8 (e)).

3.5 Results

To evaluate the proposed methods, we have applied the agglomerative and divisive methods to synthetic and photographic images. Image quantization consists in the reduction of the image intensities. This technique may be used to efficiently compress

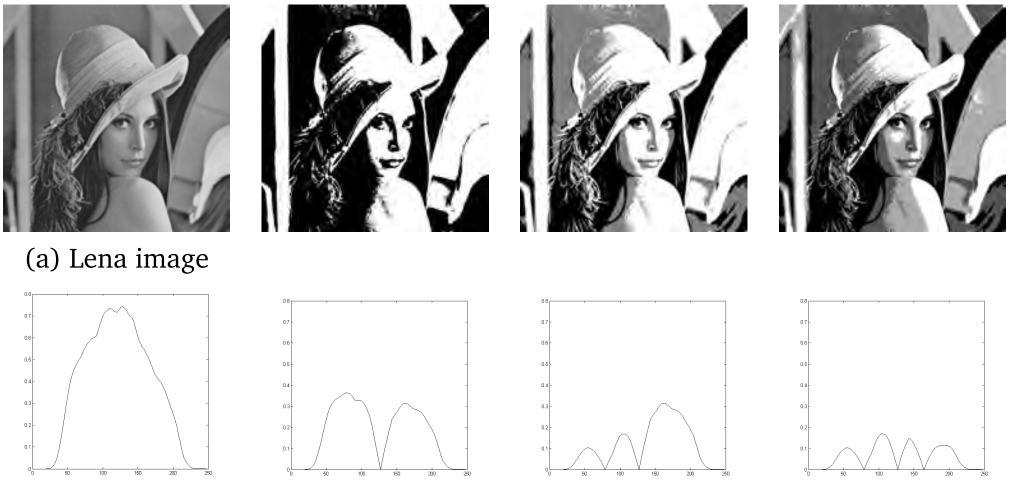
```
Input
   Joint probability distribution: $p(X_t,X_{t+1})$
   Number of clusters: $k \in \{1..n\}$
Output
   A partition of $X$ into $k$ clusters
Computation
   $\widehat{X} \leftarrow U$
   for $a = 1$ to $n-1$
      $i \leftarrow \{1..a\}$.
      $j \leftarrow \{a+1..n\}$.
      compute($\delta I_{ij}(\widehat{\mathbf{X}})$ (see Eq. (3.8))
   end for
   while $|\widehat{X}| < k$ do
      $s \leftarrow \max_a(\delta I_{ij}(\widehat{\mathbf{X}}))$
      $\{x_s, x_{s+1}\} \leftarrow$ divide($\hat{x}, s$)
      $\widehat{X} \leftarrow (\widehat{X} - \hat{x}) \bigcup \{\{x_s, x_{s+1}\}\}$
      Update $\delta I^s_{ij}(\widehat{\mathbf{X}})\ \forall i,j \in \{1..n\} | i,j \neq s$ (see Eq. (3.12))
      Update $\delta I(\widehat{\mathbf{X}})$ for $x_s$ and $x_{s+1}$
   end while
   return $\widehat{X}$
```

Figure 3.7: The divisive clustering algorithm

images and also to display images on devices that support a limited number of colors. During the quantization process, it is important to preserve the main structures or objects in the image. Quantization can be seen as a clustering problem, since the original intensity values are grouped into a short number of intensity bins. We propose to use the Markov process given by a random walk on the clustered image as the control process of the proposed algorithm. In this case, the method attempts to obtain a quantization with approximately equal area for each final intensity (to maximize the information content) and to get a high degree of correlation between neighbouring pixels (to maximize the mutual information of the channel). Both features are reasonable for a general image quantization scheme.

3.5.1 Application

To apply the proposed approach to image quantification, we need to establish how to compute the joint probability matrix of a random walk on an image. With this purpose, for each pixel at spatial coordinates (a, b) with intensity value specified by $f(a, b)$, we have considered its nearest four neighbour pixels at locations $(a-1, b)$, $(a+1, b)$, $(a, b-1)$ and $(a, b+1)$, that is, the 4-adjacency (see the book of Gonzalez and Woods [Gonzalez 2002]). After processing all pixels, we obtain the joint probability $p(x_i, x_j)$ as the number of pairs of adjacent pixels with intensity values i and j, respectively, divided by twice the total number of pixel pairs, since each pixel pair is counted twice. Note the symmetry of the joint probability matrix as $p(x_i, x_j) = p(x_j, x_i)$. From $p(x_i, x_j)$, the marginal probability is given by $p(x_i) = \sum_{j \in \mathscr{X}} p(x_i, x_j)$ and the condi-



(a) Lena image

(b) iteration 1 (c) iteration 2 (d) iteration 3 (e) iteration 4

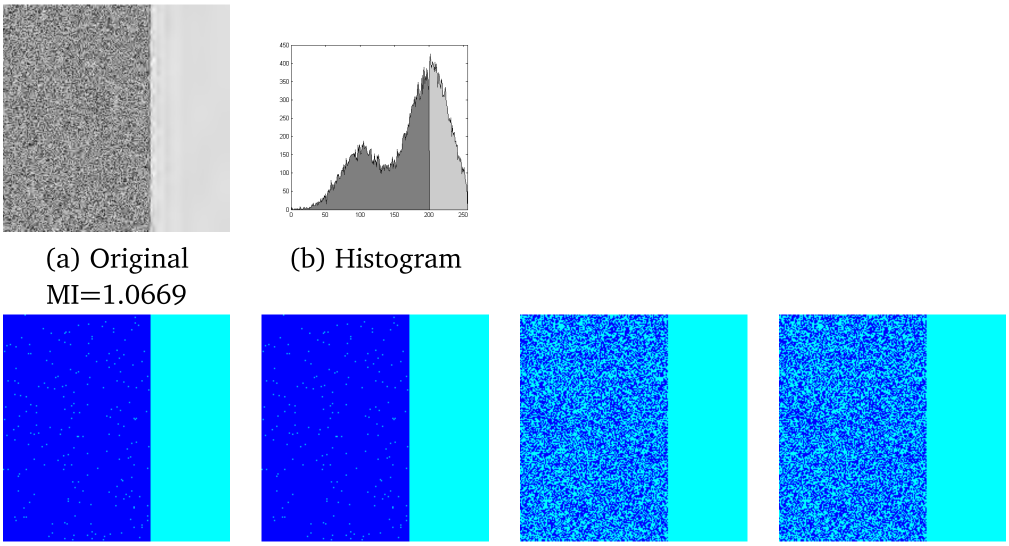
Figure 3.8: Clustering of Lena image using the divisive technique. The plots represent, for the first four iterations of the method, the gain of mutual information (y-axis) according to the threshold value (x-axis)

tional probability is given by $p(x_j^{t+1}|x_i^t) = \frac{p(x_i,x_j)}{p(x_i)}$. Both versions of the proposed approach, the agglomerative and divisive algorithms, can be used for image quantization since the image intensities follow a pre-defined order and, hence, a single threshold can separate the intensity values in clusters. These clusters preserve the correlation between neighbour intensity values and, thus, the spatial coherence of the resulting segmented image. Moreover, we assume that the intensities of a same object are more likely in neighbouring positions. In our approach, contrary to the method proposed by Bardera et al. [Bardera 2009], both variables X and Y of the information channel represent the image intensities. To carry out the tests, we have considered synthetic, photographic, and medical images.

3.5.2 Experiments

To show the effectiveness of the method, we have created a synthetic image (see Figure 3.9 (a)) with two regions, the left one with pixels between 0 and 199 and the right one between 200 and 255. The image was generated according to the histogram in Figure 3.9 (b) composed of two Gaussian distributions, one with mean 100, std. deviation 30 and weight 0.3 and another with mean 200, std. deviation 30 and weight 0.7. The dark grey area corresponds to the left side pixels, while the light grey area to the right side ones. Figure 3.9 shows, for the synthetic image of Figure 3.9 (a), the results and the MI value obtained with the agglomerative and the divisive methods considering 2 clusters. For comparison purposes, we show in Figure 3.9 (e) the results obtained with the classic Otsu's method [Otsu 1979] and in Figure 3.9 (f) with the classic k-means method [MacQueen 1967], both reporting the final MI value.

To compare our method with classic ones, we consider a first set of photographic im-



(a) Original
MI=1.0669

(b) Histogram

(c) Agglomerative
MI=0.84004

(d) Divisive
MI=0.84004

(e) Otsu
MI=0.0485

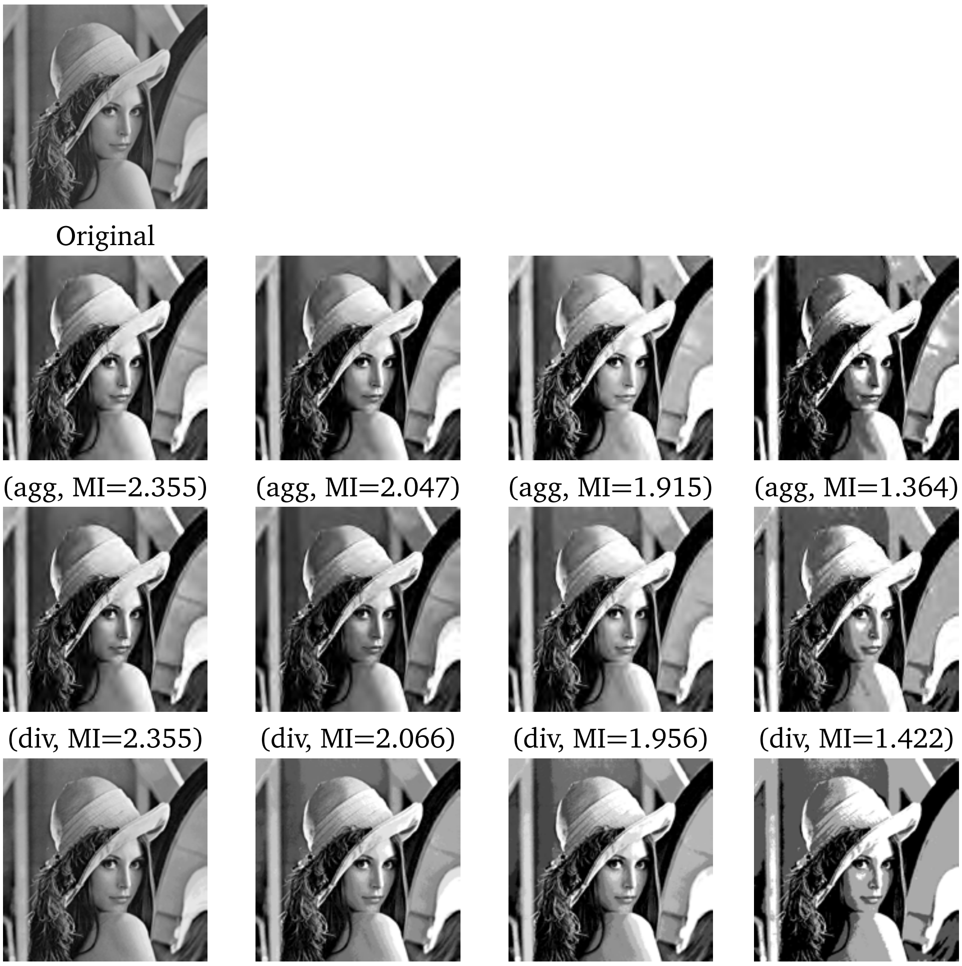
(f) *k*-means
MI=0.0503

Figure 3.9: From left to right, the original synthetic image, the proposed agglomerative and divisive clustering strategies, and other segmenation methods, Otsu and *k*-means, applied to the original synthetic image

ages comprising the well-known baboon, Lena, and peppers images. All of these have a resolution of 512×512 pixels. Figure 3.10 shows the original Lena image and the clustered images obtained with the proposed agglomerative and divisive, and the Otsu's method, considering 20, 10, 8, and 4 clusters. For each image we show the corresponding MI value. Additionally, as explained in Section 3.4.1, the proposed method can also be stopped when an MIR ratio is achieved. We use the same set of images to illustrate in Figure 3.11 the results with the divisive method considering an MIR value of 0.7, 0.8, and 0.9, respectively. The number of clusters and the MI are reported for each image.

