AN FMRI SEGMENTATION METHOD UNDER MARKOV RANDOM FIELDS FOR BRAIN DECODING

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

BY

EMRE AKSAN

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

SEPTEMBER 2015

Approval of the thesis:

AN FMRI SEGMENTATION METHOD UNDER MARKOV RANDOM FIELDS FOR BRAIN DECODING

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

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

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

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

Examining Committee Members:

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

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

Assoc. Prof. Dr. Ahmet Oğuz Akyüz
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Murat Manguoğlu
Computer Engineering Department, METU _____

Assist. Prof. Dr. Tolga Çukur
Dept. of Electrical and Electronics Engineering, Bilkent Uni. _____

Date: _____

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

Name, Last Name: EMRE AKSAN

Signature :

ABSTRACT

AN FMRI SEGMENTATION METHOD UNDER MARKOV RANDOM FIELDS FOR BRAIN DECODING

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

September 2015, 85 pages

In this study, a specially tailored segmentation method for partitioning the fMRI data into a set of "homogenous" regions with respect to a predefined cost function is proposed. The proposed method, referred as *f-MRF*, employs univariate and multivariate fMRI data analysis techniques under Markov Random Fields to estimate the segments by resolving a mixture density. The univariate approach helps identifying activation pattern of a voxel independently from other voxels. In order to capture local interactions among the voxels, pairwise functional similarity is used across a neighborhood. By incorporating both the unary and pairwise features of the voxels into the MRF energy function, we achieve to cluster the voxels in the brain into functionally homogeneous and spatially coherent segments. In the proposed study, voxel space is modeled with a Gaussian Mixture Model (GMM) over the univariate activation patterns, while the cluster labels are modeled as discrete Markov Random Field over the pairwise interactions. For estimation of the latent cluster labels, a two-step iterative approach is followed. Accordingly, given the current estimate of the model parameters, cluster labels are computed by using a graph-cut algorithm. In turn, the cluster labels are used to estimate the model parameters by employing *maximum likelihood estimation* (MLE). The final labeling result generally consists of few large clusters involving the non-activated voxels, and isolates the activated voxels into smaller-sized clusters. By partitioning the voxel space into functionally homogeneous parcels, we expect to increase representative power of the data. Thus, we propose using the *f-MRF* segmen-

tation in brain decoding tasks where the segments are employed in voxel selection or feature extraction steps. In the experiments that are conducted on the real fMRI data of visual object recognition, *f-MRF* outperforms compared segmentation methods. Moreover, the results indicate that *f-MRF* has potential to boost the performance in brain decoding studies.

Keywords: fMRI, Segmentation, Clustering, Markov Random Fields, Functional Similarity, Univariate Analysis, MVPA

ÖZ

ZİHİNSEL AKTİVİTELERİN ÇÖZÜMLENMESİ AMACIYLA MARKOV RASGELE ALANLARI ÜZERİNDE GELİŞTİRİLEN FMRG BÖLÜTLEME YÖNTEMİ

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

Eylül 2015 , 85 sayfa

Bu çalışmada fMRG voksellerini "türdeş" bölgelere ayırmak amacıyla f-MRA isimli yeni bir bölütleme metodu önerilmiştir. f-MRA, türdeş voksel bölütlerini bir karışım yoğunluğu çözümleyerek kestirmektedir. Bu amaçla, fMRG çalışmalarında kullanılan tek değişkenli analiz ve çoklu değişkenli analiz teknikleri Markov Rasgele Alan (MRA) aracılığı ile bir arada uygulanmıştır. Tek değişkenli analiz kullanılarak, her bir vokselin diğer voksellerden bağımsız tekil aktivasyon düzeni kestirilmektedir. Vokseller arasındaki yerel etkileşimleri yakalayabilmek için voksellerin komşuları ile olan ikili fonksiyonel benzerlikleri kullanılmıştır. Voksellerin aktivasyon düzenleri tekil özniteliklere karşılık gelmekteyken, vokseller arasındaki fonksiyonel benzerlikler ise ikili öznitelikler olarak tanımlanmıştır. f-MRA, enerji fonksiyonunda her iki öznitelik uzayını bir arada kullanarak, beyindeki vokselleri fonksiyonel olarak türdeş ve uzamsal olarak bütünleşik kümelere ayırabilmektedir. Voksel uzayı, tekil aktivasyon düzenleri üzerinde Gaussian karışım modeli ile modellenmişken; küme etiketleri, ikili fonksiyonel benzerlikler üzerinde Markov Rasgele Alanı olarak tanımlanmıştır. Her bir vokselin küme etiketinin bulunması amacıyla iki adımlı bir yineleme yaklaşımı izlenmiştir. Buna göre, karışım modelinin parametreleri kullanılarak MRA enerji fonksiyonunun en düşük değeri aldığı küme etiketleri kestirilir. Bu küme etiketleri kullanılarak bir sonraki adım için modelin parametreleri tekrar hesaplanır. Yinelemeler so-

nucunda elde edilen sonuç genellikle, aktiflik göstermeyen vokselleri içeren az sayıda büyük kümeden oluşmaktadır. Aktif vokseller ise çok daha küçük boyutlardaki kümelerde toplanmıştırlar. Voksel uzayının fonksiyonel olarak türdeş parçalara bölünmesi ile datanın temsil gücünün artmasını beklemekteyiz. Bu sebeple, *f-MRA* yönteminin zihinsel aktivite çözümlenmesi probleminde kullanılmasını önermekteyiz. Buna göre elde edilen türdeş bölgeler, voksel seçimi veya öznitelik çıkarımı adımlarında kullanılabilir. Yapılan karşılaştırmalı testlerde, *f-MRA* yönteminin diğer yöntemleri sınıflandırma performansı anlamında geçtiği görülmüştür. Ayrıca test sonuçları, zihinsel aktivite çözümlenmesi problemlerinde *f-MRA* yönteminin performans arttırıcı bir potansiyele sahip olduğunu ortaya koymaktadır.

Anahtar Kelimeler: fMRG, Bölütleme, Kümeleme, Markov Rasgele Alan, Fonksiyonel Benzerlik, Tek Değişkenli Analiz, Çoklu Voksel Örüntü Analizi

To my familiy

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Prof. Dr. Fatoş Tünay Yarman Vural for her guidance and encouragement throughout this research. She has always been a source of inspiration and motivation for me. I feel very privileged to work with her.

I wish to offer my special thanks to Dr. Mete Özay and Dr. Özge Öztimur Karadağ who share precious suggestions on this work.

I owe special thanks to my colleagues Burak Velioğlu, Itır Önal, Orhan Fırat, Hazal Moğultay, Barış Nasır, Arman Afrasiyabi, Sarper Alkan and Güneş Sucu for their valuable comments and friendship.

I would like to extend my deepest gratitude to my family who always believed in me. I am very grateful to my dearest mom for her never-ending support, and my dearest sister for her always being a source of joy for me. I should also thank my cousin, Murat, who has always been supportive to me all my life.

Last but surely not the least, my deepest appreciation goes to my dearly beloved Arzu for her presence beside me and for her endless patience during this study. Once again, I feel myself so lucky to have her.

I acknowledge support of TÜBİTAK (The Scientific and Technological Research Council of Turkey) within the scope of project 112E315 during my M.Sc. research.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

fMRI	functional Magnetic Resonance Imaging
rs-fMRI	resting-state fMRI
BOLD	Blood Oxygenation Level–Dependent
HRF	Hemodynamic Response Function
Hb	oxygenated Hemoglobin
dHb	(deoxygenated) Hemoglobin
MVPA	Multi-Voxel Pattern Analysis
EEG	Electroencephalography
MEG	Magnetoencephalography
PET	Positron emission tomography
GLM	Generalized Linear Model
SPM	Statistical Parametric Map/Mapping
SNR	Signal-to-Noise Ratio
ICA	Independent Component Analysis
PCA	Principle Component Analysis
vMF	von Mises-Fisher (Distribution)
FC	Functional Connectivity
MRF	Markov Random Field
MAP	Maximum a posteriori
MCMC	Markov chain Monte Carlo
BP	Belief Propagation
ICM	Iterated conditional modes
SA	Simulated annealing
MLE	Maximum-Likelihood Estimation
knn	*k-Nearest Neighbor*
GNB	Gaussian Naive Bayes
SVM	Suppor Vector Machine
LDA	Linear Discriminant Analysis

RBM	Restricted Boltzman Machines
SCSC	Spatially Constrained Spectral Clustering
GMM	Gaussian Mixture Model
SKL	Selection by Kullback-Leibler divergence
SCV	Selection by Cross-Validation

CHAPTER 1

INTRODUCTION

1.1 Problem Definition and Rationale

Starting from the early 19th century, researchers have been in search of an answer to the question: "how does the brain work?". Understanding the human brain requires identification of its functional subdivisions, and revealing the relationship between the neural code and the underlying mental states, which have not been fully accomplished yet. It is too early to talk about how information is encoded and decoded by the neurons, yet current studies focus on identifying the activation patterns at higher abstraction levels.

With the advancements in the neuroimaging technology, researchers are able to conduct in vivo experiments on humans. As a non-invasive neuroimaging techniques, functional magnetic resonance imaging (fMRI), the positron emission tomography (PET) and electroencephalography (EEG) make major contributions to the quest of discovering the human brain. Over the last decade, fMRI has become the dominant technique mainly due to its high-resolution activation data, which can be used to localize brain functions.

Data-driven approaches in machine learning help researchers reveal the activation patterns and make connections between the neural code and the real world events. Spatially high resolution data of the fMRI brings us vast amount of fine-grained brain activity to discover. However, small number of samples compared to high dimensional feature space, i.e., large number of voxels, noisy measurements and the redundancy in the data require extensive and elaborative efforts to extract useful informa-

tion [7, 8]. In fMRI studies, mainly voxel selection or feature extraction techniques are applied in order to get rid of intrinsic problems of the data, and hence increase the signal-to-noise ratio (SNR) prior to further analysis and inference steps. Univariate analysis and region of interest (ROI) approaches are the two well-accepted techniques for voxel selection in fMRI literature [9, 10]. In univariate analysis, time series of a voxel is compared with the theoretical signal by using statistical hypothesis tests. A voxel is determined to be activated if the test score achieves a predefined threshold [11]. However, univariate analysis has a major assumption that a voxel give a response to the experimental conditions independently from other voxels. This approach ignores any kind of multivariate patterns in the data. In the ROI approach, by selecting an anatomical region completely, spatial patterns can be captured. Nevertheless, it requires expert knowledge about anatomy and physiology of the brain in order to determine the anatomical regions that are expected to be activated under the experimental conditions. This approach is apparently prone to errors. Moreover, the anatomical regions barely reflect structure of the interest which is usually in a small portion of the regions (subregion) [10].

On the other hand, cluster-based analysis of the fMRI data, reveals groups of the voxels that give similar responses. Unlike ROI approach, clustering yields data-driven parcellations, hence provides a better representation for the observations, which makes the clustering a good candidate for discovery of the functional subdivisions in the brain. By capturing the distinctive activation patterns via clusters, cluster-based analysis can serve as a tool for identification of the activation patterns, which can be followed by a set of analysis routines such as noise elimination, dimensionality reduction or feature agglomeration on the clusters. Moreover, voxel selection by means of identifying the clusters of activated voxels is able to exploit multivariate patterns in the data. In contrast to ROI approach, data-driven clusters has potential to isolate fine-grained activation patterns.

Cluster-based analysis has gained popularity in recent years [12, 13]. Well-accepted clustering techniques from the image processing and computer vision domain and their variations are applied on the fMRI data under various motivations such as locating the activated voxels or generating data-driven brain atlases. Hence, in order to avoid scattering and ensure spatial continuity, it is common to impose spatial con-

straints. For example, Vincent et al. [14] employ a spatial regularization on the cluster labels under Markov Random Fields. In the studies of Craddock et al. [15] and Heller et al. [9], the spatial constraints are quantified by the functional similarities. Likewise, Michell et al. [16] employ hierarchical clustering on the functional similarities in order to identify activated voxels for brain decoding task. In addition to the use of functional similarities, Woolrich et al. [17] apply clustering after the General Linear Model (GLM) analysis while Ryali et al. [18] directly model voxel time courses.

1.2 Contributions

In this thesis, we propose a segmentation algorithm, called *f-MRF*, that particularly considers assumptions and constraints of the fMRI data. Hence, it is expected that *f-MRF* is able to capture natural structure of the fMRI data. More specifically, the uninformative voxels that correspond to large quantities in a standard fMRI experiment are isolated from the activated voxels by simply collecting them into a few large clusters. Estimation of the cluster labels is formulated as energy minimization under Markov Random Fields. The major contributions of this study can be listed as follows:

- In the fMRI literature, there exist pioneering studies that employ Markov Random Fields in voxel clustering, where MRFs are defined as the spatial prior on cluster labels [17, 18, 19, 20]. Unlike previous studies, in this thesis, the local interactions between the neighboring voxels are incorporated into the MRF model. For this purpose, an additional energy term based on the functional connectivity concept of fMRI is introduced. This term, which is called functional energy, enables homogeneity of the regions with respect to statistical similarity of voxel time series.

- In the previous studies, clustering algorithms are applied on either univariate or multivariate features. Our proposed method, *f-MRF*, exploits two different feature sets simultaneously by defining unary potential of the MRF on univariate features, and by incorporating the functional energy term. Hence, by design, *f-MRF* is able to find both functionally homogeneous and spatially continuous

clusters.

- In order to employ the clustering result of the *f-MRF* in brain decoding tasks, we propose a heuristic for cluster selection. By using the symmetric Kullback–Leibler divergence, representative power of the clusters are estimated.
- Finally, the test results indicate that partitioning of the *f-MRF* yields better representation of the fMRI data compared to the well-accepted clustering algorithms, which is evaluated under the brain decoding task.

1.3 Organization of the Thesis

In Chapter 2, a literature survey on fMRI and the common fMRI analysis techniques are provided. Moreover, existing studies regarding fMRI clustering are overviewed. Then, as the backbone of this study, the theory of the Markov Random Fields and MAP-MRF framework are explained.

Chapter 3, introduces the proposed clustering method, *f-MRF*. First, an overview on the method is provided. Later, the energy terms and how we employ existing fMRI techniques to construct the corresponding energy terms are presented. Moreover, the algorithm for estimation of the labels and complexity analysis of the overall method are provided. Finally, we explain how *f-MRF* can be employed in brain decoding problems, and provide example use cases.

In Chapter 4, on an fMRI dataset, analysis of the *f-MRF* and comparative test results are presented. First, effect of the parameters on the energy function and final partitioning are analyzed. Later, the resulted clusters are employed in voxel selection and voxel agglomeration tasks in order to construct a feature matrix for classification of the underlying cognitive states. Clustering results and classification performance of the *f-MRF* is compared with K-Means, Normalized Cut (nCut) and Gaussian Mixture Model (GMM) methods.

In the final chapter, Chapter 5, outcomes of the study are discussed and future plans are provided.

CHAPTER 2

BACKGROUND FOR FMRI SEGMENTATION AND MARKOV RANDOM FIELDS

In this chapter, a survey about main topics of this study is presented with the purpose of providing the reader a background. Firstly, working mechanism of the functional magnetic resonance imaging (fMRI) and the data acquisition procedures are described. For the reader to gain an expression of how the fMRI measurements are employed in activity detection and connectivity analysis, popular approaches are presented. Moreover, existing studies on brain partitioning and clustering of the fMRI voxels are reviewed. Finally, as the backbone of this work, Markov Random Fields theory is overviewed and how it is used in a general clustering/segmentation problems is explained.

2.1 Functional Magnetic Resonance Imaging (fMRI)

In the brain, activity of the neurons differs based on the underlying tasks from simpler actions, such as controlling a body-part, to sophisticated cognitive activities such as reasoning. The brain consists of several specialized sub-regions, and with all those particular parts, it shows diverse patterns of activations. Hence, neuroimaging methods have a substantial role in understanding physiology of the brain and mappings between the brain regions and the cognitive functions. In this context, functional magnetic resonance imaging (fMRI) techniques, as a well accepted neuroimaging procedures, are commonly used to understand how different parts of the brain respond to external stimuli. The concept of functional MRI is built on the earlier MRI

scanning technology and discovery of the relation between neuronal activity and the blood-oxygenation level.

Magnetic resonance imaging (MRI) is a widely-used medical imaging technique to examine physiology and anatomy of the body. It works by using the phenomenon of nuclear magnetic resonance (NMR) according to which nuclei in a magnetic field absorb and re-emit electromagnetic radiation so that the atoms behave like a small magnet. Due to the fact that human body consists of large amount of water, MRI machine makes use of nuclei of the hydrogen atoms (a single proton). The hydrogen nuclei can be manipulated under the strong magnetic field of the MRI machine so that the generated signals can be captured and used to create the MR images [21, 22]. In the absence of significant magnetic field, i.e., under normal circumstances, the hydrogen nuclei, most of which are in the water or fat molecules, point randomly in different directions. However, in the MRI machine, the hydrogen nuclei tend to align with the direction of the strong magnetic field that they are exposed to, which is called as equilibrium state [23]. In addition to the continuous magnetic field, the MRI machine applies an energy in the form of radio waves to deflect nuclei of the hydrogen and perturb the equilibrium state, which is named as resonance. Under the effect of radio-frequency pulse, the atoms absorbs the additional energy. When the radio-frequency source is switched off, to return the equilibrium state, the atoms release the energy, which causes a brief and faint signal to be emitted. It is the MR signal which is used to create images. By the help of gradient coils, additional magnetic fields can be generated by small increments so that different slices resonate with different frequencies, which provides spatial information of the signal during imaging phase. An MR image is, briefly, a map of signal distribution which is collected by continuously manipulating the atoms. MRI, itself, is a sensitive tool for detecting structure of the brain.

However, mapping the structure is not the same as mapping functionality of the brain. Functional Magnetic Resonance Imaging (fMRI) is an MRI procedure, which measures signal changes in the brain that are due to changing neural activity. In other words, the MRI provides images of the anatomy of the brain while fMRI measures the functional activity within the anatomic structure of the brain. However, fMRI does not directly capture the neural activity. Instead, the fluctuations in the MR sig-

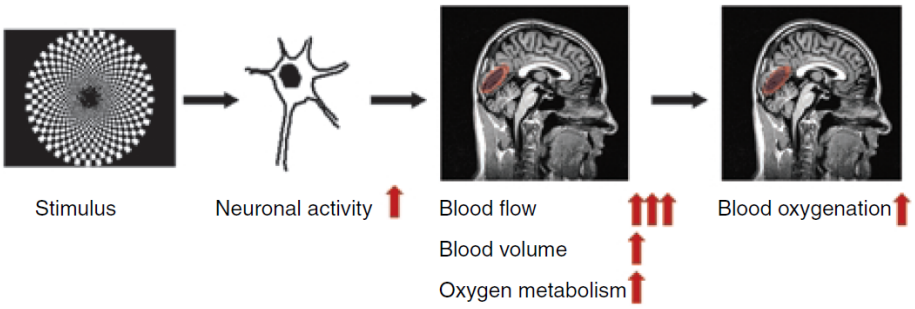


Figure 2.1: Path of the changes during a neural activity under the BOLD effect [1].

nal are due to an indirect, yet correlated effect of the changes in the blood flow which is triggered by the neural activity. The origin of this effect is that the oxygen carrier hemoglobin molecule shows different magnetic characteristics depends upon its state of oxygenation. In other words, the change in oxygen saturation of the hemoglobin molecules causes small alterations in the local MR signal, which is referred as the blood oxygenation level–dependent (BOLD) effect [1, 24].

In 1990, the discovery of blood-oxygen-level dependent MRI revolutionized the field of brain imaging in identification of the activated brain regions [25, 26]. Although the first imaging studies started in 1990s [27, 28], the theory of BOLD effect has a much longer history. Since 1890s, it has been known that the cerebral blood flow could reflect the underlying neural activities [24, 29]. Later, in 1936, Pauling and Coryell discovered how to measure the blood flow by means of the BOLD effect [30]. Specifically, oxygenated hemoglobin (Hb) has diamagnetic characteristics while deoxygenated hemoglobin (dHb) is paramagnetic. Hb in the arterials causes insignificant effects to the magnetic fields, whereas dHb in the capillary and veins tends to reduce strength of the MR signal in the neighborhood by causing distortions to the magnetic field. This phenomenon has little effect if the ratio of the Hb to dHb always remains constant. However, local neural activities trigger much more cerebral blood flow than the metabolic rate (see Fig. 2.1). As a result, proportion of Hb and dHb changes in favor of Hb. With the raised amount of oxygenated blood and Hb, the local MR signal increases.

