Unsupervised Connectivity-based Cortex Parcellation using the Information Bottleneck Method

A dissertation submitted to the

UNIVERSITY OF CAPE TOWN

Division of Biomedical Engineering

for the degree of

MASTER OF SCIENCES

in Biomedical Engineering

presented by

NICO STEPHAN GORBACH
BSc in Electromechanical Engineering, UCT
born 8th of March 1986

Thesis Advisors:

Prof. Dr. Tania Douglas
Dr. Marc Tittgemeyer

in collaboration with

Max Planck Institute for Neurological Research
Cologne, Germany

2011

Declaration

I, Nico Stephan Gorbach, herewith declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: .....

Nico Stephan Gorbach

Date: .....

Abstract

One of the most promising avenues for compiling connectivity data originates from the notion that individual brain regions maintain individual connectivity profiles; the functional repertoire of a cortical area (the functional fingerprint) is closely related to its anatomical connections (the connectional fingerprint) and, hence, a segregated cortical area may be characterized by highly coherent connectivity patterns. Connectivity fingerprints may therefore be distinguished by homogenous or modular connectivity patterns, which motivates a clustering approach for cortex parcellation.

Despite the relative success of clustering in producing anatomically sensible results, existing clustering techniques in the context of connectivity-based parcellation typically depend on several nontrivial assumptions. In this dissertation, we embody an information-theoretic framework to compress and therefore cluster anatomical connectivity data that avoids many assumptions and drawbacks imposed by previous methods. That is, the information bottleneck method distinguishes connectivity fingerprints by compressing distributional connectivity data guided by preserving anatomical connectivity information.

The unsupervised framework is based upon the notion that noise limits the amount of information and therefore the number of clusters we can resolve from the data.

Parcellation results for the inferior frontal gyrus together with the precentral gyrus reveal modular cortical areas consistent with the results of previous parcellaton studies, including cytoarchitectonic maps and results gained from fMRI. The proposed method also provides further insight into the hierarchically modular architecture of cortical areas.

Acknowledgements

I would like to express my sincere gratitude to the following people and organizations:

To my supervisors, Prof. Tania Douglas and Dr. Marc Tittgemeyer for their patience, guidance and for allowing me the opportunity to undertake this project at the Max Planck Institute for Neurological Research in Cologne, Germany. To Dr. Jenia Jitsev for his guidance and constructive criticism. To Corina Melzer for her assistance in solving tedious problems. To my parents for their continuous support without whom this project would not have been possible.

I would also like to thank the National Research Foundation, the University of Cape Town in South Africa and the Max Planck Institute for Neurological Research in Cologne, Germany for providing funding and facilities.

Contents

List of Figures

1

Introduction

Subdividing the cerebral cortex into structurally and functionally distinct areas, known as cortex parcellation, arises from the notion that cortical structure reflects function. Successful mapping of structure to function is achieved based upon functional properties of cortical elements within a region of interest. While many factors such as cytoarchitecture, myeloarchitecture and receptor architectonics reflect the functionality of such regions, evidence suggests a close relationship between anatomical connectivity and functional localization within the cortex (1). Moreover, anatomical connectivity is thought to constrain functionality and thus offers a suitable measure for differentiating between functionality of different cortical subunits. For example, structural elements of a distinct cortical region share highly coherent connectivity patterns, which are dissimilar to those of other cortical regions and therefore determine, to some extent, the functional properties of that region (2, 3). Subsequent grouping or clustering of structural elements with similar anatomical connectivity aims to segregate a cortical region of interest into functionally distinct subunits.

Early attempts at studying anatomical connectivity have been mostly revealed from post-mortem and animal studies. With the advent of diffusion magnetic resonance imaging (dMRI), *in vivo* and non-invasive characterization of long-range connectivity patterns became feasible (4). Diffusion MRI is typically used to infer information about the underlying fiber tract direction and therefore anatomical connectivity by modelling the direction-dependent mobility of water molecules (i.e. diffusion). Such water diffusion is influenced by the microscopic architecture of brain tissue and can be used to infer information about fiber bundle orientation. Diffusion is measured

in several directions yielding a 3D diffusivity profile for each voxel of interest, which reflects the orientation of fiber bundles (11).

However, noise and artifacts present in the MR scan introduce uncertainty pertaining to fiber tract direction. Further uncertainty in model parameters is caused by using simple models to describe the complex nature of the diffusion signal (5). Associating anatomical connectivity with a probability of connectivity (thereby taking into account the fore-mentioned uncertainty), as performed by probabilitic tractography, therefore offers an appropriate means for characterizing anatomical connectivity using dMRI. This ultimately opened the possibility to probe the white matter structure in the human brain (4): A convenient way to characterize anatomical connectivity of small brain areas (usually individual MRI voxels) to the entire brain is the computation of probabilistic tractograms, which can be seen as an approximation (with some reservation, see Jones (6)) to the connectivity pattern representing this brain area.

Different types of clustering algorithms have been used to perform cortex parcellation. Central and pairwise clustering methods, such as K-means clustering and spectral reordering, employ correlation as a predefined similarity measure and thus explicitly rely on the strength of linear dependency between tractograms in order to form clusters. The Dirichlet process mixture model is another technique that embodies a mixture of likelihood functions as a statistical model to describe the data. The suitability of representing probabilistic tractograms as vectorial data as well as using a Gaussian likelihood function in the Dirichlet process mixture model to perform cortex parcellation (7) remains unjustified. The question of whether or not tractograms should be grouped according to the strength of linear dependency between tractograms is also debatable.

Passingham et al. (1) describes each cortical area, denoted as a cortical subunit, as having a unique pattern of connections (connectivity fingerprint). The purpose of this dissertation is to demonstrate an information-theoretic framework to distinguish the unique pattern of connections underlying a cortical area. Clustering of probabilistic tractograms and therefore cortex parcellation arises as a consequence; structural elements within a cortical subunit belong to the same cluster, because their collective connectivity pattern distinguishes the connectivity fingerprint underlying the cortical subunit. More precisely, the information bottleneck method used in this dissertation makes use of the distributional nature of probabilistic tractograms to compress connectivity information such that the connectivity fingerprints reveal as much anatomical

connectivity information as possible. In this setting probabilistic tractograms contain information about anatomical connectivity, which is relevant for compression.

Previous parcellation attempts tend to neglect the proposed nested hierarchical architecture of cortical subunits. Actually, brain networks are more appropriately conceived of as building modules, each with a characteristic connectivity; i.e., modular hierarchies (1). The notion of a hierarchically modular organisation of cortical subunits stems from the idea that cortical subunits themselves are nested into further modular structures due to their similarity to one another with respect to anatomical connectivity. This dissertation therefore aims to investigate further properties of cortical areas such as their possible modular hierarchy.

Parcellation of the inferior frontal gyrus (IFG) together with the precentral gyrus (PCG) demonstrates a proof of concept of our approach. These gyri contain brain regions for which the anatomical segregation has been relatively well established (8, 9, 10) and have been investigated by previous connectivity-based parcellation studies (11, 12, 13, 14). Moreover, the modular hierarchy describing areas such as the primary motor, premotor, and pre-frontal regions is well established (15, 16).

Chapter 2 gives a more detailed description of the problem followed by an overview on clustering principles and cortex parcellation in chapter 3. The information-theoretic framework for unsupervised connectivity-based cortex parcellation is described in chapter 4. Chapter 5 presents an application of the framework to synthetic data and to anatomical connectivity data for the IFG and PCG. The performance of the method for connectivity-based cortex parcellation is discussed in chapter 6 followed by conclusions and proposed future work given in chapter 7.

Conference and Journal Publications

A significant portion of the work presented in this dissertation was presented at the Human Brain Mapping conference in 2010 (17) and has been accepted for publication *Frontiers in Neuroinformatics* (18).

2

Problem Setting

Relation between connectivity signal and connectivity observations: The aim of this dissertation is to identify the unique pattern of connections that distinguish the connectivity fingerprint of structural elements within a cortical area of interest. As with any other real world problem, we seek to identify the signal underlying the measurements. That is, connectivity observations gained from diffusion measurements allow us to infer the signal from which they were sampled. In order to infer the hidden structure in connectivity data we have to decide upon the relation between the signal and the connectivity observations. In our case, we have prior knowledge that the connectivity pattern of individual structural elements within the brain possess a modular architecture (1). In other words, we have prior knowledge that the connectivity patterns of structural elements are arranged in groups, whereby each group contains homogenous (i.e. similar) connectivity patterns. The connectivity pattern of a structural element denotes the collective connectivity observations from that structual element to the white matter volume. We even have prior knowledge that connectivity patterns may in fact be much more organized than that; evidence suggests that the connectivity patterns underlying a cortical area may be hierarchically modular (19).

The modularity property of connectivity patterns justifies the hypothesis that the signal partitions connectivity patterns of individual structural elements into modules or groups containing structural elements with homogenous or similar connectivity patterns. In the context of cortex parcellation such groups are referred to as cortical subunits. The collective connectivity patterns of structural cortical elements within a

cortical subunit therefore forms the connectivity fingerprint attributed to that cortical subunit.

Connectivity fingerprint: As already suggested, the most important hypothesis to consider in this dissertation is given by Passingham et al. (1): cortical subunits each possess a unique pattern of connections (i.e. connectivity fingerprint). That is, no two areas share identical patterns. It is, however, not clear what the uniqueness of connectivity patterns implies. The uniqueness of connectivity patterns does not necessarily have to be defined on a "black- and white-scale" (i.e. connectivity patterns are either unique or not). Connectivity patterns of cortical subunits may instead be defined on a "grey-scale" whereby the distinguishability of connectivity patterns measures the uniqueness of the connectivity pattern. Notice that in the latter case we still remain consistent with Passingham et al. (1) hypothesis; no two areas share identical patterns since, by definition, connectivity patterns are distinguishable from each other. It is only when connectivity patterns are indistinguishable from each other that they are no longer unique. Conversely, if connectivity patterns are maximally distinguishable from each other, they are maximally unique.

Clustering imaging data: Furthermore, due to the complex nature of connectivity within the brain, it is not clear as to what constitutes a structural cortical element possessing such anatomical connectivity information. Since we are limited to the resolution of diffusion weighted (DW) images and are only capable of measuring extrinsic anatomical connectivity without being able to differentiate between afferent and efferent connections, we will maintain a rather abstract definition of a structural cortical element throughout this dissertation: a structural cortical element is an area within the cortex at the cortical boundary connected to an individual white matter fiber following a particular pattern of connections within the white matter volume.

Since DW imaging techniques are not yet capable of capturing connectivity information pertaining to individual fibers within the human brain it is reasonable to assume that individual imaging voxels contain several cortical elements. Lets assume that the connectivity patterns each defining a cortical subunit are as distinguishable as possible which implies that each cortical element belongs to only one cortical subunit (i.e. single assignment hard clustering); and that we know exactly to which cortical subunit each cortical element belongs to. DW imaging will only capture cortical elements in terms of

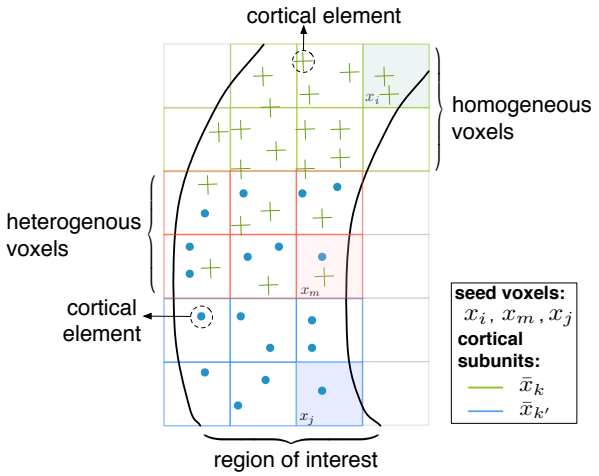

Figure 2.1: Cortical elements within the region of interest - Given a grouping of cortical elements in cortical subunits (blue and green), voxels will either contain cortical elements belonging to only one cortical element (i.e. homogenous voxels will capture properties of only one cortical subunit) or they will capture cortical elements belonging to multiple cortical subunits (i.e. heterogeneous voxels capture the properties of multiple cortical subunits).

voxels, more precisely termed seed voxels. As illustrated in figure 2.1, homogenous voxels containing cortical elements that only belong to the same cortical subunit will only contribute to the connectivity pattern defining that cortical subunit. Conversely, heterogeneous voxels containing cortical elements belonging to multiple cortical subunits will contribute to the connectivity pattern of multiple cortical subunits. Regardless of the initial assumption, that the connectivity patterns are as distinguishable as possible, the resolution of DWI data, nonetheless, coarsens the distinguishability of connectivity patterns, which therefore limits the amount of detail we can reveal in the partitioning of cortical subunits.

In the context of *in vivo* connectivity-based parcellation using DW-images, we are limited to grouping seed voxels and not the cortical elements themselves. Despite the possible hard grouping of cortical elements into cortical subunits (i.e. cortical elements possibly belong to only one cortical subunit), the limited resolution of imaging data forces us to consider multi-assignment clustering solutions since heterogeneous voxels should belong to multiple cortical subunits. Ideally, such multi-assignments should reflect the ratio of different cortical elements inside heterogenous voxels. Decreasing

the resolution, increases the number of heterogenous voxels which in turn decreases the distinguishability of connectivity patterns that we can hope to obtain through the measurement process.

Outlining the region of interest: Notice that the region of interest is clearly outlined in figure 2.1. For cortex parcellation the region of interest comprises white matter seed voxels next to the grey-white matter border. In practice, however, outlining the region of interest for cortex parcellation introduces unwanted bias by using a grey-value intensity threshold to identify the cortical boundary and therefore the cortical seed voxels. Such seed voxels at the cortical boundary are assumed to be connected to their neighboring grey matter voxels.

Quantifying anatomical connectivity *in vivo*: How does one quantify the connectivity patterns of individual seed voxels *in vivo*? Diffusion MRI (dMRI) is typically used to infer information about the underlying fiber tract direction and therefore anatomical connectivity by modelling the diffusion measurement process. However, noise and artifacts present in the MR scan introduce uncertainty pertaining to fiber tract direction. Further uncertainty in model parameters is caused by using simple models to describe the complex nature of the diffusion signal (5). Associating anatomical connectivity with a probability of connectivity (thereby taking into account the forementioned uncertainty), as performed by probabilistic tractography, therefore offers an appropriate as well as convenient means for characterizing anatomical connectivity using dMRI.

Hierarchical organization of connectivity patterns: As mentioned previously, evidence suggests that brain networks are organized as building modules thereby forming nested hierarchies. Once more, we have to consider different levels of distinguishability among the connectivity patterns of cortical subunits in order to understand their nested structure: finer connectivity patterns that are grouped at higher levels of the hierarchy are considered less distinguishable from each other, whereby coarser connectivity patterns that are grouped at lower levels are more distinguishable from each other. Put differently, given a means to quantify distinguishability in terms of similarity, cortical subunits grouped at higher levels in the hierarchy should be more similar to one another in terms of their connectivity than those grouped at lower levels. The simplified model in figure 2.2 demonstrates how a nested structure of cortical subunits expresses different levels of similarity among their connectivity profiles.

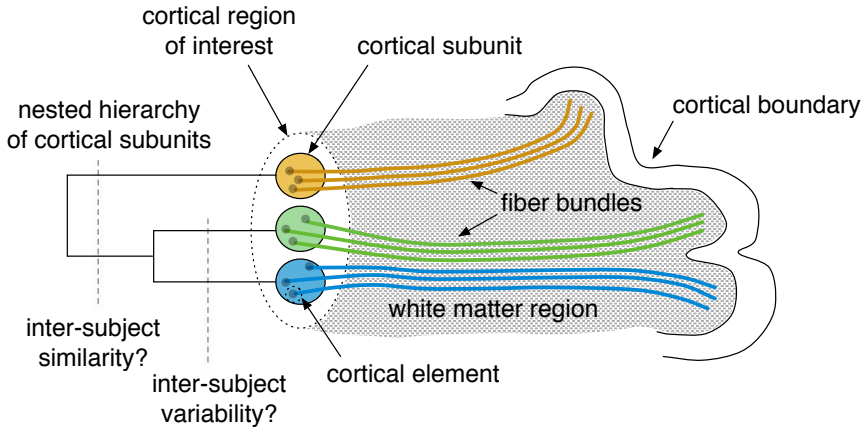

Figure 2.2: Simplified nested hierarchy of cortical subunits - Green and blue cortical subunits possess more similar connectivity patterns and are therefore less distinguishable from each other when compared to the connectivity pattern of the cortical area illustrated in yellow. Accordingly, green and blue cortical subunits are grouped at a higher level in the hierarchy. A coarser grouping of cortical subunits revealing gross anatomical connectivity patterns is hypothesized to be more similar among subjects whereas higher-level grouping may be more variable across subjects.

One must bear in mind that figure 2.2 is a simplified model to give an intuitive idea of what a nested hierarchy of extrinsic connectivity profiles signifies in terms of their distinguishability or similarity to one another. The actual extrinsic connectivity pattern is much more complex involving intermediate connections, among other things. An interesting question is which level of the hierarchy, if any, is common among subjects and which finer level connectivity patterns demonstrate significant inter-subject variability. Investigating such connectivity similarities and variability across subjects is, however, not the focus of this project.

Assessment of results: Needless to say, assessing connectivity-based cortex parcellation against other techniques is of utmost importance. However, connectivity-based cortex parcellation using dMRI is a relatively novel approach and there is limited prior knowledge with respect to the delineation of areas within the cortex. Indeed, such prior knowledge can be gained from functional magnetic resonance imaging (fMRI) studies and cytoarchitectonic maps. However, one must bear the following in mind:

- **Functional MRI studies** obtain functional areas in the cortex based upon a functional paradigm. Although functionality of a cortical area is hypothesized to

have a strong link to anatomical connectivity (1), fMRI studies do not measure anatomical connectivity directly. Moreover, such studies are useful for revealing the core areas of functional units and are not precise in delineating functional units in the cortex (i.e. borders between such functional units are not precise).