For the tests, we consider a second set of photographic images from the Berkeley Segmentation Dataset and Benchmark [Martin 2001]. The BSDS500 dataset contains 200 test images with a total of 1063 different hand labelled clustered images. We segment the images using the agglomerative and divisive methods and compare the result with the 1063 ground-truth images. As stopping criterion, we consider the number of clusters of the corresponding ground-truth image.

In addition, we compare the result with the hierarchical clustering method available in the Matlab Statistics Toolbox$^{\text{TM}}$ [Mat 2013] using the Euclidean distance option. We also test with Non-negative Matrix Factorization methods (NMF) [Paatero 1994], including the recent implementations NeNMF [Guan 2012b] and MahNMF [Guan 2012a], that has been reported to surpass classic methods as *k*-means [Wang 2013b]. These methods require as input the histogram of the probabilities for each image bin and the number of clusters and return a cluster for each bin. Using these clusters, the final image is obtained.



Original

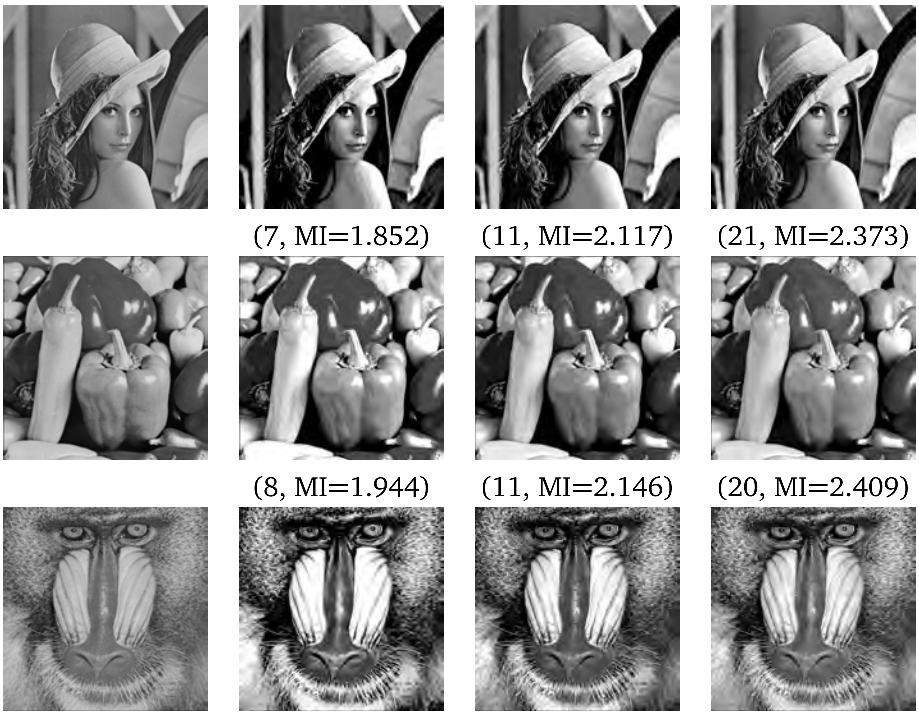
(agg, MI=2.355) (agg, MI=2.047) (agg, MI=1.915) (agg, MI=1.364)

(div, MI=2.355) (div, MI=2.066) (div, MI=1.956) (div, MI=1.422)

(Otsu, MI=2.312) (Otsu, MI=2.054) (Otsu, MI=1.948) (Otsu, MI=1.39)

(a) 20 clusters (b) 10 clusters (c) 8 clusters (d) 4 clusters

Figure 3.10: The original Lena image and, from top to bottom, the clustering obtained with the agglomerative (agg), divisive (div), and Otsu's methods considering, from left to right, 20, 10, 8, and 4 clusters, respectively. For each image we also report the MI value



(7, MI=1.852) (11, MI=2.117) (21, MI=2.373)

(8, MI=1.944) (11, MI=2.146) (20, MI=2.409)

(5, MI=0.690) (8, MI=0.798) (14, MI=0.878)

(a) Original (b) MIR=0.7 (d) MIR=0.8 (e) MIR=0.9

Figure 3.11: The original test images (Lena, peppers, and baboon) and the clustering obtained with the divisive method considering, from left to right, MIR values of 0.7, 0.8, and 0.9, respectively. We report the number of clusters and the MI value

Table 3.2: Average MI and Normalized Variation of Information values using the proposed methods on the BSDS500 Berkeley dataset, compared with the hierarchical clustering using Euclidean distance, NMF, NeNMF and MahNMF methods

Method	MI	NVI
Agglomerative	**1.83**	0.89
Divisive	1.68	**0.88**
Hierarchical	1.82	0.93
NMF	1.62	0.89
NeNMF	0.77	0.94
MahNMF	1.05	0.91

The variation of information (VI) was proposed as a measure to calculate the distance between two clustered images using the sum of conditional entropies [Meila 2005]. This measure is defined as

$$VI(X,Y) = H(X) + H(Y) - 2I(X;Y) = H(X|Y) + H(Y|X). \tag{3.17}$$

The segmented images in the Berkeley dataset differ in the number of clusters leading to a wide range of image entropy values. Since VI is an absolute measure that strongly depends on the original entropy values of the clustered images, we propose using a normalized measure to get a more accurate comparison. The normalized variation of information (NVI) is defined by

$$NVI(X,Y) = \frac{H(X|Y) + H(Y|X)}{H(X,Y)} = 1 - \frac{I(X;Y)}{H(X,Y)}. \tag{3.18}$$

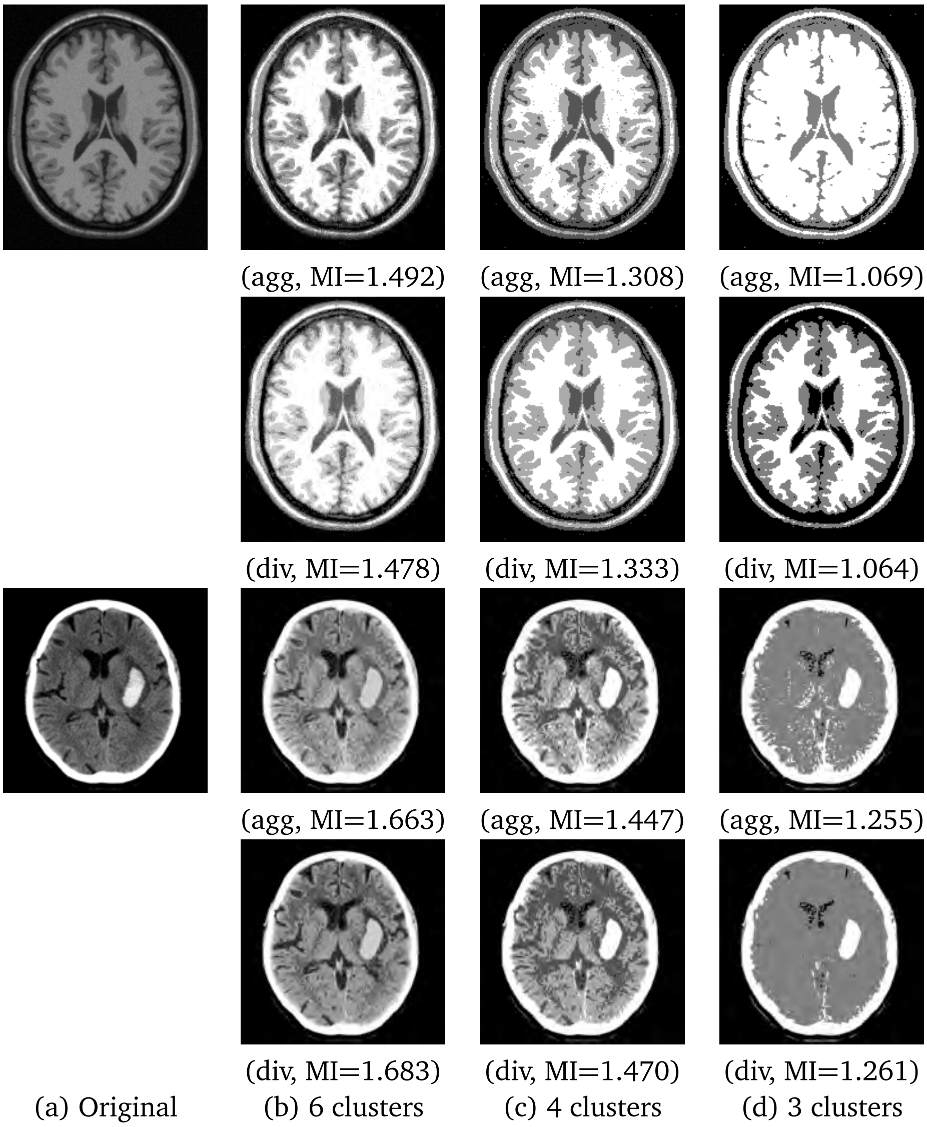
This measure takes values between 0 (when images are equal) and 1 (when images are independent).

Table 3.2 shows the averages of MI and NVI for all tested methods on the BSDS500 dataset.

The last set of images used for testing is composed of a computed tomography (CT) brain medical image (see Figure 3.12) of 420×420 pixels and a synthetic magnetic resonance (MR) brain image of 181×217 pixels from the Brainweb database. Figure 3.12 shows, for the CT and MR images, the obtained results considering 6, 4, and 3 clusters using both agglomerative and divisive methods with the corresponding MI.

To illustrate the MI variation in each partition, Figure 3.13 (a) presents the MI value with respect to the number of partitions for the agglomerative and divisive methods and Figure 3.13 (b) the difference between MI value of these methods.

Finally, Table 3.3 collects the agglomerative and divisive computation time using Lena image. Both approaches have been implemented using Matlab on a PC equipped with an Intel XEON E5 CPU and 16 GB of RAM.