The small BOLD signal change is not a direct measure of the neural activity. Instead, it is sensitive to physiological responses such as blood flow, blood volume and oxygen

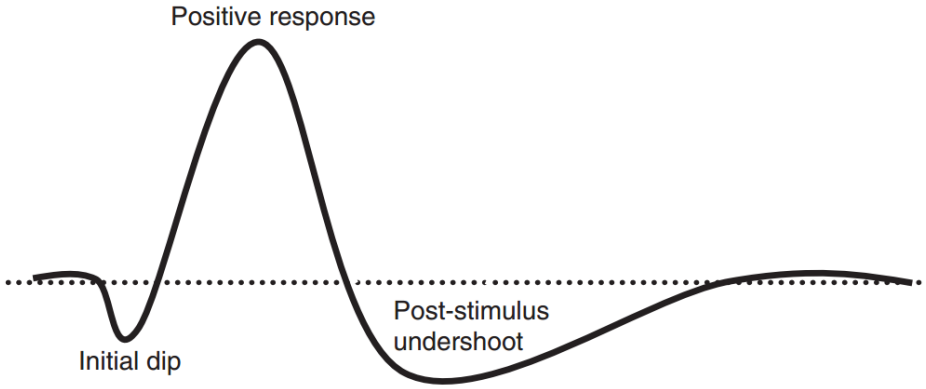


Figure 2.2: Temporal dynamics of the experimental hemodynamic response to a very short stimulus [1, 2].

rate which are collectively referred as hemodynamic response [31]. The BOLD signal in response to the neural activation is parametrized by a Hemodynamic Response Function (HRF) [1, 32]. As it is depicted in Fig. 2.2, immediately after onset of the stimulus, the activity of neurons extracts the oxygen out of local capillary, causing a momentary fall referred as *initial dip*. Shortly after, the blood flow increases and overcompensates for the initial demand. The blood flow peaks around 4-6 seconds, and the BOLD signal reaches its highest value referred as *peak* or *positive response*. When the neural activation finalizes, the hemodynamic response returns slowly to baseline, accompanied by a *post-stimulus undershoot*.

Since the early studies employing the fMRI, it has drawn great attention to become a powerful and standard neuroimaging tool to measure the brain activity [1, 33]. Compared to MEG and EEG, which are direct measures of the neural activity, fMRI and PET lacks temporal resolution due to their indirect measurement mechanisms [34]. However, fMRI distinguishes itself from other functional neuroimaging tools by its very good spatial resolution (see Fig. 2.3 for comparison). Moreover, fMRI is a non-invasive technique, hence, the subjects are not exposed to any radiation or surgical intervention. Its superior spatial resolution, also, enables mapping the functional activations to sub-regions of the brain [35].

fMRI measurements are recorded as 2-dimensional slices of the brain which are later collected into a 3-dimensional brain image representing an instance. Resolution of the fMRI does not allow to sample at the level of a cell. Instead, the 3-dimensional image

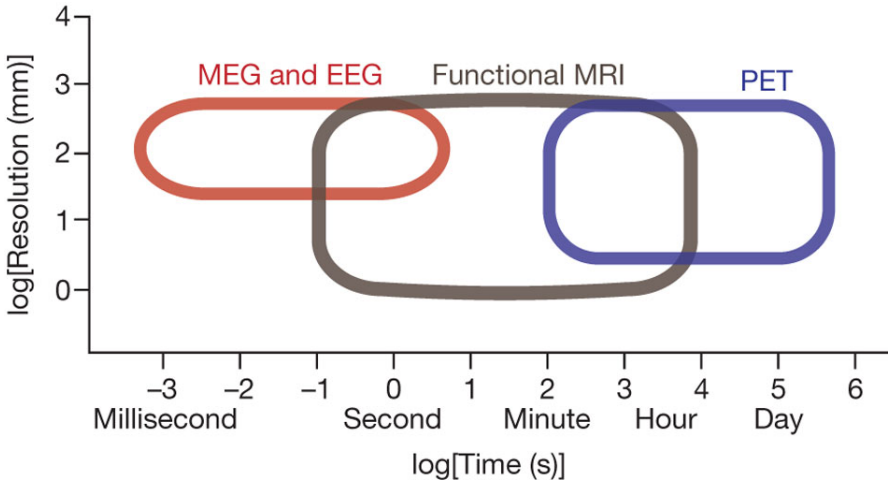


Figure 2.3: Functional neuroimaging tools with their temporal and spatial resolution comparison, from [3].

is built up in units called voxels, or referred as volumetric pixels [32]. A voxel, the smallest units of fMRI, contains thousands to millions of brain cells depending on the resolution of the MRI machine, and represents the activity of a particular coordinate in 3-dimensional space.

2.2 Univariate Analysis: Statistical Parametric Mapping

Univariate analysis of the fMRI data stands for examination of each voxel's time series independently, i.e., having an assumption that neighboring voxels are not informative about the underlying cognitive task. Statistical Parametric Mapping (SPM) [36] is the most common univariate analysis technique, which is offered to test significance of the BOLD response. SPM employs general linear model (GLM) in order to resolve a linear combination of explanatory variables, and incorporates a number of statistical models, such as t-test and F-test in order to assess voxel activation by using coefficients of the explanatory variables.

Unfortunately, source of the raw fMRI signal is not only the BOLD response. In addition to the BOLD response, some noise factors such as head motion, scanner noise and physiological effects make contribution to the measured signal. Therefore, in advance of the univariate analysis, a preprocessing step is required in order to increase the signal quality. A standard SPM framework is composed of image preprocessing,

GLM analysis and inference steps, which will be summarized below:

Image Preprocessing aims to minimize the various kinds of artifacts in voxel time series, and hence maximize contribution of the experimental conditions. The fMRI slices of a volume is acquired in slightly differing times. To make sure that all voxels of a volume are sampled at the same time, *slice-timing correction* is applied. Even small head motion of a subject may result in measurements derived from several voxel locations. In *motion correction* or *realignment* step, voxels are adjusted so that every voxel represents the same location at different instances, i.e., each volume is aligned. *Spatial smoothing* is also applied by most of the fMRI practitioners in different levels fundamentally to reduce noise, hopefully without losing fine-grained information.

General linear model (GLM) defines the observed time series of a voxel as linear combinations of several explanatory variables. It is expressed in matrix formulation by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{2.1}$$

where $\boldsymbol{y} = [y_1, \ldots, y_M]^T$ is a column vector corresponding to measured signal (consists of M time samples) of a single voxel. ϵ is the error vector following Normal distribution with zero mean and a predefined standard deviation. $\boldsymbol{X}$ is the design matrix containing explanatory variables that model the hypothesized changes in BOLD response. Each column of the $\boldsymbol{X}$ represents a presumptive signal arising from the experimental conditions. A column of the design matrix $\boldsymbol{X}$ can be generated by convolving a predefined hemodynamic response function (HRF) with the stimulus waveform which is defined with values of 1 when the experimental condition is turned on and values of 0 in other cases (see figure 2.4). In addition to the theoretical responses, vectors of artifacts such as noise, heartbeat and scanner drift can also be inserted into the design matrix [1]. $\boldsymbol{X}$ is an MxN matrix, where N is number of the explanatory variables. The parameters $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]$ indicate contribution of each explanatory variable to the measured signal $\boldsymbol{y}$.

The parameter vector $\boldsymbol{\beta}$ are estimated using Least Squares approach, which computes the estimation $\hat{\boldsymbol{\beta}}$ by minimizing the sum of squares of the residual error $\hat{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$, i.e., sum of squared differences between the actual and fitted values of the signal. $\hat{\boldsymbol{\beta}}$

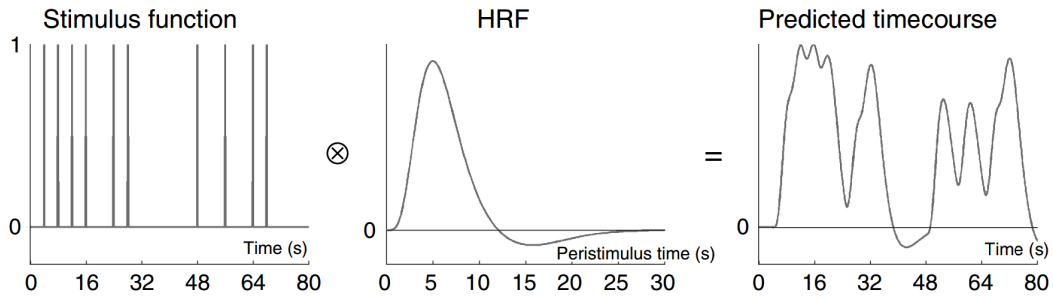


Figure 2.4: Convolution of the impulse function representing stimulus onsets (left) with theoretical BOLD response (middle) to get expected signal corresponding to a column of the design matrix $\boldsymbol{X}$ (right), from [4].

parameters can be obtained with

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{2.2}$$

Inference about the contributions of the stimulus responses to the observed signal can be made by using statistical methods to determine whether a voxel is activated or not. For this purpose, signal contribution due to the stimulus response and noise level are compared. This can be quantified in terms of *t-statistics* by testing the null hypothesis that all the estimates are occurred by noise level. Additionally, different cases can also be examined. For example, one can reveal the voxels that show significantly different activation under a specific condition by testing against a null hypothesis that all of the $\hat{\beta}$ parameters are equal.

If the null hypothesis is supported, then the *t-statistics* follows a known distribution, and hence an error probability (*p-value*) can be estimated. For example, a *p-value* of 0.05 means that probability of observing the null hypothesis is 5%. The simplest method to select the voxels showing expected behavior is applying a threshold, possibly on *p-values* of every voxel. Ideally, a *p-value* of 0.01 or 0.05 is selected to discriminate the working voxels from others.

An alternative way of identifying the activated voxels is to use pairwise correlation between observed time courses of a voxel and the model stimulus, i.e., theoretical response. The correlation coefficient ρ takes values between -1 and 1, with the value of 1 corresponding to a perfect correspondence and -1 being an inverse relation. The value of 0 indicates no correlation.

Univariate techniques offer an intuitive and practical approach to fMRI analysis. They have been the most popular analysis approach and pioneered many significant findings about the cognitive process in the brain. However, there exist potential problems with the univariate techniques deriving from the assumptions which may not be necessarily valid. First, for every voxel (and every subject), the same model of stimulus is specified. Hemodynamic response may vary between subjects and even across cortical regions of a subject due to the anatomical and physiological differences [37]. Another major assumption is that the signals from different voxels are independent. Finally, in univariate analysis a signal with low amount of noise is expected. The higher the signal-to-noise ratio, the more accurately the activated voxels are identified. Any of these assumptions is likely to be invalid, which may lead to imperfect results.

2.3 Brain Connectivity

It is well known that brain is made of a massively connected network of neurons. Assuming that every neuron or a voxel which corresponds to a group of neurons acts independently from each other would be fallacious. Instead, brain cells are in a dynamic interaction with each other, and work in harmony in order to perform a cognitive tasks [38]. Recent efforts focus on discovering the coordination between different parts of the brain, and possibly the information flow. Those studies can be gathered under a more inclusive title, called brain connectivity and categorized into three types, namely, structural, effective and functional connectivity.

Structural connectivity, also known as anatomical connectivity, refers to the physical links between neuronal elements. On the other hand, functional connectivity is defined as the statistical association or correlation among two or more anatomically distinct time-series obtained in voxels from different locations [39]. Effective connectivity, on the contrary, is based on causality, and characterizes the influence of one neural element over another.

FC is a statistical concept in which functional similarity of the units can be estimated by using different metrics, usually in the granularity of voxels or regions. One of the

important issue to take into consideration is that input data is essentially the same: voxel time series and the experimental design in particular variations. In order to measure the pairwise similarity between two units, their representative time series are compared. In the event related fMRI studies, similarity is calculated as the correlation between peak values of the BOLD responses of voxels or regions [40]. In [41], functional connectivity is defined as the similarity between $\boldsymbol{\beta}$ parameters of the univariate analysis. Friston et al. [42] measure the similarity over a set of voxel-wise components that are created by using PCA. In this study, functional similarity is calculated by using the correlation coefficient by

$$\rho_{jk} = \frac{cov(\tilde{v}_j, \tilde{v}_k)}{\sqrt{var(\tilde{v}_j)var(\tilde{v}_k)}} \tag{2.3}$$

where ρ_{jk} represents the zero-order correlation coefficient between time series of the voxels $\tilde{v}_j$ and $\tilde{v}_k$, cov and var are covariance and variance operators respectively. Unlike from [40], none of the temporal samples are discarded, and the correlation is calculated using all samples.

2.4 Segmentation of fMRI Data

Segmentation studies on the fMRI data aim partitioning the brain into a set of regions with some degree of homogeneity with respect to the information provided in the time series of voxels. It is the problem of assigning every single voxel with a label so that voxels sharing the same cluster label show similar activation patterns.

Although the fMRI studies based on segmentation focus on the very same problem, i.e. partitioning the brain into regions, they rely on different motivational aspects. In order to deal with the main challenges of the fMRI data - the curse of dimensionality and intrinsically low signal-to-noise (SNR) ratio - pipeline of the standard pattern recognition is initialized with clustering step. In brain decoding studies, for example, segments are used for voxel selection, feature agglomeration [16], or defining the feature subspaces for ensemble learning approach [Ref]. Moreover, clustering has been commonly applied on both activation (fMRI) [43, 44] and resting-state (rs-fMRI) [45, 15, 46, 10] data to define data-driven parcellations. Unlike the brain atlases, which own a predefined ontology, a data-driven parcellation models the mea-

sured signal, and provides a better fit to the data, where the resulted parcels consist of similarly activated voxels.

Clustering-based analysis, also, serves as an alternative tool for activity map generation. The standard approach is based on univariate analysis of the voxels, which is inherently limited by SNR ratio of individual voxels. In the proposed scheme, partitions consisting of similarly activated voxels are represented by the average time course of its constituent voxels, which is reported to increase SNR [9]. Accordingly, the voxels in a region are collectively regarded as activated if their representative time series is found to be active under statistical hypothesis tests. The same approach is also accepted in connectivity analysis and detection studies by averaging the time series [47] and pairwise voxel correlations [48, 49, 50]. As such spatial averages may cause loss of the fine-grained information which may be essential in the further analysis steps, clustering is also applied in a hierarchical manner [46, 51], where in the higher levels the clusters becomes larger. Therefore, representative signals at different resolutions can be calculated.

Most unsupervised exploratory methods start with the aim of partitioning the set of voxels by employing clustering on raw fMRI time courses [52, 53], or applying Independent Component Analysis (ICA) to find a decomposition of the data into a set of independent spatial [54] or temporal components [55, 56]. However, those methods do not take into consideration behavior of the voxels under the experimental conditions, or pairwise relations between voxels. Although using the raw voxel time series has potential to cover cognitive effects in the data, low SNR and increasing dimensionality of the temporal data, i.e., length of the time series may cause practical difficulties. In order to cope with the temporal dimension, in their study, Goutte et al. [57, 58], Thirion and Faugeras [59] projected the original high-dimensional time series onto a lower-dimensional space of new features which defines new measures for the similarity between voxels by exploiting the experimental conditions.

Various clustering techniques and their variants that have gained popularity in computer vision and image segmentation literature are also applied on the fMRI data. The most popular clustering techniques are mixture models [60, 61, 62], k-means [43], fuzzy clustering [52, 53, 63], hierarchical clustering [16, 46], spectral clustering

[15, 64] and a variant of edge-detection technique to detect boundaries between functionally different brain regions [65]. In order to incorporate the spatial information within voxel-based analysis and get spatially connected components, some of these approaches impose spatial constraints by adding spatial regularization terms [66], by keeping only the neighboring voxels for the model [15, 9] or by iteratively merging together only the neighboring components [16, 46].

2.5 Markov Random Fields for fMRI Analysis

Markov Random Fields have previously been used in fMRI analysis (mainly clustering and activation map generation tasks) to model spatial context information embedded in fMRI data. In the study of Descombes et al. [67], Markov Random Fields (MRF) are used for both signal restoration and activation map generation in two steps. Gaussian smoothing is a well-accepted preprocessing step to eliminate noise and increase SNR with the disadvantage of blurring and loss of fine-grained structure if the parameters are not selected wisely. In order to provide an alternative, Descombes et al. firstly reconstruct the fMRI signals "intelligently" under MRF's locality property. By using the very same architecture with the signal restoration step, and the restored signals, they group the voxels into experimental categories. Ng et al. [20] propose "Group MRF" method to exploit multi-subject information in fMRI analysis. They model activation map of all subjects with a single MRF extending the neighborhood system to all by adding edges directly between the subjects in addition to the within-subject edges. Differently from [20], Liu et al. [68], model within-subject spatial coherence by using different MRFs for each subject, which is followed by a separate group-layer with the aim of identifying group-level functional networks.

In [69, 18], Markov Random Fields are used within Bayesian Framework (see section 2.6.1 for details). The goal is estimating the unknown cluster labels given the measured fMRI data by maximizing the posterior probability. In both of the studies, the cluster labels are defined to be an MRF on the voxel grid, which represents the prior term in the Bayes rule. However, the likelihood terms have methodological differences although both of the studies employ mixture model approach. Ryali et al. [18] propose a method in order to investigate functional subdivisions of the anatomical

brain regions using resting-state fMRI. Raw time series are modeled with von Mises-Fisher (vMF) distribution without applying any dimensionality reduction. Moreover, a label cost is imposed to force the model using less number of labels by penalizing use of every distinct label [70], which is expected to uncover optimal number of clusters from data. Model parameters and cluster labels are estimated iteratively. Initializing the model with more than intended number of clusters, the model eventually converges to spatially contiguous and functionally homogeneous parcellations. Unlike [18], Woolrich et al. [69] apply a spatial mixture model on statistical parametric maps instead of the raw time courses. They propose a clustering scheme to split the brain into three distinct partitions. Given the cluster label, SPM feature set of a voxel is assumed to be following Normal distribution for non-activated and Gamma distributions for positively and negatively activated clusters.

In addition to the hard clustering methods, He et al. [71] propose a spatially regularized fuzzy clustering algorithm which employs MRF to incorporate spatial constraints. Unlike from non-spatial fuzzy clustering, membership values are weighted by the energy function of the MRF, which is referred as spatial membership.

2.6 Foundations of Markov Random Fields

In image processing and computer vision literature, Markov Random Fields have been widely used for a variety of problems such as segmentation, de-noising, stereo-matching and texture synthesis. A Markov Random Field is an undirected graph of random variables, over which inference is carried out by minimization of a predefined energy function. MRFs enable researchers to formulate their problem as energy minimization on a rectangular grid of pixels for many vision tasks, and on three dimensional grid of voxels for fMRI studies. Moreover, the energy terms of the MRF make it possible to incorporate both low-level and high-level context information about the data.

Formally, an MRF is defined by the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1 \dots N\}$ is set of nodes each of which is associated with a random variable x_i for $i = \{1 \dots N\}$, and $\mathcal{E} = \{e_{i,j}\}$ is set of links each of which connects a pair of nodes $i, j \in \mathcal{V}$. Let $\mathcal{N}_i \subset \mathcal{V}$

be the neighborhood of node i, and $j \in \mathcal{N}_i$ if and only if the edge $e_{i,j} \in \mathcal{E}$.

A joint probability distribution $P(x_1 \dots x_N)$ over the random variables $X = \{x_i\}_{i=1}^N$ can be defined as the probability of a particular configuration. The random field $X = \{x_i\}_{i=1}^N$ is an MRF on $\mathcal{G}$ with respect to a neighborhood system if and only if [72]

$$P(X) > 0, \forall X \tag{2.4}$$

$$P(x_i|\{x_j\}_{j\in\mathcal{V}\setminus i} = P(x_i|\{x_j\}_{j\in\mathcal{N}_i}) \tag{2.5}$$

where Eq. (2.5) implies that a node in the graph is conditionally independent of rest of the graph given its immediate neighbors, which is also referred as Markov blanket [5]. Note that $\mathcal{V} \setminus i$ is the set difference.