- **Cytoarchitectonic maps** are formed based upon cell density and type, but not connectivity. The apertures of both data sets (i.e. data obtained by probabilistic tractography and data obtained by cell density and type) are therefore inherently different, and hence it is difficult to assess one method by the other. The overlap of connectivity-based parcellation results and cytoarchitectonic maps, however, is interesting as it reveals something fundamental about brain cortex formation. Another issue is that connectivity-based parcellation results and cytoarchitectonic maps each describe the segregation of cortical areas in different brains. For reasons owing to the limited knowledge with respect to the variability of such areas among subjects it is desirable to compare results of both methods for the same brain. However, such studies have not been done yet.

3

Background

Conclusively, it has been shown in the mammalian brain that structural elements of a distinct cortical region share homogeneous connectivity patterns (20, 21), which are dissimilar to those of other cortical regions and therefore determine, to some extent, the functional repertoire of that region (3). These findings justify using clustering of connectivity patterns as a means to perform connectivity-based parcellation with the aim of segregating a cortical region of interest into functionally distinct subunits: Distinct homogenous connectivity patterns imply that structural elements within such regions share similar anatomical connectivity patterns. Knösche and Tittgemeyer (in press) (22) provide a current review on this topic.

As stated previously, connectivity-based cortex parcellation is formulated as a clustering problem. Section 3.1 gives an overview of clustering principles followed by previous applications of cortex parcellation and their use of clustering techniques given in section 3.2.

3.1 Clustering Principles

Clustering forms part of a cohort of methods designed to extract hidden structure in data sets (23). A basic definition of clustering is that clusters should contain homogenous (i.e. similar) objects. This definition is very general which allows for many interpretations of the clustering problem and therefore many hypotheses upon which clustering should be based. A quality criterion that quantifies the quality of a clustering solution encompasses a wide range of clustering hypotheses. Common types of

quality criteria such as cost functions, the expected distortion or a likelihood function are described in this section.

3.1.1 Modelling Cluster Structure

3.1.1.1 Clustering Definitions and Notations

Formally, clustering can be formulated in terms of grouping a set of objects $x \epsilon X$ into subsets or clusters $\bar{x} \epsilon \bar{X}$ based upon measurements $y \epsilon Y$. Clustering in the traditional sense is constrained to singleton assignments (i.e. objects belong to only one cluster), although multi-assignment clustering methods have already been proposed (24). The assignments may either be deterministic (i.e. hard clustering) or stochastic (i.e. soft clustering). For simplification in this section, we will differentiate stochastic assignments with $p(\bar{x}|x)$ from deterministic assignments with $x_i \leftarrow c_i$, where c_i is the cluster label given to object x_i and c denotes the set of labels given to all objects X.

It is important to understand the difference between hard and soft clustering: Both assume that objects belong to only *one* cluster (ground truth). In contrast to deterministic assignments, stochastic assignments take uncertainty in cluster assignments into account. Note that stochastic assignments do *not* imply that an object belongs to multiple clusters since this does not agree with ground truth. Instead they reflect uncertainty that an object belongs to *one* of multiple clusters.

3.1.1.2 Data Representation

Data representation is crucially important since it predetermines what kind of cluster structure can be discovered from the data (25). The relationship between objects and measurements is defined by the relation between a design space $\mathbb{X}$, $X \epsilon \mathbb{X}$, and a measurement space $\mathbb{Y}$, $Y \epsilon \mathbb{Y}$. Specifically such relations may be given by a functional dependency, $y : x \rightarrow y(x)$, or a stochastic dependency, $p(y|x)$, between objects and measurements. The latter defines y as a measurement sample. Such definitions provide a framework for describing the following common data types (25):

- **Vectorial data**: measurements are categorized as features and ordered into a d-dimensional feature vector. Each feature vector defines an object.

- **Proximity data**: proximity measurements provide comparisons (usually pairwise) between objects. Such proximity measurements between objects can be summarized in a similarity matrix as shown in figure 3.2C.

- **Distributional data**: Objects are described by probability distributions $P(X|Y)$. Measurements provide samples from those probability distributions. Such samples can be used to infer the probability distribution from which they were sampled. For finite data, bins forming the conditional probability distribution, $p(y|x)$, can be interpreted as features of an object x.

Note that other common data types may also be described by a mixture of the data types mentioned above. For example, the nature of data arising from a multivariate probability distribution is both vectorial and distributional. Specifically, distributions are categorized into a d-dimensional feature vector. Once more, measurements provide samples from those distributions. The following data types impose additional constraints on the data:

- **Pre-grouped data**: Observations y are already organized in specific groups prior to any data analysis or manipulation. Such cases arise in neuroscience when combining observations made within several subjects (i.e. observations are already grouped in individual subjects). Stochastic models propose a hierarchical prior that governs the clustering within subjects while combining observations made across subjects. Such applications include multi-subject analyses of functional units within the brain (7, 26).

- **Hierarchically nested or topologically structured data**: Subsets of the data can be decomposed into further subsets which in turn are decomposed into further subsets etc. That is, the subsets form a nested branching structure.

3.1.2 Clustering based upon Cost Functions

Mathematically, hard clustering can quite simply be formulated as minimizing a non-negative cost function. The cost function takes as input the data X and their hard clustering assignments c. The rationale is that cluster assignments c are given costs and that the cost function quantifies the quality of cluster assignments based upon a

hypothesis. The following section discuss several different cost functions that can be used to perform clustering.

3.1.2.1 Cost Functions

Central clustering: The hypothesis here is that clusters should have centers $\bar{x}_\perp$ and that the distance of objects to their centers should be minimized. The representation of clusters by their centers causes distortion costs $d(x, \bar{x}_\perp)$ due to information loss. The cost function can therefore simply be formulated as a sum over all distortion costs (25):

$$R(c, X) = \sum_{\bar{x}\epsilon\bar{X}} \sum_{x\epsilon\bar{x}} d(x, \bar{x}_\perp) \tag{3.1}$$

K-means clustering (27) is a special case of central clustering where the data is represented as vectors and the dissimilarity measure is accordingly given by the euclidian distance between objects and their centroids: $d(x, \bar{x}_\perp) = |x - \bar{x}_\perp|^2$.

Affinity propagation can be viewed as minimizing a similar cost function but instead summing over any predefined similarity measure $s(x, \bar{x}_\perp)$, including non-metric similarities, between objects and centroids (i.e. $R(c, X) = -\sum_{\bar{x}\epsilon\bar{X}} \sum_{x\epsilon\bar{x}} s(x, \bar{x}_\perp)$) (28). However, the cost function related to affinity propagation does not preserve the non-negativity constraint.

Pairwise clustering: Similar to central clustering, pairwise clustering cost functions favour compactness by minimizing intra-cluster dissimilarities. In particular, graph optimization problems provide a means to perform pairwise clustering. Objects represent vertices of the graph and the magnitude of the edges between objects $\mathcal{E}$ encode the pairwise similarities. The cost of assigning objects to clusters is given by (25):

$$R(c, X) = \sum_{\bar{x}\epsilon\bar{X}} \frac{N_{\bar{x}}}{N_{\mathcal{E}(\bar{x})}} \sum_{i,j\epsilon\mathcal{E}(\bar{x})} d(x_i, x_j), \tag{3.2}$$

where we sum over the magnitude of all edges in the same cluster (i.e. $\mathcal{E}(\bar{x})$). $N_{\bar{x}}$ denotes the number of objects in cluster $\bar{x}$ and $N_{\mathcal{E}(\bar{x})}$ denotes the number of edges in cluster $\bar{x}$.

Normalized cuts (29) and clustering by minimum spanning tree (30) are examples of pairwise clustering algorithms that use graph optimization techniques. As with

central clustering, pairwise clustering once more favours compactness by maximizing intra-cluster similarities.

Path-based clustering: Quite often the cluster structure may be irregular or may not possess a compact spherical structure in the feature space which is often assumed. Instead, the density of the objects plays a much more important role. Fischer et al. (31) proposes a path-based pairwise clustering cost function for such cases where a particular grouping of objects is favoured when the paths through objects in the same cluster is minimized. The rationale is that the paths through objects in the same cluster is minimal when they are densely distributed in the feature space. The cost function will not be shown here for reasons owing to simplicity.

3.1.2.2 Cluster Optimization

The optimal clustering solution, c^*, is one that minimizes the cost function. Practically, computing the cost function for every possible cluster assignment, $c \epsilon C$, is infeasible. Searching for the optimal cluster assignments c^* is a combinatorial problem for which there are various optimization techniques. The following section describes cluster optimization by Gibbs sampling and simulated annealing (25, 32):

Gibbs distribution: The probablity of sampling a feasible clustering solution c based upon the cost function is given by the joint Gibbs distribution:

$$\Pi(c) = \frac{\exp(-\beta R(c, X))}{\sum_{(c' \epsilon C)} \exp(-\beta R(c', X))}, \tag{3.3}$$

where the denominator serves as the partition function and β is the inverse computational temperature that controls the probability of sampling a cluster assignment c with low costs. In particular, for $\beta \to \infty$, we are guaranteed to sample a clustering solution c^* that (globally) minimizes the cost function.

However, the approach is still infeasible since the partition function requires summing over all possible clustering solutions, $c \epsilon C$. Gibbs sampling offers a solution to the problem by sampling "locally" using the following conditional distribution:

$$\Pi(c_i) = \frac{\exp(\beta R(c, X))}{\sum_{(1 \leq c_i' \leq K)} \exp(-\beta R(c', X))}, \tag{3.4}$$

where c_i denotes the clustering assignment given to object x_i and K denotes the number of clusters. Notice that the denominator in equation 3.4 requires summing over

only possible cluster assignments $1 \leq c_i \leq K$ for object x_i and *not* over all possible combinations of cluster assignments (i.e. all possible clustering solutions), $c \epsilon C$. More precisely, the denominator is summed over cluster assignments, c_i, while keeping all other cluster assignments, c_{-i}, fixed.

Simulated annealing: The Gibbs sampler requires initializing cluster assignments c_{-i}. Simulated annealing overcomes this problem by following an annealing schedule whereby the inverse computational temperature β is increased from $\beta = 0$ to $\beta \to \infty$. Clustering solutions for a particular β serve as intialization for clustering solutions given a slightly increased β. The rationale is that at $\beta = 0$, c_i can be intialized randomly since the probability of drawing a clustering solution in 3.3 is *not* governed by the cost function and is therefore entirely random for any intialization.

3.1.3 Clustering based upon Rate Distortion Theory

Rate distortion theory tells us that clustering can be interpreted as a form of data compression (33). More precisely, soft clustering is achieved by compressing the data $x \epsilon X$ to form compact representations $\bar{x} \epsilon \bar{X}$ (i.e. clusters). Controlling the compression of the data is simply done by minimizing a complexity term given by the mutual information between objects and their compact representations, $I(\bar{X}, X)$. However, in doing so we distort the data. Similar to the cost function principle introduced above we want our data compression (i.e. clustering) to be governed by a quality measure so that the data distortion is minimized. The rate distortion functional minimizes $I(\bar{X}, X)$ under the constraint of a quality measure given by the expected distortion $\langle d(x, \bar{x}) \rangle$:

$$\mathcal{L}[p(\bar{x}|x), p(\bar{x})] = I(\bar{X}, X) + \lambda \langle d(x, \bar{x}) \rangle, \tag{3.5}$$

where $\mathcal{L}$ is minimized with respect to assigment probabilities $p(\bar{x}|x)$ and marginal probabilities $p(\bar{x})$. λ controls the tradeoff between data compression $I(\bar{X}, X)$ and expected distortion $\langle d(x, \bar{x}) \rangle = \sum_{\bar{x} \epsilon \bar{X}} \sum_{x \epsilon \bar{x}} p(x, \bar{x}) d(x, \bar{x})$. Assuming differentiability, minimizing the rate distortion functional 3.5 is simply done by taking its derivative and equating it to zero which yields the following self-consistent equations:

$$\begin{cases} p_t(\bar{x}|x) = \frac{p_t(\bar{x})}{Z(x,\lambda)} \exp(-\lambda d(x, \bar{x})) \\ p_{t+1}(\bar{x}) = \sum_x p(x) p_t(\bar{x}|x), \end{cases} \tag{3.6}$$

where $Z(x, \lambda)$ is a normalization constant and t denotes the iteration sequence. Blahut (34) demonstrates that iterating through the self-consistent equations yields the minimum of the rate distortion functional $\mathcal{L}$. Notice that, similar to cost functions, the self-consistent equations require initializations. Once more, simulated annealing can be used as an optimization technique to obtain the global minimum of $\mathcal{L}$ and is explained in section 4.3.

Rate distortion theory gives rise to a wide framework of clustering methods since its derivation is applicable for any distortion cost $d(x, \bar{x})$. The following clustering techniques are all based upon rate distortion theory but differ only in their distortion measures.

3.1.3.1 Prototype-based Clustering

How do we define the distortion cost $d(x, \bar{x})$ or dissimilarity between object x and cluster $\bar{x}$? If we can assume that clusters have a prototypical characteristic we can compute the distortion cost as a dissimilarity between objects and prototypes. For example, given that clusters are formed based upon central tendencies in the data we can assume that the center of the cluster $\bar{x}_{\perp}$ is protoypical for that cluster $\bar{x}$. Computation of the distortion cost is therefore given by:

$$d(x, \bar{x}) = d(x, \bar{x}_{\perp}) \tag{3.7}$$

Notice that the same type dissimilarity measure is used in central clustering (equation 3.1). It should, however, be noted that this type of clustering method assumes that we have access to the distortion between objects and prototypes $d(x, \bar{x}_{\perp})$.

3.1.3.2 Information-based Clustering using Collective Similarities

Slonim et al. (35) formulates the"information-based clustering" method by computing the distortion cost between object x_i and cluster $\bar{x}$ as a sum over any pairwise similarity measures between object x_i and all other objects weighted by their cluster membership:

$$d(x_i, \bar{x}) = -s(x_i, \bar{x}) = \sum_{j \epsilon \bar{x}} p(x_j|\bar{x}) s(x_i, x_j) \tag{3.8}$$

Substituting the distortion cost into the expected distortion yields:

$$\langle d(x, \bar{x}) \rangle = -\langle s(x, \bar{x}) \rangle = -\sum_{\bar{x} \epsilon \bar{X}} \sum_{i \epsilon \bar{x}} p(x_i, \bar{x}) s(x_i, \bar{x}) \tag{3.9}$$

Slonim et al. (35) extends the notion of pairwise similarities to a more general idea of collective similarities among objects (i.e. $s(x_1, x_2, x_3, \ldots)$).

3.1.3.3 Distributional Clustering and the Information Bottleneck Method

As is often the case, the relationship between objects and measurements may be of distributional nature and therefore involve stochastic dependencies (i.e. $p(y|x)$). Each object is thus characterized by a conditional probability distribution, $p(Y|x)$, over measurement samples. In the case of finite samples grouping objects is done by grouping their histograms. The distortion between distributional data is naturally interpreted by the Kullback-Leibler distance between objects and clusters (36):

$$\langle d(\bar{x}, x)\rangle = \langle D_{KL}[p(y|x)||p(y|\bar{x})]\rangle = I(X,Y) - I(\bar{X},Y), \tag{3.10}$$

where $I(X,Y)$ is the mutual information between objects X and measurement samples Y and $I(\bar{X},Y)$ is the mutual information between compact representations of objects $\bar{X}$ and measurements samples Y. The derivation of equation 3.10 is given in chapter 8. Substituting the expected distortion by the expected Kullback-Leibler distance in the rate distortion functional $\mathcal{L}$ gives:

$$\mathcal{L} = I(\bar{X},X) + \lambda\left((I(X,Y) - I(\bar{X},Y)\right) \tag{3.11}$$

$$F = I(\bar{X},X) - \lambda I(\bar{X},Y) \propto \mathcal{L}, \tag{3.12}$$

where F is precisely the cost functional minimized in the information bottleneck method. Using the Kullback-Leibler distance as the distortion measure has several important consequences:

1. $I(X,Y)$ does not depend on $p(\bar{x},x)$ and therefore remains a constant in the rate distortion functional $\mathcal{L}$. Minimization of the rate distortion functional $\mathcal{L}$ can therefore be simplified to minimizing the cost functional F which yields a new set of self-consistent equations (equations 4.5).

2. The (mutual) information that clusters $\bar{X}$ provide about measurement samples Y in the cost functional F (equation 3.12) replaces the expected distortion $\langle d(\bar{x},x)\rangle$ in the rate distortion functional $\mathcal{L}$ (equation 3.5).

3. Consequently, in contrast to most clustering methods, minimization of the cost functional F does *not* require explicitly defining a distortion measure *a priori*. Instead, clustering is simply based on the intuitive notion that the data compression $\bar{X}$ should capture most of the information with respect to measurement samples Y. A distortion measure given by the Kullback-Leibler distance in equations 4.5 arises implicitly by minimizing the cost functional F.

Tishby et al. (33) provide an alternative perspective to deriving the cost functional F and is described in section 4.3.

3.1.4 Clustering based upon Sufficient Statistics

Another means to perform clustering is by means of achieving sufficient statistics in a stochastic model. That is, parameters $\theta \epsilon \Theta$ are sufficient to describe the distribution of objects within clusters. The functional form of the distribution of objects within clusters is given by the likelihood function $\mathcal{F}(X|\Theta)$ which has to be predefined.