(agg, MI=1.492) (agg, MI=1.308) (agg, MI=1.069)

(div, MI=1.478) (div, MI=1.333) (div, MI=1.064)

(agg, MI=1.663) (agg, MI=1.447) (agg, MI=1.255)

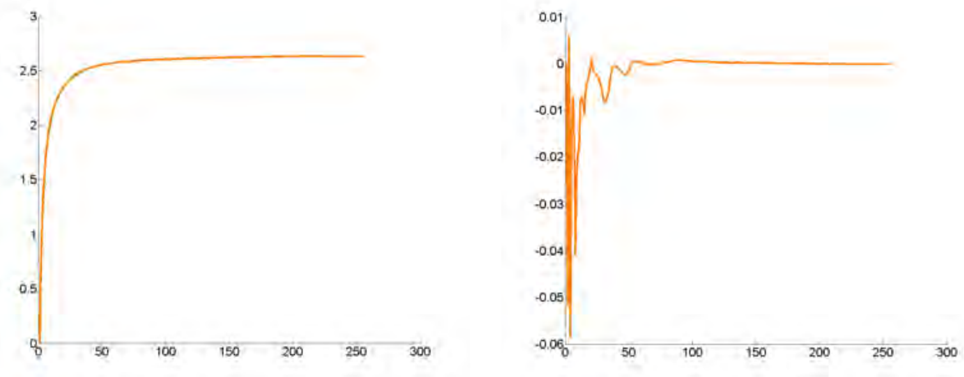
(div, MI=1.683) (div, MI=1.470) (div, MI=1.261)

(a) Original (b) 6 clusters (c) 4 clusters (d) 3 clusters

Figure 3.12: The proposed agglomerative (agg) and divisive (div) clustering strategies applied to segment (first and second rows) a synthetic MR brain image and (third and fourth rows) a CT brain. From left to right, the original images and the final clustering considering 6, 4, and 3 clusters, respectively. For each image we also report the gain of MI

Table 3.3: Computation time in seconds

Num. of clusters	Agglomerative	Divisive
20	25.27	5.5
10	25.36	2.80
8	25.47	2.37
4	25.61	1.32



(a) (b)

Figure 3.13: (a) Mutual information value (y-axis) with respect to the number of partitions (x-axis) for the agglomerative and divisive methods and (b) difference between agglomerative and divisive mutual information values.

3.6 Discussion

In this chapter, we have proposed a new hierarchical clustering method by extending the applicability of the agglomerative information bottleneck algorithm. The main feature of our approach is that instead of adopting a control variable, the different states of a stationary Markov process are clustered by maximally preserving the mutual information between two consecutive states of this process.

The proposed approach has been used to image quantization and has been tested on synthetic, photographic and medical images, including a benchmark dataset. It has also been compared with other methods in the literature. With the synthetic image we have seen that on maximizing MI, the obtained result is more representative than the ones obtained with other methods such as Otsu and k-means (see Figure 3.9). Note that the Markov process that leads the clustering algorithm is based on a random walk on the quantized image. From the fact that mutual information is the marginal entropy minus the conditional entropy (see Equation 2.12), by maximizing mutual information, we would expect a final clustered image with a similar amount of pixels for each cluster (high marginal entropy) and with a high correlation between neighboring pixels (low conditional entropy). Notice that the original image has two big different parts (dark blue and blue) and the result with the proposed method differentiates these two parts and only few pixels are misclassified. On the contrary, the Otsu's and k-means methods only cluster the intensities of the image depending on the histogram and, in this case, the results are not satisfactory. Notice that Otsu's and k-means methods do not take into account the neighboring information.

When it is applied to photographic images (see Figure 3.10), the higher the MI the better is the clustering since it will be closer to the optimal MI. We also observe that the highest MI values are achieved with the divisive method. Note the good separation between skin, hat, and hair, and the homogeneity of the spatial regions. When the results of the proposed approaches are compared with the Otsu results, we observe that, in general, the obtained MI values are higher with our approaches. This is the expected behavior since our approaches maximize the MI value, while the Otsu's approach optimizes other measures (minimize the intra-class variance and, consequently, maximize inter-class variance).

An interesting feature of our method is the stopping criterion of MIR. In this case, the algorithm is not stopped when a certain number of clusters is reached, but when a certain ratio of mutual information is obtained. This gives us a criterion that uses an adaptive number of clusters. In Figure 3.11, if we compare the number of clusters and the MI value for the different MIR, we observe that peppers and Lena images have a similar behaviour while the baboon image is decomposed into a lower number of clusters. This is due to the features of the original images. In the case of peppers and Lena images, illumination effects cause soft intensity variations with gradual transitions which are difficult to represent with a low number of clusters. In the baboon image, there are low-intensity variations due to illumination and, although there are highly-textured regions, it is easier to achieve a given MIR value with less number of clusters.

Testing with hand-labelled images and methods such as hierarchical clustering using Euclidean distance, NMF, NeMF and ManhNMF, on the Berkeley segmented dataset, we observe that the resulting images have significantly higher MI with the proposed methods. This behavior is expected since MI is the optimization criterion of the proposed hierarchical clustering methods. We also notice that the agglomerative MI value is slightly higher than the divisive MI, that indicates that the agglomerative strategy leads to a better measure optimization. With regard to the normalized variation of information measure, it can be seen that the proposed methods obtain the lowest values, i.e. there is more similarity between the manually labelled images and the resulting ones. The difference of NVI between the agglomerative and the divisive version is not significant.

When it is applied to medical images, we observe that both methods perfectly separate foreground from background in the case of MR images (see first and second rows of Figure 3.12). This fact is of special interest in many applications since it can enhance the interpretation process giving insights for diagnosis. In the case of six clusters, both methods perfectly delineate the main structures of the brain including: skull, white matter, grey matter, ventricles, and cerebrospinal fluids. With four clusters, background, skull, white matter and grey matter, and cerebrospinal fluid are very well separated. With 3 clusters, background, soft tissue and cerebrospinal fluids are easily distinguished. If we compare both agglomerative and divisive methods, the divisive one leads to better results. For instance, using the agglomerative method, pixels corresponding to other clusters can be observed in the white matter area. This kind of problem is not observed using the divisive method. In the case of CT images (third and fourth rows in Figure 3.12), we observe that the brain structures are not as well delineated as in the MR images. This is due to the features of the original images, as MR captures better soft tissues than CT and this affects the final clustering. However, observe how the damaged area is very well detected in all the cases. Note that with six clusters, bone and lesion are in different clusters although they have high intensity values. Comparing the agglomerative and divisive results, we observe that similar results are achieved in all the cases, except with three clusters. In this case, the divisive approach presents better results since soft tissue, cerebrospinal fluid and bone are well delineated, while with the agglomerative approach misclassified pixels corresponding to bone or lesions appear in the soft tissue cluster.

If we evaluate the MI value with respect to the number of partitions for the agglomerative and divisive methods (see Figure 3.13), both methods behave similarly although the divisive method reaches higher values for few partitions than the agglomerative one. From the difference between the MI value of the agglomerative and divisive methods, the negative values of the MI difference denote the better behavior of the divisive method.

With respect to computation time (see Table 3.3), although a more efficient GPU-based implementation could be designed, we observe that the agglomerative algorithm spends a lot of time in the first iterations, due to the huge size of the joint probability matrix, while in the last iterations the cost is minimal. On the contrary, in the divisive algorithm, while the cost of the first iterations is low, the computation time increases very quickly according to the number of clusters.

3.7 Conclusions

The agglomerative and divisive versions of a new hierarchical clustering approach that extends the information bottleneck method by substituting the control variable by a stationary Markov process that controls the clustering have been proposed. These algorithms extend the application fields of information bottleneck-based methods. In particular, a framework for image quantization based on a random walk on the image has been introduced. This application has been tested on different datasets and compared with other methods such as k-means, Otsu, hierarchical clustering, and NMF. The obtained results demonstrate the good performance of the method considering different quality measures. The experimental results are just a demonstration of the advantages of the proposed method and its applications.

The immediate future work is to perform an extensive evaluation of the agglomerative approach as a brain parcellation method, which is described in the following chapter. Other future work will be focused on an extensive evaluation of the proposed method on other possible application fields. We will also investigate the application of other clustering strategies similar to the sequential information bottleneck clustering [Slonim 2002] to the proposed approach. In addition, to improve the agglomerative algorithm performance, we will investigate some methods introduced to speed-up the pairwise nearest neighbor method [Fränti 2000, Virmajoki 2004].

CHAPTER 4

Brain Parcellation Based on Information Theory

Contents

4.1 Introduction

The connectome models the brain as a graph, where nodes represent brain areas and edges represent structural or functional connections. The first requirement in order to create this graph consists in applying a parcellation method to subdivide the brain cortex into different subregions according to a predefined criteria (i.e., anatomical, structure, function...) (see Section 2.2.2). The main motivation for creating a brain parcellation method is the strong restriction that the structural paths reflect in the brain functional localization. Originally, brain parcellations were created by using ex-vivo architectonic characteristics leading to the creation of anatomical atlas. Broadmann's atlas [Brodmann 1909] is the most popular so far, although it does not takes into account any structural or functional connectivity information.

In this chapter, we consider the agglomerative hierarchical method presented in Chapter 3 to parcellate the brain. Our approach models the brain functions as a random walk on the connectome network by using the connectivity matrix. This interpretation

Chapter 5

Complexity Measures Based on Information Theory

Contents

5.1 Introduction

Complexity measures are relevant for the analysis of the brain network, specially to describe topological features of the brain structure that may help to improve the understanding of the functionality of the system. Several measures have been proposed and showed successfully that are capable to associate disruptions with different diseases [van den Heuvel 2010, Meskaldji 2013, Sporns 2013, Crossley 2014]. However, it is still unknown which are the measures that describe best the brain network. For this reason, novel network measures are needed in order to better understand the brain structure and functioning [Papo 2014].

In this chapter, we present novel measures based on information theory to characterize weighted brain networks. Instead of measuring correlations between subsets to study the centrality and segregation (see section 2.2.3), as it has been done previously, we define the brain network as a stochastic process where neuronal impulses

CHAPTER 6

Conclusions

Contents

The human brain is a complex system formed by a massive network of neurons sharing information, whose behavior is challenging to characterize. The study of the brain network, also denoted as a connectome, is a growing research field which aims to understand the structure and the function of this fascinating organ. Although the large amount of methods, measures and ongoing projects studying the connectome, the complete understanding of the brain functioning is still far to be clear.

In this thesis, we have centered our interest on two main focus of research. First, on brain parcellation, which is a key step to perform brain studies since defines the regions to be analyzed. Second, on the definition of complex brain measures to better characterize brain properties. Below, a detailed description of the main contributions of this thesis as well as the publications related to each contribution are given.

6.1 Contributions

The aim of this work has been to investigate and provide new methods to improve the understanding of the human brain complexity at different scales by using information theory.