By the Hammersley-Clifford theorem, a joint probability distribution over an MRF that satisfies Eqs. (2.4) and (2.5) can take the form of a product of potential functions on maximal cliques of $\mathcal{G}$ [72, 73, 5]:

$$P(x_1 \dots x_N) = \frac{1}{Z} \prod_{c\in\mathcal{C}} \phi_c(S_c) \tag{2.6}$$

where $\mathcal{C}$ is set of maximal cliques. A maximal clique is a fully connected subgraph of $\mathcal{G}$ where it is not possible to insert any other nodes without breaking the full connectivity. Each potential function $\phi_c, c \in \mathcal{C}$ returns a positive value, and it is defined on a subset of random variables $S_c \subset \mathcal{V}$. Obviously, the probability increases when the clique potentials take higher values. In other words, each of these functions adjusts the tendency for the variables to adopt a certain configuration. Z is the partition function normalizing the product of potential functions so that the joint probability sums to one. Mathematically speaking, the normalization function is defined as;

$$Z = \sum_{x_1 \dots x_N} \prod_{c\in\mathcal{C}} \phi_c(S_c). \tag{2.7}$$

Calculating the normalization term Z by using Eq. (2.7) is intractable and one of the main limitations of MRF. It has the complexity of exponential in the size of the model. Having a model with N nodes and L discrete states, the calculation requires summing over L^N states. Since any parameter controlling the potential functions is

involved in the partition function Z, it is essential to calculate Z for parameter learning. Generally, approximation techniques are followed to overcome this problem.

For Eq. (2.4) to be hold, the potentials ϕ_c should be strictly positive. Therefore, it is convenient to express the potential functions as exponentials,

$$\phi_c(S_c) = \exp\left[-E_c(S_c)\right] \tag{2.8}$$

so that the joint distribution in Eq. (2.6) can equivalently be written as Gibbs distribution,

$$P(x_1 \dots x_N) = \frac{1}{Z} \exp\left[-E_s(x_1 \dots x_N)\right], \tag{2.9}$$

$$E_s(x_1 \dots x_N) = \sum_{c \in \mathcal{C}} E_c(S_c). \tag{2.10}$$

The term $E_s(x_1 \dots x_N)$ is referred as the energy, and finding the best configuration of random variables X, i.e., increasing the probability, can be referred as energy minimization [5, 73].

2.6.1 MAP-MRF Framework

Markov Random Fields enable us modeling the priori probability of context-dependent patterns. In an MRF model, a node favors patterns of its own class by associating them with higher probability, i.e., lower energy values than other patterns. It is common to use MRFs in conjunction with statistical decision or estimation models so as to incorporate context information inherently involved in the data. *Maximum a posteriori* (MAP) has been one of the most popular optimality criteria in modeling with MRFs since 1984 - pioneer study of Geman and Geman [74]. In the MAP-MRF framework, the objective is the joint posterior probability of the random variables that are involved in the MRF. The joint priori distribution of the variables and likelihood of the observed data controls the posterior by means of the Bayes formula.

Segmentation problem can be formalized as an unsupervised learning or clustering where the task is to assign a set of labels to a set of sites of homogeneous clusters.

Mathematically speaking, let V be the set of sites to be labeled so that

$$V = \{v_1 \dots v_N\}, \tag{2.11}$$

where each v_i is also the observed data sampled from the corresponding site. And, X is set of random variables

$$\begin{aligned} X &= \{x_1 \dots x_N\} \\ x_i \in \mathcal{L} &= \{\ell_1 \dots \ell_L\} \end{aligned} \tag{2.12}$$

where each x_i assigns a label for the site v_i among L possible discrete values from the label set $\mathcal{L}$. The labeling process is referred as a *configuration* in the terminology of the random fields. And, it can be regarded as a function with domain V and range $\mathcal{L}$. In other words, in the context of clustering task, a configuration is equivalent to the assignment of the possibly best labels that fulfills an objective function.

The process of estimating the cluster labels can be formulated by the Bayes Theorem so that *maximum a posteriori* estimation (MAP) yields a configuration of the latent labels:

$$\hat{x}_1 \dots \hat{x}_N = \underset{x_1 \dots x_N}{\operatorname{argmax}} \left[P(x_1 \dots x_N | v_1 \dots v_N) \right]. \tag{2.13}$$

And, the formulation of the problem by using the Bayes rule as follows:

$$P(x_1 \dots x_N | v_1 \dots v_N) = \frac{P(v_1 \dots v_N | x_1 \dots x_N) P(x_1 \dots x_N)}{P(v_1 \dots v_N)}, \tag{2.14}$$

where $P(X)$ is the prior probabilities of the configuration X, and $P(V|X)$ is likelihood of the data V. By making an assumption that the conditional probability $P(V|X)$ factorizes into product of individual terms [5], Eq. (2.14) can be rewritten as

$$P(x_1 \dots x_N | v_1 \dots v_N) = \frac{\prod_{i=1}^{N} P(v_i | x_i) P(x_1 \dots x_N)}{P(v_1 \dots v_N)}. \tag{2.15}$$

By reorganizing the Eqs. (2.13) and (2.15), the MAP solution can be expressed as

$$\begin{aligned}
\hat{x}_1 \dots \hat{x}_N &= \underset{x_1 \dots x_N}{\operatorname{argmax}} \left[P(x_1 \dots x_N | v_1 \dots v_N) \right] \\
&= \underset{x_1 \dots x_N}{\operatorname{argmax}} \left[\prod_{i=1}^{N} P(v_i|x_i) P(x_1 \dots x_N) \right] \\
&= \underset{x_1 \dots x_N}{\operatorname{argmax}} \left[\sum_{i=1}^{N} log[P(v_i|x_i)] + log[P(x_1 \dots x_N)] \right] \\
&= \underset{x_1 \dots x_N}{\operatorname{argmin}} \left[\sum_{i=1}^{N} -log[P(v_i|x_i)] + \sum_{c \in \mathcal{C}} E_c(S_c) \right] \\
&= \underset{x_1 \dots x_N}{\operatorname{argmin}} \left[E_d(x_1 \dots x_N) + E_s(x_1 \dots x_N) \right]
\end{aligned} \tag{2.16}$$

where Eq. (2.14) is transformed to the log domain. Note that denominator of the Bayes formula, i.e., $P(v_1 \dots v_N)$, and the normalization term Z are discarded since they do not make any influence on the MAP inference. $E_s(x_1 \dots x_N)$ is same as the Eq. (2.10) and corresponds to the clique potentials from the MRF prior, and it is referred as smoothing energy. It penalizes different labeling of adjacent sites. $E_d(x_1 \dots x_N)$ is the unary energy term that is the negative log likelihood of sites given their cluster labels. This is a cost for observing a site v_i with state, i.e., label, x_i.

A fully four-connected subgraph contains cliques of size 2, 3 and 4 (see Fig. 2.5 right). In the formulation above, the potential functions are assumed to be formulated in terms of the maximal cliques only, which can be in the form of either pairwise or higher order cliques. If the maximal cliques only connect pairs of nodes, those models are called as pairwise Markov Random Fields. If the number of the constituent nodes of a clique is more than two, then we talk about high-order Markov Random Fields. Order of an MRF model has a key role in determining shape of the smoothing energy term $E_s(x_1 \dots x_N)$ and the optimization approach.

2.6.2 Pairwise Markov Random Fields

In order to foster simplicity and make the inference efficient, the neighborhood is set between only pairs of nodes, hence assuming a conditional independence of all other nodes, given direct neighbors. On a two-dimensional grid, it corresponds to

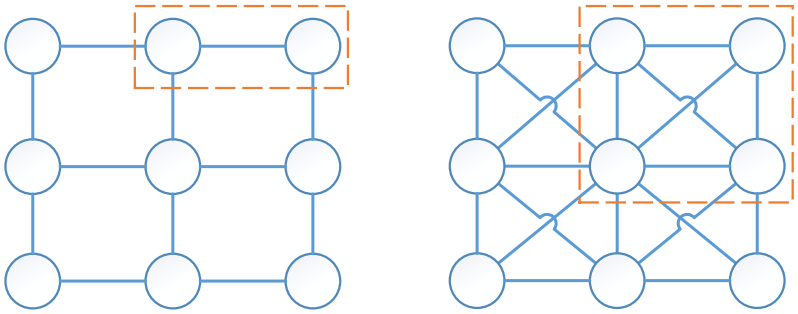


Figure 2.5: MRF neighborhood structures in two dimensional grid where the dashed lines denote an example of maximal cliques. (Left) Pairwise MRF (Right) Example for an higher-order MRF.

4-connected neighborhood (see Fig. 2.5 left), and on a three-dimensional grid the model consists of 6-connected neighborhood - face touching. In pairwise MRFs, it is common to accept homogeneous model. In other words, the potential functions between each pair of nodes are assumed to be same for all. Although this neighborhood structure seems to be very simple, by means of transitivity, it indirectly links all nodes.

The smoothing energy of a pairwise MRF model has the following form:

$$E_s(x_1 \dots x_N) = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \beta_s U_s(x_i, x_j), \tag{2.17}$$

where β_s is the hyper-parameter adjusting smoothing penalty. Choosing a potential function f is crucial to accurately model neighboring relations. The standard Potts model is a well-studied and commonly applied smoothing function [73, 5] as follows

$$U_s(x_i, x_j) = \begin{cases} 1, & \text{if } x_i \neq x_j \\ 0, & \text{otherwise.} \end{cases} \tag{2.18}$$

Moreover, in variants of the Potts model, modifications are made so that the function U_s takes values based on the observations. While this can improve the performance, some might argue that the pairwise MRFs are conceptually limited due to their minimal cliques, which restrict expressive power of the model [75, 76].

2.6.3 High-Order Markov Random Fields

Higher-order MRFs are a generalization of pairwise MRFs, where the cliques encompass more than two nodes (see Fig. 2.5 right). The smoothing term is defined over the maximal cliques so that

$$E_s(x_1 \dots x_N) = \sum_{c \in \mathcal{C}} f_c(S_c), \tag{2.19}$$

where f_c is the potential function of the clique S_c. As it is in the pairwise MRFs, such models assume the spatial homogeneity. In other words, the potential functions are the same for all cliques because choosing a suitable function is even more difficult due to the larger size of the cliques.

Introducing a higher-order neighborhood system increases expressive power of the MRFs and enables to model the long range dependencies. However, with a few exceptions only, pairwise MRFs are preferred. One of key challenges with respect to the high-order MRFs is the efficient inference. The computational cost of the model makes the optimization algorithms that even yield suboptimal solutions impractical. There exist inference techniques for a few higher-order MRFs that have special structure [77, 78, 79]. Converting higher-order MRFs into pairwise MRFs by introducing additional variables is another approach [80]. However, it is only applicable on the small size of MRFs since number of the auxiliary variables grows exponentially with the size of cliques [76].

2.6.4 Inference: Energy Minimization

Inference in the graphical models is the task of discovering the hidden information for the random variables X given the observed variables V. In the context of clustering problems, inference is equivalent to estimation of the cluster labels. This can be achieved by finding configuration of the latent variables X that maximizes the posterior probability - *maximum a posteriori* estimation. As it is expressed in section 2.6.1, minimization of the energy gives the same result.

Having a solution in closed form is the most desirable. However, the complexity caused by interactions between the nodes reveals model specific constraints, and

makes it very difficult to express the solution in closed form. Therefore, exact inference in graphical models is generally very hard and computationally intractable. In practice, approximate inference techniques are employed by means of iterative search. Four popular approaches of approximate inference are Iterated Conditional Modes (ICM), Simulated Annealing (SA), Belief Propagation (BP) and Graph cuts, which are described below:

Iterated Conditional Modes (ICM), proposed by Besag [81], is one of the most popular optimization algorithms. It is a deterministic algorithm based on optimizing the local energy, iteratively. It follows a greedy strategy and maximizes the probability of each node conditioned on the rest (only the neighbors of the node itself due to the Markovianity property 2.5) and the observed data. After updating every node separately, a cycle of the ICM concludes. The iterations continues until a predefined convergence criterion, which guarantees the convergence [81]. It has been reported that ICM finds the global minimum if the search space is convex, which is an optimistic assumption for MRF inference. MRF energy functions are generally do not end up with convex search spaces. Therefore, result of the ICM heavily depends on the initial labeling. ICM may return a local solution that is far from the optimal [73].

Simulated Annealing (SA) deals with the local minimum problem by accepting different configurations, which helps the method to avoid getting stuck in a local minimum. A time-varying parameter T, referred as temperature in the Gibbs distribution, controls the probability of accepting worse states. The method is initialized with very high temperature and a random configuration, i.e. labeling. Simulated annealing is a variant of sampling based algorithms, particularly Markov chain Monte Carlo (MCMC) sampling. At every possible values of T, probability of jumping another less optimal state is determined by using the Gibbs sampling. As the temperature decreases, the solution space becomes more explored. Therefore, probability of choosing a worse configuration also decreases, and the algorithm yields optimal -or near optimal- solution [73, 74]. Note that in the limiting case where T is zero, SA behaves like a greedy algorithm such as ICM. Theoretically, if SA is provided with enough time and proper parameters, it finds the global optimum. However, size of the search space and computationally expensive sampling steps make the SA algorithm prohibitive for real world problems.

Belief Propagation (BP) formulates the inference problem for pairwise MRFs as maximizing the marginal posterior probabilities over individual variables. The BP algorithm works by iteratively propagating the local messages across nodes of the graph. At every iteration, each node of the MRF graph sends a message for each label to its neighboring nodes and accepts incoming messages from the neighbors until there is no change in the messages. A message from node p to node q about the label ℓ indicates how likely node p and its neighbors other than q support that node q should be assigned label ℓ. After the message passing system reaches convergence, the so-called belief of each node is computed by considering the observed data of the node and incoming messages from its neighbors. The belief is an approximation of the marginal probability of the node [73]. Finally, the label that maximizes the belief, i.e., the marginal probability, or equivalently minimizes the negative log probability is assigned to the node. On the tree-structured graphs, the BP algorithm is able to make exact inference while it returns suboptimal solutions for general graphs.

Graph cuts is a class of algorithms that can be efficiently employed to solve MAP inference tasks by translating the energy minimization problem into a maximum-flow (or max-flow) / min-cut problem. It has been reported that exact inference, i.e. global solution, is tractable for binary problems with pairwise MRFs [82]. Later, this has been extended to solve multi-label MRF problems that has convex smoothness term. For a more general energy function, i.e., a non-convex model, it has been shown that good approximate solutions near the global optimum, and even exact inference in some cases can be achieved by using graph cuts [5, 83, 84].

In order to transform the energy minimization task into a max-flow/min-cut problem, two auxiliary vertices, namely s (source) and t (sink), are introduced (see Fig. 2.6). Let the resulted graph be $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, where $\tilde{\mathcal{V}} = \mathcal{V} \cup \{s, t\}$. Now, the task is finding maximum amount of flow from source to sink nodes. This is equivalent to finding a *cut* which is defined as a minimal subset of edges that separates the source from the sink. In the context of binary labeling, the vertices that are connected to the source are labeled as 1 and the vertices that are attached to the sink are given label 0. In the multi-label case, however, label of each node is defined by the edge that cuts its chain (see Fig. 2.7).

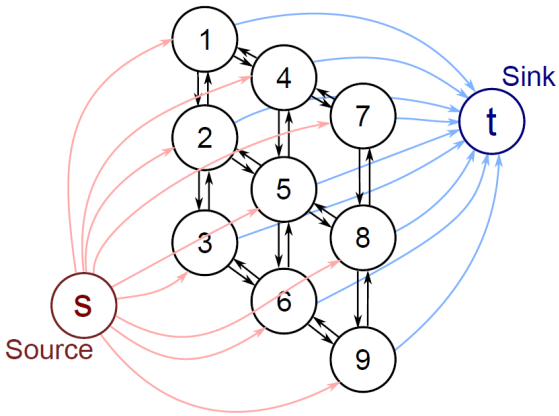


Figure 2.6: Reorganization of a 3x3 binary MRF as a max-flow problem. In addition to the source and sink nodes, every site is represented in the final graph (black nodes) [5].

If the pairwise potentials are non-convex (like Potts model, Eq. 2.18), then exact inference can not be achieved, and it is NP-hard. Boykov et al. [83, 85] proposed good approximate methods, namely α-*expansion* and α-β *swap* that can achieve solutions within a known factor of the global minimum. The α-*expansion* algorithm can only be employed if the pairwise cost is a metric. And, the pairwise cost satisfies at least semi-metric conditions, then α-β *swap* algorithm can be used as an alternative. Both of them works by reducing the multi-label problem into a series of binary problems that can be solved exactly. In this study, we minimize the MRF energy by using α-*expansion* algorithm. Therefore, details of the α-β *swap* algorithm is not provided.

At each iteration, a label α is selected and the nodes with labels other than α are given the label $\bar{\alpha}$ (non-alpha). A max-flow graph is generated dynamically with respect to the current configuration at each iteration. The source and sink nodes are represented with α and $\bar{\alpha}$ respectively. In other words, α-*expansion* algorithm considers either keeping $\bar{\alpha}$ label or switching it to α for each node by finding the minimum cut. The algorithm repeatedly cycles through all possible labels as α until no further improvement is possible.

One important advantage of the α-*expansion* algorithm is that it enables the simultaneous changes on set of nodes while ICM and SA switch one node label at a time. The larger number of changes at every iteration keeps the search from getting stuck in local minima, and hence guarantees an energy decrease.

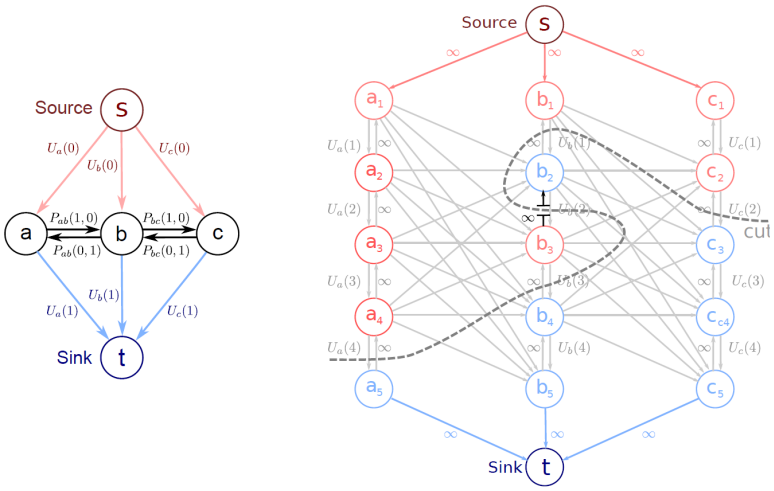


Figure 2.7: Reorganization as max-flow problem. (Left) Binary pairwise MRF. Edge costs are defined by the unary (U_i) and pairwise (P_{ij}) energy terms. (Right) Multi-label pairwise MRF (4 labels). There are $|\mathcal{L}| + 1$ nodes for sites (one for each label) and fully connected neighboring edges (one for each pairwise label assignment) in the graph [5].

2.7 Summary

In this chapter, firstly, background information about the fMRI data acquisition is presented. Secondly, widely accepted analysis techniques of fMRI data that are used for inference about the voxel activation, and the connectivity types in the brain are overviewed. Later, from a large family of clustering techniques, exemplar studies on the fMRI data are reviewed. The theory of Markov Random Fields and MAP-MRF framework are explained. Finally, energy minimization techniques for MRF models are compared qualitatively, and details of the α-*expansion* algorithm is provided.

CHAPTER 3

F-MRF: A BRAIN SEGMENTATION METHOD BASED ON MARKOV RANDOM FIELDS

In the previous chapter, a background about the fMRI data analysis techniques and a clustering approach that employs MRFs are provided. In this chapter, we introduce a segmentation method for the fMRI data by using a clustering algorithm based on the Markov Random Fields. Our proposed method, *f-MRF* employs fMRI data analysis techniques under the *maximum a posteriori*-Markov Random Fields (MAP-MRF) framework.