Bayes' rule gives the probability of assigning parameters to the data:

$$P(\Theta|X) \propto \mathcal{F}(X|\Theta)P(\Theta), \tag{3.13}$$

where the prior $P(\Theta)$ induces the partition of objects into clusters. The use of the prior $P(\Theta)$ makes the distinction between parametric and non-parametric Bayesian models. For example, a Gaussian mixture model is considered a parametric clustering technique since the functional form of the prior contains parameters that have to be predefined. A Bayesian model is non-parametric if a conjugate prior is used to achieve sufficient statistics in the posterior distribution $P(\Theta|X)$. Clustering by sufficient statistics requires that the posterior distribution $P(\Theta|X)$ be multinomial. Accordingly, the conjugate prior of a multinomial distribution is given by a Dirichlet distribution (37):

$$P(\Theta) = \mathrm{Dir}(\alpha, G_0), \tag{3.14}$$

where α is the concentration parameter and G_0 is the base distribution of the Dirichlet distribution. The non-parametric Bayesian clustering model is therefore known as the Dirichlet mixture model. Note that a non-parametric Bayesian model does not mean to say that the model involves no parameters that have to be predefined. In particular, the Dirichlet mixture model replaces parameters by hyperparameters that have to be

predefined. Similar to selecting cost functions, the likelihood function, $\mathcal{F}(\theta|x)$, determines the quality of clustering and should be chosen based upon data representation. For example, vectorial data may be modelled by multivariate distributions. A multivariate Gaussian distribution is typically used in the Dirichlet process mixture model. Orbanz (38) proposes an alternative likelihood function for distributional data.

Directly sampling a set of sufficient statistics Θ from a joint posterior distribution $P(\Theta|X)$ is infeasible. Notice that clustering based upon cost functions poses the same problem because it is infeasible to directly sample a clustering solution c from the joint Gibbs distribution (equation 3.3). Once more, Gibbs sampling can be used to sample "locally" using the conditional distribution $\theta_i|\Theta_{-i}$ instead of the joint distribution, where Θ_{-i} denotes the set of all sufficient statistics except for θ_i.

Teh et al. (39) extend the Dirichlet mixture model to the hierarchical Dirichlet mixture model to cluster pre-grouped data (section 3.1.1). Blei et al. (40) use a similar hierarchical Dirichlet prior to cluster hierarchically nested data.

3.1.5 Model Validation

An indispensable requirement of models used to describe the data is that the model must be generalizable across different sample data sets under the influence of noise. That is, *the model should be reproducible under the influence of noise.* Here we assume that the sample data sets are sampled from the same probability distribution. Although the purpose behind minimizing cost functions or the expected distortion is to quantify the quality of the solution, they cannot be used as model selection criteria because they are sensitive to noise in the data. As suggested previously, clustering forms part of a cohort of methods designed to extract the signal (i.e. hidden structure) from the data. In the context of clustering the hypothesis is that the signal partitions the data set into clusters that contain similar elements. In practice, however, we sample noisy versions of the signal which gives rise to the under- and overfitting dilemma shown in figure 3.1.

Noise in the data can be characterized by comparing two sample data sets. Noise causes uncertainty in measurements which translate to uncertainty in the partitioning (i.e. the solution space). Using quality criteria that do not account for noise cause overfitting as shown in figure 3.1. Needless to say, identifying the partitioning signal is the ultimate task in clustering. Model validation poses the following question: How can

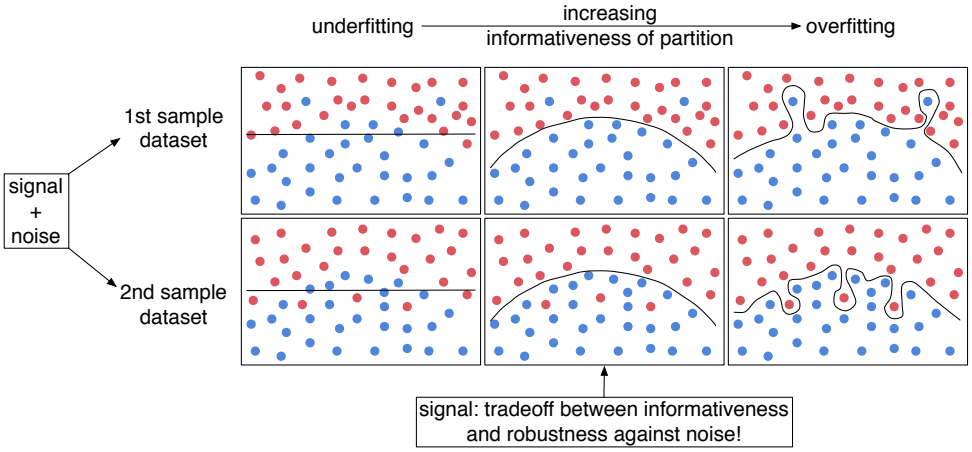

Figure 3.1: Model validation: the under- overfitting dilemma - Two data sets, sampled from the same probability distribution, are clustered (clusters are labelled red and blue) according to maximal fit given by optimizing a suitable quality criterion. Model validation discriminates signal from noise by maximizing the informativeness of the partition while maintaining robustness against noise and therefore generalizability of the solution. Such a tradeoff avoids under- and overfitting.

we characterize the relationship between the quality criterion and the partitioning signal? Indeed figure 3.1 demonstrates the relationship: The generalizability requirement demands that the partitioning and therefore the quality criteria be insensitive to noise; the partitioning signal is therefore one for which the quality criterion is suboptimal. That is, the partitioning signal is identified by the most informative partitioning that is still robust against noise in the quality criterion.

The task of identifying the signal is therefore still done by optimizing a *suitable* quality criterion, but however, accounting for noise such that the generalizability property is maintained. Note that only a suitable hypothesis (i.e. suitable quality criterion) for approximating the signal will maintain the generalizability property at suboptimal solutions. Henceforth, solutions can be validated according to their generalizability performance at suboptimal solutions. Two major schools of thought are used to perform model selection:

- **Correcting for finite samples**: Approximate the uncertainty induced in measurements due to finite samples and investigate how such uncertainty propagates through to the quality criterion and therefore the solution. Correcting the qual-

ity criterion to account for noise yields a reguralized quality criterion term that should be insensitive to noise. Correcting for finite samples (41) will be used as a model validation technique in this dissertation and is explained in greater detail in section 4.4.

- **Two-sample scenario**: Quantify the generalizability and the informativeness of the solution by comparing the solutions obtained for different sample data sets as shown in figure 3.1. Statistical analyses quite often use two samples given by training and test data sets to measure the agreeability between them. Clearly, two samples provide by far too little information to characterize the uncertainty in measurements. However, we are interested in the uncertainty in the solution. The rationale is that the uncertainty in the solution space is much smaller than the uncertainty in the measurements space. Two large enough samples therefore contain sufficient information to characterize the uncertainty in the solution space (i.e. clustering solutions) (42). Buhmann (42) proposes an information-theoretic model validation technique utilizing an approximation set coding (ASC) scenario that quantifies the tradeoff between the informativeness of the solution and its robustness against noise between training and test solutions.

3.2 Connectivity-based Parcellation

For the purpose of cortex parcellation, probabilistic tractography does not necessarily have to accurately reflect the connectivity pattern of an individual area. The sensitivity of probabilistic tractography to differences in connectivity of cortical areas plays a much more important role (22). This motivates the application of tractography for connectivity-based parcellation: When each cortical area is characterised by unique cortico-cortical connections (i.e. connectivity fingerprint), then tractograms within an area should be similar (22).

Recently, tractography-based parcellation has been applied to a variety of sub-cortical and cortical areas, in the macaque as well as in the human brain. These areas include the thalamus (5, 43, 44, 45) basal ganglia (46, 47, 48), amygdala (49) and midbrain (50) and cortical regions including inferior frontal cortex (11, 12, 51), premotor cortex (13, 14), cingulate cortex (52), medial frontal (43, 53) and insula cortex (54) as well as postcentral gyrus (55).

The fore-mentioned attempts at clustering probabilistic tractograms, however, impose several nontrivial assumptions about the underlying structure of the data. To date, two different types of clustering algorithms have been used to perform tractography-based parcellation:

1. **Clustering based upon cost functions** that relies on pairwise similarity measures: K-means clustering (11, 12, 54) or spectral reordering (56), employ correlation as a predefined similarity measure and thus rely on the strength of linear dependency between tractograms in order to form clusters. The similarity between any two tractograms is summarized in a symmetric similarity matrix, also referred to as a "connectivity correlation matrix" (11). The methodology followed by such methods is summarized in figure 3.2.

   Entries in the connectivity correlation matrix define the cross correlation between tractograms. The objective is to cluster or group seed voxels with similar tractograms and separate them from seed voxels with dissimilar tractograms. Johansen-Berg et al. (56) as well as Klein et al. (12) used spectral reordering to rearrange the similarity matrix in such a way to allow the user to visually identify clusters from the rearranged matrix as shown in figure 3.2C. Klein et al. (12) and Anwander et al. (11) proposed a k-means clustering strategy using the "connectivity correlation matrix" as input. Both k-means clustering and spectral reordering were nevertheless subject to user input (i.e. choosing the number of clusters). The criteria used by Anwander et al. (11) to choose the number of clusters were as follows:

   - Subdivisions should be consistent across subjects and rely on a one-to-one correspondence between subjects.
   - Each subdivision must be characteristic of a single coherent volume within the region of interest (i.e. clusters must be spatially compact).

   In other words, the choice of the number of cortical subunits was subject to forming representative, meaningful cortical regions while still maintaining relative consistency across subjects. It is often difficult to justify such an approach to select the number of clusters for regions where anatomical segregation may be highly variable across subjects and may contain clusters that are *not* spatially

compact. Additionally, it is debatable whether similarity between tractograms should be defined by their pairwise linear dependency to one another.

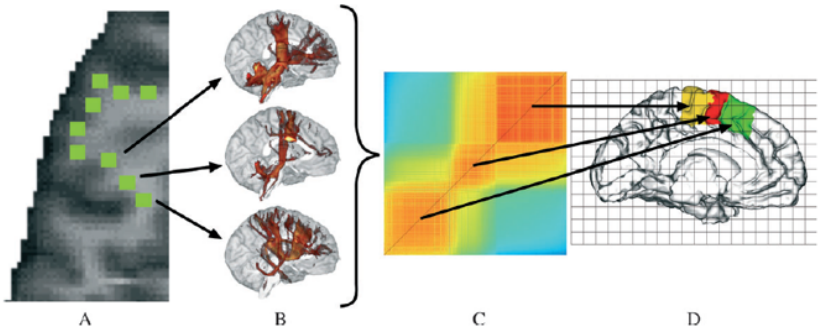

Figure 3.2: Cortex parcellation based upon pairwise similarity measures - **A**: The region of interest contains white matter voxels at the white and grey matter interface. **B**: Probabilistic tractograms quantify the anatomical connectivity pattern of each seed voxel. **C**: The rearranged connectivity correlation matrix captures the pairwise similarities between probabilistic tractograms. **D**: Segregation of the cortical region of interest based upon the connectivity correlation matrix. Reproduced from Anwander et al. (11).

2. **Clustering based upon sufficient statistics** that implicitly relies on similarity measures: Dirichlet process mixture models (DPMM) embody a Bayesian nonparametric model for clustering of probabilistic tractograms. Such stochastic processes typically assume data to be generated from a mixture of Gaussian distributions. In this setting, two tractograms are similar if they arise from the same Gaussian distribution. In an application to multiple-subject parcellation of the thalamus, Jbabdi et al. (7) represented tractograms as vectorial data and grouped them based upon a Gaussian likelihood function. Note that such stochastic models determine the number of clusters automatically given a choice of likelihood function. However, whether or not individual tractograms can be interpreted as vectors or that clusters assume a Gaussian form is debatable.

3. **Clustering based upon rate distortion theory** that relies on distortion costs between objects and prototypes (17, 18): As with k-means clustering and the Gaussian likelihood function used in the DPMM, this method assumed that clusters are formed on the basis of central tendencies in the data. Exemplars found by affinity propagation were used to express such central tendencies in the data

and therefore served as prototypes for clusters. Variation of information between connectivity patterns was explicitly used as the distortion measure between seed voxels and exemplars. The information bottleneck method applied in this dissertation is a variation of this method that does not rely on prototypes and implicitly uses the Kullback-Leibler distance as the distortion measure.

4

Methodology

4.1 Cortex Parcellation based upon Distributional Connectivity Data

As mentioned previously, cortical subunits are defined by their unique connectivity patterns (i.e. connectivity fingerprints) (1). How does one define the unique connectivity pattern underlying the cortical area? Since we advocate a probabilistic approach to quantifying anatomical connectivity using dMRI we formally define the connectivity pattern of a cortical area as a probability distribution of connectivity among voxels within the white matter volume. Such connectivity distributions are thus said to be unique and therefore distinguishable from connectivity probabilities defining different cortical areas. However, in practice, the connectivity distribution is unknown. As with any other application we have to rely on connectivity measurements, gained from dMRI, to obtain connectivity observations.

It should, however, be noted that we do not directly sample from the true connectivity signal (figure 4.1A) but instead sample from the noisy connectivity measurement signal (figure 4.1C) gained from dMRI. The connectivity measurements contain incomplete information about the true connectivity of microstructures but are assumed to contain sufficient connectivity information about the macroscopical connectivity patterns with limitations mentioned in section 4.2. The connectivity measurement signal is therefore assumed to approximate the true connectivity signal to some extent. The term "connectivity distribution" is used to refer to the noisy distibution of connectivity measurements in this dissertation. Connectivity samples therefore refer to the

observations of noisy connectivity measurements.

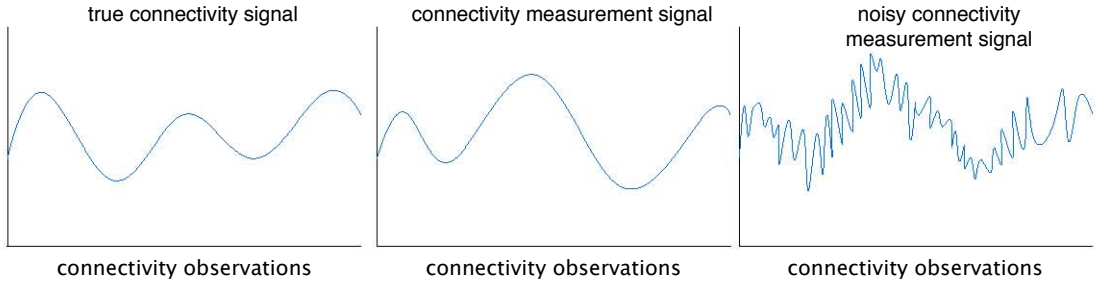

Figure 4.1: Synthetic connectivity signals - The connectivity measurement signal models the true connectivity signal incompletely, but is, however, assumed to approximate the true connectivity signal. In practice, however, we only sample from the noisy connectivity signal obtained by dMRI.

The task is thus to use a quality criterion to infer the unknown connectivity distribution (i.e. connectivity fingerprint) from the finite connectivity observations made by seed voxels. Given prior knowledge of the modular architecture of cortical subunits, the quality criterion should group connectivity patterns in order to infer the unknown connectivity distribution. The general methodology followed to perform cortex parcellation using a distributional interpretation of connectivity data is given in figure 4.2. The first step is to outline the region of interest from which to sample seed voxels x (figure 4.2A and B). As with many problems lacking prior knowledge, each seed voxel is equally likely to be sampled. The subsequent task is to sample white matter voxels y to which the seed voxel x is connected. Obtaining such samples from the conditional distribution as shown in figure 4.2C requires connectivity measurements, typically obtained from dMRI data as discussed in the next section. The quality criterion (figure 4.2D) serves to infer the underlying probability distribution using only connectivity samples taken from that distribution. The inferred discrete probability distributions (figure 4.2E) represent the connectivity fingerprints of individual cortical subunits. The contribution of connectivity patterns of sampled seed voxels to the connectivty fingerprint of cortical subunits achieves clustering of seed voxels and therefore cortex parcellation (figure 4.2E).

The following sections discuss the sampling technique used to obtain connectivity samples using probabilistic tractography (section 4.2) and the information bottleneck

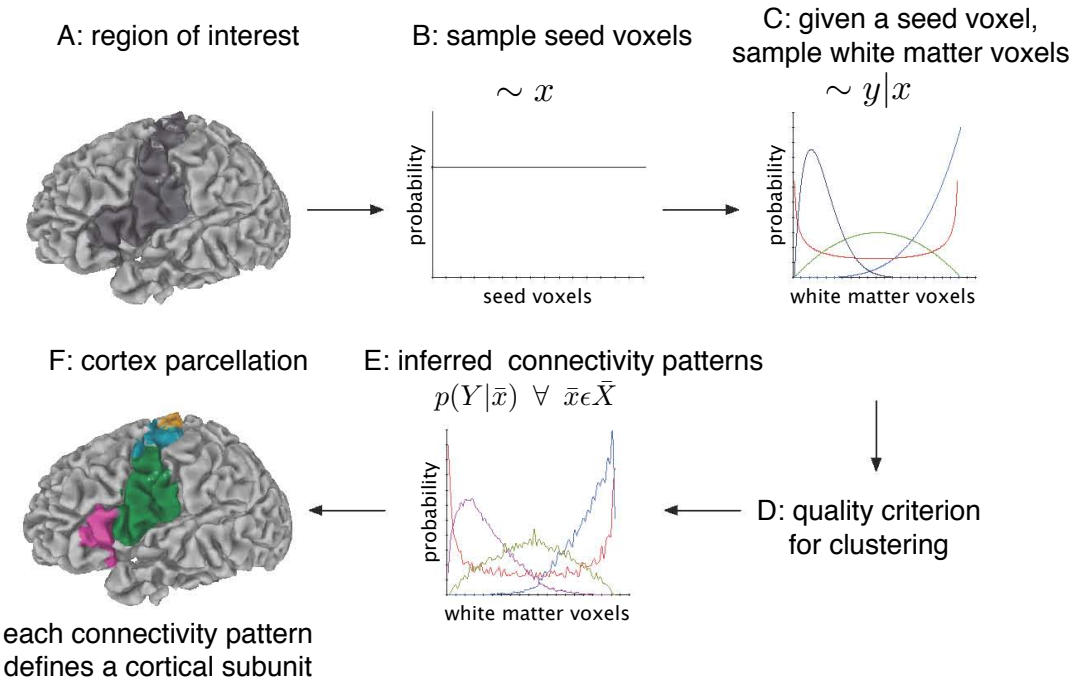

Figure 4.2: Cortex parcellation based upon distributional connectivity data - **A and B**: A region of interest is selected (A) from which seed voxels are sampled using a uniform prior distribution (B). The hypothesis is that there is an unknown connectivity distribution characterizing the connectivity of seed voxels to the white matter volume. **C**: Given a seed voxel, connectivity measurements in the form of dMRI allow us to obtain sample white matter voxels from that unknown distribution. Such samples reflect information about the connectivity between the seed voxel and the sample white matter voxels. **D and E**: A suitable clustering quality criterion D is used to infer the unknown connectivity distribution (i.e. connectivity fingerprints) underlying the region of interest (E). Each connectivity fingerprint defines a cortical subunit. **F**: The contribution of each seed voxel to the connectivity fingerprint defining cortical subunits achieves cortex parcellation. Note that the distributions shown in C and E above are synthetic and shown solely for illustration purposes.

method as a quality criterion (section 4.3) used to identify the connectivity fingerprints underlying the cortical area of interest.