This aim has been achieved with the following proposals:

- New clustering method based on the information bottleneck

  Clustering techniques organize elements into groups (or clusters) whose members are similar and dissimilar to elements belonging to other clusters. A key element of these techniques is the definition of a similarity measure. The information bottleneck method provides a full solution of the clustering problem with no need to define a similarity measure. We have exploited this advantage to propose a new hierarchical clustering method. The main feature of our approach is that, instead of adopting a control variable, the different states of a stationary Markov process are clustered by maximally preserving the mutual information between two consecutive states of this process. We have presented both versions of the algorithm, the agglomerative and the divisive. The agglomerative approach, at each step, merges the pair of elements with a minimum loss of mutual information until the number of predefined clusters is reached. The divisive approach, at each step, divides the pair of elements with a higher gain of mutual information. These algorithms extend the application of the information bottleneck-based methods. The main advantage of this method is that by maximizing the MI, the obtained result is more representative than using other methods. An interesting feature of our method is the stopping criteria when a certain ratio of mutual information is obtained, which eliminates the requirement of defining a specific number of clusters a priori. This method has been tested on synthetic, photographic and medical images. The well-performance of the approach encouraged us to further investigate the method as a new brain parcellation technique.

This work has lead to the publication titled *Hierarchical clustering based on the information bottleneck method using a control process*, which has been published in Pattern Analysis and Applications, vol. 17, no. 3, pages 619-637, March 2015 [Bonmati 2015].

- New brain parcellation technique based on the information bottleneck clustering approach

  Brain parcellation is a fundamental procedure that consists in dividing the brain into smaller meaningful areas to define the regions of study. This procedure is usually done by an unsupervised clustering method or by registering with an atlas. In this thesis, we have considered the agglomerative clustering approach based on the information bottleneck to propose a new method to parcellate the brain connectome at different scales. We have proposed a brain model that allows the applicability of the clustering method by interpreting the brain as a stochastic process. The method is capable to cluster the brain regions while preserving the maximum information about the connectivity structure. Our approach takes into account the global connectivity pattern of the regions instead of similarity between fiber tracts as has been done previously. Using these approach, there is no need to define the number of clusters in advance. The method has been tested on synthetic model networks, functional and anatomical brain connectivity data considering different scales. The obtained parcellation preserves the main properties of the original network, with a higher value of mutual information and a lower clustering coefficient. The consistency across subjects demonstrates the robustness and the well-performance of our proposal.

  This work has lead to the publication titled *Brain parcellation based on information theory*, which has been submitted to the journal Computer Methods and Programs in Biomedicine.

- Novel complexity measures based on information theory

  Classical approaches model the brain network as a graph at which different information theory measures can be applied. In this work, we have considered the brain network as a stochastic process where neuronal impulses have been modeled as a random walk. Such a new interpretation has provided us a solid theoretical framework from which global and local measures have been derived. Global measures provide quantitative values for the whole-brain network characterization while local measures quantify the informativeness associated to each node. The proposed measures have been evaluated and compared with standard measures considering synthetic, structural and functional human networks at different scales. The obtained results have shown the uncertainty in predicting the next node, and how unique is the path that a node belongs to. The consistency across healthy subjects has demonstrated the robustness of the proposed measures.

This work, titled *Novel brain complexity measures based on information theory*, has been submitted to the Medical Image Analysis journal.

6.2 Future work

The work presented in this thesis can be extended and further investigated in different directions. With respect to the proposed clustering approach, the current implementation can be improved by speeding up the pairwise nearest neighbor method [Fränti 2000, Virmajoki 2004]. Additionally, it is worth mentioning that the method can be potentially used and applied in other fields such as medical imaging or feature extraction.

Referring to the actual approach of the parcellation method, future work will be focus on evaluating the benefits of clustering the regions with the same minimum loss of information at the same step, or adding new restrictions such as neighborhood information in the merging stage. On the other hand, we aim to further investigate the brain parcellation method by using different atlas to evaluate the consistency. An interesting aspect consist in extending the applicability at a voxel level, in this case, the aim would be to offer a novel hierarchical atlas that may improve the study of brain complexity by taking into account the pattern of the connections.

Regarding to the complexity measures, the human brain connectome can be further analyzed to detect specific disruptions due to a particular diseases using the proposed measures. In this case, measures could be used as a biomarkers, which is an ambitious application that would require the help of neurologists. In this work, we have focus on the whole-brain complexity, however, it would be interesting to focus the analysis on specific structures. Finally, these measures can be used to improve the visualization of the brain highlighting the properties presented.

Bibliography

[Achard 2006] S Achard, R Salvador, B Whitcher, J Suckling and E T Bullmore. *A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs*. Journal of Neuroscience, vol. 26, no. 1, pages 63–72, jan 2006. (Cited on pages IX, 18, 24 and 65.)

[Alexander 2001] A L Alexander, K M Hasan, M Lazar, J S Tsuruda and D L Parker. *Analysis of partial volume effects in diffusion-tensor MRI*. Journal of Magnetic Resonance, vol. 45, no. 5, pages 770–780, 2001. (Cited on page 14.)

[Anwander 2007] A Anwander, M Tittgemeyer, D Y von Cramon, A D Friederici and T R Knösche. *Connectivity-based parcellation of Broca's area*. Cerebral cortex, vol. 17, no. 4, pages 816–25, apr 2007. (Cited on page 61.)

[Baldassano 2015] C Baldassano, D M Beck and L Fei-Fei. *Parcellating connectivity in spatial maps*. PeerJ, vol. 3, no. e784, pages 1–24, jan 2015. (Cited on pages 62 and 64.)

[Bardera 2009] A Bardera, J Rigau, I Boada, M Feixas and M Sbert. *Image segmentation using the information bottleneck method*. IEEE Transactions on Image Processing, vol. 18, no. 7, pages 1601–1612, jul 2009. (Cited on page 49.)

[Basser 1994] P J Basser, J Mattiello and D LeBihan. *Estimation of the effective self-diffusion tensor from the NMR spin echo*. Journal of Magnetic Resonance, vol. 103, no. 3, pages 247–254, mar 1994. (Cited on page 10.)

[Basser 1996] P J Basser and C Pierpaoli. *Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI*. Journal of Magnetic Resonance, vol. 111, no. 86, pages 209–219, 1996. (Cited on page 10.)

[Baumgartner 1997] R Baumgartner, G Scarth, C Teichtmeister, R Somorjai and E Moser. *Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part I: reproducibility*. Journal of Magnetic Resonance Imaging, vol. 7, no. 6, pages 1094–101, jan 1997. (Cited on page 62.)

[Beckmann 2005] C F Beckmann, M DeLuca, J T Devlin and S M Smith. *Investigations into resting-state connectivity using independent component analysis*. Philosophical Transactions of the Royal Society, vol. 360, no. 1457, pages 1001–13, may 2005. (Cited on page 24.)

[Behrens 2007] T E J Behrens, H J Berg, S Jbabdi, M F S Rushworth and M W Woolrich. *Probabilistic diffusion tractography with multiple fibre orientations: what can we gain?* NeuroImage, vol. 34, no. 1, pages 144–155, jan 2007. (Cited on page 14.)

[Behrens 2014] T E J Behrens, S N Sotiropoulos and S Jbabdi. *MR diffusion tractography*. In Diffusion MRI, pages 429–451. Academic Press, San Diego, 2014. (Cited on pages IX, 14 and 15.)

[Bellec 2006] P Bellec, V Perlbarg, S Jbabdi, M Pélégrini-Issac, J-L Anton, J Doyon and H Benali. *Identification of large-scale networks in the brain using fMRI*. NeuroImage, vol. 29, no. 4, pages 1231–43, feb 2006. (Cited on page 62.)

[Berenschot 2003] G Berenschot. Visualization of diffusion tensor imaging. Master's thesis, Technische Universiteit Eindhoven, may 2003. (Cited on pages IX and 14.)

[Berres 2012] A Berres, M Goldau, M Tittgemeyer, G Scheuermann and H Hagen. *Tractography in context: multimodal visualization of probabilistic tractograms in anatomical context*. In Eurographics Workshop on Visual Computing for Biology and Medicine. The Eurographics Association, 2012. (Cited on pages IX and 9.)

[Björnemo 2002] M Björnemo, A Brun, R Kikinis and C-F Westin. *Regularized stochastic white matter tractography using diffusion tensor MRI*. In Medical Image Computing and Computer-Assisted Intervention, pages 435–442, Tokyo, Japan, 2002. (Cited on page 14.)

[Blumensath 2013] T Blumensath, S Jbabdi, M F Glasser, D C van Essen, K Ugurbil, T E J Behrens and S M Smith. *Spatially constrained hierarchical parcellation of the brain with resting-state fMRI*. NeuroImage, vol. 76, pages 313–24, aug 2013. (Cited on page 62.)

[Bonmati 2015] E Bonmati, A Bardera, I Boada, M Feixas and M Sbert. *Hierarchical clustering based on the information bottleneck method using a control process*. Pattern Analysis and Applications, vol. 18, no. 3, pages 619–637, mar 2015. (Cited on pages 63 and 104.)

[Böttger 2014] J Böttger, A Schäfer, G Lohmann, A Villringer and D S Margulies. *Three-dimensional mean-shift edge bundling for the visualization of functional connectivity in the brain*. IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 3, pages 471–80, mar 2014. (Cited on pages IX and 2.)

[Brodmann 1909] K Brodmann. Vergleichende lokalisationslehre der grosshirnrinde in ihren prinzipien dargestellt auf grund des zellenbaues. J. A. Barth, Leipzig, 1909. (Cited on pages 8, 18 and 59.)

[Bullmore 2009] E T Bullmore and O Sporns. *Complex brain networks: graph theoretical analysis of structural and functional systems*. Nature Reviews Neuroscience, vol. 10, no. 3, pages 186–198, mar 2009. (Cited on pages 24, 61, 65 and 86.)

[Bullmore 2011] E T Bullmore and D S Bassett. *Brain graphs: graphical models of the human brain connectome*. Annual Review of Clinical Psychology, vol. 7, pages 113–140, jan 2011. (Cited on page 20.)

[Burbea 1982] J Burbea and C R Rao. *On the convexity of some divergence measures based on entropy functions*. IEEE Transactions on Information Theory, vol. 28, no. 3, pages 489–495, may 1982. (Cited on pages 32 and 33.)

[Butts 2003] D A Butts. *How much information is associated with a particular stimulus?* Network: Computation in Neural Systems, vol. 14, no. 2, pages 177–187, 2003. (Cited on pages 30, 31 and 32.)

[Cai 2014] R Cai, Z Zhang, A K H Tung, C Dai and Z Hao. *A general framework of hierarchical clustering and its applications*. Information Sciences, vol. 272, no. 10, pages 29–48, 2014. (Cited on page 36.)

[Cammoun 2012] L Cammoun, X Gigandet, D E Meskaldji, J-P Thiran, O Sporns, K Q Do, P Maeder, R Meuli and P Hagmann. *Mapping the human connectome at multiple scales with diffusion spectrum MRI*. Journal of Neuroscience Methods, vol. 203, pages 386–397, jan 2012. (Cited on pages 65 and 87.)