First, an overview of the proposed method and motivation of our study is presented. Second, *f-MRF* method is decomposed into individual energy terms and explained in details. Later, we present the algorithm that is used for estimation of the cluster labels and computational complexity analysis of the *f-MRF*. Finally, the approaches that are employed in performance evaluation are provided.

3.1 Overview of f-MRF Segmentation

f-MRF is a segmentation method specially tailored for fMRI by pursuing the fMRI data analysis techniques. *f-MRF* principally aims to partition the brain into a set of disjoint regions that are functionally homogeneous and spatially coherent. Accordingly, the voxels having similar time series are collected into the same region. Thus, each segment in the brain is expected to represent a distinctive activation pattern. Considering distributed nature of the brain, however, there may exist some remote regions showing similar activation patterns. By using a clustering based approach,

f-MRF finds the clusters that consist of functionally similar regions.

Given the fMRI data, recorded under a predefined cognitive stimulus, let $\tilde{V} = \{\tilde{\boldsymbol{v}}_{\boldsymbol{i}}\}_{i=1}^{N}$ represent the set of voxel time series $\tilde{v}_i = \{\tilde{v}_i(t)\}$ for $1 < t < T$, where N is the number of voxels and T is the number of fMRI observations. A partition at the output of a segmentation algorithm creates a set of homogeneous regions $\mathcal{R} = \{\mathcal{R}_i\}_{i=1}^{S}$ where S is the number of regions. The partitioning $\mathcal{R}$ satisfies the following conditions:

1. $\mathcal{R}_i \cap \mathcal{R}_j = \varnothing$,
2. $\bigcup_{i=1}^{S} \mathcal{R}_i = \tilde{V}$,
3. For a homogeneity predicate $\mathcal{P}$,
   $\mathcal{P}(\tilde{v}_m, \tilde{v}_n) =$ TRUE , if $\tilde{v}_m, \tilde{v}_n \in \mathcal{R}_i$
   $\mathcal{P}(\tilde{v}_m, \tilde{v}_n) =$ FALSE , if $\tilde{v}_m \in \mathcal{R}_i$ and $\tilde{v}_n \in \mathcal{R}_j$ $\forall m \neq n, i \neq j$.

The homogeneity predicate in this study is defined by the minimum energy function of the MRF model which is introduced in section 3.3.

3.2 Motivation

In the task-based fMRI experiments where the individuals are exposed to external stimuli or asked to perform a task, not all of the voxels are correlated with the underlying experimental conditions. Rather, majority of the voxels are noisy, inactive or redundant for the underlying task [8, 86, 7, 87]. Hence, these voxels are non-informative for the further analysis steps. Moreover, small number of samples compared to high dimensionality requires extensive and elaborative efforts to extract useful information.

In the fMRI studies, it is common to apply voxels selection and feature extraction techniques in order to alleviate the dimensionality problem and increase signal-to-noise ratio (SNR). Having a mapping of voxels that reflects activation patterns in the data would enhance precision of the further dimensionality reduction steps. In this study, we propose a new segmentation technique, called *f-MRF*, in order to group the

non-informative voxels under the same cluster and reveal groups of the informative voxels. For this purpose, we design the *f-MRF* in a way that it gathers segments of the mostly inactive and negligible voxels in a few large clusters, and it isolates the informative voxels into relatively small and functionally homogeneous clusters.

The prevalent techniques in dimensionality reduction mainly rely on (1) univariate analysis, (2) region of interest (ROI), and (3) data-driven parcellations [10]. Recall from chapter 2.2, in the *univariate analysis* approach, time series of each individual voxel is compared with the theoretical blood oxygenation level–dependent (BOLD) response under a statistical hypothesis test [11, 88]. Although the univariate analysis is intuitive and practical, underlying assumptions driving the model may not be necessarily valid. Firstly, co-activation patterns of multi voxels are simply ignored. In other words, although two different voxels are not informative individually, the linear or non-linear relation between them might carry information [8]. Secondly, in fMRI studies, for every voxel, the same or minimally varied hemodynamic response functions (HRF) are employed although it is reported that BOLD response may vary due to the anatomical and physiological differences [37, 89]. Finally, because univariate approaches require higher SNR, it is common to apply spatial smoothing in order to increase the sensitivity [90]. However, this idea is debated in fMRI literature [88, 13] since smoothing cause distortions on spatial patterns that might be informative.

In the *region of interest (ROI)* approach, on the other hand, regions that are expected to be activated under the experimental conditions are selected by using a predefined anatomical atlas [10, 9]. Although this approach provides functional and biological homogeneity, recent studies reveal that the anatomical regions defined by atlases can be coarse [9]. More specifically, depending the underlying task, there might be subregions showing different behavior.

Data-driven parcellations provide better representations for the voxel space by collecting similarly activated voxels into the same parcels. For this purpose, well-accepted clustering methods in the vision domain are employed on fMRI data as well as the algorithms that are dedicatedly crafted for fMRI. Each method, intrinsically, has assumptions and constraints such as underlying distribution or structure of the data. For example, the K-Means algorithm which minimizes the within cluster vari-

ance, and Spectral Clustering method that minimizes between-cluster similarity are strongly biased towards a result with uniformly sized clusters due to their objective functions [15, 46]. Mixture models is another commonly used technique that has the potential to capture various-sized clusters [60, 61, 62]. It makes an assumption that the voxels are sampled from a mixture of distributions, and estimates the clustering by assigning the voxels to the most likely mixture components. Heller et al. [9] propose a clustering method based on functional connectivity of the voxels, which is used in estimation of the representative signals. Michel et al. [16] employ clustering in order to reveal a subset of informative voxels for classification task. First, they apply Ward hierarchical clustering to construct a clustering tree. Later, the tree is pruned with respect to a prediction score in order to discard non-informative clusters.

The existing studies in the literature are based on either unary features that simply involve individual voxel properties or multivariate features which reveal the relationship of a voxel with others. In this study, we employ these two different feature sets simultaneously under MAP-MRF framework. By using both univariate properties and local functional similarities of the voxels, we aim to acquire both functionally homogeneous and spatially continuous clusters. From a different point of view, the *f-MRF* algorithm can be considered as a hybrid between the clustering approaches that employ unary and pairwise features separately. While unary feature set determines characteristics of a cluster, i.e. consist of informative or non-informative voxels, the pairwise feature set enforces both spatial coherence and functional homogeneity.

3.3 Proposed Method: f-MRF

Let $V = \{\boldsymbol{v_i}\}_{i=1}^{N}$ be the set of random variables where each $\boldsymbol{v_i}$ corresponds to a $d-$dimensional feature vector of the i^{th} voxel, and $\boldsymbol{v_i}$ is generated from the voxel's time-series $\tilde{\boldsymbol{v}}_{\boldsymbol{i}}$. It is referred as *unary features*. Let $X = \{x_i\}_{i=1}^{N}$ be the set of random variables where each x_i makes a label assignment for the voxel $\boldsymbol{v_i}$ from the set of cluster labels such that $x_i \in \mathcal{L} = \{\ell_1 \dots \ell_L\}$ assuming that there are L clusters.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the graph that defines the Markov Random Field. Nodes $\mathcal{V} = \{1 \dots N\}$ of the graph $\mathcal{G}$ is associated with the above-mentioned random variables x_i

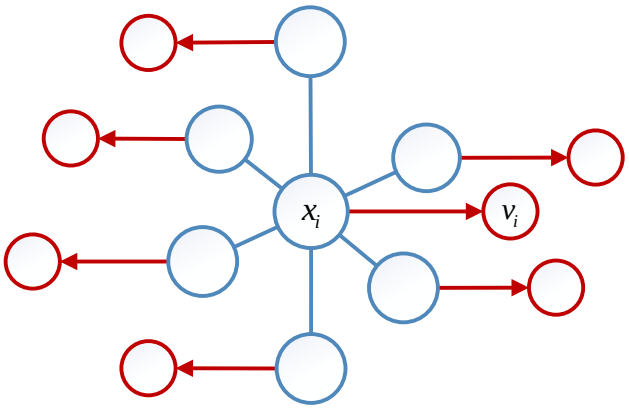


Figure 3.1: MRF model on a 3-dimensional lattice. (Blue nodes and blue undirected edges) The latent label node x_i of the i^{th} voxel and its neighborhood $\mathcal{N}_i$. (Red nodes and red directed edges) The observed data $\boldsymbol{v}_i$ is conditionally dependent on the associated cluster label x_i.

(see Fig. 3.1). Let $\mathcal{N}_i \subset \mathcal{V}$ be the neighborhood of node i, and $j \in \mathcal{N}_i$ if and only if there exist an edge $e_{i,j} \in \mathcal{E}$ between them.

Neighbors $\mathcal{N}_i$ of the node i are selected with respect to the Euclidean distance between positions of the voxel i and the voxels around. The distance threshold is chosen to include 6 spatially nearest neighbors of the voxel $\boldsymbol{v_i}$ in the $3-$dimensional voxel grid, i.e., face touching. Note that voxels that are located on the surface of the cortex may have less than 6 neighbors.

The latent cluster labels are estimated by using *maximum a posteriori* estimation (MAP) such that,

$$\hat{x}_1 \dots \hat{x}_N = \underset{x_1 \dots x_N}{\operatorname{argmax}} \left[P(x_1 \dots x_N | \boldsymbol{v_1} \dots \boldsymbol{v_N}) \right]. \tag{3.1}$$

Recall from chapter 2.6.1, under the MAP-MRF framework, the estimation can be

expressed as an energy minimization problem:

$$\begin{aligned}
\hat{x}_1 \dots \hat{x}_N &= \underset{x_1 \dots x_N}{\operatorname{argmax}} \left[P(x_1 \dots x_N | \boldsymbol{v_1} \dots \boldsymbol{v_N} \right] \\
&= \underset{x_1 \dots x_N}{\operatorname{argmax}} \left[\prod_{i=1}^{N} P(\boldsymbol{v_i} | x_i) P(x_1 \dots x_N) \right] \\
&= \underset{x_1 \dots x_N}{\operatorname{argmax}} \left[\sum_{i=1}^{N} log[P(\boldsymbol{v_i} | x_i)] + log[P(x_1 \dots x_N)] \right] \\
&= \underset{x_1 \dots x_N}{\operatorname{argmin}} \left[E_d(x_1 \dots x_N) + E_s(x_1 \dots x_N) \right],
\end{aligned} \tag{3.2}$$

where the solution hinges upon the unary and smoothing energy terms, E_d and E_s, respectively. In this study, we formulate the clustering problem as a pairwise Markov Random Field, where cliques of the graph $\mathcal{G}$ consist of pairs of nodes. The unary energy is defined as negative log-likelihood of the voxels. In other words, the less the distance between a voxel and its cluster, the lower energy it emits. The smoothing term, on the other hand, can be regarded as a constraint that forces the model to assign the same cluster labels to the neighboring voxels.

The main contribution of this study is that we incorporate an additional energy term in order to ensure both spatial regularization and functional homogeneity. For this purpose, we benefit from pairwise voxel similarity which is expected to reveal the multivariate activation patterns in the fMRI data. Accordingly, the total energy E can be decomposed into the following terms:

$$E(x_1 \dots x_N) = E_d(x_1 \dots x_N) + E_p(x_1 \dots x_N) + E_f(x_1 \dots x_N), \tag{3.3}$$

where E_d and E_p are the unary and Potts energy terms, respectively. E_f is the new energy term that employs functional similarity. Hence, it is referred as *functional energy*. In this setting, both E_p and E_f apply a penalty if any pair of voxels are assigned to different clusters.

3.4 Unary Energy Term

The voxel space is summarized by the mixture of L models where each model represents a cluster. Assume that d-dimensional feature vector measured at each voxel is

sampled from a density $P(\boldsymbol{v})$, where $P(\boldsymbol{v})$ is a finite mixture model with L components. A voxel $\boldsymbol{v_i}$ follows the distribution:

$$P(\boldsymbol{v_i}; \boldsymbol{\Theta}) = \sum_{\ell=1}^{L} \lambda_\ell P_\ell(\boldsymbol{v_i}|x_i = \ell; \boldsymbol{\theta_\ell}), \tag{3.4}$$

where $P_\ell(\boldsymbol{v_i}|x_i = \ell; \boldsymbol{\theta_\ell})$ is the mixture component for $1 \leq \ell \leq L$, and each component is defined as a d-variate distribution with parameters $\boldsymbol{\theta_\ell}$. λ_ℓ is the mixture weights for $1 \leq \ell \leq L$ that correspond to the probability of a voxel $\boldsymbol{v_i}$ to be sampled from the component ℓ. Note that,

$$\sum_{\ell=1}^{L} \lambda_\ell = 1. \tag{3.5}$$

$\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_1 \dots \boldsymbol{\Theta}_L\}$ is the set of parameters for our mixture model with L components:

$$\boldsymbol{\Theta_\ell} = \{\lambda_\ell, \boldsymbol{\theta_\ell}\}, \tag{3.6}$$

where $\boldsymbol{\theta_\ell}$ is parameters of the underlying mixture distribution. Using the above formalism, a partition $C = \{C_1 \dots C_L\}$ can be defined in terms of the mixture components, where each cluster C_ℓ is represented by a component of the mixture and its corresponding parameters $\boldsymbol{\theta_\ell}$.

Recall from Eq. (3.2) that the unary energy term is defined as negative log-likelihood of a voxel $\boldsymbol{v_i}$, given its cluster label,

$$E_d(x_1 \dots x_N) = \beta_d \sum_{i=1}^{N} -log[P(\boldsymbol{v_i}|x_i)], \tag{3.7}$$

where β_d is the weight parameter adjusting the influence of the data cost. The log-likelihood is defined as,

$$P(\boldsymbol{v_i}|x_i = \ell) = \frac{\lambda_\ell P_\ell(\boldsymbol{v_i}|x_i = \ell; \boldsymbol{\theta_\ell})}{\sum_{n=1}^{L} \lambda_n P_n(\boldsymbol{v_i}|x_i = n; \boldsymbol{\theta_\ell})}. \tag{3.8}$$

Note that,

$$\sum_{\ell=1}^{L} P(\boldsymbol{v_i}|x_i = \ell) = 1. \tag{3.9}$$

Although we have an assumption that clusters consist of different number of voxels, and even some of the clusters gather majority of the voxels, the mixture weight λ of

every cluster is initialized as the same. And, we let the MRF model determine shape and size of the clusters, by optimizing the energy function.

In this study, a voxel $\boldsymbol{v_i} \in \mathbb{R}^d$ given its cluster label x_i follows d-variate Normal distribution such that the parameter set $\boldsymbol{\theta_\ell}$ of the mixture component is defined as $\boldsymbol{\theta_\ell} = \{\boldsymbol{\mu_\ell}, \boldsymbol{\Sigma_\ell}\}$. The cluster component is approximated by a Normal distribution:

$$P(\boldsymbol{v_i}|x_i = \ell; \boldsymbol{\theta_\ell}) \sim \mathcal{N}_d(\boldsymbol{\mu_\ell}, \boldsymbol{\Sigma_\ell}), \text{ and} \tag{3.10}$$

$$P(\boldsymbol{v_i}|x_i = \ell; \boldsymbol{\theta_\ell}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma_\ell}|}} \exp\left(-\frac{1}{2}(\boldsymbol{v_i} - \boldsymbol{\mu_\ell})^{\mathrm{T}} \boldsymbol{\Sigma_\ell}^{-1} (\boldsymbol{v_i} - \boldsymbol{\mu_\ell})\right). \tag{3.11}$$

Note that in the above formulation, each mixture component consists of a set of "homogeneous" regions distributed over three dimensional brain volume. Accordingly, a cluster may involve one or more regions as follows:

$$\underbrace{\mathcal{R}_1 \dots \mathcal{R}_m}_{C_1}, \underbrace{\mathcal{R}_{m+1} \dots \mathcal{R}_n}_{C_2} \dots \underbrace{\mathcal{R}_p \dots \mathcal{R}_S}_{C_L}$$

3.4.1 Unary Features

Performing clustering on raw intensity measurements $\tilde{\boldsymbol{v}}_{\boldsymbol{i}}$ has a potential of covering the underlying cognitive effects, and no additional information such as experimental design and BOLD response is incorporated. However, low signal-to-noise ratio and increasing dimensionality of the temporal data, i.e., length of the time series, may cause practical difficulties in modeling the fMRI data [58, 51]. Instead of using the time series of a voxel $\tilde{\boldsymbol{v}}_{\boldsymbol{i}} \in \mathbb{R}^T$, Goutte et al. [57, 58] propose performing the clustering on a new feature space $\boldsymbol{v_i} \in \mathbb{R}^d$ where $d \ll T$, with an argument that "notion of distance becomes counterintuitive" in high-dimensional spaces. Moreover, estimation of the model parameters, such as covariance of the Normal distribution Σ becomes impractically large in the high-dimensional spaces.

In the fMRI literature, univariate analysis techniques have been applied in order to project the high-dimensional time series of the voxels onto a new feature space [58, 51, 10, 17]. Univariate analysis helps us making an inference about how experimental conditions affect response of the individual voxels. Since the experimental design is incorporated into the analysis, these features become directly associated with the

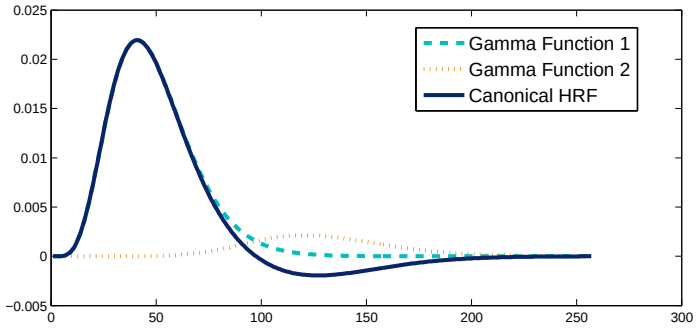


Figure 3.2: Canonical HRF function and its constituent Gamma functions where "Gamma Function 1" and "Gamma Function 2" correspond to the first and second terms of the equation 3.13 respectively.

underlying conditions, and hence become more descriptive of the cognitive tasks.