4.2 Quantifying Anatomical Connectivity *in vivo*

A well established notion among neuroscientists is that water is more likely to diffuse along brain fibers than across them, which makes dMRI useful for inferring the orientation of fibers *in vivo*. Note that such diffusion information cannot differentiate between afferent and efferent connections. The MR signal contains limited information on the diffusion and therefore mictrostructure due to the finite resolution of imaging data, the limited sampling of diffusion direction, the diffusion time and diffusion length determined by the gradient strength. Consequently, the MR signal can only infer the orientation of fiber bundles and *not* individual fibers. In particular, such limitations lead to poor measures of fiber orientation in areas containing fiber crossings. The following section describes the data aquisition used to obtain dMRI measurements followed by a means to obtain connectivity samples from those measurements using probabilistic tractography.

4.2.1 dMRI Data Acquisition

Diffusion-weighted data and high-resolution 3-dimensional (3D) T1- and T2-weighted images were acquired on a Siemens 3T Trio scanner with an 8-channel array head coil and maximum gradient strength of 40 mT/m. The diffusion-weighted data were acquired using spin-echo planar imaging (EPI) (TR=12 s, TE=100 ms, 72 axial slices, resolution $1.72 \times 1.72 \times 1.72$ mm, no cardiac gating). A GRAPPA technique (reduction factor 2.0) was chosen as parallel imaging scheme. Diffusion weighting was isotropically distributed along 60 directions (b-value=1000 s/mm2). Additionally, seven data sets with no diffusion weighting were acquired initially and interleaved after each block of 10 diffusion weighted images as anatomical reference for motion correction. The high angular resolution of the diffusion weighting directions improves the robustness of the tensor estimation by increasing the signal-to-noise ratio (SNR) and reducing directional bias. To further increase SNR, scanning was repeated three times for averaging, requiring a total scan time for the dMRI protocol of approximately 45 min. dMRI data were acquired after the T2-weighted images in the same scanner reference system.

As a first step in preprocessing the data, the 3D T1-weighted (MPRAGE; TR=1300 ms, TI=650 ms, TE=3.97 ms, resolution $1.0 \times 1.0 \times 1.0$ mm, flip angle 10, 2 acquisitions) images were reorientated to the sagittal plane through the anterior and posterior commissures. Upon reorientation, the 3D T2-weighted images (RARE; TR=2 s, TE=355 ms, resolution $1.0 \times 1.0 \times 1.0$ mm, flip angle 180) were co-registered to the reorientated 3D T1-weighted images using rigid-body transformations (57), implemented in FSL (http://www.fmrib.ox.ac.uk/fsl). The images without diffusion weightings were used to estimate motion correction parameters with the same registration method. The motion correction for the dMRI data was combined with the global registration to the T1 anatomy. The gradient direction for each volume was corrected using the rotation parameters. The registered images were interpolated to an isotropic voxel resolution of 1 mm and the three corresponding acquisitions were averaged. Finally, a diffusion tensor was fitted to the dMRI data for each voxel. For presentation purposes, cortical surfaces were rendered on basis of the T1-weighted images by using Freesurfer (58).

4.2.2 Probabilistic Tractography

The purpose of probabilistic tractography is to characterize the connectivity pattern of seed voxels utilizing the orientation dependence of water within fiber bundles. The three-dimensional random walk method developed by Anwander et al. (11) samples a sequence of connected white matter voxels using diffusion tensor images. As the name implies the random walk method describes a random path taken by a particle starting from a given seed voxel and transitioning through target voxels within the white matter volume based upon local diffusivity measurements (i.e. local diffusivity measurements in the form of local diffusion tensors determine the transition probability from voxels to neighboring voxels). The random walk is terminated once the particle leaves the white matter volume. At each step a white matter voxel neighboring the voxel location of the particle is sampled from a probability distribution based upon local DTI data. The sampled white matter voxel determines the subsequent voxel location of the particle. The particle thus follows a Markov chain whereby each white matter voxel sample is dependent upon previous samples (i.e. the current location of the particle is dependent

upon its previous locations):

$$\begin{cases} \sim y(2)|y(1), x_i \\ \quad \vdots \\ \sim y(\eta)|y(\eta-1), \ldots, y(2), y(1), x_i, \end{cases} \tag{4.1}$$

where η denotes the sampling iteration sequence and x_i denotes the cortical seed voxel of interest with neighboring white matter voxel $y(1)$ from which the particle starts its random walk. The random path taken by the particle is thus constructed by a sequence of white matter voxel samples.

Samples allow us to gain information about the probability distribution from which they were sampled. In particular, for reasons owing to simplicity, we consider the independently and identically distributed (i.i.d.) connectivity probability from a seed voxel to any white matter voxel $p(y|x_i)$, where $y \epsilon Y$. Obtaining a probability of connectivity from a seed voxel x_i to a particular white matter voxel y_l requires marginalizing over all possible paths pa from the seed voxel leading to the white matter voxel:

$$p(y_l|x_i) = \sum_{pa} p(y_l, pa|x_i), \tag{4.2}$$

where pa is a sequence of white matter voxels, excluding y_l, for a path originating from x_i leading to y_l. Summing over all such paths, however, is infeasible. An approximation is given by repeating the random walk taken by a particle for many trials (i.e. 10 000) each starting from the same seed voxel x_i. By law of large numbers, the frequency at which y_l is sampled over all random walk trials taken by the particle originating from x_i, $f(y_l|x_i)$, give an approximation to equation 4.2:

$$p(y_l|x_i) \approx \frac{f(y_l|x_i)}{Z}, \tag{4.3}$$

where Z is a normalization constant. The joint conditional distribution, $p(Y|x)$, thus conveniently expresses the i.i.d. connectivity probability of a seed voxel to the white matter volume without having to deal with the more complex sequential connectivity data. The connectivity pattern given by the i.i.d. connectivity probability is illustrated in figure 4.3.

Note that, given a seed voxel, the white matter voxel is the only random variable in the connectivity distribution. The joint conditional distribution, $p(Y|x)$, is thus a discrete, univariate and i.i.d..

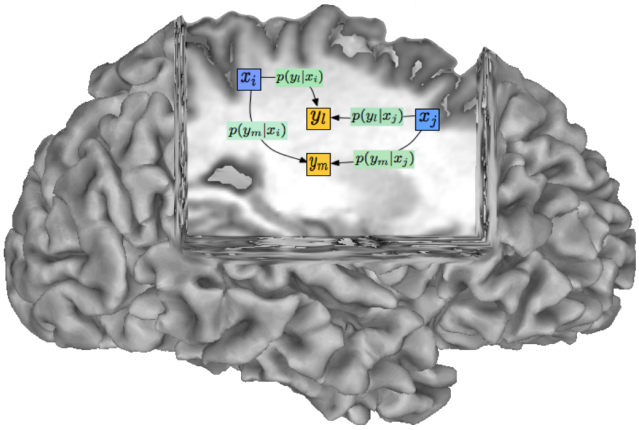

Figure 4.3: Distribution of connectivity from seed voxels to the white matter volume - The i.i.d. conditional connectivity probability gives the probability of finding a connection between the seed voxel (blue) and a white matter voxel (yellow) (i.e. connectivity pattern of seed voxels to the whole white matter volume, $p(y|x)\ \forall\ y\epsilon Y$).

4.3 Information Bottleneck Method

Probabilistic tractography allows us to characterize the connectivity pattern of individual seed voxels. However, we wish to quantify the connectivity pattern of cortical subunits containing seed voxels. The connectivity patterns of cortical subunits are therefore compact representations of the connectivity patterns of seed voxels. Within an information theoretic framework, the information that the connectivity patterns of seed voxels provide about the connectivity pattern of the entire region of interest is given by the mutual information between compact representations of seed voxels $\bar{X}$ and connectivity observations Y, $I(\bar{X}, Y)$. Forming such compact representations (i.e. clusters), $\bar{x}\epsilon\bar{X}$, is done by minimizing a complexity term given by the mutual information between objects X and their compact representations $\bar{X}$, $I(\bar{X}, X)$. However, in order to reveal the collective connectivity pattern of seed voxels, the complexity term is minimized while preserving as much information with respect to anatomical connectivity as possible (i.e. $I(\bar{X}, Y) \lessapprox I(X, Y)$). In other words, the compressed data, given in the form of clusters, should express as much information about anatomical connectivity as possible as shown in figure 4.4.

Tishby et al. (33) propose a variational principle reminiscent of rate distortion theory that quantizes the tradeoff between compressing the data and preserving infor-

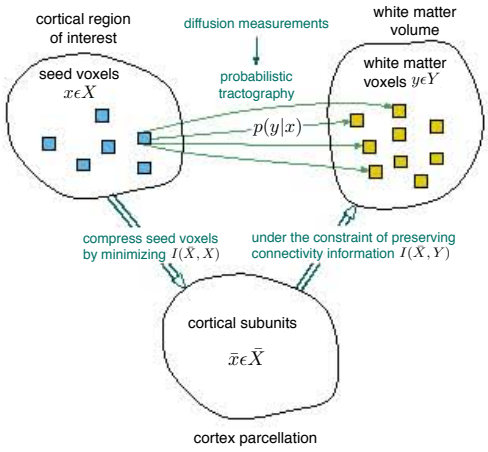

Figure 4.4: Information bottleneck method applied to cortex parcellation - Probabilistic tractography characterizes the connectivity pattern of a cortical region by a distribution of connectivity from seed voxels to white matter voxels $y\epsilon Y$. The notion of cortex parcellation is to compress seed voxels to form cortical subunits by minimizing $I(\bar{X},X)$. Compression is guided by preservation of anatomical connectivity information, $I(\bar{X},Y)$, such that the connectivity fingerprint of cortical subunits is as expressive as possible of the collective connectivity pattern of seed voxels.

mation relevant for anatomical connectivity:

$$F[p(\bar{x}|x),p(y|\bar{x}),p(\bar{x})] = I(\bar{X},X) - \lambda I(\bar{X},Y), \tag{4.4}$$

where λ is the Lagrange multiplier that controls the tradeoff. The cost functional functional F is minimized with respect to variables $p(\bar{x}|x),p(y|\bar{x}),p(\bar{x})$. $p(\bar{x}|x)$ quantifies the probability of assigning seed voxel x to cluster $\bar{x}$ and $p(y|\bar{x})$ is the probability of finding a connection between cluster $\bar{x}$ and white matter voxel y. Minimizing the cost functional (equation 4.4) yields the following set of self-consistent equations:

$$\begin{cases} p_t(\bar{x}|x) = \frac{p_t(\bar{x})}{Z_t(x,\beta)}\exp(-\lambda D_{KL}[p(y|x)||p(y|\bar{x})]) \\ p_{t+1}(\bar{x}) = \sum_x p(x)p_t(\bar{x}|x) \\ p_{t+1}(y|\bar{x}) = \sum_x p(y|x)p_t(x|\bar{x}), \end{cases} \tag{4.5}$$

where t denotes the iteration sequence and $Z_t(x,\beta)$ serves as a normalization constant. It is easily verified that iterating over convex sets given above guarantees convergence of the cost functional F (33). The cluster assignment probabilities $p(\bar{x}|x)$ quantifies the contribution of the connectivity of seed voxel x to the connectivity fingerprint of cortical subunit $\bar{x}$, which results in cortex parcellation.

It is important to appreciate the advantage of the information bottleneck principle to most clustering algorithms; the dissimilarity between the distributional nature of tractograms x and clusters $\bar{x}$ is captured by the Kullback-Leibler distance (i.e. $D_{KL}[p(y|x)||p(y|\bar{x}])$) in equation 4.5 which arises as a result of minimizing the cost functional F. The dissimilarity measure is *not* explicitly defined. Furthermore, the information bottleneck method does not make any assumptions about the shape of clusters. More importantly, this clustering method seeks to identify the connectivity fingerprints, $p(Y|\bar{x})$, for each cortical subunits (i.e. cluster) $\bar{x}$ that together preserve most of the connectivity information of individual seed voxels (i.e. $p(Y|x)$).

The self-consistent equations 4.5 require initialisation of variables $p(\bar{x}|x)$, $p(\bar{x})$ and $p(y|\bar{x})$. Different initializations can lead to different partitionings which correspond to different local minima of F (33). Tishby et al. (33) propose a simulated annealing procedure (section 3.1.2.2) that incrementally increases λ from $\lambda = 0$ in order to "track" the changes in the effective partitioning as the system shifts its preferences from compression to preservation of anatomical connectivity. Notice the case where $\lambda = 0$, the $I(\bar{X}, Y)$ term vanishes and there is no preservation of connectivity information and the resulting compression of seed voxels forms one effective cluster that minimizes the complexity term $I(\bar{X}, X)$. That is the intialisation at $\lambda = 0$ has no influence on the clustering solution. Gradually increasing λ and iterating over self-consistent equations 4.5 until convergence results in the optimal connectivity-compression tradeoff and thus the global minimum F for each λ. Moreover, the connectivity-compression tradeoff demonstrates a sequence of phase transitions at critical λ values that explores a hierarchy of effective partitionings, $\mathcal{K}$, as shown in figures 5.4 and 5.5 (33).

In contrast to stochastic models, the advantage of using a cost or expected distortion function for clustering is that they are more tractable. In other words, their clustering behaviour is more easily understood and managed. For example, prior to clustering, we can rank the contribution of individual white matter voxels on the clustering solution by computing their contribution to the anatomical connectivity information (59):

$$I(X, y) = p(y) \sum_{x \epsilon X} p(x|y) \log \frac{p(x|y)}{p(x)} \tag{4.6}$$

Another interesting application of the information bottleneck method is to form compact representations of white matter voxels $\bar{Y}$ prior to forming clusters of seed voxels

$\bar{X}$ (59). The resulting compact representation of the data $\{\bar{X}, \bar{Y}\}$ should therefore capture as much information about the original data $\{X, Y\}$ as possible:

$$I(\bar{X}, \bar{Y}) \lessapprox I(X, \bar{Y}) \lessapprox I(X, Y) \tag{4.7}$$

Forming $\bar{Y}$ as well as computing $I(X, y)$ have the useful advantage of allowing for feature reduction since anatomical connectivity data given by probabilistic tractograms are often difficult to process due to their large data size.

4.4 Model Validation

Given the hierarchy of effective partitionings the following question arises: What is the upper limit on the number of effective clusters $\mathcal{K}$? As λ increases the number of effective clusters $\mathcal{K}$ rise and we resolve more structure from the data until $\mathcal{K} = K$ at $\lambda \to \infty$. In practice, λ should be set sufficiently high such that stochastic assignments, $p(\bar{x}, x)$, effectively reach the deterministic limit ($\lambda \to \infty$). The upper limit is achieved at $\lambda \to \infty$ where the number of effective partitionings equals the predefined number of clusters (i.e. $\mathcal{K} = K$). As is the case with most clustering algorithms the information bottleneck method requires defining the number of clusters K *a priori*. More precisely, λ controls the number of effective clusters $\mathcal{K}$, however, the upper boundary on the number of effective clusters (i.e. $\max(\mathcal{K}) = K$) is predefined.

This section follows the intuitive notion that noise in the data limits the amount of meaningful structure that can be resolved from the data (41). That is, resolving finer partitionings beyond a limit would not yield more information since one would simply be fitting the sampling noise in the data. In the context of connectivity-based cortex parcellation noise in the diffusion images is caused by the MR scan, physiological as well as motion affects, etc. The tractogram and consequently the connectivity distribution $p(Y|x)$ is therefore subject to fluctuations due to noise in the diffusion images which in turn causes fluctuations in the solution space, $p(Y|\bar{x})$ and $p(\bar{x}|x)$. The task of finding the upper limit on the number of clusters is therefore to maximize the informativeness of the partitioning while remaining robust against noise in the data.