[Campbell 2014] J S W Campbell, P MomayyezSiahkal, P Savadjiev, I R Leppert, K Siddiqi and G B Pike. *Beyond crossing fibers: bootstrap probabilistic tractography using complex subvoxel fiber geometries*. Frontiers in Neurology, vol. 5, page 216, jan 2014. (Cited on pages IX and 15.)

[Caspers 2013] S Caspers, S B Eickhoff, K Zilles and K Amunts. *Microstructural grey matter parcellation and its relevance for connectome analyses*. NeuroImage, vol. 80, pages 18–26, 2013. (Cited on page 18.)

[Cloutman 2012] L L Cloutman and M A Lambon Ralph. *Connectivity-based structural and functional parcellation of the human cortex using diffusion imaging and tractography*. Frontiers in Neuroanatomy, vol. 6, pages 1–18, 2012. (Cited on page 61.)

[Cohen 2008] A L Cohen, D A Fair, N U F Dosenbach, F M Miezin, D Dierker, D C van Essen, B L Schlaggar and S E Petersen. *Defining functional areas in individual human brains using resting functional connectivity MRI*. NeuroImage, vol. 41, no. 1, pages 45–57, may 2008. (Cited on page 62.)

[Colizza 2006] V Colizza, A Flammini, M A Serrano and A Vespignani. *Detecting rich-club ordering in complex networks*. Nature Physics, no. 2, pages 110–115, jan 2006. (Cited on page 26.)

[Conturo 1996] T E Conturo, N F Lori, T S Cull, E Akbudak, A Z Snyder, J S Shimony, R C McKinstry, H Burton and M E Raichle. *Tracking neuronal fiber pathways in the living human brain*. Proceedings of the National Academy of Sciences of the United States of America, vol. 35, no. 18, pages 10422–10427, aug 1996. (Cited on pages 10 and 13.)

[Cordes 2002] D Cordes, V Haughton, J D Carew, K Arfanakis and K Maravilla. *Hierarchical clustering to measure connectivity in fMRI resting-state data*. Magnetic Resonance Imaging, vol. 20, no. 4, pages 305–317, may 2002. (Cited on page 62.)

[Cover 1991] T M Cover and J A Thomas. Elements of information theory. Wiley Series in Telecommunications, 1991. (Cited on pages 27, 30, 31, 32, 33, 46, 63 and 79.)

[Craddock 2012] R C Craddock, G A James, P E Holtzheimer, X P Hu and H S Mayberg. *A whole brain fMRI atlas generated via spatially constrained spectral clustering*. Human brain mapping, vol. 33, no. 8, pages 1914–1928, aug 2012. (Cited on pages 61 and 62.)

[Craddock 2013] R C Craddock, S Jbabdi, C-G Yan, J T Vogelstein, F X Castellanos, A Di Martino, C Kelly, K Heberlein, S Colcombe and M P Milham. *Imaging human connectomes at the macroscale*. Nature Methods, vol. 10, no. 6, pages 524–39, jun 2013. (Cited on pages IX and 19.)

[Crossley 2014] N A Crossley, A Mechelli, J Scott, F Carletti, P T Fox, P McGuire and E T Bullmore. *The hubs of the human connectome are generally implicated in the anatomy of brain disorders*. Brain, vol. 137, no. 8, pages 2382–2395, 2014. (Cited on pages 24 and 77.)

[Crutchfield 1983] J P Crutchfield and N H Packard. *Symbolic dynamics of noisy chaos*. Physica 7D, vol. 7, pages 201–223, may 1983. (Cited on page 33.)

[Dai 2011] D Dai and H He. *VisualConnectome: Toolbox for brain network visualization and analysis*. In Human Brain Mapping, 2011. (Cited on page 94.)

[Damoiseaux 2009] J S Damoiseaux and M D Greicius. *Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity*. Brain Structure and Function, vol. 213, no. 6, pages 525–33, oct 2009. (Cited on pages 18, 20 and 24.)

[de Reus 2013] M A de Reus and M P van den Heuvel. *The parcellation-based connectome: limitations and extensions*. NeuroImage, vol. 80, pages 397–404, 2013. (Cited on pages 60 and 61.)

[Dennis 2012] E L Dennis, N Jahanshad, A W Toga, K McMahon, G I de Zubicaray, N G Martin, M J Wright and P M Thompson. *Test-retest reliability of graph theory measures of structural brain connectivity*. In Medical Image Computing and Computer-Assisted Intervention, volume 7512, pages 305–312. Springer, 2012. (Cited on page 87.)

[Desikan 2006] R S Desikan, F Ségonne, B Fischl, B T Quinn, B C Dickerson, D Blacker, R L Buckner, A M Dale, R P Maguire, B T Hyman, M S Albert and R J Killiany. *An automated labeling system for subdividing the human cerebral cortex on MRI scans* *into gyral based regions of interest*. NeuroImage, vol. 31, no. 3, pages 968–80, jul 2006. (Cited on page 61.)

[DeWeese 1999] M R DeWeese and M Meister. *How to measure the information gained from one symbol*. Network: Computation in Neural Systems, vol. 10, no. 4, pages 325–340, 1999. (Cited on pages 30, 31, 32, 83 and 84.)

[Dhillon 2003] I S Dhillon, S Mallela and D S Modha. *Information-theoretic co-clustering*. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 89–98, New York (NY), USA, 2003. ACM Press. (Cited on pages 38 and 39.)

[Duda 2001] R Duda, P Hart and D Stork. Pattern classification. Wiley, 2001. (Cited on page 35.)

[Edelman 2001] G M Edelman and J A Gally. *Degeneracy and complexity in biological systems*. Proceedings of the National Academy of Sciences, vol. 98, no. 24, pages 13763–13768, nov 2001. (Cited on page 79.)

[Everitt 2001] B Everitt, S Landau, M Leese and D Stahl. Cluster analysis. John Wiley and Sons Inc., 5th edition, 2001. (Cited on pages 35 and 36.)

[Feldman 1997] D P Feldman. *A brief introduction to: information theory, excess entropy and computational mechanics*, 1997. (Cited on page 34.)

[Feldman 1998] D P Feldman and J P Crutchfield. *Discovering noncritical organization: statistical mechanical, information theoreticand computational views of patterns in one-dimensional spin systems*. Working paper, Santa Fe Institute, Santa Fe (NM), USA, apr 1998. (Cited on page 33.)

[Feldman 2003] D P Feldman and J P Crutchfield. *Structural information in two-dimensional patterns: entropy convergence and excess entropy*. Physical Review E, vol. 67, no. 5, page 051104, may 2003. (Cited on page 81.)

[Felleman 1991] D J Felleman and D C Van Essen. *Distributed hierarchical processing in the primate cerebral cortex*. Cerebral Cortex, pages 1–47, 1991. (Cited on page 17.)

[Filzmoser 1999] P Filzmoser, R Baumgartner and E Moser. *A hierarchical clustering method for analyzing functional MR images*. Magnetic Resonance Imaging, vol. 17, no. 6, pages 817–826, jul 1999. (Cited on page 62.)

[Fornito 2010] A Fornito, A Zalesky and E T Bullmore. *Network scaling effects in graph analytic studies of human resting-state fMRI data*. Frontiers in Systems Neuroscience, vol. 4, page 22, jan 2010. (Cited on page 67.)

[Fornito 2013] A Fornito, A Zalesky and M Breakspear. *Graph analysis of the human connectome: promise, progress, and pitfalls*. NeuroImage, vol. 80, pages 426–444, 2013. (Cited on page 86.)

[Fränti 2000] P Fränti, T Kaukoranta, D-F Shen and K-S Chang. *Fast and memory efficient implementation of the exact PNN*. IEEE Transactions on Image Processing, vol. 9, no. 5, pages 773–777, may 2000. (Cited on pages 57 and 105.)

[Glasser 2013] M F Glasser, S N Sotiropoulos, J A Wilson, T S Coalson, B Fischl, J L R Andersson, J Xu, S Jbabdi, M Webster, J R Polimeni, D C van Essen and M Jenkinson. *The minimal preprocessing pipelines for the Human Connectome Project*. NeuroImage, vol. 80, pages 105–124, 2013. (Cited on page 88.)

[Golland 2008] Y Golland, P Golland, S Bentin and R Malach. *Data-driven clustering reveals a fundamental subdivision of the human cortex into two global systems*. Neuropsychologia, vol. 46, no. 2, pages 540–553, jan 2008. (Cited on page 62.)

[Gonzalez 2002] R Gonzalez and R Woods. Digital image processing. Prentice Hall, Upper Saddle River (NJ), USA, 2002. (Cited on page 48.)

[Gorbach 2011] N Gorbach. *Hierarchical information-based clustering for connectivity-based cortex parcellation*. Frontiers in Neuroinformatics, vol. 5, no. September, pages 1–13, 2011. (Cited on page 61.)

[Goutte 1999] C Goutte, P Toft, E Rostrup, F Å Nielsen and L K Hansen. *On clustering fMRI time series*. NeuroImage, vol. 9, no. 3, pages 298–310, mar 1999. (Cited on page 62.)

[Grassberger 1986] P Grassberger. *Toward a quantitative theory of self-generated complexity*. International Journal of Theoretical Physics, vol. 25, no. 9, pages 907–938, 1986. (Cited on page 33.)

[Guan 2012a] N Guan, D Tao, Z Luo and J Shawe-Taylor. *MahNMF: Manhattan non-negative matrix factorization*. Journal of Machine Learning Research, 2012. (Cited on pages 39 and 50.)

[Guan 2012b] N Guan, D Tao, Z Luo and B Yuan. *NeNMF: an optimal gradient method for nonnegative matrix factorization*. IEEE Transactions on Signal Processing, pages 2882–2898, 2012. (Cited on pages 39 and 50.)

[Hagmann 2003] P Hagmann, J-P Thiran, L Jonasson, P Vandergheynst, S Clarke, P Maeder and R Meuli. *DTI mapping of human brain connectivity: statistical fibre tracking and virtual dissection*. NeuroImage, vol. 19, no. 3, pages 545–554, jul 2003. (Cited on page 14.)

[Hagmann 2005] P Hagmann. *From diffusion MRI to brain connectomics*. PhD thesis, EPFL, Lausanne, 2005. (Cited on page 17.)

[Hagmann 2007] P Hagmann, M Kurant, X Gigandet, P Thiran, V J Wedeen, R Meuli and J-P Thiran. *Mapping human whole-brain structural networks with diffusion MRI*. PLoS One, vol. 2, no. 7, page e597, jan 2007. (Cited on pages 17 and 65.)

[Hagmann 2008] P Hagmann, L Cammoun, X Gigandet, R Meuli, C J Honey, V J Wedeen and O Sporns. *Mapping the structural core of human cerebral cortex*. PLoS Biology, vol. 6, no. 7, page e159, 2008. (Cited on pages IX, 18, 20, 21 and 23.)