In this study, the unary feature set $\boldsymbol{v_i} = \{\rho_n\}_{n=1}^{d}$ is defined as the statistical similarity of the i^{th} voxel with the theoretical BOLD response under various conditions. And, the statistical similarity is calculated by using Pearson correlation coefficient between the time series of the i^{th} voxel $\tilde{\boldsymbol{v}}_{\boldsymbol{i}}$ and the time series of the theoretical response $\boldsymbol{u_n}$ such that

$$\rho_n = \frac{cov(\tilde{\boldsymbol{v}}_{\boldsymbol{i}}, \boldsymbol{u_n})}{\sqrt{var(\tilde{\boldsymbol{v}}_{\boldsymbol{i}})var(\boldsymbol{u_n})}}. \tag{3.12}$$

It has been reported that if enough separation is provided between consecutive stimuli -at least 5 seconds-, the BOLD response can be regarded as linear with respect to the stimulus [87, 91]. Therefore, the input-output relationships between the stimuli and the corresponding BOLD signal can be modeled as a linear time-invariant system, so that the theoretical BOLD response can be estimated by convolving the stimulus function with the theoretical form of the hemodynamic response function (see 2.4). Accordingly, we estimate the theoretical BOLD response, $\boldsymbol{u_n}$, by convolving the *canonical HRF* function of SPM toolbox [92, 93, 94] with the stimulus function under various conditions. The *canonical HRF* is modeled as a mixture of two Gamma distributions (see Fig. 3.2) such that

$$h(t) = \left(\frac{t^{\alpha_1 - 1} \beta_1^{\alpha_1} e^{-\beta_1 t}}{\Gamma(\alpha_1)} - c \frac{t^{\alpha_2 - 1} \beta_2^{\alpha_2} e^{-\beta_2 t}}{\Gamma(\alpha_2)} \right), \tag{3.13}$$

where $\Gamma(\cdot)$ is the Gamma function and the parameters $\alpha_1, \alpha_2, \beta_1$ and β_2 are set to their

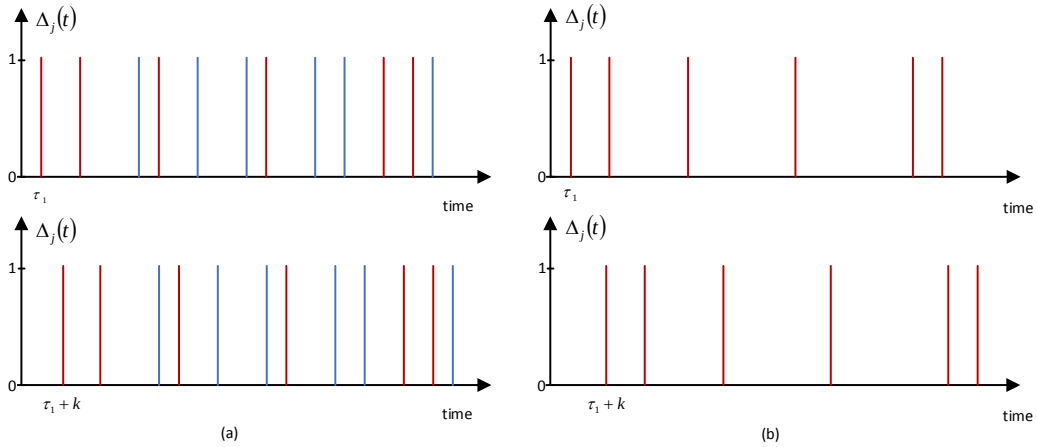


Figure 3.3: Examples of the stimulus function $\boldsymbol{\Delta}_j$ from a two-class experiment where red and blue colors represent different class conditions. (a) Stimulus function is constructed by using both of the conditions. (b) Stimulus function is constructed by only using one of the class conditions. (Top) No delay is applied. (Bottom) Stimulus onsets are shifted in time in k units of TR.

default values $6, 16, 0.125$ and 0.125, respectively. Time-invariance property of the model enables us to represent different delay conditions in the response. Moreover, in order to consider the behavior of a voxel under various stimulus conditions, the stimulus function $\boldsymbol{\Delta}_j$ is constructed with j^{th} condition-induced regressors. Suppose the stimuli of j^{th} condition are presented at times $\tau_1 \dots \tau_M$, and the stimulus at time τ_m is represented with the Dirac delta function $\delta(t - \tau_m)$. Then, the stimulus function $\boldsymbol{\Delta}_j$ can be defined as,

$$\boldsymbol{\Delta}_j(t) = \sum_{m=1}^{M} \delta(t - (\tau_m + k)), \tag{3.14}$$

where k controls amount of the shift in time, and it is in the unit of time repeat (TR) value of the MRI machine (see Fig. 3.3). Thus, the theoretical BOLD response $\boldsymbol{u_n}$ is defined as,

$$\boldsymbol{u}_n(t) = \boldsymbol{\Delta}_j(t) * h(t), \tag{3.15}$$

where the $*$ indicates the convolution operator. It can be concluded that each feature ρ_n reveals the characteristics of a voxel under different class conditions. First, the stimulus function $\boldsymbol{\Delta}_j$ can be constructed by using different sets of stimulus onsets, which implies coherence of a voxel with the corresponding tasks. Second, by shifting

the stimulus onsets in time, theoretical responses under various delay conditions can be estimated, which is expected to help covering the delayed activation patterns of voxels.

3.5 Potts Energy Term

We keep the Potts energy term E_p in the model for the purpose of spatial smoothness. In other words, E_p helps the model providing spatial continuity. The Potts energy term E_p is defined by using the standard Potts model which is a well-studied and commonly applied smoothing function [5, 73]:

$$E_p(x_1 \dots x_N) = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \beta_p U_p(x_i, x_j), \tag{3.16}$$

where β_p is the hyper-parameter adjusting the penalty. U_p is the potential function that enforces the model to give the same cluster labels to he neighboring voxels, and it is defined as

$$U_p(x_i, x_j) = \begin{cases} 1, & \text{if } x_i \neq x_j \\ 0, & \text{otherwise.} \end{cases} \tag{3.17}$$

3.6 Functional Energy Term

Differently from the Potts energy E_p, our functional energy term E_f incorporates data-dependent voxel similarity into the model. If any neighboring voxels are not assigned with the same labels, E_f term applies a penalty in proportion to functional similarity of the voxels. In other words, the model tends to gather voxels that give correlated responses into the same cluster. Hence, E_f ensures both spatial continuity and functional homogeneity simultaneously.

Functional connectivity, is defined as the statistical similarity between time-series of a voxel pair. More specifically, it is the temporal correlation that reflects co-activation of the distinct voxels [95]. Kriegeskorte et al. [96] state that spatially closer voxels are more likely to give similar responses under the same stimuli. Moreover, Both Firat et al. [97] and Onal et al. [98], in the brain decoding studies, employ features

based on the local voxel interactions, and report that local relations are more informative compared to the unary features. Although HRF varies across individuals and even brain regions [99, 89, 100], Bazargani and Nosratinia [101] state that BOLD response across a local neighborhood remains constant. Therefore, based upon the empirical evidences found in neuroscience literature, we may define a homogeneity predicate for brain regions, when the voxels in the same neighborhood are statistically correlated, regardless of their positions in the brain. More specifically, the local interactions of a voxel with its immediate neighbors are independent from behavior or analytic form of the BOLD response. Since these local interactions are based on the functional connectivity concept, we refer them as *functional texture* in the fMRI data. Note that local interactions are not informative about the type of the voxel behavior, i.e., activated or non-activated. Hence, the unary features and these pairwise features are complementary to each other.

Based upon the above discussion, the functional energy term E_f is defined as,

$$E_f(x_1 \dots x_N) = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \beta_f U_f(x_i, x_j), \tag{3.18}$$

where β_f is the hyper-parameter that controls contribution of the functional energy to the total energy. Lower values of the β_f make the model more tolerant towards missing the local interactions, and hence the clustering result becomes more scattered. The potential function U_f is defined as

$$U_f(x_i, x_j) = \begin{cases} |\rho_{ij}|, & \text{if } x_i \neq x_j \\ 0, & \text{otherwise} \end{cases} \tag{3.19}$$

where ρ_{ij} is the Pearson correlation coefficient. It is a measure of linear dependence between time series of the i^{th} and j^{th} voxels. It is given by

$$\rho_{ij} = \frac{cov(\tilde{\boldsymbol{v}}_{\boldsymbol{i}}, \tilde{\boldsymbol{v}}_{\boldsymbol{j}})}{\sqrt{var(\tilde{\boldsymbol{v}}_{\boldsymbol{i}}) var(\tilde{\boldsymbol{v}}_{\boldsymbol{j}})}}. \tag{3.20}$$

The correlation coefficient ρ_{ij} takes values between $[-1, 1]$ where values around zero indicate that voxels behave independently while a correlation value of -1 and $+1$ are an indicator of negative and positive statistical similarity respectively. In this study, negative and positive coefficients are given the same emphasize by considering the absolute value of ρ_{ij}.

Table 3.1: Triangle inequality condition. Possible configurations $(\alpha, \gamma, \varsigma)$ of a pair of random variables x_i and x_j, and corresponding energy (penalty) assignments under potential function U_f of the *functional energy* E_f are listed.

Configuration	Triangle inequality Eq. (3.23)
$\alpha = \gamma = \varsigma$	$0 \leq 0 + 0$
$\alpha \neq \gamma, \gamma \neq \varsigma, \alpha \neq \varsigma$	$\rho_{ij} \leq \rho_{ij} + \rho_{ij}$
$\alpha = \gamma, \gamma \neq \varsigma$	$0 \leq \rho_{ij} + \rho_{ij}$
$\alpha \neq \gamma, \alpha = \varsigma$	$\rho_{ij} \leq 0 + \rho_{ij}$
$\alpha \neq \gamma, \gamma = \varsigma$	$\rho_{ij} \leq \rho_{ij} + 0$

3.7 f-MRF Energy Minimization Algorithm

The configuration that minimizes the total energy E, defined in Eq. (3.3), is estimated by using the α-*expansion* energy minimization algorithm [83, 85]. It requires the pairwise potential functions to be a metric on the cluster label space $\mathcal{L}$. Potential function U is a metric if it satisfies

$$U(x_i, x_j) = 0 \Leftrightarrow x_i = x_j, \tag{3.21}$$

$$U(x_i, x_j) = U(x_j, x_i) > 0, \tag{3.22}$$

$$U(x_i = \alpha, x_j = \gamma) \leq U(x_i = \alpha, x_j = \varsigma) + U(x_i = \varsigma, x_j = \gamma), \tag{3.23}$$

for any cluster labels x_i, x_j and label instantiations $\alpha, \gamma, \varsigma \in \mathcal{L}$. It has been previously shown that one of the pairwise potential functions, U_p of the Potts model, is a metric [83]. The potential function U_f of the *functional energy* that has the same structure with Potts' model, is controlled by the pairwise correlation ρ_{ij} of the voxels. Since we employ absolute value of the Pearson correlation, and the Pearson correlation is inherently symmetric, the conditions given by 3.21 and 3.22 are satisfied. For the third condition, all possible configurations of the label instantiations are provided in table (3.1). By inspecting the table, it can be seen that potential function U_f of the *functional energy* term does not violate any of three metric conditions. Hence, α-*expansion* algorithm can be employed for minimization of the energy $E = E_d + E_p + E_f$ of our method.

3.8 Estimation of Labels

The cluster labels can be estimated by maximizing the posterior probability or by minimizing the total energy equivalently. In both settings, formulation of the problem is crucial. In terms of energy minimization, even if the optimization algorithm is able to find the global solution, the final labeling result could still be unfavorable. More specifically, the energy terms are inadequate in defining the problem. Therefore, it is important to build a decent setup, which provides a reliable search space for the energy minimization algorithm. In this study, an E-M like iterative approach [18, 102] is adopted where at every iteration the model is expected to find a better representation for the data. The cluster labels are estimated iteratively in two steps such that:

1. Given the current estimate of parameters of the mixture model $\hat{\boldsymbol{\Theta}}^{\boldsymbol{t}}$, unary energy term ${E_d}^t$ is calculated. By using the α-*expansion* algorithm, best configuration of the labels $\hat{X}^t$ that minimizes the total energy is determined.
2. Given the current estimate of cluster labels $\hat{X}^t$, for the next iteration, parameters of the mixture model $\hat{\boldsymbol{\Theta}}^{\boldsymbol{t+1}}$ is estimated.

The iterations continue until there is no difference between $\hat{X}^t$ and $\hat{X}^{t+1}$. At every iteration, the aim is tailoring the mixture components in accordance with the labeling so that the clusters become more homogeneous. In return, the MRF model quickly converges as the unary energy terms become more stable and more precise. Note that pairwise energy terms E_p and E_f are not modified during the iterations.

We also apply a threshold to put a constraint on minimum size of the clusters. During the iterations, if the number of voxels in a cluster C_ℓ becomes less than an empirically defined threshold, then probability of voxels having label ℓ are set to zero such that

$$P(\boldsymbol{v_i}|x_i = \ell) = 0, \forall i. \tag{3.24}$$

In other words, a cluster that is smaller than the intended size is disbanded. In the labeling step, MRF model ignores the cluster C_ℓ in order to avoid the large penalty. Thus, the voxels assigned with cluster label ℓ prefer clusters other than C_ℓ. This practice noticeably prevents the model from yielding scattered clustering results. Moreover, it enables the model to favor a subset of the mixture components. Different

from the mixture models, *f-MRF* does not necessarily find an estimation for all components. In fact, *f-MRF* has potential to uncover optimal number of clusters in the data.

At the second step of the aforementioned approach, parameters of the mixture components, i.e., parameters of the clusters, $\hat{\boldsymbol{\theta}}_{\boldsymbol{\ell}} = \{\hat{\boldsymbol{\mu}}_{\boldsymbol{\ell}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\ell}}\}$ are calculated by using the *maximum-likelihood estimation* (MLE):

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\ell}} = \frac{1}{N_\ell} \sum_{i \in C_\ell} \boldsymbol{v_i}, \tag{3.25}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\ell}} = \frac{1}{N_\ell} \sum_{i \in C_\ell} (\boldsymbol{v_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\ell}})(\boldsymbol{v_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\ell}})^T, \tag{3.26}$$

where C_ℓ is the set of voxels having cluster label ℓ, and the partitioning C can be written as $C = \bigcup_{\ell=1}^{L} C_\ell$.

3.9 Computational Complexity Analysis

Computational complexity of the proposed segmentation method *f-MRF* is determined by the energy minimization steps, given in lines 11-12 of the algorithm (1). In a cycle, the α-*expansion* algorithm iterates for each cluster label $\ell \in \mathcal{L}$, in total of $|\mathcal{L}|$ times. Recall from chapter 2.6.4 that at every iteration a max-flow graph is constructed followed by the min-cut operation. It was reported that the worst-case computational complexity of this operation is $O(mn^2|C|)$ [85] where $|C|$ is the cost of the minimum cut, n is the number of nodes and m is number of edges in the max-flow graph. Note that n and m include auxiliary nodes and edges, and bounded by $O(|V|)$ and $O(|\mathcal{E}|)$, respectively. The algorithm is guaranteed to terminate in $O(|V|)$ cycles under the assumptions that energy terms are constant, thus independent of the graph size. However, it is shown that it takes a few cycles before the termination, in practice [83]. Therefore, computational complexity of the lines 11-12 can be reported as $O(|\mathcal{L}|mn^2|C|)$.

The iterative approach between lines 7-14 continues for a predefined number of iterations. If it is not specified it halts when the convergence criterion is met (line 14). In the experiments, number of iterations t that the algorithm requires to converge was quite reasonable and insignificant relative to number of nodes $|V|$. In fact, at every

Algorithm 1 Steps of *f-MRF* Clustering Method

Require: Voxel time series $\tilde{V}$

Ensure: Label estimation $\hat{X}$.

1: Construct the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ on $3-$dimensional voxel grid.
2: Calculate *unary features* by using Eq. (3.12)
3: Calculate smoothing energy terms E_p and E_f by using Eqs. (3.17 and 3.18), respectively.
4: Initialize a spatially constrained random labeling $\hat{X}^1$.
5: $\mathcal{L}^1 \leftarrow \mathcal{L}$
6: $t \leftarrow 0$.
7: **do**
8: $t \leftarrow t + 1$
9: Estimate $\hat{\Theta}^t$ based on the labeling $\hat{X}^t$ by using Eqs. (3.25 and 3.26).
10: Calculate unary energy term ${E_d}^t$ by using Eq. (3.7) for all voxels and cluster labels. Meanwhile, identify the clusters C_ℓ, where $\ell \in \bar{\mathcal{L}}$ that consist of less number of voxels than the cluster size threshold.
11: $\mathcal{L}^{t+1} \leftarrow \{\mathcal{L}^t\} - \{\bar{\mathcal{L}}\}$.
12: Initialize α-*expansion* algorithm on graph $\mathcal{G}$.
13: $\hat{X}^{t+1} \leftarrow$ Run α-*expansion* algorithm with parameters $\{{E_d}^t, E_p, E_f, \mathcal{L}^{t+1}\}$.
14: **while** $\hat{X}^t \neq \hat{X}^{t+1}$
15: **return** $\hat{X}^{t+1}$

iteration, since some of the clusters are ignored due to the size threshold ($\bar{\mathcal{L}}$), the α-*expansion* algorithm runs on a subset of labels ($\mathcal{L}^{t+1} \subset \mathcal{L}$), which dramatically decreases the run time of the energy minimization step.

3.10 Evaluation of the Output of f-MRF Segmentation

The output of the *f-MRF* algorithm yields a partition of the brain into a set of homogeneous regions. We expect that some of the active regions are responsible of generating the underlying cognitive process. We also expect that majority of the voxels which belongs to large clusters do not contribute to the cognitive process. In order

to measure degree of validity of the *f-MRF* segmentation, we need to define some measures. In this study, the validity of the segmentation is evaluated by the classification performance of the resulted segments. In order to identify informative voxels, we employ the *f-MRF* as initial step of the classification pipeline. More specifically, voxel selection and feature extraction steps are performed on top of the rigorously generated clusters. Then, the voxels in the active segments are used as the feature vectors of a classification algorithm. Note that a cluster may consists of several functionally similar regions. Hence, instead of using the regions individually, we employ the cluster with its constituent segments.

Redundant and non-informative voxels are eliminated from the data by means of selecting the most informative clusters. In order to select the clusters of activated voxels which contribute to the underlying cognitive stimuli, we propose two different greedy approaches, namely, **s**election by **c**ross **v**alidation (**SCV**) and **s**election by **K**ullback **L**eibler divergence (**SKL**). After selecting the active clusters by using the training data, we compute the classification performance on the test data.

Let $\boldsymbol{F_{tr}}$ and $\boldsymbol{F_{te}}$ be our training and test feature matrices of size N_{tr}xM and N_{te}xM where N_{tr} and N_{te} are number of training and test samples, respectively, and M is the number of features. Let $\boldsymbol{c_{tr}} = \{c_1 \dots c_{N_{tr}}\}$ be the vector of class labels for training data and $\boldsymbol{c_{te}} = \{c_1 \dots c_{N_{te}}\}$ be the ground truth vector of class labels for test data, respectively. Note that each class label c_i represents category of the stimulus (sample) in the fMRI experiment.

After training a classifier with $\boldsymbol{F_{tr}}$ and $\boldsymbol{c_{tr}}$, and asking class labels for the unseen test data $\boldsymbol{F_{te}}$, the classifier yields a vector of estimated class labels $\hat{\boldsymbol{c}}_{\boldsymbol{te}} = \{\hat{c}_1 \dots \hat{c}_{N_{te}}\}$. Accuracy of the classifier, acc, is calculated by using

$$acc = \frac{1}{N} \sum_{i=1}^{N_{te}} \delta(c_i, \hat{c}_i), \tag{3.27}$$

where $\delta(c_i, \hat{c}_i) = 1$ if $c_i = \hat{c}_i$ and $\delta(c_i, \hat{c}_i) = 0$ otherwise.

Note that in the voxel selection tasks **SCV** and **SKL**, columns of the feature matrices $\boldsymbol{F_{tr}}$ and $\boldsymbol{F_{te}}$ consist of intensity values of the selected voxels. Therefore, the number of features M corresponds to the number of selected voxels. Performance of a cluster is measured by applying K-fold cross validation on the training data, leaving out equal

Algorithm 2 Steps of cluster selection by cross-validation (**SCV**)

Require: Labeling X and the corresponding partitioning $C = \{C_1 \dots C_L\}$, training dataset D_{tr} divided into folds, vectors of the class labels for training folds $\boldsymbol{c_{tr}}$ and test folds $\boldsymbol{c_{te}}$, threshold T that determines number of selected clusters.

Ensure: A subset of clusters $\tilde{C} \subset C$ that yields highest cross-validation performance.

1: **for** $\ell = 1 \rightarrow L$ **do**
2: $\{\boldsymbol{F_{tr}^{\ell}}, \boldsymbol{F_{te}^{\ell}}\} \leftarrow$ Construct training and test feature matrices of the cluster C_ℓ on the training data D_{tr} by concatenating intensity values of its constituent voxels.
3: $\Omega_\ell \leftarrow$ Perform classification on $\{\boldsymbol{F_{tr}^{\ell}}, \boldsymbol{c_{tr}}, \boldsymbol{F_{te}^{\ell}}, \boldsymbol{c_{te}}\}$ in order to calculate the selection criterion, i.e., the classification accuracy, of cluster C_ℓ.
4: **end for**
5: $\hat{C} \leftarrow$ Sort clusters in descending order with respect to the selection criterion Ω.
6: **for** $t = 1 \rightarrow T$ **do**
7: $C_\ell \leftarrow \hat{C}(t)$,
8: $\{\boldsymbol{F_{tr}^{\ell}}, \boldsymbol{F_{te}^{\ell}}\} \leftarrow$ Construct training and test feature matrices of the cluster C_ℓ,
9: $\boldsymbol{F_{tr}^{(t)}} \leftarrow \boldsymbol{F_{tr}^{(t)}} \cup \boldsymbol{F_{tr}^{\ell}}$ and $\boldsymbol{F_{te}^{(t)}} \leftarrow \boldsymbol{F_{te}^{(t)}} \cup \boldsymbol{F_{te}^{\ell}}$,
10: $\psi^{(t)} \leftarrow$ Find the classification accuracy of $\{\boldsymbol{F_{tr}^{(t)}}, \boldsymbol{c_{tr}}, \boldsymbol{F_{te}^{(t)}}, \boldsymbol{c_{te}}\}$.
11: **end for**
12: $t \leftarrow \underset{t}{\operatorname{argmax}}\, \psi^{(t)}$
13: $\tilde{C} \leftarrow \{\hat{C}(1) \dots \hat{C}(t)\}$.
14: **return** $\tilde{C}$

number of samples per class on each fold. Hence, $\boldsymbol{F_{te}}$ is composed of the fold that is separated for the test.