4.4.1 Correcting for Finite Samples

The preservation of anatomical connectivity information, $I(\bar{X},Y)$, used in the information bottleneck method measures the informativeness of the partition. However, the reason why $I(\bar{X},Y)$ cannot be used as a model selection criterion is because it does *not* include a reguralization term that avoids overfitting the data (i.e. fitting the sampling noise). That is, the formulation of the information bottleneck method assumes that we have access to the true continuous distribution $p(Y|x)$. The amount of connectivity information that we resolve, $I(\bar{X},Y)$, will therefore strictly increase for increasing λ and increasing number of clusters K. Clearly the distribution $p(Y|x)$ is subject to errors and therefore noise since $p(Y|x)$ is derived from finite samples due to the finite resolution of diffusion images as well as finite samples obtained by probabilistic tractography. Such errors limits λ and therefore the amount of anatomical connectivity information we can preserve. Correcting for errors in the distribution (i.e. $\hat{p}(y|x) = p(y|x) + \delta p(y|x)$) yields an upper ($UB$) and lower bound ($LB$) of $I(\bar{X},Y)$) (41):

$$\begin{cases} I(\bar{X},Y)^{\text{reg}}_{UB} = I(\bar{X},Y) - \frac{2^{I(\bar{X},Y)}}{2\ln(2)N} \\ I(\bar{X},Y)^{\text{reg}}_{LB} = I(\bar{X},Y) - \frac{2^{I(\bar{X},Y)}}{2\ln(2)N} N_{bins}, \end{cases} \tag{4.8}$$

where $I(\bar{X},Y)^{\text{reg}}_{UB/LB}$ is a reguralized mutual information term. N_{bins} is the number of white matter voxels (i.e. $N_{bins} = |Y|$) used to characterize the connectivity probability and is controlled by the finite resolution of the dMRI data. N is the total number of observation of seed voxels x and white matter voxels y (i.e. $f(X) \times f(Y)$, where f denotes the frequency of observations). N is therefore determined by the number of times we sample seed voxels and the number of trials used per seed voxel to perform probabilistic tractography. Note that changing the resolution affects both the number of white matter voxels N_{bins} and the total number of observations N.

In the deterministic limit (i.e. $\lambda \to \infty$) both $I(\bar{X},Y)^{\text{reg}}_{UB}$ and $I(\bar{X},Y)^{\text{reg}}_{LB}$ coincide. Henceforth, we can correct $I(\bar{X},Y)_{\lambda\to\infty}$:

$$I(\bar{X},Y)^{\text{reg}}_{\lambda\to\infty} = I(\bar{X},Y)_{\lambda\to\infty} - \frac{N_{bins}}{2\ln(2)N}K, \tag{4.9}$$

By correcting for finite sample effects as proposed by Still and Bialek (41) we can resolve the maximum number of clusters from the data at which $I(\bar{X},Y)^{\text{reg}}_{\lambda\to\infty}$ yields a maximum or at least a platform.

4.5 Hierarchical Organization of Cortical Subunits

Our approach attempts to reveal the possible hierarchically modular architecture of cortical subunits. Although, the information bottleneck method used in a simulated annealing procedure already reveals a hierarchy of effective partitionings, we wish to investigate the possible nested structure of partitionings by comparing the overlap of parcellation results in the deterministic limit ($\lambda \to \infty$) at different model orders (i.e. different K) that have been obtained independently of each other. That is, the partitioning of cortical subunits at any level in the hierarchy does not depend upon the partitioning of cortical subunits at any other level. In this manner, we do not enforce a hierarchy of partitionings on the data as demonstrated by common hierarchical clustering methods, but instead aim to investigate the possibility of an existing nested structure in anatomical connectivity data. This method is inspired by Slonim et al. (35) who used the same method to investigate the relation between hard clustering solutions.

5

Results

The purpose of this dissertation is to demonstrate a proof of concept. The regions of interest considered, namely the IFG and IFG+PCG, are therefore regions for which the anatomical segregation has been relatively well established based upon cytoarchitectonic maps as well as previous connectivity-based cortex parcellation studies. The IFG and the IFG+PCG regions were parcellated seperately in order to investigate whether or not parcellation of a particular region will produce similar delineation of cortical areas to independently obtained parcellation results of only a subset of that region. The following section demonstrates the performance of the information bottleneck method on clustering synthetic data followed by an application on cortex parcellation.

5.1 Compressing Synthetic Data

In the interest of demonstrating the usefulness of the information bottleneck method in compressing anatomical connectivity data, the synthetic probability distributions were chosen to represent the nature of connectivity probability distributions as close as possible. We therefore generate our samples from a mixture of discrete, univariate and i.i.d. probability distributions. The nature and shape of the synthetic distribution is analogous to the connectivity patterns within the white matter volume since both data sets are univariate, discrete and i.i.d.. Note that the information bottleneck method imposes no constraints on the shape of the distributions. Several beta distributions with different shape parameters were therefore chosen to explore the method's effectiveness at resolving any shape as shown in figure 5.1.

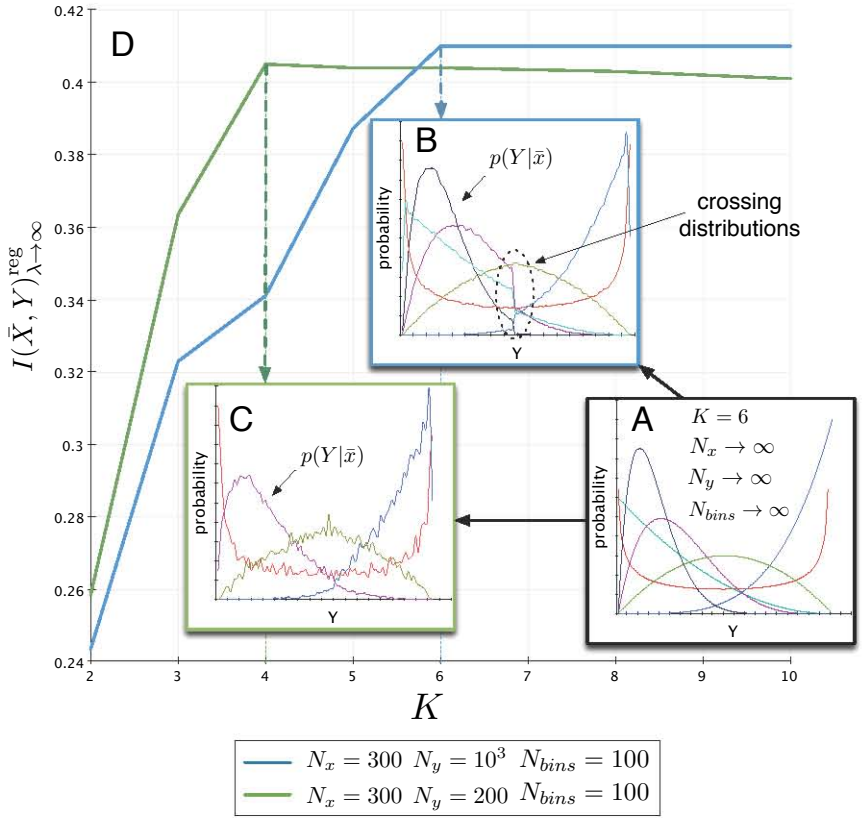

Figure 5.1: Demonstration of the information bottleneck framework on synthetic data - **A**: The synthetic probability distribution consists of 6 different beta distributions with different shape parameters. **B and C**: Data sets B and C were sampled as follows: A distribution in A is chosen randomly for each object x. Each object x is defined by a stochastic dependency, $p(Y|x)$, on features Y that are sampled from that distribution in A and then binned in equal sized bins. The information bottleneck method compresses objects, each defined by a sample distribution $p(Y|x)$, so that the representative distributions $p(Y|\bar{x})$ of clusters x reveal the probability distributions from which features Y were sampled. In order to investigate the influence of noise in the solution both sample data sets, B and C, were assigned equal bins (i.e. $N_{bins} = 100$) but varying number of samples N_y (i.e. B: $N_y = 10^3$ and C: $N_y = 200$). Data set sample C therefore contains more noise than data set sample B due to fewer samples N_y taken from A. **D** As expected, in the hard clustering limit (i.e. at $\lambda \to \infty$), the reguralized, $I(\bar{X}, Y)^{\text{reg}}_{\lambda\to\infty}$, resolves fewer distributions in C ($K = 4$) than in B (i.e. $K = 6$) due to greater noise in C. Reconstruction of the distributions appear to be poor in regions where there are many distribution crossings as shown in B.

The objective of the information bottleneck method in cortex parcellation is to identify the probability distributions from which the samples are taken. Noise in the data is controlled by the number of samples taken from the distributions together with their binning (i.e. fewer bin observations results in greater noise in the data and vice versa). Two sample data sets that keep the number of bins fixed and vary only the number of samples taken were generated from the synthetic probability distributions. As expected by keeping the number of bins fixed and increasing the number of samples observed, thereby reducing noise, we should be able to resolve more structural information from the data (i.e. resolve more clusters from the data) as shown in figure 5.1. Notice, however, that the method capability at resolving representative distributions is limited in regions where the distributions cross as shown in figure 5.1B. Such distribution crossings are analogous to fiber crossings in the context of connectivity-based cortex parcellation.

Another interesting behaviour of the information bottleneck method is the occurrence of phase transitions by varying the tradeoff parameter λ. Figure 5.2 illustrates phase transitions in the synthetic data for increasing λ.

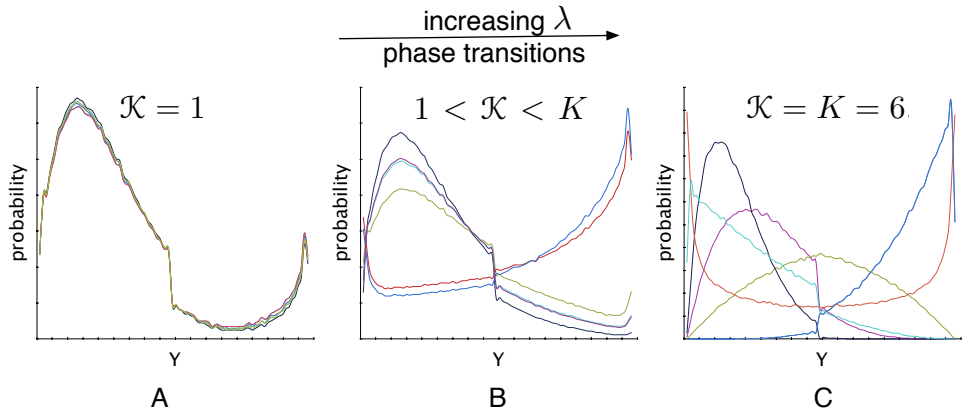

Figure 5.2: Phase transitions in synthetic data - Each graph contains $K = 6$ distributions, however, their distinguishability varies with λ. Increasing λ causes phase transitions that increases the number of distinguishable distributions and therefore the number of effective clusters $\mathcal{K}$. **A**: Distributions are indistinguishable. **B**: Distributions are more distinguishable from each other. **C**: Distributions are maximally distinguishable from each other.

At low λ (figure 5.2A) the representative distributions $p(Y|x)$ are indistinguishable

from each other. In other words, the information bottleneck method has produced $K = 6$ copies of the same probability distribution. All objects are therefore equally distributed among all representative probability distributions and the number of effective clusters is thus $\mathcal{K} = 1$. At a higher λ (figure 5.2B) the probability distributions become more distinguishable, however, some copies still remain. This implies that only a subset of the objects are distributed equally among the copies and the number of effective partitionings is $1 < \mathcal{K} < K$. At $\lambda \to \infty$ the distributions become maximally distinguishable and no copies remain (figure 5.2C). The number of clusters is therefore $\mathcal{K} = K$.

A phase transition is said to occur at any finite or critical λ where a copy or copies are removed and the representative probability distributions, $p(Y|x)$, become more distinguishable from one another. A hierarchy of effective partitionings is established because dissimilar representative probability distributions tend to distinguish between themselves already for greater data compression (i.e. at lower λ) than similar representative distributions. This property agrees with our intuitive understanding of a hierarchy of connectivity patterns outlined in the problem setting (chapter 2): connectivity patterns grouped at lower levels of the hierarchy (i.e. higher data compression) should be more distinguishable from each other than those grouped at higher levels (i.e. less data compression).

5.2 Connectivity-based Cortex Parcellation

5.2.1 Feature Reduction

In order to reduce computational effort we reduced the number of white matter voxels used as features in the information bottleneck method by ranking them according to their contribution to connectivity information $I(X, y)$ (equation 4.6). Figure 5.3 plots the ranked cummulative connectivity information contribution. A cut-off at 90% of the overall information $I(X, Y)$ reduced the number of features to 300 000 white matter voxels.

5.2.2 Connectivity-compression Plane

As stated previously, the objective of the information bottleneck method is to form compact representations of seed voxels, $x \epsilon X$, given in the form of cortical subunits, $\bar{x} \epsilon \bar{X}$

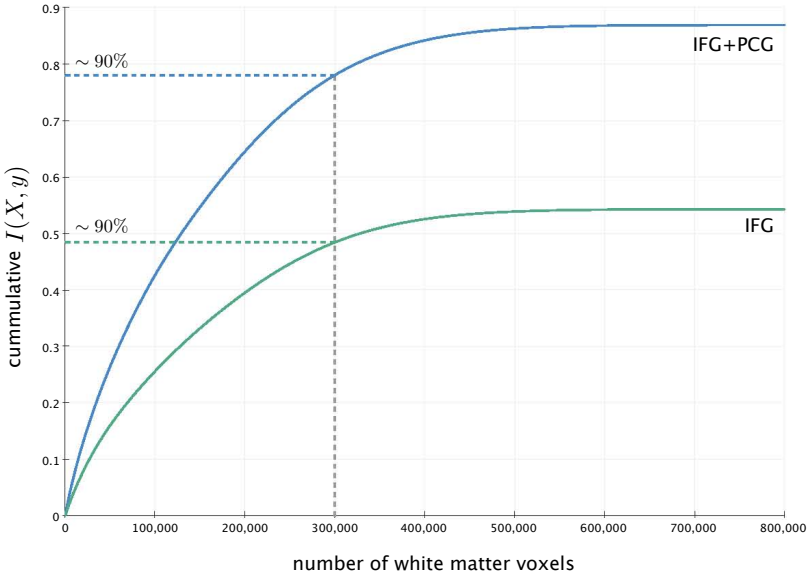

Figure 5.3: Feature reduction for the IFG and IFG+PCG - The ranked cummulative information contribution $I(X, y)$ of white matter voxels is plotted against the number of white matter voxels. 90% of the overall information $I(X, Y)$ for each region of interest reduces the number of features to 300 000 white matter voxels.

based upon their connectivity information to the whole white matter volume, $I(X, Y)$. The tradeoff between compression $I(\bar{X}, X)$ and connectivity information $I(\bar{X}, Y)$ is shown in the connectivity-compression plane, also termed the information plane. Figures 5.4 illustrates the connectivity-compression curves in the information plane for the IFG+PCG region of interest.

Minimizing the cost functional F (equation 4.4) constructs the optimal connectivity-compression curve that quantifies the optimal tradeoff between compression and anatomical connectivity. Such an optimal curve seperates two regions in the information plane:

- The region below the curve for which the connectivity-compression tradeoff is achievable but suboptimal.
- The region above the curve for which the connectivity-compression tradeoff is not achievable.

Given an upper boundary on the number of representative distributions, the monotonical behaviour of the connectivity-compression curves in figure 5.4 is explained as follows:

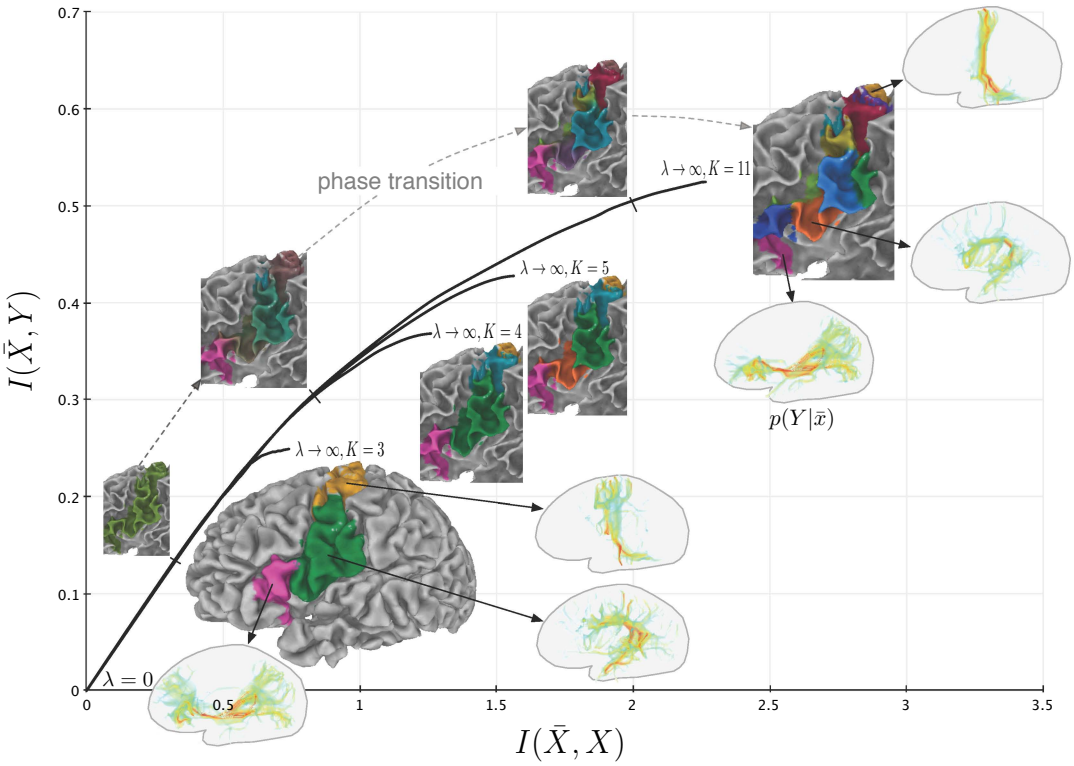

Figure 5.4: Connectivity-compression plane for the IFG+PCG - The tradeoff between compression, $I(\bar{X}, X)$, and anatomical connectivity information, $I(\bar{X}, Y)$, is controlled by gradually increasing λ from 0 to ∞ in a simulated annealing procedure that constructs the optimal connectivity-compression curves for $K = 3, 4, 5, 11$ clusters. Each curve demonstrates a sequence of phase-transitions that explores a hierarchy of effective partitionings for a given number of clusters K. White matter surfaces showing connectivity-based cortex parcellation results as well as the connectivity fingerprints, $p(Y|\bar{x})$, of individual cortical subunits are also shown. Red regions in the depicted connectivity fingerprints indicate areas to which the associated cortical subunit has a relatively high probability of connectivity.