[Hagmann 2010] P Hagmann, L Cammoun, X Gigandet, S Gerhard, P E Grant, V J Wedeen, R Meuli, J-P Thiran, C J Honey and O Sporns. *MR connectomics: principles and challenges*. Journal of Neuroscience Methods, vol. 194, no. 1, pages 34–45, 2010. (Cited on page 17.)

[Hansen 1997] P Hansen and B Jaumardi. *Cluster analysis and mathematical programming*. Mathematical Programming, vol. 79, pages 191–215, oct 1997. (Cited on page 35.)

[Harriger 2012] L Harriger, M P van den Heuvel and O Sporns. *Rich club organization of macaque cerebral cortex and its role in network communication*. PLoS One, vol. 7, no. 9, page e46497, 2012. (Cited on page 26.)

[Hartigan 1975] J Hartigan. Clustering algorithms. Wiley, 1975. (Cited on page 35.)

[He 2007] Y He, Z J Chen and A C Evans. *Small-world anatomical networks in the human brain revealed by cortical thickness from MRI*. Cerebral cortex, vol. 17, no. 10, pages 2407–19, oct 2007. (Cited on page 65.)

[Heller 2006] R Heller, D Stanley, D Yekutieli, N Rubin and Y Benjamini. *Cluster-based analysis of fMRI data*. NeuroImage, vol. 33, no. 2, pages 599–608, nov 2006. (Cited on page 62.)

[Hodge 2015] M R Hodge, W Horton, T Brown, R Herrick, T Olsen, M Hileman, M McKay, K A Archie, E Cler, M P Harms, G C Burgess, M F Glasser, J S Elam, S W Curtiss, D M Barch, R Oostenveld, L J Larson-Prior, K Ugurbil, D C Van Essen and D S Marcus. *ConnectomeDB: sharing human brain connectivity data*. NeuroImage, 2015. (Cited on page 88.)

[Honey 2007] C J Honey, R Kötter, M Breakspear and O Sporns. *Network structure of cerebral cortex shapes functional connectivity on multiple time scales*. Proceedings of the National Academy of Sciences of the United States of America, vol. 104, no. 24, pages 10240–5, jun 2007. (Cited on page 20.)

[Horn 2014] A Horn, D Ostwald, M Reisert and F Blankenburg. *The structural-functional connectome and the default mode network of the human brain*. NeuroImage, vol. 102, Part, no. 0, pages 142–151, 2014. (Cited on page 20.)

[Hsu 2007] C-C Hsu, C-L Chen and Y-W Su. *Hierarchical clustering of mixed data based on distance hierarchy*. Information Sciences, vol. 177, no. 20, pages 4474–4492, 2007. (Cited on page 36.)

[Huang 2014] K Huang, N D Sidiropoulos and A Swami. *Non-negative matrix factorization revisited: uniqueness and algorithm for symmetric decomposition*. IEEE

Transactions on Signal Processing, vol. 62, no. 1, pages 211–224, 2014. (Cited on page 39.)

[Jain 1988] A K Jain and R C Dubes. Algorithms for clustering data. Prentice-Hall, 1988. (Cited on pages 35 and 36.)

[Jain 1999] A K Jain, M Murty and P Flynn. *Data clustering: a review*. ACM Computing Surveys, vol. 31, no. 3, pages 264–323, 1999. (Cited on pages 35 and 36.)

[Jbabdi 2011] S Jbabdi and H Johansen-Berg. *Tractography: where do we go from here?* Brain Connectivity, vol. 1, no. 3, pages 169–83, jan 2011. (Cited on page 14.)

[Johansen-Berg 2004] H Johansen-Berg, T E J Behrens, M D Robson, I Drobnjak, M F S Rushworth, J M Brady, S M Smith, D J Higham and P M Matthews. *Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex*. Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 36, pages 13335–13340, sep 2004. (Cited on page 60.)

[Kaiser 2011] M Kaiser. *A tutorial in connectome analysis: topological and spatial features of brain networks*. NeuroImage, vol. 57, no. 3, pages 892–907, aug 2011. (Cited on page 24.)

[Kaufman 1990] L Kaufman and P J Rousseeuw. Finding groups in data: an introduction to cluster analysis. John Wiley, 1990. (Cited on page 36.)

[Kennedy 2013] H Kennedy, K Knoblauch and Z Toroczkai. *Why data coherence and quality is critical for understanding interareal cortical networks*. NeuroImage, vol. 80, pages 37–45, 2013. (Cited on pages 24, 26 and 86.)

[Kersting 2010] K Kersting, M Wahabzada, C Thurau and C Bauckhage. *Hierarchical convex NMF for clustering massive data*. In Proceedings of the 2nd Asian Conference on Machine Learning, pages 253–268, 2010. (Cited on page 36.)

[Kim 2010] J-H Kim, J-M Lee, H J Jo, S H Kim, J H Lee, S T Kim, S W Seo, R W Cox, D L Na, S I Kim and Z S Saad. *Defining functional SMA and pre-SMA subregions in human MFC using resting state fMRI: functional connectivity-based parcellation method*. NeuroImage, vol. 49, no. 3, pages 2375–86, feb 2010. (Cited on page 62.)

[Kindlmann 2004] G Kindlmann. *Superquadric tensor glyphs*. In Proceedings of the 6th Joint Eurographics - IEEE TCVG Symposium on Visualization, VISSYM'04, pages 147–154, Aire-la-Ville, Switzerland, Switzerland, 2004. Eurographics Association. (Cited on pages IX and 11.)

[Kiviniemi 2003] V Kiviniemi, J-H Kantola, J Jauhiainen, A Hyvärinen and O Tervonen. *Independent component analysis of nondeterministic fMRI signal sources*. NeuroImage, vol. 19, no. 2 Pt 1, pages 253–60, jun 2003. (Cited on page 16.)

[Klein 2007] J C Klein, T E J Behrens, M D Robson, C E Mackay, D J Higham and H Johansen-Berg. *Connectivity-based parcellation of human cortex using diffusion MRI: establishing reproducibility, validity and observer independence in BA 44/45 and SMA/pre-SMA*. NeuroImage, vol. 34, no. 1, pages 204–211, jan 2007. (Cited on page 61.)

[Knösche 2011] T R Knösche and M Tittgemeyer. *The role of long-range connectivity for the characterization of the functional-anatomical organization of the cortex*. Frontiers in Systems Neuroscience, vol. 5, page 58, jan 2011. (Cited on page 60.)

[Lam 2014] D Lam and D C Wunsch. Academic press library in signal processing. Elsevier, 2014. (Cited on page 35.)

[Latora 2001] V Latora and M Marchiori. *Efficient behavior of small-world networks*. Physical Review Letters, vol. 87, no. 19, page 198701, oct 2001. (Cited on page 26.)

[Lazar 2003a] M Lazar. *White matter tractography: an error analysis and human brain fiber tract reconstruction study*. PhD thesis, University of Utah, 2003. (Cited on page 13.)

[Lazar 2003b] M Lazar, D M Weinstein, J S Tsuruda, K M Hasan, K Arfanakis, M E Meyerand, B Badie, H A Rowley, V Haughton, A Field and A L Alexander. *White matter tractography using diffusion tensor deflection*. Human Brain Mapping, vol. 18, no. 4, pages 306–321, apr 2003. (Cited on page 14.)

[Lee 2012] M H Lee, C D Hacker, A Z Snyder, M Corbetta, D Zhang, E C Leuthardt and J S Shimony. *Clustering of resting state networks*. PLoS One, vol. 7, no. 7, page e40370, jan 2012. (Cited on page 62.)

[Leergaard 2012] T B Leergaard, C C Hilgetag and O Sporns. *Mapping the connectome: multi-level analysis of brain connectivity*. Frontiers in Neuroinformatics, vol. 6, no. May, pages 1–6, 2012. (Cited on page 60.)

[Liu 2012] X Liu, X H Zhu, P Qiu and W Chen. *A correlation-matrix-based hierarchical clustering method for functional connectivity analysis*. Journal of Neuroscience Methods, vol. 211, no. 1, pages 94–102, 2012. (Cited on page 62.)

[Lu 2003] Y Lu, T Jiang and Y F Zang. *Region growing method for the analysis of functional MRI data*. NeuroImage, vol. 20, no. 1, pages 455–65, sep 2003. (Cited on page 62.)

[MacQueen 1967] J MacQueen. *Some methods for classification and analysis of multivariate observations*. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–296, 1967. (Cited on pages 35 and 49.)

[Maggioni 2014] E Maggioni, M G Tana, F Arrigoni, C Zucca and A M Bianchi. *Constructing fMRI connectivity networks: a whole brain functional parcellation method for node definition*. Journal of Neuroscience Methods, vol. 228, pages 86–99, 2014. (Cited on page 62.)

[Margulies 2013] D S Margulies, J Böttger, A Watanabe and K J Gorgolewski. *Visualizing the human connectome*. NeuroImage, vol. 80, no. 0, pages 445–461, 2013. (Cited on pages IX and 11.)

[Marrelec 2008] G Marrelec, P Bellec, A Krainik, H Duffau, M Pélégrini-Issac, S Lehéricy, H Benali and J Doyon. *Regions, systems, and the brain: hierarchical measures of functional integration in fMRI*. Medical Image Analysis, vol. 12, no. 4, pages 484–496, 2008. (Cited on page 78.)

[Martin 2001] D Martin, C Fowlkes, D Tal and J Malik. *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics*. In IEEE International Conference on Computer Vision, volume 2, pages 416–423, jul 2001. (Cited on page 50.)

[Mat 2013] *Matlab statistics toolbox*. http://uk.mathworks.com/help/stats/hierarchical-clustering.html, 2013. (Cited on page 50.)

[McIntosh 1994] A R McIntosh, C L Grady, L G Ungerleider, J V Haxby, S I Rapoport and B Horwitz. *Network analysis of cortical visual pathways mapped with PET*. Journal of Neuroscience, vol. 14, no. 2, pages 655–66, feb 1994. (Cited on pages IX and 18.)

[McKeown 1998] M J McKeown, S Makeig, G G Brown, T P Jung, S S Kindermann, A J Bell and T J Sejnowski. *Analysis of fMRI data by blind separation into independent spatial components*. Human Brain Mapping, vol. 6, no. 3, pages 160–88, jan 1998. (Cited on pages 17 and 62.)

[Meila 2005] M Meila. *Comparing clusterings: an axiomatic view*. In Luc D Raedt and Stefan Wrobel, editeurs, Proceedings of the 22nd International Conference on Machine Learning, pages 577–584, 2005. (Cited on page 53.)

[Meskaldji 2013] D E Meskaldji, E Fischi-Gomez, A Griffa, P Hagmann, S Morgenthaler and J-P Thiran. *Comparing connectomes across subjects and populations at different scales*. NeuroImage, vol. 80, pages 416–425, 2013. (Cited on pages 24 and 77.)

[Messé 2015] A Messé, D Rudrauf, A Giron and G Marrelec. *Predicting functional connectivity from structural connectivity via computational models using MRI: an extensive comparison study*. NeuroImage, vol. 111, no. 0, pages 65–75, 2015. (Cited on page 87.)