Mitchell et al. [7] estimate discriminating power of a voxel by using accuracy of a single-voxel classifier over the training data of only the corresponding voxel. Likewise, in the **SCV** approach, a separate classifier is trained for each cluster using only the observations of its constituent voxels, and accuracy of the classifier is considered as discriminating power of the cluster. Starting from the cluster with highest classification accuracy, clusters are added iteratively in a descending order of the classifica-

tion accuracy. Hence, at each iteration t, performance on the expanding training and test feature matrices $\boldsymbol{F}_{tr}^{(t)}$ and $\boldsymbol{F}_{te}^{(t)}$ is computed. Note that a feature matrix $\boldsymbol{F}$ consists of the voxel intensity values which comes from the clusters of highest recognition accuracy. Finally, the clusters that give the highest classification performance together are selected. Steps of the **SCV** is provided in the algorithm (2).

In the second approach, **SKL**, we assess the informative power of a cluster by means of the discrepancy between class conditional densities the cluster yields. In an ideal clustering, similarly activated voxels are expected to be fallen into the same clusters. Based upon this fact, we can build a heuristic in order to estimate informative power of the clusters. We make an assumption that entries of the feature matrix F_ℓ of a cluster C_ℓ, i.e., intensity values, follows the same distribution. More specifically, sources of the observations, i.e., the voxels, are assumed to be the same. By doing so, we measure how successfully the cluster, with its constituent regions, discriminates samples of different categories. Accordingly, the class conditional densities of the samples, i.e., intensity values, are estimated by using Parzen Window approach. In order to calculate the difference between two class conditional distributions, symmetric Kullback-Leibler divergence [103] is used as follows

$$D(P,Q) = \frac{D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)}{2}, \tag{3.28}$$

where P and Q are discrete class conditional distributions, and D_{KL} is Kullback-Leibler divergence. It is defined as a non-symmetric measure of difference between discrete distributions P and Q:

$$D_{KL}(P \parallel Q) = \sum_i P(i) ln \frac{P(i)}{Q(i)}, \tag{3.29}$$

In **SKL**, clusters are greedily accumulated in the order of discrepancy scores, i.e., symmetric KL-divergence. Cluster selection is handled similar to the **SCV** approach. As it is in the **SCV**, the selected clusters are determined by the cross-validation accuracy. Note that **SCV** and **SKL** approaches only differ in the cluster selection criteria Ω. See the steps of the **SKL** in algorithm (3).

By using **SCV** and **SKL** criteria in cluster selection, we aim to see how well the clusters decompose the feature space into coherent partitions. More specifically, the clustering result is expected to increase the representative power of the data. By

Algorithm 3 Steps of cluster selection by Kullback-Leibler divergence (**SKL**)

Require: Labeling X and the corresponding partitioning $C = \{C_1 \dots C_L\}$, training dataset D_{tr} divided into folds, vectors of the class labels for training folds $\boldsymbol{c_{tr}}$ and test folds $\boldsymbol{c_{te}}$, threshold T that determines number of selected clusters.

Ensure: A subset of clusters $\tilde{C} \subset C$ that yields highest cross-validation performance.

1: **for** $\ell = 1 \rightarrow L$ **do**
2: $\Omega_\ell \leftarrow$ Estimate the selection criterion, i.e., symmetric KL-divergence, of cluster C_ℓ by using Eq. (3.28).
3: **end for**
4: $\hat{C} \leftarrow$ Sort clusters in descending order with respect to the selection criterion Ω.
5: **for** $t = 1 \rightarrow T$ **do**
6: $C_\ell \leftarrow \hat{C}(t)$,
7: $\{\boldsymbol{F}^{\boldsymbol{\ell}}_{\boldsymbol{tr}}, \boldsymbol{F}^{\boldsymbol{\ell}}_{\boldsymbol{te}}\} \leftarrow$ Construct training and test feature matrices of the cluster C_ℓ,
8: $\boldsymbol{F}^{(\boldsymbol{t})}_{\boldsymbol{tr}} \leftarrow \boldsymbol{F}^{(\boldsymbol{t})}_{\boldsymbol{tr}} \cup \boldsymbol{F}^{\boldsymbol{\ell}}_{\boldsymbol{tr}}$ and $\boldsymbol{F}^{(\boldsymbol{t})}_{\boldsymbol{te}} \leftarrow \boldsymbol{F}^{(\boldsymbol{t})}_{\boldsymbol{te}} \cup \boldsymbol{F}^{\boldsymbol{\ell}}_{\boldsymbol{te}}$,
9: $\psi^{(t)} \leftarrow$ Find the classification accuracy of $\{\boldsymbol{F}^{(\boldsymbol{t})}_{\boldsymbol{tr}}, \boldsymbol{c_{tr}}, \boldsymbol{F}^{(\boldsymbol{t})}_{\boldsymbol{te}}, \boldsymbol{c_{te}}\}$.
10: **end for**
11: $t \leftarrow \underset{t}{\operatorname{argmax}}\, \psi^{(t)}$
12: $\tilde{C} \leftarrow \{\hat{C}(1) \dots \hat{C}(t)\}$.
13: **return** $\tilde{C}$

partitioning the feature space into functionally homogeneous and spatially coherent regions, we expect to get more informative feature subsets compared to all features.

We have conducted another test in order to compare the performance of the clustering under a simple, yet popular feature extraction task. In the fMRI literature, representing a group of voxels by their average time series is a widely accepted step in data analysis. It is applied by carrying various motivations such as noise elimination, dimensionality reduction or feature extraction [16, 10, 9]. Accordingly, we can represent each cluster with average of the intensity values of its constituent voxels. In the context of classification, the cluster-based averages are concatenated in order to construct a new feature matrix $\boldsymbol{F}$. This test is referred with the abbreviation **AVG**. Steps of the **AVG** approach is provided in algorithm (4).

Algorithm 4 Steps of feature matrix construction under **AVG** approach

Require: Labeling X and the corresponding partitioning $C = \{C_1 \dots C_L\}$, dataset D.

Ensure: Feature matrix $\boldsymbol{F}$.

1: **for** $\ell = 1 \rightarrow L$ **do**
2: $\quad \boldsymbol{F}^{\ell} \leftarrow$ Find the representative signal of cluster C_ℓ by averaging time series of its constituent voxels.
3: $\quad \boldsymbol{F} \leftarrow \boldsymbol{F} \cup \boldsymbol{F}^{\ell}$.
4: **end for**
5: **return** $\boldsymbol{F}$

3.11 Summary

In this chapter, the proposed segmentation method, *f-MRF* which partitions the brain into a set of homogeneous regions is explained. Main purpose of the *f-MRF* is gathering the segments of similarly activated voxels into the same clusters by using activity patterns in the fMRI data, hence providing clusters of informative voxels as well as redundant (non-informative) voxels. Differently from the existing solutions in the fMRI literature, *f-MRF* employs two different feature sets, namely *unary features* and *pairwise features* simultaneously. While the *unary features* reveals activation patterns of the voxels, the *pairwise features* ensures functional homogeneity and spatial continuity. *f-MRF* formulates the clustering problem as energy minimization under Markov Random Fields. Energy function of the *f-MRF* consist of three different terms, namely, unary, potts and functional energy terms, where we incorporate the functional energy term into the MRF model in order to exploit local functional interactions in the fMRI data. The method starts by modeling the voxel space with a mixture model over the *unary features*. By iteratively estimating the cluster labels given the mixture model and parameters of the mixture model given the current configuration of labels, *f-MRF* yields the labeling that minimizes the energy. *f-MRF* is proposed as an initial step for classification of the cognitive states. Hence, we evaluate performance of the final labeling by applying voxel selection and feature extraction operations on top of the clusters.

CHAPTER 4

EXPERIMENTS TO ANALYZE VALIDITY OF THE F-MRF METHOD

In this chapter, behavior of the *f-MRF* and compared algorithms are examined on a real fMRI data of visual object recognition. In the analysis section, we observe effects of the energy weight parameters of *f-MRF* on the energy function during the iterative solution. Moreover, the *unary* and *pairwise features* of the *f-MRF* are analyzed, and contribution of the functional energy term is presented.

In the comparative results section, *f-MRF* is compared with three well-known algorithms, namely, K-Means, Gaussian Mixture Model (GMM) and Normalized Cuts (nCut). We propose partitioning the high-dimensional voxel space, i.e., feature space, into a set of homogeneous segments in order to increase representative power of the data for brain decoding tasks. Therefore, quality of the segmentation results is evaluated by using classification performance. Accordingly, by constructing feature matrices under different voxel selection and voxel agglomeration routines which are performed on the clusters, classification performance is computed for each segmentation algorithm. Moreover, we provide illustrations of example segmentation results, and make an evaluation about size and activation patterns of the clusters.

4.1 fMRI Data Acquisition

In this study, we have conducted validation tests on a real fMRI dataset that consists of the neural activations during a visual object recognition experiment. fMRI samples are acquired under a one-back repetition detection task [6]. The subjects are presented

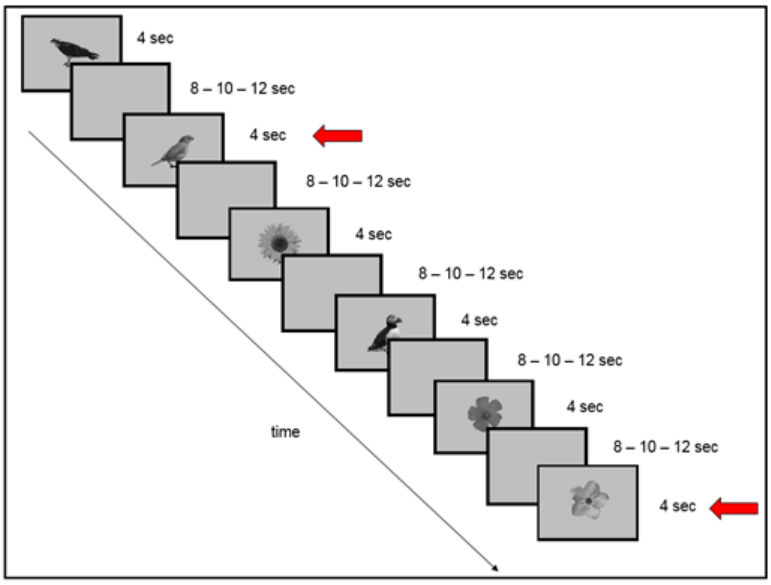


Figure 4.1: A sample sequence of the visual recognition experiment. After presentation of the stimulus image for 4 seconds, a rest period of 8,10 or 12 seconds follows [6].

visual stimuli for 4 seconds and asked whether category of the current stimulus is same with the previous one's. In order to clearly separate the responses given to the stimuli, the stimuli onsets are followed by 8, 10 or 12 seconds of rest periods [104]. The visual stimuli consist of gray-scale images that belong to two categories, namely birds and flowers. The images are randomly selected for each run and are mutually exclusive across runs. fMRI measurements are recorded by using 3T Siemens MRI scanner with a TR of 2 seconds. SPM8 toolbox (http://www.fil.ion.ucl.ac.uk/spm/) is used for pre-processing of the images which are realignment of the functional images and co-registration to the anatomical image.

fMRI data acquisition experiment is conducted in 6 runs on 5 participants where one of the participants is discarded due to an error during the fMRI experiment. At each run, measurements from 36 stimuli are recorded in a total duration of 252 samples, where each class has equal number of samples. In order to preserve continuity of the time courses, an entire run is selected for either training or test dataset. The fMRI dataset is divided into two equal parts. Training data consists of samples from odd numbered runs (1,3,5) while even numbered runs (2,4,6) are reserved as test data. Thus, the training and test data consist of 54 samples for each class. Time series of every voxel are first detrended to account for baseline shifts and scanner drift across

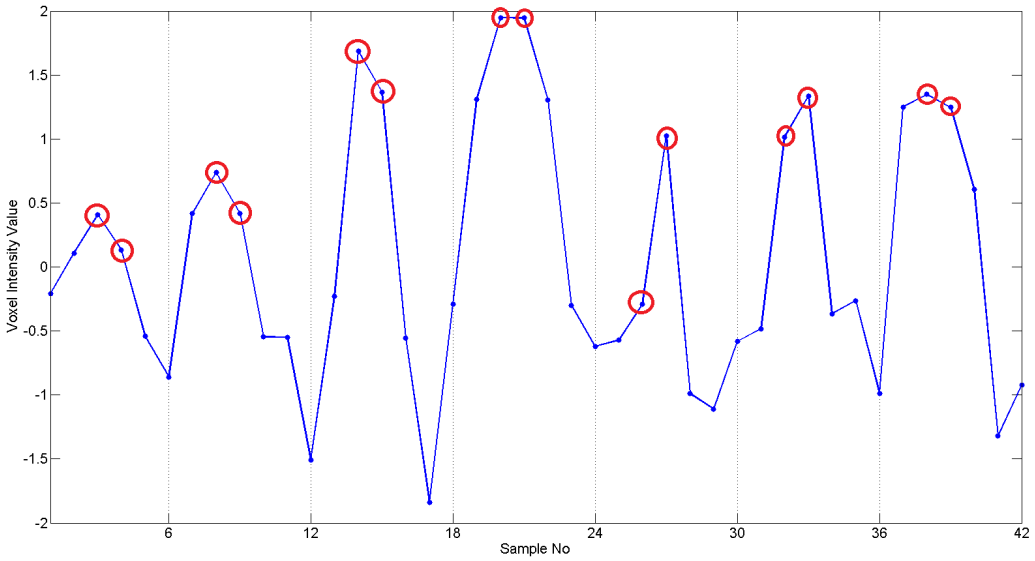


Figure 4.2: Time series of a voxel. Every 6-sample period belongs to a stimulus. Average of 2nd and 3rd observations after the stimulus onset (marked with red circles) is used in classification.

the runs, and then normalized so that each voxel has mean 0 and standard deviation 1 in a run.

A voxel $\tilde{v}_i$ consists of the fMRI measurements collected at every 2 seconds (TR) where the measurements include both the responses to the stimuli and the rest periods (see Fig. 4.1). Let $\boldsymbol{F_{tr}}$ and $\boldsymbol{F_{te}}$ be our training and test feature matrices of size N_{tr}xM and N_{te}xM, which are generated by using voxel intensity values. N_{tr} and N_{te} are the number of training and test samples, respectively. In this study, since we evenly split the data by means of odd and even runs, N_{tr} and N_{te} are 54. Note that size of the voxel observations $\tilde{v}_i$ is larger than $N_{tr} + N_{te}$. In the feature matrices, there exist one entry for each stimulus while we have 6 to 8 observations in the voxel time series. Accordingly, entries of the feature matrices $\boldsymbol{F_{tr}}$ and $\boldsymbol{F_{te}}$ correspond to average of 2nd and 3rd observations after the stimulus onset (see Fig. 4.2). On the other hand, dimension of the feature space M varies with respect to the voxel selection and feature extraction approaches. Let $\boldsymbol{c_{tr}} = \{c_1 \dots c_{N_{tr}}\}$ be the vector of class labels for training data and $\boldsymbol{c_{te}} = \{c_1 \dots c_{N_{te}}\}$ be the ground truth vector of class labels for the test data, respectively. Each entry of these vectors corresponds to the category of the stimulus presented, i.e., bird or flower.

4.2 Experimental Setup

In this study, we perform brain decoding in order to evaluate quality of the segmentation results. Hence, category of the stimulus, i.e., class labels, for the fMRI observations are required. After the fMRI data is divided into training and test data, the routines of segmentation and feature matrix design for brain decoding are performed on the training data. The test data is remained unseen until the classification.

After training a classifier with $\boldsymbol{F_{tr}}$ and $\boldsymbol{c_{tr}}$, and asking class labels for the unseen test data $\boldsymbol{F_{te}}$, the classifier yields a vector of estimated labels $\hat{\boldsymbol{c}}_{\boldsymbol{te}} = \{\hat{c}_1 \dots \hat{c}_{N_{te}}\}$. We can calculate accuracy of the classifier, acc, by using

$$acc = \frac{1}{N} \sum_{i=1}^{N_{te}} \delta(c_i, \hat{c}_i), \tag{4.1}$$

where $\delta(c_i, \hat{c}_i) = 1$ if $c_i = \hat{c}_i$ and $\delta(c_i, \hat{c}_i) = 0$ otherwise. We employ *k-Nearest Neighbor* (kNN) algorithm in order to compute classification accuracy, where the k value has an essential role on performance of the classifier. Thus, optimal value of the k is determined among the candidates starting from 1 to $\sqrt{N_{tr}}$ by running a K-Fold cross-validation on the training data.

Recall from chapter 3 that, the energy function of the *f-MRF* consists of three hyper-parameters, $\{\beta_d, \beta_p, \beta_f\}$, which adjust the weight of the energy terms. Initial number of clusters, β_c, is another hyper-parameter that is common for all clustering methods. In this study, clustering results are obtained by conducting a grid search on these parameters. All clustering methods *f-MRF*, K-Means, GMM and nCut are initialized with same β_c values such that $\beta_c \in \{30, 50, 100, 200, 300, 500, 700\}$.

Moreover, *f-MRF* results in less number of clusters than the initialization. In order to see the performance of the compared clustering methods using small number of clusters, we also initialized them with the number of *f-MRF*'s resulted clusters which varies across subjects (see Table 4.1). In the following sections, these results are titled with "Same # of Clusters".

For *f-MRF*, weight of the functional energy term β_f are varied while the parameters β_d and β_p remain constant. Then, our parameter search space for the *f-MRF* becomes β_f x β_c, where $\beta_f \in \{0.05, 0.3, 0.5, 1, 2, 2.5, 3, 3.5, 4, 5\}$. Note that the range of pa-

Table 4.1: Number of clusters in *f-MRF* results, which are employed as the initial number of clusters, β_c, for K-Means, GMM and nCUT algorithms. Note that *f-MRF* is initialized with the number of clusters $\beta_c \in \{30, 50, 100, 200, 300, 500, 700\}$.

Participants	# of *f-MRF*'s resulted clusters						
Participant 1	13	22	31	40	49	58	67
Participant 2	11	19	23	32	36	47	60
Participant 3	7	11	17	21	27	38	53
Participant 4	13	22	32	41	50	60	69

rameter sets are determined empirically.

After getting the labeling with all parameter configurations, feature matrices $\boldsymbol{F_{tr}}$ and $\boldsymbol{F_{te}}$ are constructed for each segmentation result. Recall from chapter 3.10 that three approaches, namely **SCV** and **SKL** for voxel selection and **AVG** for voxel agglomeration are proposed. In all three approaches, a cluster is taken into consideration with its all constituent segments.

In **SCV**, discriminative power of the clusters is measured by using classification accuracy. The accuracy is computed by applying cross-validation on the training data of the constituent voxels of each cluster. In **SKL** criteria, on the other hand, the clusters are evaluated with the symmetric Kullback-Leibler divergence between the class conditional densities. Both criteria select the clusters by using an iterative approach. Based on the evaluation score, i.e., accuracy or KL divergence, clusters are greedily selected. On the training data, classification performance is calculated by using constituent voxels of the selected clusters. The set of clusters that gives the highest accuracy is obtained at output of the **SCV** and **SKL** algorithms. In this study, both approaches continue until 20 iterations. In other words, outputs of **SCV** and **SKL** is allowed to consist of at most 20 clusters, where number of selected clusters is observed between 1 and 10 in the experiments. Feature matrices $\boldsymbol{F_{tr}}$ and $\boldsymbol{F_{te}}$ consists of the voxels which come from the selected clusters, where the entries correspond to average of 2$^{\text{nd}}$ and 3$^{\text{rd}}$ observations after the stimulus onset. Note that the size of the feature space varies with respect to the size and number of selected clusters.