At one extreme, $\lambda = 0$, the data compression, $I(\bar{X}, X)$, is maximal. Consequently, all representative distributions, $p(Y|\bar{x})$, of the connectivity data are indistinguishable from each other. Henceforth, there is no preservation of connectivity information (i.e. $I(\bar{X}, Y) = 0$). Accordingly, cortical subunits are indistinguishable from each other within the cortical region of interest because their connectivity patterns are indistinguishable from each other. The number of effective clusters thus equals one (i.e. $\mathcal{K} = 1$) because all seed voxels are equally distributed among the cortical subunits.

As λ gradually increases, the connectivity information constraint becomes more demanding. At some finite or critical λ the system undergoes a phase transition where preference shifts from compression to preservation of connectivity information. Consequently, some representative connectivity patterns $p(Y|\bar{x})$ become distinguishable from each other which translates into distinguishable cortical subunits within the region of interest. In other words, cortical subunits are distinguishable from each other within the region of interest because their representative connectivity patterns are distinguishable from each other.

At the other extreme, where $\lambda \to \infty$, the system shifts its preference entirely on the preservation of connectivity information at the cost of minimal compression. All K representative distributions are consequently distinguishable from each other and, hence, the number of distinguishable cortical subunits (i.e. number of effective clusters $\mathcal{K}$) is thus equal to the predefined upper boundary (i.e. $\mathcal{K} = K$). Figure 5.4 depicts a connectivity-compression curve for each predefined K.

Notice that the representative connectivity patterns of cortical subunits (i.e. connectivity fingerprint) are plotted in a three dimensional space given by the white matter volume (figure 5.4) as opposed to the one dimensional space given to synthetic data (figure 5.1). However, bear in mind that the nature of both data sets is the same: They are both i.i.d univariate probability distributions. That is, the connectivity distribution contains only one random variable: the white matter voxel to which the cortical area of interest is connected. In the context of connectivity analysis, the distribution of connectivity is only sensible when plotted in a three-dimensional volume within the brain. The same method was applied to a subset of the IFG+PCG region, namely the IFG region only (figure 5.5), in order to investigate whether or not we would obtain the same delineation of cortical areas as those discovered in the IFG+PCG region of interest.

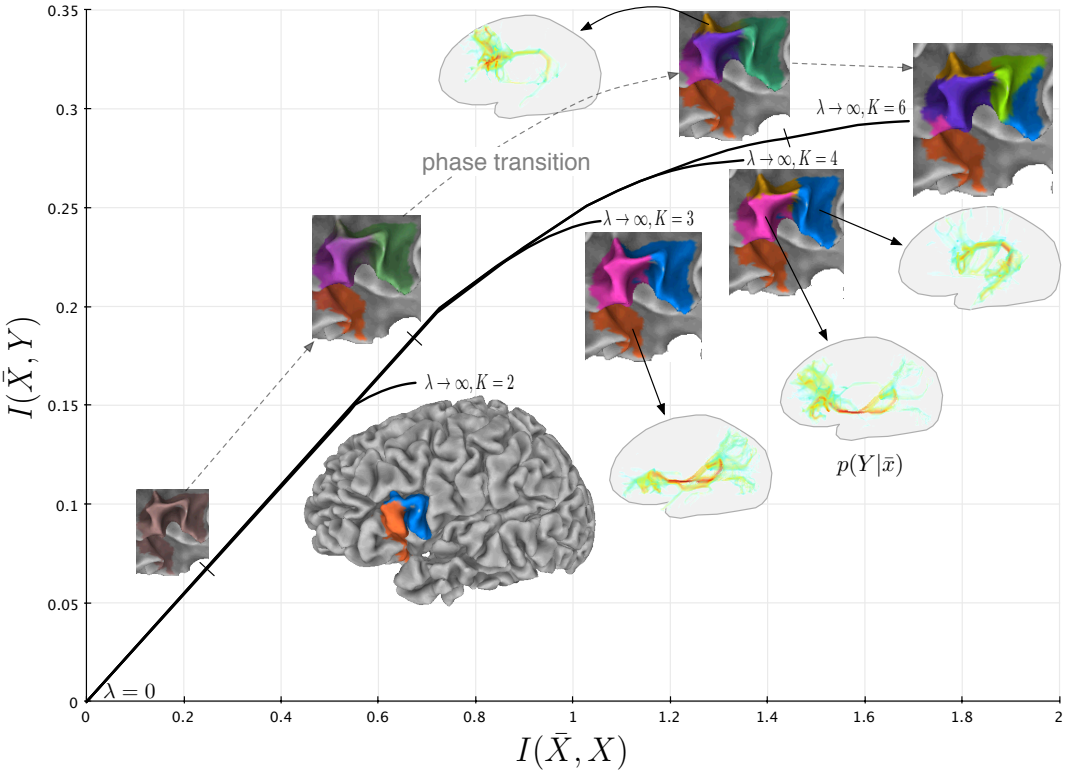

Figure 5.5: Connectivity-compression plane for the IFG - The tradeoff between compression, $I(\bar{X}, X)$, and anatomical connectivity information, $I(\bar{X}, Y)$, is controlled by gradually increasing λ from 0 to ∞ in a simulated annealing procedure that constructs the optimal connectivity-compression curves for $K = 2, 3, 4, 6$ clusters. Each curve demonstrates a sequence of phase-transitions that explores a hierarchy of effective partitionings for a given number of clusters K. White matter surfaces showing connectivity-based cortex parcellation results as well as the connectivity fingerprints, $p(Y|\bar{x})$, of individual cortical subunits are also shown. Red regions in the depicted connectivity fingerprints indicate areas to which the associated cortical subunit has a relatively high probability of connectivity.

5.2.3 Model Validation

Increasing the number of clusters K and increasing λ resolves more connectivity information, $I(\bar{X}, Y)$, from the data as shown in figures 5.4 and 5.5. Clearly, the resolution in the imaging data limits the amount of connectivity-information, $I(\bar{X}, Y)$, we can resolve: Finite resolution together with the finite number of trials used in probabilistic tractography results in finite connectivity samples which is the source for noise in the data. The question is therefore: How much connectivity information, $I(\bar{X}, Y)$, can we resolve from the imaging data without fitting the sampling noise? By correcting $I(\bar{X}, Y)$ for finite samples we obtain a reguralized connectivity information term, $I(\bar{X}, Y)^{\text{reg}}_{\lambda\to\infty}$, (equation 4.9) at $\lambda \to \infty$ that should yield a maximum or at least a platform for increasing K. If $I(\bar{X}, Y)^{\text{reg}}_{\lambda\to\infty}$ does not increase with K we are simply fitting the sampling noise and will therefore not resolve any more meaningful connectivity structure from the data. Figure 5.6 plots $I(\bar{X}, Y)^{\text{reg}}_{\lambda\to\infty}$ versus K for the IFG and IFG+PCG.

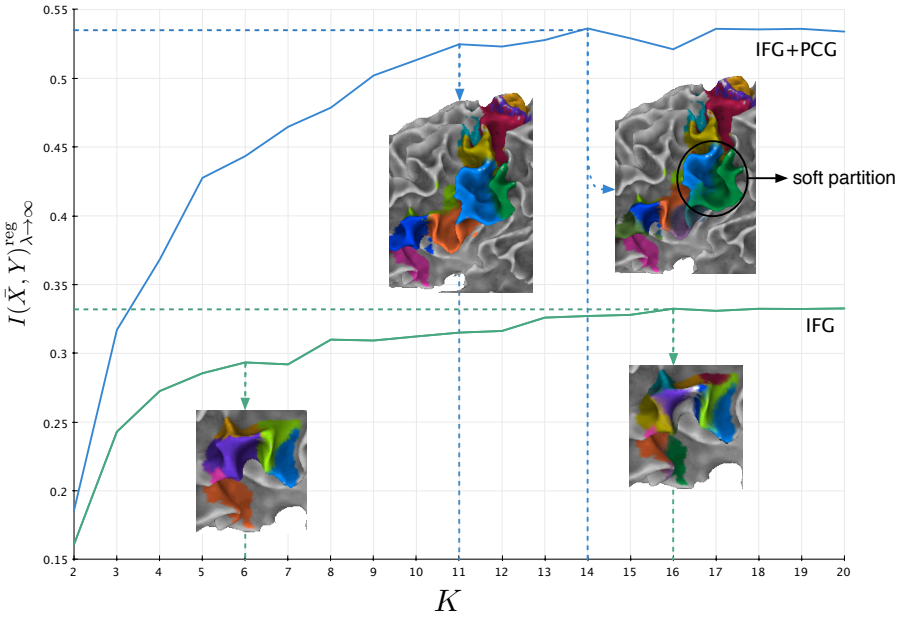

Figure 5.6: Model validation for the IFG and IFG+PCG - The reguralized connectivity information, $I(\bar{X}, Y)^{\text{reg}}_{\lambda\to\infty}$, reaches a maximum at $K = 16$ for the IFG and $K = 14$ for the IFG+PCG. Violations of the strictly monotonically increasing behaviour (before reaching a maximum or platform) of $I(\bar{X}, Y)^{\text{reg}}_{\lambda\to\infty}$ are observed beyond $K = 6$ for the IFG and $K = 11$ for the IFG+PCG.

Note that $I(\bar{X}, Y)^{\text{reg}}_{\lambda\to\infty}$ is supposed to be strictly monotonically increasing for increasing number of clusters K until a maximum or platform is reached as demonstrated in figure 5.1. However, for large data sets used in connectivity-based cortex parcellation, increasing the number of clusters K introduces greater computational complexity which may lead to poor convergence of self-consistent equations 4.5. Moreover, such increased complexity may result in λ not reaching the deterministic limit as shown in figure 5.6 (i.e. λ was not set sufficiently high such that stochastic assignments effectively reach the deterministic limit, $\lambda \to \infty$). Consequently, the cost functional F (equation 4.4) requires more computational effort to reach the global minimum at $\lambda \to \infty$ for increased number of clusters K. Violations of the strictly monotonical behaviour of $I(\bar{X}, Y)^{\text{reg}}_{\lambda\to\infty}$ are observed in figure 5.6 for the number of clusters beyond $K = 6$ for the IFG and $K = 11$ for the IFG+PCG. Such violations indicate that the cost functional F did not reach the global miniumum at $\lambda \to \infty$. The maximum number of clusters for the IFG and IFG+PCG were therefore chosen as $K = 6$ and $K = 11$, respectively.

Connectivity fingerprints of the IFG+PCG are shown in chapter 9 (figure 9.1) for each cortical subunit for $K = 5$ and $K = 11$ clustering solutions at $\lambda \to \infty$.

5.2.4 Hierarchical Organization of Cortical Subunits

Instead of imposing a hierarchy on the connectivity data, we want to investigate whether or not there is evidence of a hierarchically modular organization of cortical subunits and their connectivity fingerprints. Using several *independent* partitionings of the cortical region of interest at different model orders (i.e. different K) we can investigate their overlap with one another. In figure 5.7 we see that an approximate hierarchy of cortical subunits emerges in the IFG+PCG region showing coarser cortical structures attributed to primary as well as pre-motor and those associated with pre-frontal areas. The levels of clustering in figure 5.7 were selected according to maximal overlap of clusters at different model orders K. A description of individual cortical subunits within the IFG+PCG according to previous parcellation studies is also given in figure 5.7D.

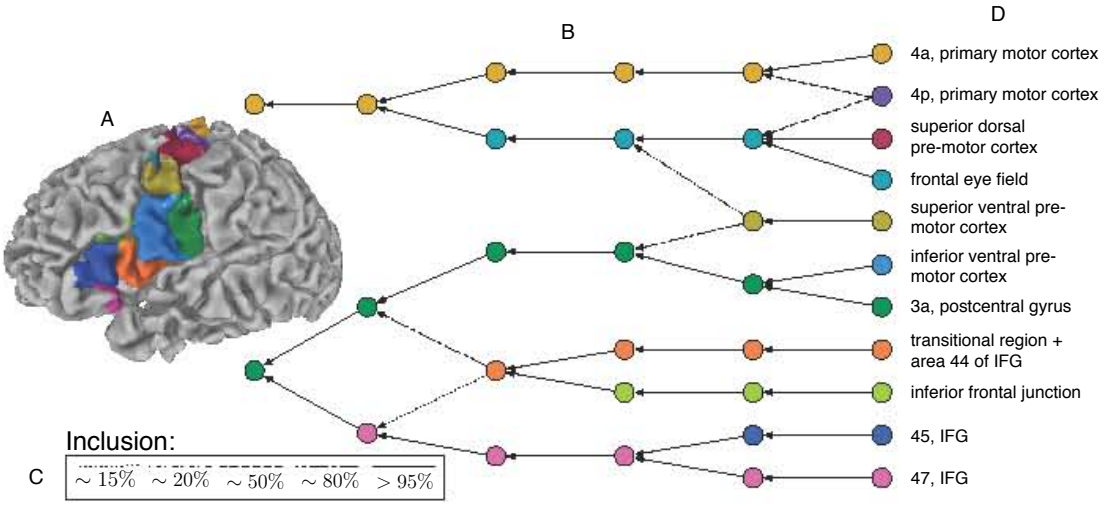

Figure 5.7: Approximate hierarchy of the IFG+PCG - **A**: Gray matter surface showing connectivity-based cortex parcellation of the IFG+PCG region into $K = 11$ cortical subunits (i.e. clusters). **B**: Approximate hierarchy showing the overlap of several independent hard clustering solutions ($\lambda \to \infty$) at $K = 2, 3, 5, 6, 8, 11$ within the IFG+PCG region. The hierarchy of partitionings at $K = 5$ reveals coarser cortical structures that differentiate between primary as well as pre-motoric areas and those associated with prefrontal areas. **C**: Legend indicating the magnitude of cluster overlaps in the hierarchy. **D**: A comparison to previous parcellation studies provides a description of individual cortical areas.

6

Discussion

6.1 Anatomical Interpretation

Parcellation of the IFG+PCG at K=3 clusters: The precentral gyrus (PCG) is primarily divided into two global structures, namely a dorsal area (yellow) and a ventral area (green) together with a transition into the posterior inferior frontal gyrus (IFG) at the ventral tip of the PCG. Concerning the convexity of the PCG, the average Talairach coordinate of the border between ventral and dorsal areas was 49 which is consistent with other reports from functional imaging studies (60) and previous connectivity-based parcellation studies (13, 14) (figure 6.1). The IFG is also primarily separated into two areas, namely the aforementioned ventral transition of the PCG (green) and the group containing the pars opercularis of the IFG as well the pars triangularis of the IFG together with the deep frontal operculum (pink). The separation between these two areas, demonstrated in figures 5.5 and 5.7, confirms the distinction of areas in the ventral PCG and the posterior ventral precentral cortex.

Parcellation of the IFG+PCG at K=11 clusters: The dorsal PCG can be separated into the primary motor cortex (yellow and purple) (8), a premotor area (red) and the frontal eye field at the rostral bank of the precentral sulcus and the ventral branch of the posterior superior frontal sulcus (61). Moreover, the ventral PCG is further subdivided into a superior-rostral (brown) and an inferior-caudal area (light blue). For validation purposes a part of the (inferior) postcentral gyrus (green) was included within IFG+PCG region of interest which was correctly separated from the PCG. The delineation of the sub-areas in the posterior ventral precentral cortex

resembles results from cytoarchitectonic and multireceptor studies in Amunts et al. (10). This includes previously unknown areas such as the anterior and posterior areas 45a (dark blue) and 45p (light green) as well as the area in the frontal operculum op9 (pink) (figure 6.1). Such delineation is also evident when parcellating only the IFG as shown in figure 5.5.

Anatomically disjoint areas were distinguished, consistent with Amunts et al. (10), one being located in the depths of the inferior frontal sulcus, the other immediately rostrally to the ventral premotor area. Both were found at the junction of the inferior frontal and the precentral sulcus and therefore may correspond to the inferior frontal junction (IFJ) (62, 63). Interestingly, our results accurately reflect the delineation of areas concerning the IFJ obtained by Derrfuss et al. (64). Note that the fMRI data used by Derrfuss et al. (64) were taken from the same subject (subject 2 in Derrfuss et al. (64)). Our results therefore suggest a specific connectivity underlying the IFJ, rendering this region as a distinct anatomical area (figure 6.1).

The merging of the postcentral region (green) with the ventral PCG at a rather high hierarchical level (figure 5.7) seems to be supported by findings in non-human primates, implying dense bidirectional connections between the rostral portion of the inferior parietal lobule and the adjacent opercular area (i.e. ventral premotor area 6 (65)). However, whether this suggestion is indeed evident in probabilistic tractography-based connectivity fingerprints remains to be studied in detail and requires improved visualization techniques of connectivity patterns (66).