[Meunier 2010] D Meunier, R Lambiotte and E T Bullmore. *Modular and hierarchically modular organization of brain networks*. Frontiers in Neuroscience, vol. 4, page 200, jan 2010. (Cited on page 61.)

[Mezer 2009] A Mezer, Y Yovel, O Pasternak, T Gorfine and Y Assaf. *Cluster analysis of resting-state fMRI time series*. NeuroImage, vol. 45, no. 4, pages 1117–25, may 2009. (Cited on page 62.)

[Mishra 2014] A Mishra, B P Rogers, L M Chen and J C Gore. *Functional connectivity-based parcellation of amygdala using self-organized mapping: a data driven approach*. Human Brain Mapping, vol. 35, no. 4, pages 1247–60, apr 2014. (Cited on page 62.)

[Moreno-Dominguez 2014] D Moreno-Dominguez, A Anwander and T R Knösche. *A hierarchical method for whole-brain connectivity-based parcellation*. Human Brain Mapping, vol. 35, no. 10, pages 5000–5025, oct 2014. (Cited on page 61.)

[Mori 1999] S Mori, B J Crain, V P Chacko and P C M van Zijl. *Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging*. Annals of Neurology, vol. 45, no. 2, pages 265–9, feb 1999. (Cited on page 13.)

[Mori 2001] S Mori, S Wakana, L M Nagae-Poetscher and P C M van Zijl. MRI atlas of human white matter. Elsevier, 2001. (Cited on page 12.)

[Mumford 2010] J A Mumford, S Horvath, M C Oldham, P Langfelder, D H Geschwind and R A Poldrack. *Detecting network modules in fMRI time series: a weighted network analysis approach*. NeuroImage, vol. 52, no. 4, pages 1465–76, oct 2010. (Cited on page 62.)

[Nagpal 2013] A Nagpal, A Jatain and D Gaur. *Review based on data clustering algorithms*. In IEEE Conference on Information and Communication Technologies, pages 298–303, apr 2013. (Cited on page 36.)

[Nakagawa 2013] T T Nakagawa, V K Jirsa, A Spiegler, A R McIntosh and G Deco. *Bottom up modeling of the connectome: linking structure and function in the resting brain and their changes in aging*. NeuroImage, vol. 80, pages 318–329, 2013. (Cited on page 20.)

[Nanetti 2009] L Nanetti, L Cerliani, V Gazzola, R Renken and C Keysers. *Group analyses of connectivity-based cortical parcellation using repeated k-means clustering*. NeuroImage, vol. 47, no. 4, pages 1666–1677, 2009. (Cited on page 61.)

[Newman 2003a] M E J Newman. *Fast algorithm for detecting community structure in networks*. Physical Review E, vol. 69, no. 2, page 5, sep 2003. (Cited on page 26.)

[Newman 2003b] M E J Newman. *The structure and function of complex networks*. SIAM Review, vol. 45, no. 2, pages 167–256, jan 2003. (Cited on page 26.)

[Ngan 1999] S C Ngan and X P Hu. *Analysis of functional magnetic resonance imaging data using self-organizing mapping with spatial connectivity*. Magnetic Resonance in Medicine, vol. 41, no. 5, pages 939–46, may 1999. (Cited on page 62.)

[O'Donnell 2013] L J O'Donnell, A J Golby and C-F Westin. *Fiber clustering versus the parcellation-based connectome*. NeuroImage, vol. 80, pages 283–289, 2013. (Cited on pages 60 and 61.)

[Ogawa 1990] S Ogawa, T M Lee, A R Kay and D W Tank. *Brain magnetic resonance imaging with contrast dependent on blood oxygenation*. Proceedings of the National Academy of Sciences of the United States of America, vol. 87, no. 24, pages 9868–9872, dec 1990. (Cited on page 16.)

[Otsu 1979] N Otsu. *A threshold selection method from gray-level histogram*. IEEE Transactions on Systems, Man and Cybernetics, vol. 9, no. 1, pages 62–66, 1979. (Cited on page 49.)

[Paatero 1994] P Paatero and U Tapper. *Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values*. Environmetrics, vol. 5, no. 2, pages 111–126, jun 1994. (Cited on pages 39 and 50.)

[Padula 2015] M C Padula, M Schaer, El Scariati, M Schneider, D van de Ville, M Debbané and S Eliez. *Structural and functional connectivity in the default mode network in 22q11.2 deletion syndrome*. Journal of neurodevelopmental disorders, vol. 7, no. 1, page 23, jan 2015. (Cited on page 65.)

[Papo 2014] D Papo, J M Buldú, S Boccaletti and E T Bullmore. *Complex network theory and the brain*. Philosophical Transactions of the Royal Society, vol. 369, no. 1653, pages 20130520–20130520, 2014. (Cited on pages 6, 24, 25 and 77.)

[Pauca 2004] V P Pauca, F Shahnaz, M W Berry and R J Plemmons. *Text mining using non-negative matrix factorization*. In Proceedings of the SIAM International Conference on Data Mining, pages 452–456, 2004. (Cited on page 39.)

[Peled 1998] S Peled, H Gudbjartsson, C-F Westin, R Kikinis and F A Jolesz. *Magnetic resonance imaging shows orientation and asymmetry of white matter tracts*. Brain Research, vol. 780, no. 1, pages 27–33, jan 1998. (Cited on page 10.)

[Prados 2012] F Prados, I Boada, M Feixas, A Prats-Galino, G Blasco, J Puig and S Pedraza. *Information-theoretic approach for automated white matter fiber tracts reconstruction*. Neuroinformatics, vol. 10, no. 3, pages 305–18, jul 2012. (Cited on page 14.)

[Pujol 2015] S Pujol, W Wells, C Pierpaoli, C Brun, J Gee, G Cheng, B Vemuri, O Commowick, S Prima, A Stamm, M Goubran, A Khan, T Peters, P Neher, K H Maier-Hein, Y Shi, A Tristan-Vega, G Veni, R Whitaker, M Styner, C-F Westin, S Gouttard, I Norton, L Chauvin, H Mamata, G Gerig, A Nabavi, A J Golby and R Kikinis. *The DTI Challenge: toward standardized evaluation of diffusion tensor imaging* *tractography for neurosurgery*. Journal of Neuroimaging, aug 2015. (Cited on page 13.)

[Qi 2015] S Qi, S Meesters, K Nicolay, B M ter Har Romeny and P Ossenblok. *The influence of construction methodology on structural brain network measures: a review*. Journal of neuroscience methods, vol. 253, pages 170–182, jun 2015. (Cited on page 13.)

[Ray 2015] K L Ray, D H Zald, S Bludau, M C Riedel, D Bzdok, J Yanes, K E Falcone, K Amunts, P T Fox, S B Eickhoff and A R Laird. *Co-activation based parcellation of the human frontal pole*. NeuroImage, aug 2015. (Cited on page 62.)

[Ribeiro 2015] A S Ribeiro, L M Lacerda and H A Ferreira. *Multimodal imaging brain connectivity analysis (MIBCA) toolbox*. PeerJ, vol. 3, page e1078, jul 2015. (Cited on page 87.)

[Rubinov 2010] M Rubinov and O Sporns. *Complex network measures of brain connectivity: uses and interpretations*. NeuroImage, vol. 52, no. 3, pages 1059–1069, sep 2010. (Cited on pages 24, 25, 65, 86 and 88.)

[Sato 2013] J R Sato, D Y Takahashi, M Q Hoexter, K B Massirer and A Fujita. *Measuring network's entropy in ADHD: a new approach to investigate neuropsychiatric disorders*. NeuroImage, vol. 77, no. 0, pages 44–51, 2013. (Cited on page 24.)

[Shannon 1948] C E Shannon. *A mathematical theory of communication*. The Bell System Technical Journal, vol. 27, pages 379–423,623–656, 1948. (Cited on pages 26 and 27.)

[Shattuck 2008] D W Shattuck, M Mirza, V Adisetiyo, C Hojatkashani, G Salamon, K L Narr, R A Poldrack, R M Bilder and A W Toga. *Construction of a 3D probabilistic atlas of human cortical structures*. NeuroImage, vol. 39, no. 3, pages 1064–80, mar 2008. (Cited on page 61.)

[Shaw 1984] R Shaw. The dripping faucet as a model chaotic system. Aerial Press, Santa Cruz (CA), USA, 1984. (Cited on page 33.)

[Simas 2015] T Simas, M Chavez, P R Rodriguez and A Diaz-Guilera. *An algebraic topological method for multimodal brain networks comparisons*. Frontiers in psychology, vol. 6, page 904, jan 2015. (Cited on page 65.)

[Slonim 2000a] N Slonim and N Tishby. *Agglomerative information bottleneck*. In Proceedings of the Neural Information Processing Systems, pages 617–623. MIT Press, 2000. (Cited on pages 37 and 44.)

[Slonim 2000b] N Slonim and N Tishby. *Document clustering using word clusters via the information bottleneck method*. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 208–215, Athens, Greece, 2000. ACM Press. (Cited on page 33.)

[Slonim 2002] N Slonim, N Friedman and N Tishby. *Unsupervised document classification using sequential information maximization*. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 129–136. ACM Press, 2002. (Cited on page 57.)

[Slonim 2005] N Slonim, G S Atwal, G Tkacik and W Bialek. *Information-based clustering*. Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 51, pages 18297–18302, dec 2005. (Cited on page 61.)

[Slonim 2006] N Slonim, N Friedman and N Tishby. *Multivariate information bottleneck*. Neural Computation, vol. 18, pages 1739–1789, 2006. (Cited on pages 36 and 39.)

[Smith 2013] S M Smith, C F Beckmann, J L R Andersson, E J Auerbach, J Bijsterbosch, G Douaud, E Duff, D A Feinberg, L Griffanti, M P Harms, M Kelly, T Laumann, K L Miller, S Moeller, S Petersen, J Power, G Salimi-Khorshidi, A Z Snyder, A T Vu, M W Woolrich, J Xu, E Yacoub, K Ugurbil, D C van Essen and M F Glasser. *Resting-state fMRI in the Human Connectome Project*. NeuroImage, vol. 80, pages 144–168, 2013. (Cited on pages IX and 9.)

[Sporns 2000] O Sporns, G Tononi and G M Edelman. *Connectivity and complexity: the relationship between neuroanatomy and brain dynamics*. Neural Networks, vol. 13, no. 8-9, pages 909–922, 2000. (Cited on page 78.)

[Sporns 2004] O Sporns, D R Chialvo, M Kaiser and C C Hilgetag. *Organization, development and function of complex brain networks*. Trends in Cognitive Sciences, vol. 8, no. 9, pages 418–425, sep 2004. (Cited on page 26.)

[Sporns 2005] O Sporns, G Tononi and R Kötter. *The human connectome: a structural description of the human brain*. PLoS Computational Biology, vol. 1, no. 4, page e42, sep 2005. (Cited on page 17.)