In the **AVG** approach, each cluster is simply represented with the average time series of its constituent voxels. The representative signals of the clusters are employed in order to construct $\boldsymbol{F_{tr}}$ and $\boldsymbol{F_{te}}$ feature matrices. Entries of the feature matrices

correspond to the average of 2nd and 3rd observations on the representative signal. The size of the feature space is equivalent to number of the clusters. For K-Means and GMM, it is equal to preset number of clusters β_c. Since nCut and *f-MRF* may result in empty clusters, dimension of the feature space may be less than β_c.

For every clustering result, classification accuracy under the aforementioned approaches is computed. Over all clustering results of each algorithm (7 for nCut, GMM and K-Means, 70 for *f-MRF*) average and maximum accuracy is provided in the following sections.

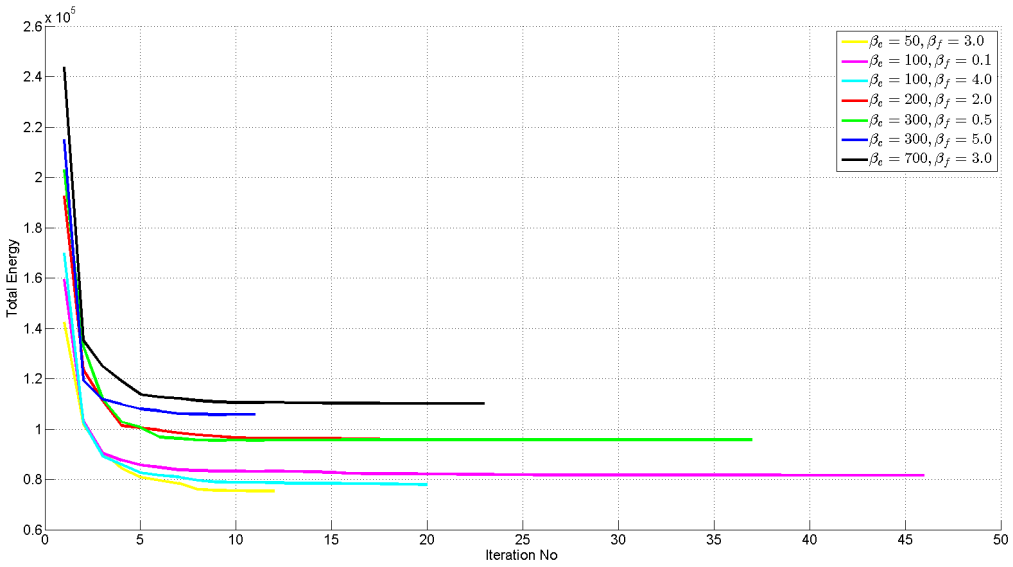
4.3 Analysis of the f-MRF

In this section, we analyze behavior of the *f-MRF* method based on the empirical evidence under different hyper-parameter settings. Recall that energy function of the *f-MRF* is controlled by three hyper-parameters, $\{\beta_d, \beta_p, \beta_f\}$, by adjusting weight of the energy terms. Moreover, we have and additional β_c parameter, initial number of clusters. In the experiments β_d and β_p are kept constant while $\beta_c \in \{30, 50, 100, 200, 300, 500, 700\}$ and $\beta_f \in \{0.05, 0.3, 0.5, 1, 2, 2.5, 3, 3.5, 4, 5\}$. Hence, we have a search space of β_c x β_f for *f-MRF*.

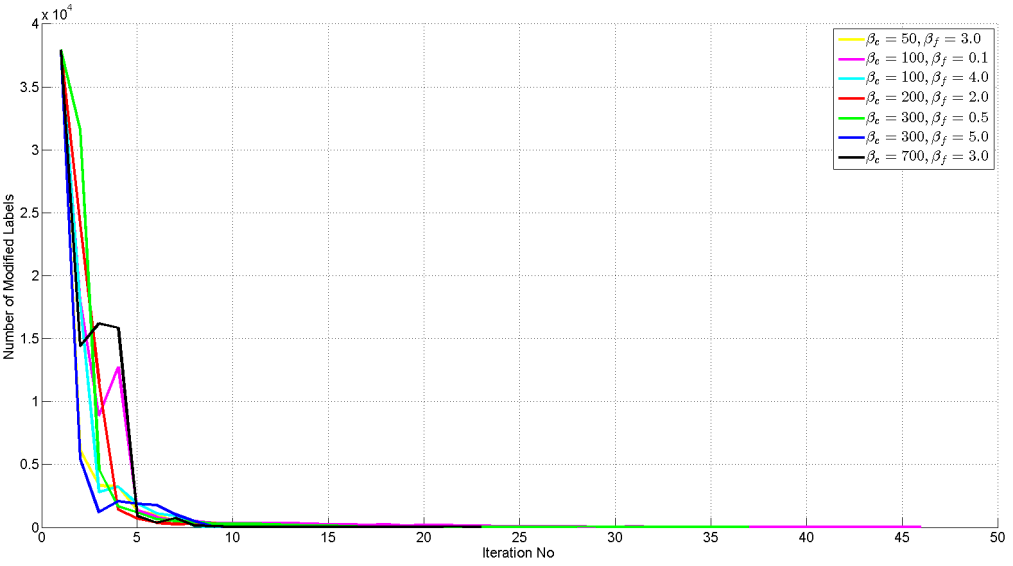
4.3.1 Effect of the hyper-parameters β_f and β_c on Convergence of the Iterative Solution

Recall from chapter 3.8 that in order to estimate latent labels, by iteratively updating the model parameters and the labeling, *f-MRF* tries to find a better fit to the data until a convergence is reached. In this study, as a convergence criterion, we prefer using number of the modified labels between consecutive iterations. In other words, the algorithm looks for the solution until any label change occurs. In [18], for a similar clustering method, Rylai et al. propose using fractional change in the energy between two iterations. Accordingly, if the fractional change is smaller than the threshold (0.1% for example), then the algorithm stops.

In Fig. (4.3), details of the iteration steps is shown for randomly selected clustering



(a)



(b)

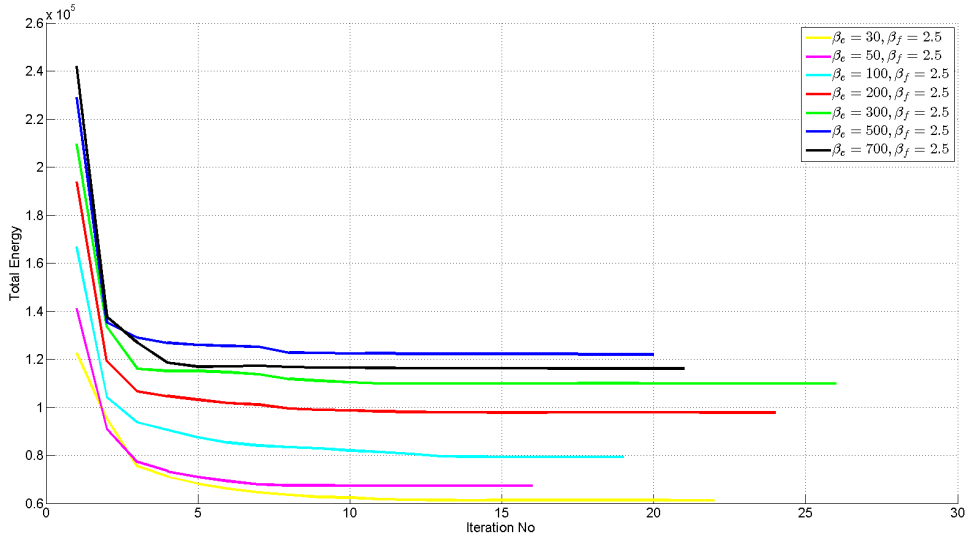
Figure 4.3: (a) Change in the total energy at every iterations of the solution and (b) total number of the modified labels with respect to the previous iteration under various β_c and β_f settings.

results. In the figure 4.3-a, total energy of the system during the iterative solution is plotted. In the first iterations of the solution (1-5), there is a dramatic decrease in the energy. Since the algorithm starts from a random configuration, in the first steps, the mixture model does not represent the data well. Besides, the unary costs which determine the unary energy, E_d, are almost uniform. Hence large penalties are

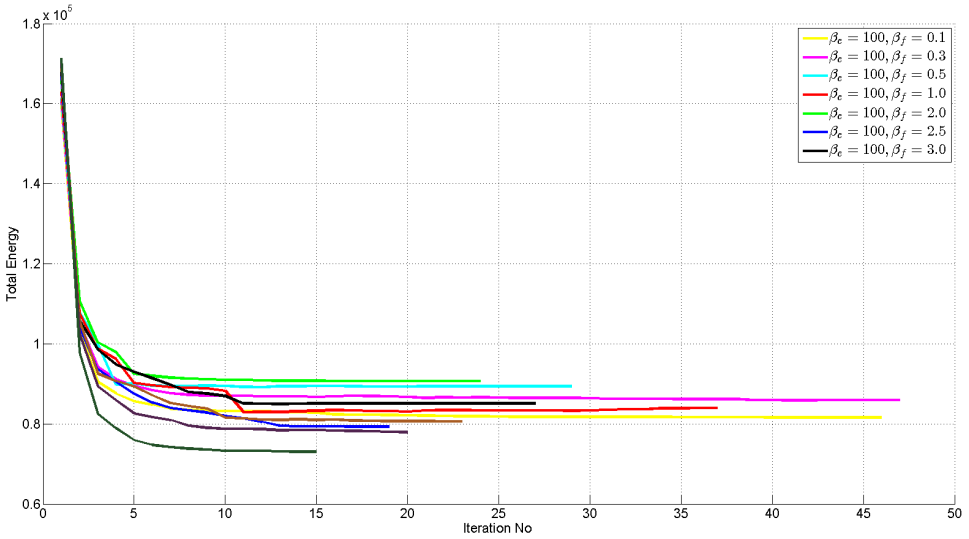
applied to the label assignments. As the iterations continue, clusters provide a better fit to the voxel space with an income of stabilized costs, thus lower unary energy. Independent from initialization of the β_c and β_f parameters, the model reaches a convergence around iteration 10, where the energy becomes steady. Note that the energy does not gradually decrease during the iterations. Small increases may be observed. However, these increments are negligible (in the order of 10^{-4} - 10^{-6} with respect to the previous iteration).

In figure 4.3-b, the number of the modified labels are illustrated during the iterations. Similar to energy curve, most of the changes occur in the first 5 iterations, and improvements for rest of the iterations are almost negligible (in the order of 10s out of $\sim$40000). Notice that the bumps before iteration 5 are due to the major modifications on the mixture model at the previous iteration, i.e., large amount of clusters are discarded in order to provide a better fit. Recall from chapter 3.4 and Eq. (3.8), large number of clusters introduces larger unary costs especially when the model is not stable during the first iterations. Hence, iterative solution of the *f-MRF* favors less number of clusters.

It can be seen from Fig. (4.4) that when we compare figures (a) and (b), initial cluster number, β_c parameter barely affects the convergence point. On the other hand, functional energy weight, β_f has a dominant role on the number of iterations required for a convergence. Notice that the higher values β_f takes, the faster the convergence is reached. This is because the fact that *f-MRF* becomes more aggressive when effect of the smoothing parameters (β_f and β_p) is larger. More specifically, large pairwise costs dominate the unary costs, and force the model to favor assignment of the same labels to the neighboring voxels. We can also observe from Fig. (4.4) that the energy mainly varies with respect to β_c parameter. This fact may essentially be attributed to the amount of unary penalty. As it is explained previously, probability of a voxel's being assigned to a cluster is much larger when number of the candidates is less, which introduces lower penalty values.



(a)



(b)

Figure 4.4: Change in the total energy at every iterations of the solution. (a) β_f is constant at 2.5 and β_c varies over all possible initial values, (b) β_c is constant at 100 and β_f varies over all possible settings.

4.3.2 Analysis of the Unary and Pairwise Features

Recall from chapter 3 that the *f-MRF* method makes use of two feature sets namely *unary* and *pairwise features*. By using a univariate approach, *unary features* of a voxel are defined as the statistical similarity between time series of the voxel and different theoretical response signals, where each feature corresponds to activity score of the

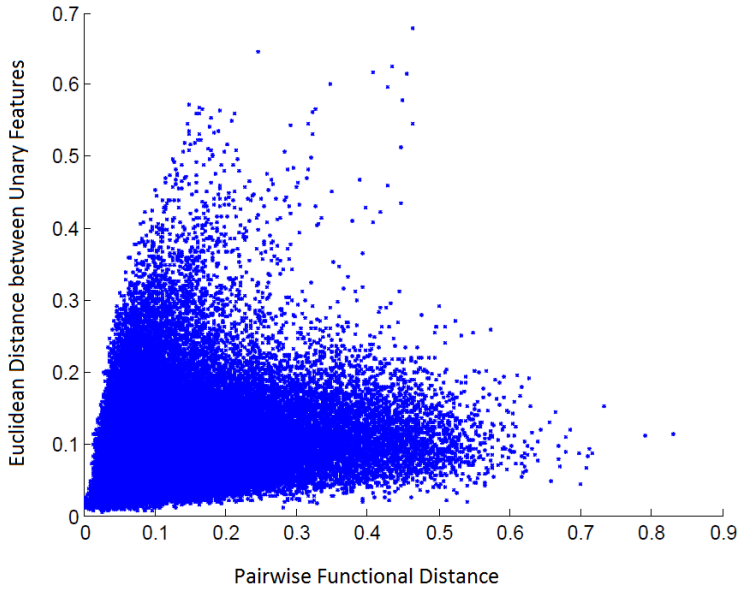


Figure 4.5: (Vertical axis - normalized between 0 and 1) Pairwise Euclidean distance between neighboring voxels i and j, $d_u(i,j)$. (Horizontal axis) Pairwise correlation distance between voxel time series of the voxels i and j, $d_p(i,j)$.

voxel under the corresponding experimental condition(s). On the contrary, *pairwise features* are measures of the functional similarity between pairs of voxels. In order to see how correlated these two feature sets, the distance between each neighboring voxels are calculated by using both *unary* and *pairwise features*.

Let $\boldsymbol{v_i}$ be the vector of $d-$dimensional *unary features*, and $\mathcal{N}_i$ be the neighborhood of the i^{th} voxel. Euclidean distance between pair of voxels i and j over the *unary features* $d_u(i,j)$ is defined by

$$d_u(i,j) = \sqrt{(\boldsymbol{v_i} - \boldsymbol{v_j})^T \cdot (\boldsymbol{v_i} - \boldsymbol{v_j})}, \tag{4.2}$$

where $j \in \mathcal{N}_i$. Since *pairwise features*, by definition, correspond to the similarity of neighboring voxels i and j, which take values between 0 and 1, the distance can be defined as

$$d_p(i,j) = 1 - \rho_{ij} \tag{4.3}$$

where ρ_{ij} is Pearson correlation between time series of the voxels i and j.

In Fig. (4.5), a scatter plot is provided. As it can be seen from the plot, these two distances are not equivalent for most of the instances. Only pairs of the highly correlated voxels give similar distance values (particularly around 0-0.2). The less correlated a

Table 4.2: Entries of the first three column are average ("Mean-Acc"), maximum ("Max-Acc") and standard deviation ("Std-Acc") of the accuracies that are obtained on the clustering results of each parameter configuration. In the fourth and fifth columns ("Mean-ClNo" and "Std ClNo"), average and standard deviation of the final cluster numbers in the clustering results are provided, respectively.

	Mean-Acc	Max-Acc	Std-Acc	Mean-ClNo	Std ClNo
f-MRF	0.86	0.93	0.04	38.43	18.53
f-MRF($\beta_f = 0$)	0.81	0.90	0.08	40	41.63

pair of voxels, the more likely it is to observe differences between *unary* and *pairwise features*. By inspecting the Fig. (4.5), we can also say that partial correlation of the time series (around 0.5) is not necessarily derived from similar activation characteristics.

4.3.3 Analysis of Contribution of the Functional Energy

In order to observe the contribution of the functional energy term to the overall performance, we have conducted **SKL**, **SCV** and **AVG** tests on the *f-MRF* results where the functional energy term E_f is simply ignored by setting the $\beta_f = 0$. In other words, *f-MRF* results only rely on unary and Potts energy terms. Since all three approaches give similar results, only performance of the **SKL** is provided in Table (4.2) and Fig. (4.6).

As it can be observed from fig. 4.6, in both settings, there is a decrease in the classification performance as weight of the pairwise cost ($\beta_f or \beta_p$) increases. Because the cost of assigning different labels to the neighboring voxels is getting larger, the MRF model becomes less tolerant to such occasions. It aggressively gathers voxels into the same cluster, and hence the final labeling results in fewer number of clusters. Yet positive effect of the functional energy can be observed from the results. By the help of functional energy, *f-MRF* applies spatial regularization selectively. More specifically, the model starts gathering the neighboring voxels into the same cluster from functionally most similar voxel pairs. Hence, functional energy is more robust to initialization of the weight parameters than the standard Potts energy.

In Table (4.2), the classification results of Fig. (4.6) is summarized. First three

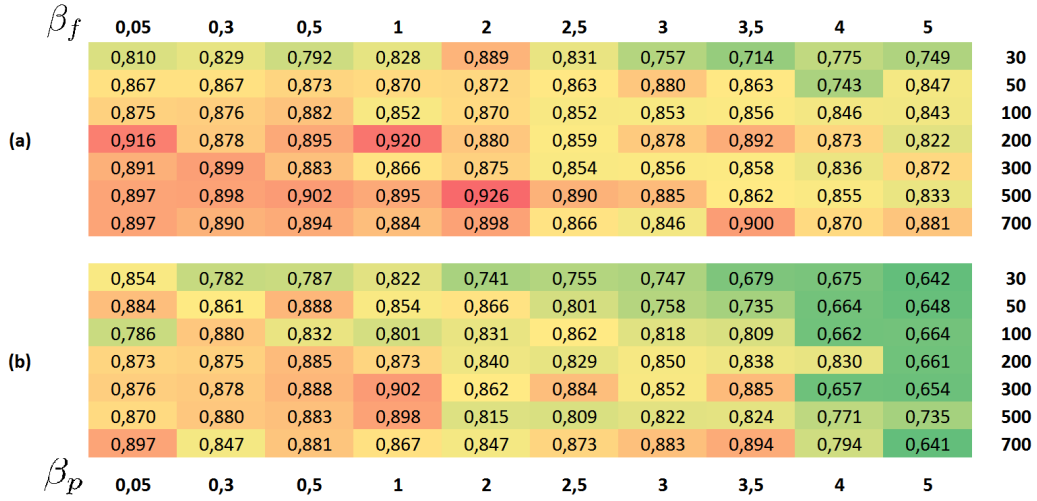


Figure 4.6: Classification accuracy of *f-MRF* results. Each entry corresponds to **SKL** performance of a clustering result. At the right most column initial number of clusters are listed. (a) Functional energy weight β_f varies while unary and Potts energy weights remain constant ($\beta_d = 1, \beta_p = 1$). (b) Potts energy weight β_p changes while unary and functional energy weights remain constant ($\beta_d = 1, \beta_f = 0$).

columns of the table shows average, maximum and standard deviation of the accuracies in Fig. (4.6). In columns fourth and five, mean and standard deviation of the final number of clusters under all parameter settings are presented. As it can be seen from the table, both of the mean and maximum accuracy values are better when functional energy is introduced to the total energy term. Moreover, variance in the number of the clusters and classification accuracy is much larger when the functional energy is not employed.