6.2 Methodology and Results

Connectivity-based parcellation using probabilistic tractograms has attracted so much interest because probabilistic tractography is a convenient means to characterize the complex nature of anatomical connectivity in the brain. In order to fully exploit probabilistic tractography to learn connectivity structure from the data we have to define the nature of the data appropriately. The interpretation of probabilistic tractograms and therefore their representation within the context of clustering is of utmost importance (6, 67) as it predetermines what cluster structure can be discovered in the data (25). In the information bottleneck framework we have interpreted the connectivity of a cortical

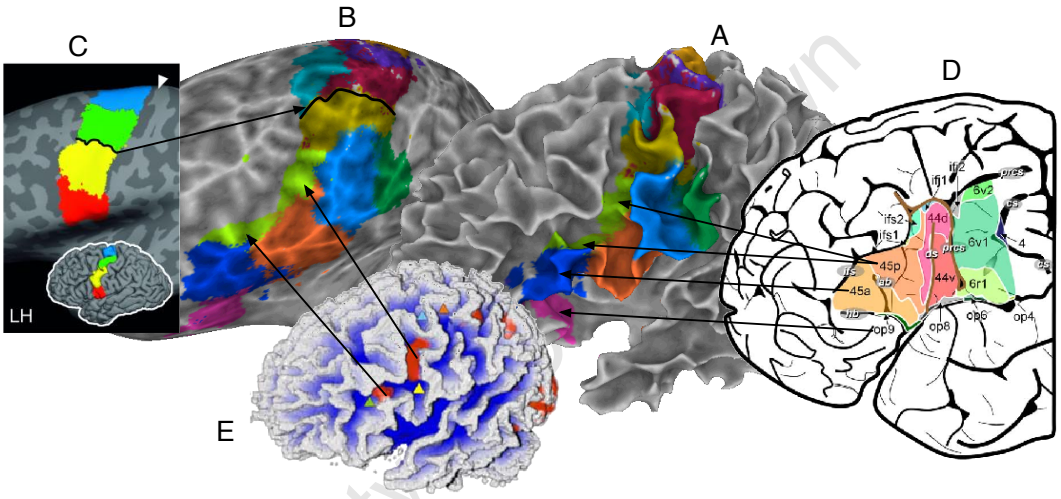

Figure 6.1: Comparison of connectivity-based parcellation results to previous studies - **A**: Parcellation results at $K = 11$ clusters depicted on the white matter surface. **B**: Parcellation results at $K = 11$ clusters on an inflated cortical surface. **C**: Connectivity-based parcellation results of the PCG taken from Schubotz et al. (14) showing a seperation between the dorsal PCG (dPCG) and ventral PCG (vPCG). **D**: Cytoarchitecture and multireceptor studies taken from Amunts et al. (10) showing the delineation of the IFG. **E**: Functional imaging maps of the same subject used in this dissertation taken from Derrfuss et al. (64) showing anatomically disjoint areas corresponding to the inferior frotal junction (IFJ).

area as being distributed among voxels within the white matter volume. Probabilistic tractography allows us to sample white matter voxels connected to those cortical elements of interest. The information bottleneck framework for connectivity-based cortex parcellation proposed in this dissertation offers an appropriate quality measure for grouping distributional data to reveal the connectivity fingerprint underlying a cortical area of interest.

This dissertation demonstrates a proof of concept of the approach which avoids many assumptions imposed on the data as shown by previous attempts and nevertheless produces parcellation results consistent with previous parcellation studies. Such assumptions include the shape of clusters, the similarity measure selected *a priori* as well as the representation of probabilistic tractograms. Essential properties of cortical subunits are revealed such as their connectivity fingerprints and their hierarchically modular architecture.

Dissimilarity between connectivity patterns: Intuitively, connectivity patterns quantified by probabilistic tractograms should be grouped based upon capturing the shape of probabilistic tractograms. In other words, probabilistic tractograms should be grouped together if they have similar shapes. Defining the shape of a probabilistic tractogram is not straightforward since a probabilistic tractogram is a volume containing connectivity probabilities for each white matter voxel. We define two tractograms as having similar shapes if their connectivity probability in corresponding white matter voxels are similar. The similarity measure should therefore involve a pairwise comparison of connectivity scores with pairs of connectivity probabilities. The information bottleneck method captures the shape similarity between tractograms and connectivity fingerprints of cortical subunits using the Kullback-Leibler distance between them. Note that the bins of the connectivity probabilities are defined by white matter voxels. Probabilistic tractography quantifies the frequency at which such bins or white matter voxels occur for each seed voxel. The Kullback-Leibler distance between probabilistic tractograms thus involves pair wise operations between correponding bins and therefore corresponding white matter voxels.

Model validation: Due to the limited prior knowledge about anatomical connectivity within the human brain, the ultimate goal is to perform automatic, unsupervised clustering of probabilistic tractograms to achieve cortex parcellation without relying on

prior assumptions about the structure of the solution. A crucial step towards unsupervised connectivity-based cortex parcellation is therefore model validation. The notion that noise limits the amount of information we can extract from the data provides the rationale behind model validation in order to resolve as many clusters from the noisy data as possible.

Model validation by correcting for finite samples as proposed by Still and Bialek (41) requires computing partitioning solutions in the deterministic limit ($\lambda \rightarrow \infty$). However, resolving more clusters from the data by increasing λ and the number of clusters K introduces more computational complexity associated with iterating over self-consistent equations 4.5 until convergence. In figure 5.6, λ was not set sufficiently high to reach the effective deterministic limit (i.e. $\lambda \rightarrow \infty$) for the number of clusters beyond $K = 6$ for the IFG and $K = 11$ for the IFG +PCG. Consequently, the cost functional F did not reach the global minimum for clusters beyond $K = 6$ for the IFG and $K = 11$ for the IFG +PCG at the effective deterministic limit. The maximum number of clusters were therefore chosen based upon the convergence performance of the cost functional F at $\lambda \rightarrow \infty$. However, such a convergence criteria may be related to computational issues and may not necessarily reflect the true number of clusters resolvable from the data. Appropriately using the model validation proposed by Still and Bialek (41) requires further computation for clusters beyond $K = 6$ for the IFG and $K = 11$ for the IFG+PCG in which case λ should be set sufficiently higher to reach the effective deterministic limit.

It should be noted that deriving the reguralized connectivity information term, $I(\bar{X}, Y)^{\text{reg}}_{\lambda\rightarrow\infty}$, involves many oversimplifications and may therefore not be truly insensitive to noise in order to establish generalizability. A possibly much more effective means to account for noise in the DW images in the context of model selection is to utilize the two sample scenario (i.e. training and test images) widely used in statistics. In this setting, training and test images constitute two parcellation solutions obtained from two seperate DW images of the same person and carried out under the same conditions. The difference between each of the DW images is therefore solely due to noise which in turn causes differences in parcellation results between training and test images. Two DW images provide by far too little information to characterize the uncertainty in diffusion measurements. However, the rationale is that two large enough parcellation results based on seperate DW images contain sufficient information to determine

the uncertainty in the solution space (42). The approximation set coding (ASC) scenario (42) provides an information-based quantity that maximizes the informativeness of the clustering solution under the constraint that the clustering model with associated cost function and input parameters (eg. number of clusters K) be robust against fluctuations in the diffusion measurements. Consequently, the ASC scenario selects the particular parcellation solution with associated cost function and number of clusters that generalizes from training DW image to test DW image under the influence of noise. The ASC scenario allows for model selection among different types of cost functions that parcellate cortical subunits. However, ASC model validation is only applicable for non-negative cost functions. The quality criterion for distributional clustering used in the information bottleneck method would therefore have to be formulated in terms of a non-negative cost function.

Hierarchically modular organization of cortical areas: Another aim of this paper was to investigate the possible hierarchically modular architecture of cortical subunits within the IFG+PCG region. The large overlap of several independently obtained parcellation results at different K confirms the preferred grouping of cortical subunits into primary as well as pre-motoric and those associated with pre-frontal areas. However, a clear limitation of the proposed method used to investigate the hierarchical relationship between cortical subunits is that the levels of partitionings included in the hierarchy were selected by the user. A further step towards allowing connectivity data to vote for its preferred nested architecture is to make use of a quality criterion that captures the quality not only of a single clustering solution but a hierarchy of clustering solutions. The subsequent task validating hierarchical models is to search for the hierarchy that is reproducible under the influence of noise.

Clustering imaging data: It should, however, be noted that the information bottleneck method does *not* address the multi-assignment problem as discussed in the problem setting (chapter 2). This is because, fundamentally, probabilistic assignments of seed voxels to clusters cannot be used to reflect the ratio of different cortical elements within heterogenous voxels. Probabilistic assignments are always associated with uncertainty in the data and can therefore only be used to reflect uncertainty in cluster assignments and not multi-assignments to clusters. Consider sampling a particular heterogenous voxel an infinite number of times. After assigning the heterogenous voxel to clusters an infinite number of times based upon its connectivity properties there should

be no uncertainty remaining as to which clusters the heterogenous voxel belongs to. However, the heterogenous voxel nevertheless belongs to mutiple clusters even in the abscence of uncertainty. The information bottleneck method therefore assumes that all seed voxels are homogenous and produces probabilistic assignments solely due to noise in the data. Although such heterogenous voxels may be given soft assignments we cannot classify them as heterogenous since probabilistic assignments are only due to uncertainty in the data. Note that in cases of poor image resolution the number of heterogenous voxels is expected to rise which further justifies using multi-assignment clustering.

Assessment of results: Ideally we would like to assess the results numerically. However, such assessment is difficult since we don't have access to ground truth and can only compare our results with results obtained by alternative methods. We have primarily used fMRI studies, cytoarchitectonic maps and previous cortex parcellation studies to assess our results. Assessing our parcellation results numerically with such studies may be inconclusive due to limited knowledge with respect to the link between the nature of different data sets as well as the variability of cortical areas across subjects as described in the problem setting (chapter 2).

Notice, however, that the connectivity fingerprints shown in figure 5.4 resemble previously established fiber bundles. Such fiber bundles include the motor tract, the arcuate fasciculus and the inferior fronto-occipital fasciculus. Figure 9.2 in chapter 9 labels selected connectivity fingerprints according to known fiber bundles.

7

Conclusions and Future Work

Based upon cortex parcellation results obtained for the IFG together with the PCG region, the information bottleneck method proves quite useful in modelling the distributional nature of data to segregate a cortical region of interest into cortical subunits consistent with previous parcellation studies without relying on the assumptions imposed by previous methods. The hierarchy of partitionings also reveals plausible results constistent with previous studies. Moreover, the connectivity fingerprints resemble known fiber bundels. Further work includes examining such connectivity fingerprints in greater detail using an improved visualization technique for probabilistic tractography (66).

However, despite the plausible results, the unsupervised framework demonstrated in this dissertation is debatable because a rather subjective assessment of the convergence performance was used to select the maximum number of clusters as opposed to using noise in the data as a model selection criterion. This is due to the violations of the strictly monotonical behaviour of $I(\bar{X}, Y)^{\text{reg}}_{\lambda\to\infty}$ prior to reaching a maximum or platform which indicates that the cost functional F did not reach the global minimum at the effective deterministic limit (i.e. $\lambda \to \infty$) for the number of clusters beyond $K = 6$ for the IFG and $K = 11$ for the IFG+PCG. Convergence of the cost functional F at the determinstic limit is necessary to appropriately perform the model validation technique (41) used in this dissertation. Furthermore, the model validation technique proposed by Still and Bialek (41) may not accurately account for noise in the data due to oversimplifications made in its derivation.

A possibly much more effective means for model selection using approximation set coding (ASC) (42) requires further investigation. Moreover, the levels of the approximate hierarchy shown in figure 5.7 were chosen based upon maximal overlap between partitionings. Model validation of the hierarchy of partitionings is needed in order to establish confidence in the results. Such model selection requires validating quality criteria that capture the quality of nested hierarchy solutions and *not* just "flat" clustering solutions considered in this dissertation.

Although the information bottleneck method adequately deals with the distributional nature of *in vivo* anatomical connectivity data given by probabilistic tractography (i.e. connectivity of cortical subunits is distributed among voxels within the white matter volume) it fails to account for the heterogenous property of the data (i.e. connectivity data may contain hetergenous voxels). Depending on the number of heterogenous voxels, multi-assignment clustering may be more suitable for cortex parcellation. Note that in cases of poor image resolution this is especially the case since the number of heterogenous voxels is expected to rise.

A further step towards understanding the organization of cortical subunits is to study the consistency or heterogeneity of hierarchically modular cortical subunits across subjects. Individual variability, however, is an important issue in anatomical studies, because any given area (even a primary sensory area) can vary in size by twofold or more (68, 69, 70) and because the consistency with which each area is located with respect to topographic boundaries has important implications for physiological and neuroimaging studies. The approximate hierarchy of cortical subunits for the IFG across subjects is shown in chapter 10 (figure 10.1).

8

Appendix A

Equation 3.10 in chapter 2 is essential to demonstrate the link between the information bottleneck method and rate distortion theory. Its derivation will be given in this chapter for completeness (36):

$$\langle d(\bar{x},x)\rangle = \langle D_{KL}[p(y|\bar{x})||p(y|x)]\rangle = \sum_{\bar{x}\epsilon\bar{X}}\sum_{x\epsilon X} p(\bar{x},x)D_{KL}[p(y|\bar{x})||p(y|x)]$$

$$= \sum_{\bar{x}\epsilon\bar{X}}\sum_{x\epsilon X}\sum_{y\epsilon Y} p(\bar{x}|x)p(x)p(y|x)\log\left(\frac{p(y|x)}{p(y|\bar{x})}\right)$$

$$= \sum_{\bar{x}\epsilon\bar{X}}\sum_{x\epsilon X}\sum_{y\epsilon Y} p(\bar{x},x,y)\left(\log\frac{p(y|x)}{p(y)} - \log\frac{p(y|\bar{x})}{p(y)}\right)$$

$$= \sum_{x\epsilon X}\sum_{y\epsilon Y}\left(\sum_{\bar{x}\epsilon\bar{X}} p(\bar{x},x,y)\right)\log\frac{p(y|x)}{p(y)} - \sum_{\bar{x}\epsilon\bar{X}}\sum_{y\epsilon Y}\left(\sum_{x\epsilon X} p(\bar{x},x,y)\right)\log\frac{p(y|\bar{x})}{p(y)}$$

$$= \sum_{\bar{x}\epsilon\bar{X}}\sum_{y\epsilon Y} p(x,y)\log\frac{p(y|x)}{p(y)} - \sum_{\bar{x}\epsilon\bar{X}}\sum_{y\epsilon Y} p(\bar{x},y)\log\frac{p(y|\bar{x})}{p(y)} = I(X,Y) - I(\bar{X},Y)$$

The Markovian independence relation $\bar{X} \leftrightarrow X \leftrightarrow Y$ is used above (i.e. variables $\bar{x}$ and x are each *only* dependent upon variable x):

$$p(\bar{x},x,y) = p(\bar{x}|x,y)p(x,y) = p(\bar{x}|x)p(x,y) = p(\bar{x}|x)p(x)p(y|x),$$

where $p(\bar{x}|x,y) = p(\bar{x}|x)$ due to the Markovian independence relation.

9

Appendix B

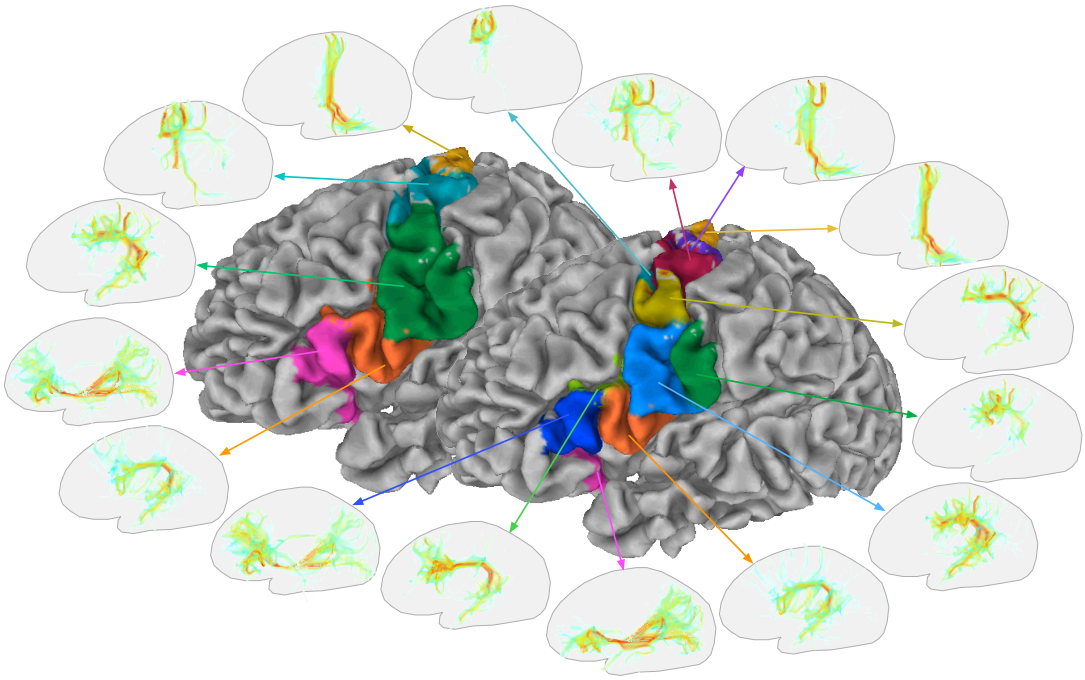

Figure 9.1: Connectivity fingerprints of the IFG+PCG - Cortex parcellation results of the IFG+PCG are shown for $K = 5$ and $K = 11$ clusters. The connectivity fingerprint of each cortical subunit is also shown. Red regions in the depicted connectivity fingerprints indicate areas to which the associated cortical subunit has a relatively high probability of connectivity.