[Sporns 2011] O Sporns. *The human connectome: a complex network*. Annals of the New York Academy of Sciences, vol. 1224, no. 1, pages 109–125, apr 2011. (Cited on page 20.)

[Sporns 2013] O Sporns. *The human connectome: origins and challenges*. NeuroImage, vol. 80, pages 53–61, 2013. (Cited on pages 17, 24 and 77.)

[Sporns 2015] O Sporns. *Cerebral cartography and connectomics*. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, vol. 370, no. 1668, pages 20140173—-, may 2015. (Cited on page 60.)

[Stam 2007] C J Stam and J C Reijneveld. *Graph theoretical analysis of complex networks in the brain*. Nonlinear Biomedical Physics, vol. 1, no. 1, page 3, 2007. (Cited on page 24.)

[Stephan 2013] K E Stephan. *The history of CoCoMac*. NeuroImage, vol. 80, pages 46–52, 2013. (Cited on page 24.)

[Strehl 2002] A Strehl, J Ghosh and C Cardie. *Cluster ensembles - a knowledge reuse framework for combining multiple partitions*. Journal of Machine Learning Research, vol. 3, pages 583–617, mar 2002. (Cited on page 64.)

[Szépfalusy 1986] P Szépfalusy and G Györgyi. *Entropy decay as a measure of stochasticity in chaotic systems*. Physical Review A, vol. 33, no. 4, page 2852, 1986. (Cited on page 33.)

[Tao 2007] D Tao, X Li, X Wu and S J Maybank. *General tensor discriminant analysis and Gabor features for gait recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 10, pages 1700–1715, oct 2007. (Cited on page 36.)

[Thirion 2006] B Thirion, G Flandin, P Pinel, A Roche, P Ciuciu and J-B Poline. *Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets*. Human Brain Mapping, vol. 27, no. 8, pages 678–93, aug 2006. (Cited on page 62.)

[Thirion 2014] B Thirion, G Varoquaux, E Dohmatob and J-B Poline. *Which fMRI clustering gives good brain parcellations?* Frontiers in Neuroscience, vol. 8, page 167, jan 2014. (Cited on page 62.)

[Tishby 1999] N Tishby, F Pereira and W Bialek. *The information bottleneck method*. In Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, pages 368–377, 1999. (Cited on pages 36, 37 and 62.)

[Tononi 1994] G Tononi, O Sporns and G M Edelman. *A measure for brain complexity: relating functional segregation and integration in the nervous system*. Proceedings of the National Academy of Sciences, vol. 91, no. 11, pages 5033–5037, may 1994. (Cited on page 78.)

[Tononi 1996] G Tononi, O Sporns and G M Edelman. *A complexity measure for selective matching of signals by the brain*. Proceedings of the National Academy of Sciences of the United States of America, vol. 93, no. 8, pages 3422–3427, apr 1996. (Cited on page 78.)

[Tononi 1998a] G Tononi, G M Edelman and O Sporns. *Complexity and coherency: integrating information in the brain*. Trends in Cognitive Sciences, vol. 2, no. 12, pages 474–484, 1998. (Cited on page 78.)

[Tononi 1998b] G Tononi, A R McIntosh, D P Russell and G M Edelman. *Functional clustering: identifying strongly interactive brain regions in neuroimaging data*. NeuroImage, vol. 7, no. 2, pages 133–149, 1998. (Cited on page 79.)

[Tononi 1999] G Tononi, O Sporns and G M Edelman. *Measures of degeneracy and redundancy in biological networks*. Proceedings of the National Academy of Sciences of the United States of America, vol. 96, no. 6, pages pp. 3257–3262, 1999. (Cited on page 79.)

[Tuch 2002] D S Tuch, T G Reese, M R Wiegell, N Makris, J W Belliveau and V J Wedeen. *High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity*. Magnetic Resonance in Medicine, vol. 48, no. 4, pages 577–582, oct 2002. (Cited on page 14.)

[Tzourio-Mazoyer 2002] N Tzourio-Mazoyer, B Landeau, D Papathanassiou, F Crivello, O Etard, N Delcroix, B Mazoyer and M Joliot. *Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain*. NeuroImage, vol. 15, no. 1, pages 273–289, 2002. (Cited on pages 61 and 65.)

[van den Heuvel 2008] M P van den Heuvel, R Mandl and H E Hulshoff Pol. *Normalized cut group clustering of resting-state fMRI data*. PLoS One, vol. 3, no. 4, page e2001, jan 2008. (Cited on page 62.)

[van den Heuvel 2010] M P van den Heuvel, H E Hulshoff Pol and H E H Pol. *Exploring the brain network: a review on resting-state fMRI functional connectivity*. European Neuropsychopharmacology, vol. 20, no. 8, pages 519–534, 2010. (Cited on pages 24 and 77.)

[van den Heuvel 2012] M P van den Heuvel, R S Kahn, J Goñi and O Sporns. *High-cost, high-capacity backbone for global brain communication*. Proceedings of the National Academy of Sciences of the United States of America, vol. 109, no. 28, pages 11372–11377, 2012. (Cited on page 26.)

[Van Essen 2012] D C Van Essen, K Ugurbil, E J Auerbach, D M Barch, T E J Behrens, R Bucholz, A Chang, L Chen, M Corbetta, S W Curtiss, S D Penna, D A Feinberg, M F Glasser, N Harel, A C Heath, L J Larson-Prior, D S Marcus, G Michalareas, S Moeller, R Oostenveld, S E Petersen, F Prior, B L Schlaggar, S M Smith, A Z Snyder, J Xu and E Yacoub. *The Human Connectome Project: a data acquisition perspective*. NeuroImage, vol. 62, no. 4, pages 2222–2231, 2012. (Cited on page 88.)

[van Horn 2012] J D van Horn, A Irimia, C M Torgerson, M C Chambers, R Kikinis and A W Toga. *Mapping connectivity damage in the case of Phineas Gage*. PLoS One, vol. 7, no. 5, page e37454, jan 2012. (Cited on pages IX, 20 and 22.)

[Verdú 2008] S Verdú and T Weissman. *The information lost in erasures*. IEEE Transactions on Information Theory, vol. 54, no. 11, pages 5030–5058, nov 2008. (Cited on page 81.)

[Versalius 1564] A Versalius. De humani corporis fabrica libri septem. Basileae, 1564. (Cited on pages IX and 7.)

[Virmajoki 2004] O Virmajoki. *Pairwise nearest neighbor method revisited*, 2004. (Cited on pages 45, 57 and 105.)

[Viviani 2005] R Viviani, G Grön and M Spitzer. *Functional principal component analysis of fMRI data*. Human Brain Mapping, vol. 24, no. 2, pages 109–29, feb 2005. (Cited on page 62.)

[Wang 2013a] Y Wang and T-Q Li. *Analysis of whole-brain resting-state fMRI data using hierarchical clustering approach*. PLoS One, vol. 8, no. 10, page e76315, 2013. (Cited on page 61.)

[Wang 2013b] Y-X Wang and Y-J Zhang. *Nonnegative matrix factorization: a comprehensive review*. IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, pages 1336–1353, jun 2013. (Cited on pages 39 and 50.)

[Watts 1998] D J Watts and S H Strogatz. *Collective dynamics of 'small-world' networks*. Nature, vol. 393, no. 6684, pages 409–410, 1998. (Cited on pages 25 and 26.)

[Wedeen 2008] V J Wedeen, R P Wang, J D Schmahmann, T Benner, W Y I Tseng, G Dai, D N Pandya, P Hagmann, H D'Arceuil and A J de Crespigny. *Diffusion spectrum magnetic resonance imaging (DSI) tractography of crossing fibers*. NeuroImage, vol. 41, no. 4, pages 1267–1277, jul 2008. (Cited on page 11.)

[Westin 1994] C-F Westin. *A tensor framework for multidimensional signal processing*. PhD thesis, Linköping University, Sweden, S-581 83 Linköping, Sweden, 1994. (Cited on page 10.)

[Westin 1997] C-F Westin, S Peled, H Gudbjartsson, R Kikinis and F A Jolesz. *Geometrical diffusion measures for MRI from tensor basis analysis*. In ISMRM '97, page 1742, Vancouver Canada, apr 1997. (Cited on page 10.)

[Wu 2013] G-R Wu, W Liao, S Stramaglia, J-R Ding, H Chen and D Marinazzo. *A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data*. Medical Image Analysis, vol. 17, no. 3, pages 365–374, 2013. (Cited on page 20.)

[Xia 2013] M Xia, J Wang and Y He. *BrainNet Viewer: a network visualization tool for human brain connectomics*. PLoS One, vol. 8, no. 7, page e68910, jan 2013. (Cited on pages 65 and 69.)

[Xu 2005] R Xu and D C Wunsch. *Survey of clustering algorithms*. IEEE Transactions on Neural Networks, vol. 16, no. 3, pages 645–678, may 2005. (Cited on page 35.)

[Xu 2014] C Xu, D Tao and C Xu. *Large-margin multi-view information bottleneck*. IEEE TTransactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 8, pages 1559–1572, aug 2014. (Cited on pages 36 and 39.)

[Yeung 2002] R W Yeung. A first course in information theory. Springer, 2002. (Cited on pages 27 and 30.)

[Yu 2011] J Yu, D Liu, D Tao and H S Seah. *Complex object correspondence construction in two-dimensional animation*. IEEE Transactions on Image Processing, vol. 20, no. 11, pages 3257–3269, 2011. (Cited on page 35.)

[Yu 2012] J Yu, M Wang and D Tao. *Semisupervised multiview distance metric learning for cartoon synthesis*. IEEE Transactions on Image Processing, vol. 21, no. 11, pages 4636–4648, 2012. (Cited on page 35.)

[Yu 2013] J Yu and D Tao. Modern machine learning techniques and their applications in cartoon animation research. Wiley-IEEE Press, 2013. (Cited on page 35.)

[Zalesky 2010] A Zalesky, A Fornito, I H Harding, L Cocchi, M Yücel, C Pantelis and E T Bullmore. *Whole-brain anatomical networks: does the choice of nodes matter?* NeuroImage, vol. 50, no. 3, pages 970–83, apr 2010. (Cited on page 62.)

[Zhao 2005] Y Zhao, G Karypis and U M Fayyad. *Hierarchical clustering algorithms for document datasets*. Data Mining and Knowledge Discovery, vol. 10, no. 2, pages 141–168, 2005. (Cited on page 36.)

[Zhao 2012] X Zhao, Y Liu, X Wang, B Liu, Q Xi, Q Guo, H Jiang, T Jiang and P Wang. *Disrupted small-world brain networks in moderate Alzheimer's disease: a resting-state fMRI study*. PLoS One, vol. 7, no. 3, page e33540, jan 2012. (Cited on page 65.)

[Zhou 2006] C Zhou, L Zemanová, G Zamora, C C Hilgetag and J Kurths. *Hierarchical organization unveiled by functional connectivity in complex brain networks*. Physical Review Letters, vol. 97, no. 23, page 238103, dec 2006. (Cited on page 21.)