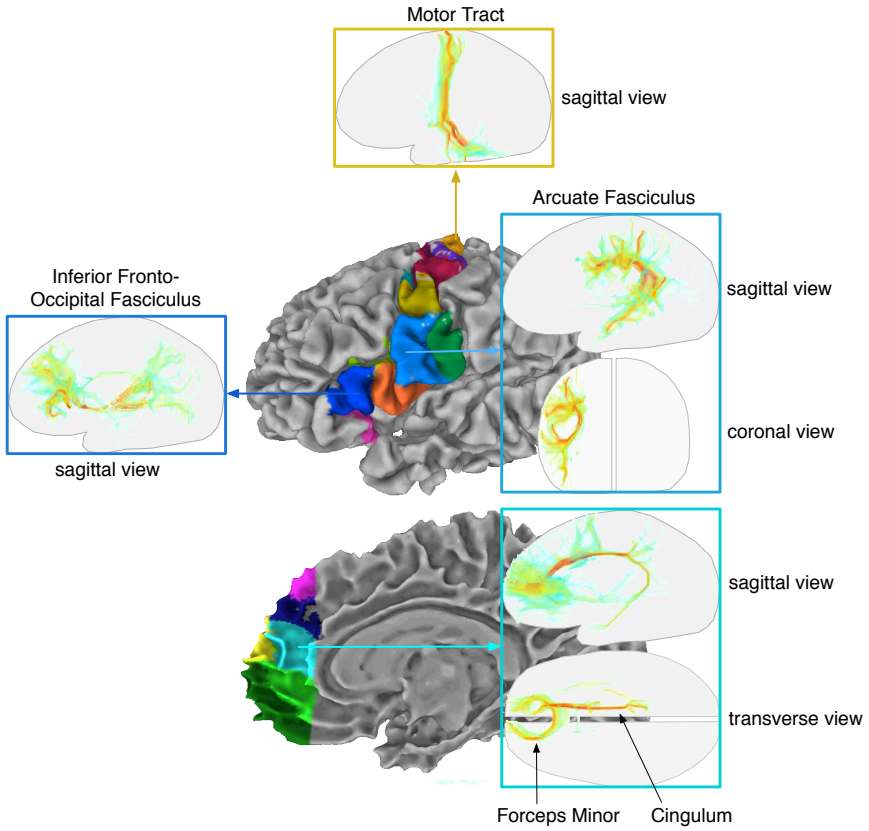

Figure 9.2: Labelled connectivity patterns of the IFG+PCG and the anterior portion of the prefrontal cortex - Cortex parcellation of the IFG+PCG at $K = 11$ and the anterior portion of the prefrontal cortex (aPFC) at $K = 12$ clusters reveals connectivity fingerprints that share resemblance with known fiber bundles such as the motor tract, the arcuate fasciculus, the inferior fronto-occipital fasciculus in the IFG+PCG and the cingulum as well as the forceps minor in the aPFC. Note that model validation was not performed for the aPFC. Ideally, model validation applied to the aPFC to resolve the maximum number of cortical subunits (i.e. clusters) should select a model that splits the cyan region into two cortical subunits such that the cingulum connectivity pattern seperates from the forceps minor.

10

Appendix C

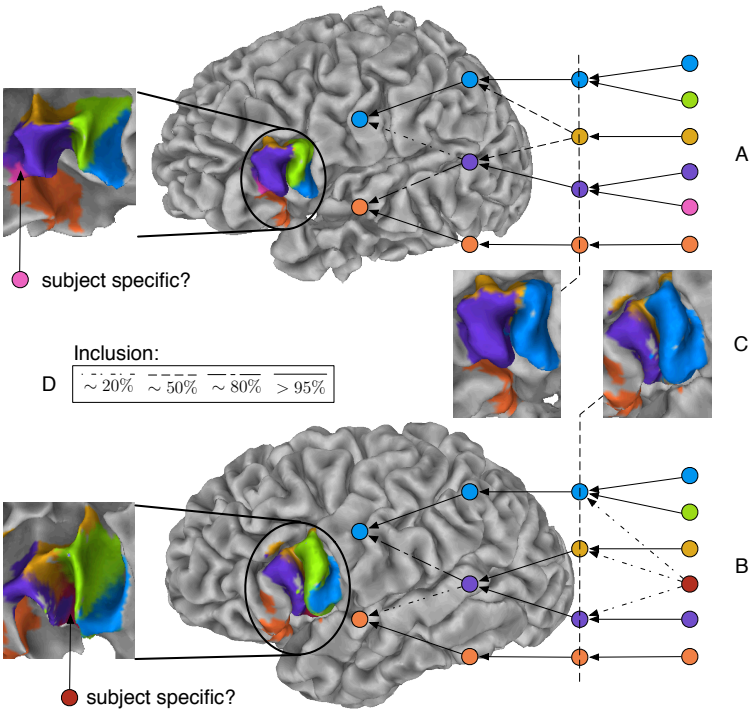

Figure 10.1: Approximate hierarchy of cortical subunits of the IFG for two subjects - The IFG is parcellated into 6 cortical subunits using the information bottleneck method. An interesting question is which level(s) of the hierarchy is/are similar across subject and which level(s) demonstrate(s) considerable variability across subjects. Model validation is crucially important in order to ensure that differences in the hierarchy are not due to noise in the data but are, instead, due to differences in the hierarchical modular architecture of connectivity fingerprints across subjects.

Bibliography

[1] R.E. Passingham, K.E. Stephan and R. Kötter. The anatomical basis of functional localization in the cortex. *Nature Reviews Neuroscience*, 3:606–616, 2002.

[2] C.C. Hilgetag, G.A. Burns, M.A. O'Neill, J.W. Scannell and M.P. Young. Anatomical connectivity defines the organization of clusters of cortical areas in the macaque monkey and the cat. *Phil Trans Royal Soc B*, 355:91–110, 2000a.

[3] K.E. Stephan, C.C. Hilgetag, G.A. Burns, M.A. O'Neill, M.P. Young and R. Kötter. Computational analysis of functional connectivity between areas of primate cerebral cortex. *Phil Trans Royal Soc B*, 355:111–126, 2000.

[4] H.J. Johansen-Berg and M. Rushworth. Using Diffusion Imaging to Study Human Connectional Anatomy. *Annual Revenue of Neuroscience*, 2009.

[5] T.E.J. Behrens, M.W. Woolrich, M. Jenkinson, H.J. Johansen-Berg, R.G. Nunes, S. Clare, P.M. Matthews, J.M. Brady and S.M. Smith. Characterization and Propagation of Uncertainty in Diffusion-Weighted MR Imaging. *Magnetic Resonance in Medicine*, 50:1077–1088, 2003.

[6] D.K. Jones. Challenges and limitations of quantifying brain connectivity in vivo with diffusion MRI. *Imaging Med.*, 2:341–355, 2010.

[7] S. Jbabdi, M. W. Woolrich and T. E. J. Behrens. Multiple-subjects connectivity-based parcellation using hierarchical Dirichlet process mixture models. *NeuroImage*, 44(2):373–384, 2009.

[8] S. Geyer, A. Ledberg, A. Schleicher, S. Kinomura, T. Schormann, U. Bürgel, T. Klingberg, J. Larsson, K. Zilles and P.E. Roland. Two different areas within the primary motor cortex of man. *Nature*, 382:805–807, 1996.

[9] S. Geyer, M. Matelli, G. Luppino and K. Zilles. Functional neuroanatomy of the primate isocortical motor system. *Anat Embryol*, 202:443–474, 2000.

[10] K. Amunts, M. Lenzen, A.D. Friederici, A. Schleicher, P. Morosan, N. Palomero-Gallagher and K. Zilles. Broca's region: novel organizational principles and multiple receptor mapping. *PLoS Biol*, 8, 2010.

[11] A. Anwander, M. Tittgemeyer, D.Y. von Cramon, A.D. Friederici and T.R. Knösche. Connectivity-based Parcellation of Broca's Area. *Cerebral Cortex*, 17: 826–825, 2007.

[12] J.C. Klein, T.E.J. Behrens, M.D. Robson, C.E. Mackay, D.J. Higham and H.J. 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*, 34:204–211, 2007.

[13] V. Tomassini, S. Jbabdi, J.C. Klein, T.E.J. Behrens, C. Pozzilli, P.M. Mathews, M.F.S. Rushworth and H.J. Johansen-Berg. Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations. *Journal of Neuroscience*, 27:10259–10269, 2007.

[14] R.I. Schubotz, A. Anwander, T.R. Knösche, D.Y. von Cramon and M. Tittgemeyer. Anatomical and functional parcellation of the human lateral premotor cortex. *NeuroImage*, 50:369–408, 2010.

[15] A.W. Toga and J.C. Mazziotta. *Brain Mapping: The Trilogy, Three-Volume Set: Brain Mapping: The Systems*. Academic Press, 2000.

[16] B.B. Averbeck, A. Battaglia-Mayer, C. Guglielmo and R. Caminti. Statistical analysis of parieto-frontal cognitive-motor networks. *Journal of Neurophysiology*, 102:1911–1920, 2009.

[17] N.S. Gorbach, C. Melzer, C. Schütte, K. Amunts, T. Douglas and M. Tittgemeyer. Hierarchical Clustering for Connectivity-based Cortex Parcellation. In *16th Annual Meeting of the Organization for Human Brain Mapping*, 2010.

[18] N.S. Gorbach, C. Schütte, C. Melzer, M. Goldau, O. Sujazow, J. Jitsev, T. Douglas and M. Tittgemeyer. Hierarchical Information-based Clustering for Connectivity-based Cortex Parcellation. *Frontiers in Neuroinformatics*, to appear, 2011.

[19] M. Kaiser and C.C. Hilgetag. Optimal hierarchical modular topologies for producing limited sustained activation of neural networks. *Frontiers in Neuroinformatics*, 4:8, 2010.

[20] C.C. Hilgetag and S. Grant. Uniformity, specificity and variability of cortico-cortical connectivity. *Phil Trans Royal Soc B*, 355:7–20, 2000.

[21] N.T. Markov, P. Misery, A. Falchier, C. Lamy, J. Vezoli, R. Quilodran, M.A. Gariel, P. Giroud, M. Ercsey-Ravasz, L.J. Pilaz, C. Huissoud, P. Barone, C. Dehay, Z. Toroczkai, D.C. van Essen, H. Kennedy and K. Knoblauch. Weight Consistency specifies Regularities of Macaque Cortical Networks. *Cerebral Cortex*, 2010.

[22] 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*, in press.

[23] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[24] A. Streich. *Multi-Label Classification and Clustering for Acoustics and Computer Security*. PhD thesis, Swiss Federal Institute of Technology, 2010.

[25] J. Buhmann. Clustering Principles and Empirical Risk Approximation. Technical report, Rheinische Friedrich-Wilhelms-Universität, Institut für Informatik.

[26] D. Lashkari, R. Sridharan, E. Vul, P. Hsieh, N. Kanwisher and P. Golland. Non-parametric Hierarchical Bayesian Model for Functional Brain Parcellation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 15–20, 2010.

[27] S.P. Lloyd. Least square quantization in PCM. *Bell Telephone Laboratories Paper*, pages 129–137, 1982.

[28] B.J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[29] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern*, 22:888–905, 2000.

[30] C. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 20:68–86, 1971.

[31] B. Fischer, T. Zoeller and J. Buhmann. Path-based pairwise data clustering with application to texture segmentation. *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 235–250, 2001.

[32] B. Fischer. *"Complex" Statistical Clustering Models in Image Analysis and Proteomics*. PhD thesis, Swiss Federal Institute of Technology, 2006.

[33] N. Tishby, F. Pereira and W. Bialek. The Information Bottleneck Method. *The 37th annual Allerton Conference on Communication, Control and Computing*, 1999.

[34] R.E. Blahut. Computation of channel capacity using rate-distortion theory. *IEEE Transactions in Information Theory*, 18:460–473, 1978.

[35] N. Slonim, G.S. Atwal, G. Tkacik and W. Bialek. Information-based clustering. *PNAS*, 102:18297–18302, 2005.

[36] N. Slonim. *The Information Bottleneck: Theory and Applications*. PhD thesis, Hebrew University, 2002.

[37] D.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[38] P. Orbanz. *Infinite-Dimensional Exponential Families in Cluster Analysis of Structured Data*. PhD thesis, Swiss Federal Institute of Technology, 2008.

[39] Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

[40] D.M. Blei, T.L. Griffiths and M.I. Jordan. The nested Chinese restaurant process and Bayesian inference of topic hierarchies. *Journal of the ACM*, 57, 2010.

[41] S. Still and W. Bialek. How Many Clusters? An Information-Theoretic Prespective. *Neural Computation*, 16:2483–2506, 2004.

[42] J.M. Buhmann. Information theoretic model validation for clustering. *International Symposium on Information Theory*, 2010.

[43] H.J. Johansen-Berg, T.E.J. Behrens, E.. Sillery, O. Ciccarelli, A. Thompson, S. Smith and P. Matthews. Functional-anatomical validation and individual variation of diffusion tractography-based segmentation of the human thalamus. *Cerebral Cortex*, 15:31–39, 2005.

[44] J.T. Devlin, E.L. Sillery, D.A. Hall, P. Hobden, T.E.J. Behrens, R.G. Nunes, S. Clare, P.M. Matthews, D.R. Moore and H.J. Johansen-Berg. Reliable identification of the auditory thalamus using multi-modal structural analyses. *NeuroImage*, 30:1112–1120, 2006.

[45] J. O'Muircheartaigh, C. Vollmar, C. Traynor, G. Barker, V. Kumari, M. Symms, P. Thompson, J. Duncan, M. Koepp and M. Richardson. Clustering probabilistic tractograms using independent component analysis applied to the thalamus. *NeuroImage*, 54:2020–2032, 2011.

[46] S. Lehericy, M. Ducros, P.F. van De Moortele, C. Francois, L. Thivard, C. Poupon, N.V. Swindale, K. Ugurbil and D. Kim. Diffusion tensor fiber tracking shows distinct corticostriatal circuits in humans. *Ann Neurol*, 55:522–529, 2004.

[47] E. Sillery, R.G. Bittar, M.D. Robson, T.E.J. Behrens, J. Stein, T.Z. Aziz and H.J. Johansen-Berg. Connectivity of the human periventricular-periaqueductual gray region. *J Neurosurg*, 103:1030–1034, 2005.

[48] B. Draganski, F. Kherif, S. Kloppel, P. Cook, D.C. Alexander, G.J.M. Parker, R. Deichmann, J. Ashburner and R.S.J. Frackowiak. Evidence for Segregated and Integrative Connectivity Patterns in the Human Basal Ganglia. *Journal of Neuroscience*, 28:7143–7152, 2008.

[49] D. Bach, S.T. Behren, L. Garrido, N. Weiskopf and R. Dolan. Deep and superficial amygdala nuclei projections revealed in vivo by probabilistic tractography. *Journal of Neuroscience*, 31:618–623, 2011.

[50] R.A. Menke, S. Jbabdi, K.L. Miller, P.M. Matthews and M. Zarai. Connectivity-based segmentation of the substantia nigra in human and its implications in Parkinson's disease. *NeuroImage*, 52:1175–1180, 2010.

[51] A. Ford, K.M. Mcgregor, K. Case, B. Crosson and K.D. White. Structural connectivity of Broca's area and medial frontal cortex. *NeuroImage*, 52:1230–1237, 2010.

[52] M. Beckmann, H.J. Johansen-Berg and M.F.S. Rushworth. Connectivity-based Parcellation of Human Cingulate Cortex and Relation to Functional Specialization. *Journal of Neuroscience*, 29:1175–1190, 2009.

[53] A. Crippa, L. Cerliani, L. Nanetti and J.B.T.M. Roerdink. Heuristics for connectivity-based brain parcellation of SMA/pre-SMA through force-directed graph layout. *NeuroImage*, 54:2176–2184, 2011.

[54] L. Nanetti, L. Cerliani, V. Gazzola, R. Renken and C. Keysers. Group analyses of connectivity-based cortical parcellation using repeated k-means clustering. *NeuroImage*, 47:1666–1677, 2009.

[55] P. Roca, A. Tucholka, D. Riviere, P. Guevara, C. Poupon and J.F. Mangin. Inter-subject connectivity-based parcellation of a patch of cerebral cortex. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2010.

[56] H.J. Johansen-Berg, T.E.J. Behrens, M.D. Robson, I. Drobnjak, M.F.S. Rushworth, 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. *PNAS*, 101:13335–13340, 2004.

[57] M. Jenkinson, P. Bannister, M. Brady and S.M. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17:825–841, 2002.

[58] A.M. Dale, B. Fischle and M.I. Sereno. Cortical surface-based analysis; segmentation and surface reconstruction. *NeuroImage*, 9:179–194, 1999.

[59] 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, 2000.

[60] G. Rizzolatti, L. Fogassi and V. Gallese. Motor and cognitive functions of the ventral premotor cortex. *Curr Opin Neurobiol*, 12:149–154, 2002.

[61] C. Amiez and M. Petrides. Anatomical organization of the eye fields in the human and non-human primate frontal cortex. *Progress Neurobiology*, 89:220–230, 2009.

[62] M. Brass, J. Derrfuss, B.U. Forstmann and D.Y. von Cramon. The role of the inferior frontal junction area in cognitive control. *Trends Cogn Sci*, 9:314–316, 2005.

[63] K. Amunts and D.Y. von Cramon. The anatomical segregation of the frontal cortex: what does it mean for function? *Cortex*, 42:525–528, 2006.

[64] J. Derrfuss, M. Brass, D.Y. von Cramon, G. Lohmann and K. Amunts. Neural activations at the junction of the inferior frontal sulcus and the inferior precentral sulcus: interindividual variability, reliability and association with sulcul morphology. *Human Brain Mapping*, 30:299–311, 2009.

[65] J.D. Schmahmann and D.N. Pandya. *Fiber Pathways of the Brain*. New York: Oxford University Press., 2007.

[66] M. Goldau, A. Wiebel, N.S. Gorbach, C.Melzer, M. Hlawitschka, G. Scheuermann and M. Tittgemeyer. Fiber Stippling: An Illustrative Rendering for Probabilistic Diffusion Tractography. In *1st IEEE Symposium on Biological Data Visualization BioVis*, to appear.

[67] D.K. Jones. Studying connections in the living human brain with diffusion MRI. *Cortex*, 44:936–952, 2008.

[68] I. Filiminoff. Über die variabilität der Grosshirnrindenstruktur. Mitteilung 2. Regio occipitalis beim erwachsenen Menschen. *J Psychol Neurol*, 44:1–96, 1932.

[69] J.H. Maunsell and D.C. van Essen. Topographic organization of the middle temporal visual area in the macaque monkey: representational biases and the relationship to callosal connections and myeloarchitectonic boundaries. *J Comp Neurol*, 266: 535–555, 1987.

[70] H. Uylings, G. Rajkowska, E. Sanz-Arigita, K. Amunts and K. Zilles. Consequences of large interindividual variability for human brain atlases: converging macroscopical imaging and microscopical neuroanatomy. *Anat Embryol*, 2005.