4.4 Comparative Results

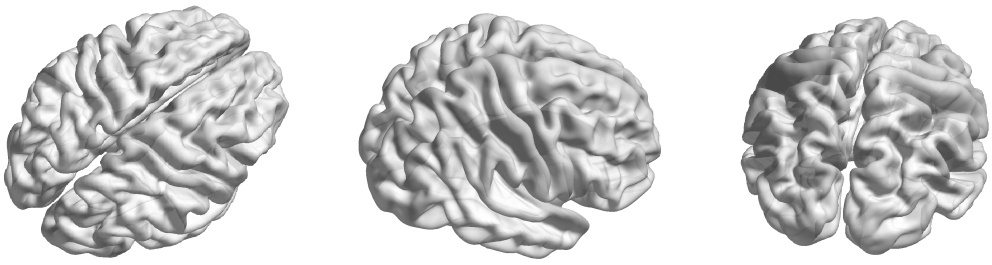
f-MRF can be considered as a hybrid model which combines the clustering approaches that employs either unary or pairwise features. Hence, we select the comparative clustering methods carefully in order to reveal advantages and disadvantages of such an hybrid approach. In this study, *f-MRF* is compared with standard K-Means, Gaussian Mixture Models (GMM) and spatially constrained spectral clustering (SCSC) [15] algorithms, where the abbreviation SCSC is used interchangeably with nCut.

Originally, Craddock et al. [15] apply the SCSC clustering algorithm on the resting-

state fMRI data in order to generate a continuous brain atlas. SCSC performs normalized cut (nCut) on graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of the MRF model where the set of egde weights $\mathcal{E}$ consist of the correlation values between neighboring voxel, i.e., pairwise features. Since SCSC and *f-MRF* operate on the very same graph, we employ SCSC in order to see effect of including the *pairwise features* only. Likewise, both K-Means and GMM methods are preferred in order to make a comparison between our hybrid approach and the approaches that are solely based on *unary features*.

In Figs. (4.8, 4.9, 4.10 and 4.11), example results for the four algorithms are illustrated from three viewpoints, and a bar plot showing size of the clusters in sorted order is provided. Since it is unclear in advance how many clusters *f-MRF* returns, first we get labeling of the *f-MRF* which consists of 60 clusters. Then, K-Means, GMM and nCut algorithms are initialized with 60 clusters. In Fig. (4.7), a 3d brain template is also illustrated as a reference by using the very same viewpoints that we visualize the clustering results.

In Fig. (4.8), *f-MRF* yields a partitioning that collects vast amount of voxels (28860 out of 37944) into a single cluster. It shows that non-activated voxels overwhelmingly spread across the brain in this visual recognition task. The MRF model, by accumulating the non-activated voxels under the same cluster, is able to capture structure of the data. The remaining clusters mostly consists of the voxels that show distinctive patterns of activation. Notice that voxels from the primary visual area are isolated



Azimuth = 50
Elevation = 80
(a)

Azimuth = 80
Elevation = 40
(b)

Azimuth = 12
Elevation = 40
(c)

Figure 4.7: Visualization of a 3d brain model as a reference in three different viewpoints which are also used in Figs. (4.8, 4.9, 4.10 and 4.11). Azimuth and elevation values set angle of the view in the horizontal coordinate system.

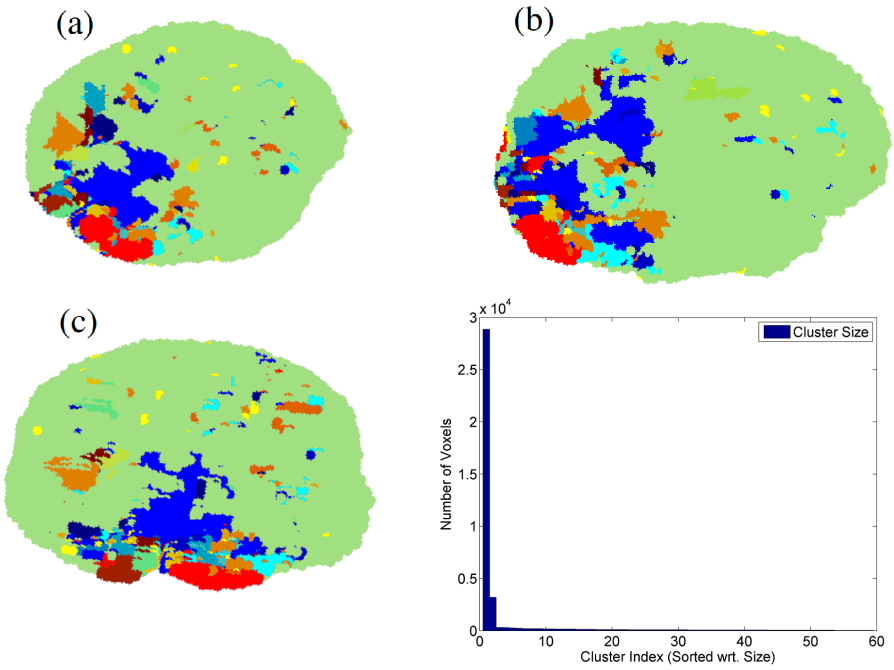


Figure 4.8: Visualization and cluster size plot (clusters are in sorted order with respect to size) for an example segmentation result of the *f-MRF*, where the parameters $\{\beta_d, \beta_p, \beta_f, \beta_c\}$ are initialized with $\{1, 1, 3.5, 700\}$.

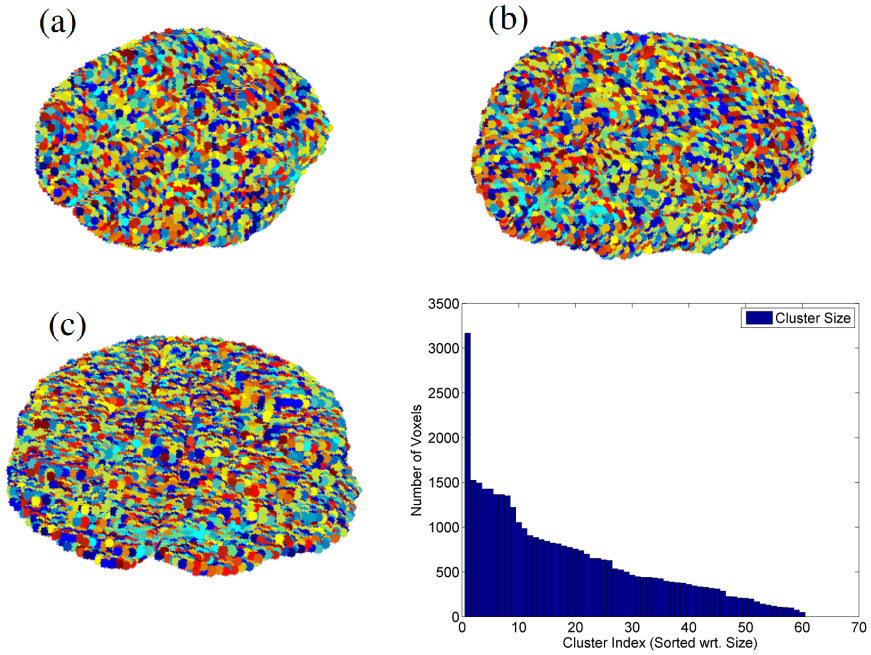


Figure 4.9: Visualization and cluster size plot (clusters are in sorted order with respect to size) for an example segmentation result of the K-Means, where the parameter $\{\beta_c\}$ is initialized with $\{60\}$.

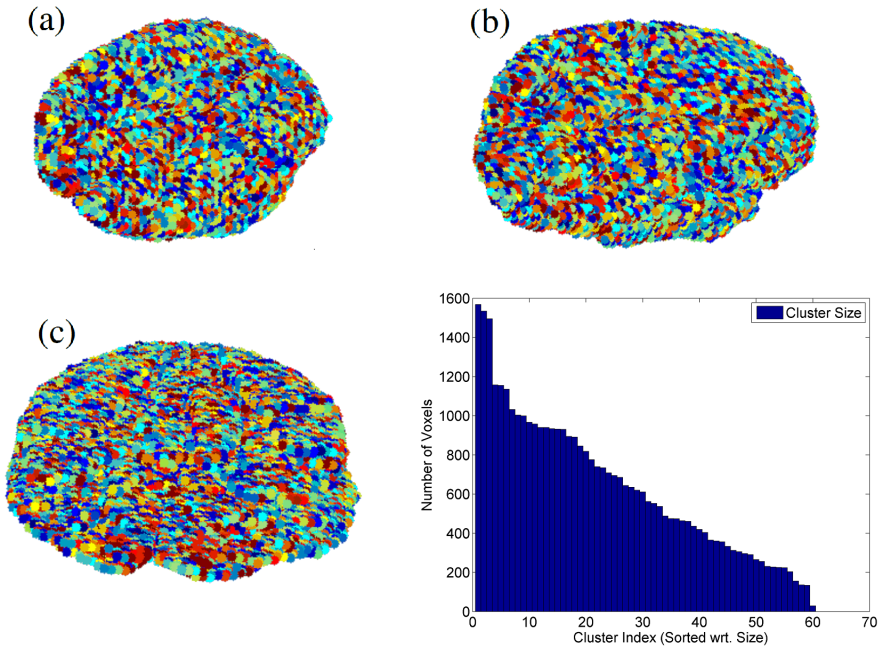


Figure 4.10: Visualization and cluster size plot (clusters are in sorted order with respect to size) for an example segmentation result of the GMM, where the parameter $\{\beta_c\}$ is initialized with $\{60\}$.

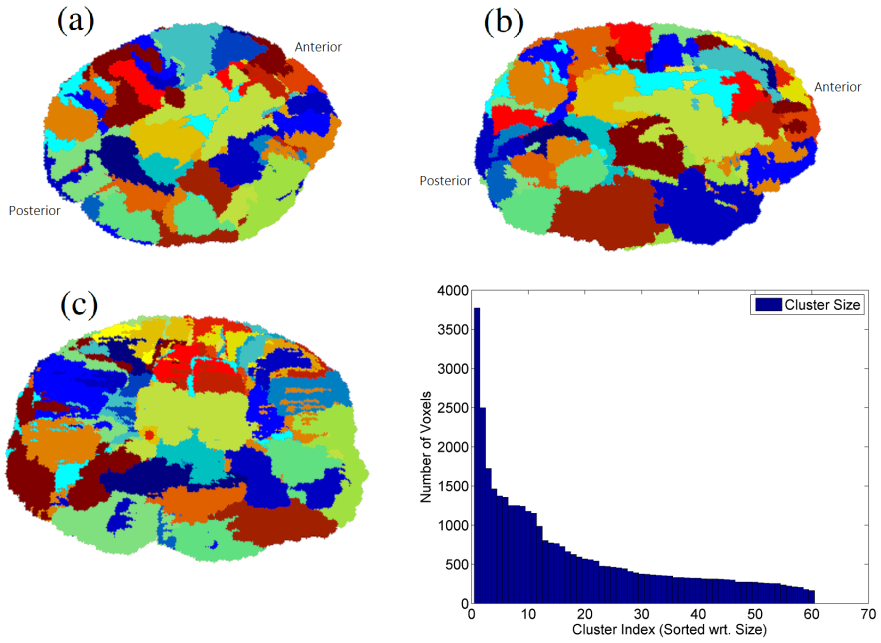


Figure 4.11: Visualization and cluster size plot (clusters are in sorted order with respect to size) for an example segmentation result of the nCut, where the parameter $\{\beta_c\}$ is initialized with $\{60\}$.

from the non-activated voxels, which is an expected outcome of the fMRI experiment. Moreover the resulted clusters are barely scattered compared to the results of K-Means and GMM algorithms due to the spatial regularization of the *f-MRF* model. Recall that clusters may consist of several continuous regions. Although the model enforces spatial continuity, the distributed patterns are represented in terms of disconnected regions of the clusters.

In order to quantitatively evaluate amount of scattering in the clustering results, we define the measure, S, for the clustering $C = \{C_1 \dots C_L\}$ as follows

$$S = \frac{1}{L} \sum_{\ell=1}^{L} \frac{|\mathcal{R}_\ell|}{|C_\ell|}, \tag{4.4}$$

where $\mathcal{R}_\ell$ is the regions of the cluster C_ℓ and $|\mathcal{R}_\ell|$ is the number of regions. Note that the regions correspond to the disconnected components on the graph of MRF. And, $|C_\ell|$ is number of the voxels in the cluster. Accordingly, if the clusters consist of one-voxel-sized regions where there is not any neighboring relationship between the constituent voxels, then S gives the maximum scattering score of 1. The values near zero indicate that the clustering is spatially continuous.

In Fig. (4.12), scattering scores of the algorithms are plotted. As it is expected, spatially constrained nCut gives continuous partitions (Fig. 4.10) while K-Means (Fig. 4.9) and GMM (Fig. 4.10) methods result in highly scattered partitions in the absence of spatial regularization. On the other hand, *f-MRF* segmentation controls the problem of scattering via the pairwise energy terms, E_f and E_p.

In Fig. (4.13), we aim to present representative power of the *f-MRF*. For this purpose, based on the *unary features*, i.e., activation statistics, starting from the cluster with minimum activation score, clusters are gathered. Recall that in the *f-MRF* method, clusters are represented with mixture components where the parameters $\boldsymbol{\theta_\ell} = \{\boldsymbol{\mu_\ell}, \boldsymbol{\Sigma_\ell}\}$ of the Gaussian component is defined on the *unary features*. Likewise, activation statistics of the cluster C_ℓ can be regarded as its mean vector $\boldsymbol{\mu_\ell}$. Moreover, each element of the unary feature vector $\boldsymbol{v_i} = \{\rho_n\}_{n=1}^{d}$ of the i^{th} voxel corresponds to the voxel's coherence with the underlying experimental condition(s). Therefore, by looking maximum value of the unary feature vector, regardless of the causing condition, we can estimate an activation score for the voxel. Similarly, by

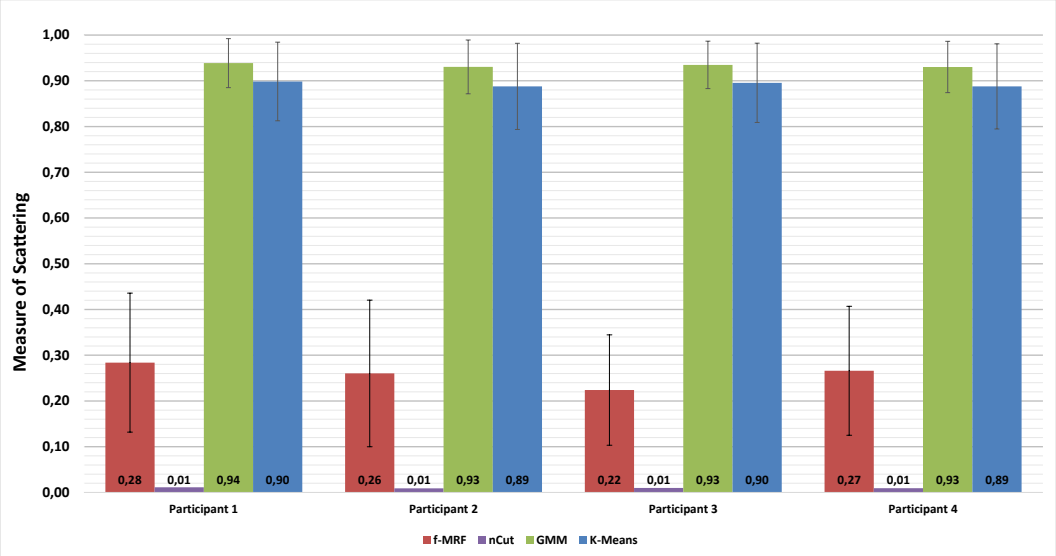


Figure 4.12: Average and standard deviation of scattering scores are calculated over all clustering results of the algorithms that are initialized with the same number of clusters, $\beta_c \in \{30, 50, 100, 200, 300, 500, 700\}$.

taking maximum value of the $\boldsymbol{\mu}_{\ell}$, activation score of the cluster C_{ℓ} can be measured. As we can see from the plot in Fig. (4.13), vast amount of the voxels (28860 out of 37944) are coming from the least-activated cluster. Moreover, the small flatness indicates that small clusters do not necessarily consist of activated voxels.

The regular or almost-regular increase in the curves of nCut, K-Means and GMM is mainly because of similarly-sized clusters compared to cluster size distribution of the *f-MRF* (see Figs. 4.8, 4.9, 4.10 and 4.11). Both K-Means and SCSC algorithms are strongly biased towards a clustering with uniformly sized clusters due to their objective functions. K-Means finds a labeling that minimizes within-cluster variance. Likewise, the SCSC aims at finding a cut so that between-cluster similarity is minimized. When a cut can not be determined uniquely, nCut method favors equal-sized clusters [15]. It can be seen from the cluster size plots that both K-Means and nCut find partitions that tend to have similar number of voxels. Although GMM is able to model various-sized clusters theoretically, it is observed that GMM fails on this dataset.

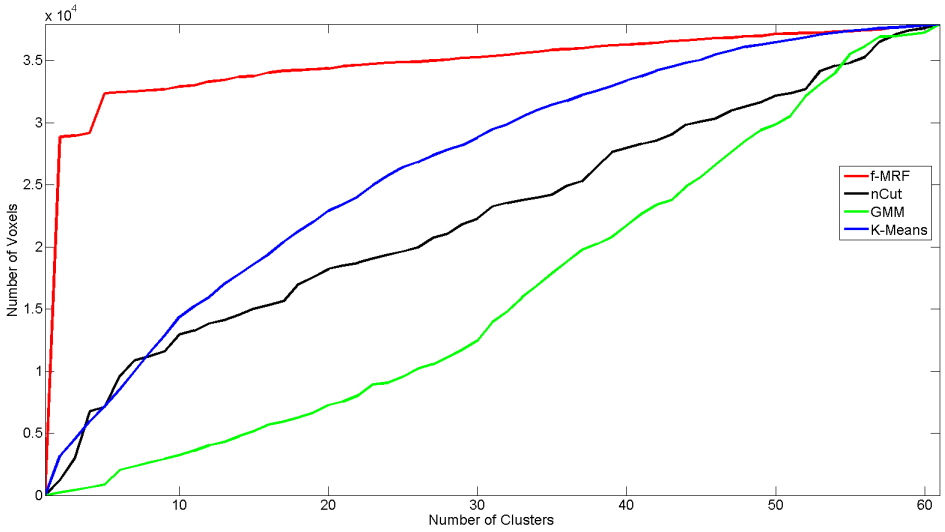
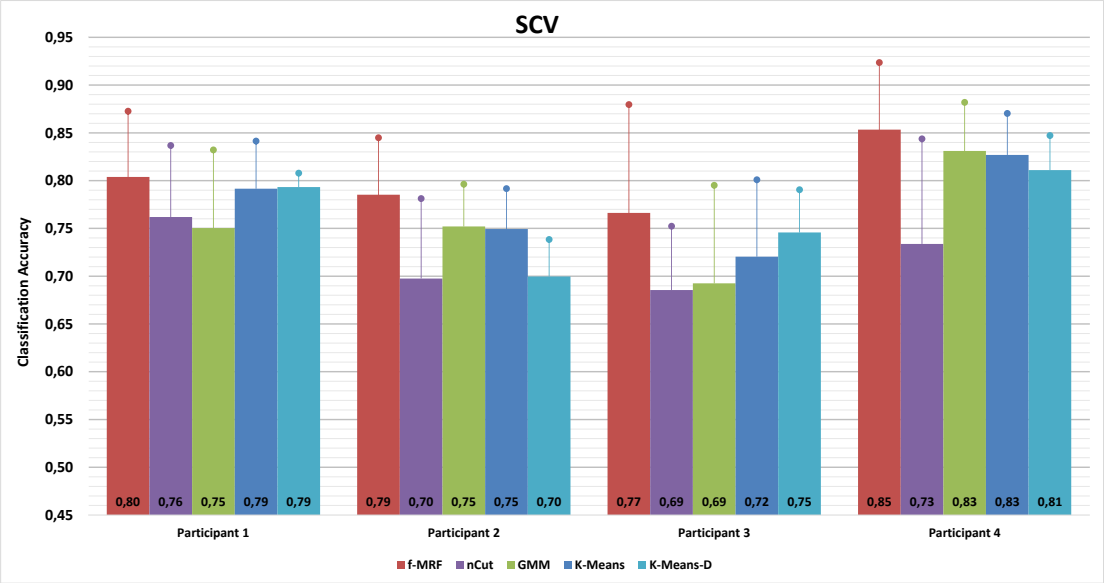


Figure 4.13: Cumulative distribution of the cluster size, where clusters are sorted in ascending order with respect to the cluster activation score. Horizontal axis is number of the clusters added, and the vertical axis corresponds to the total number of voxels in the combined set.

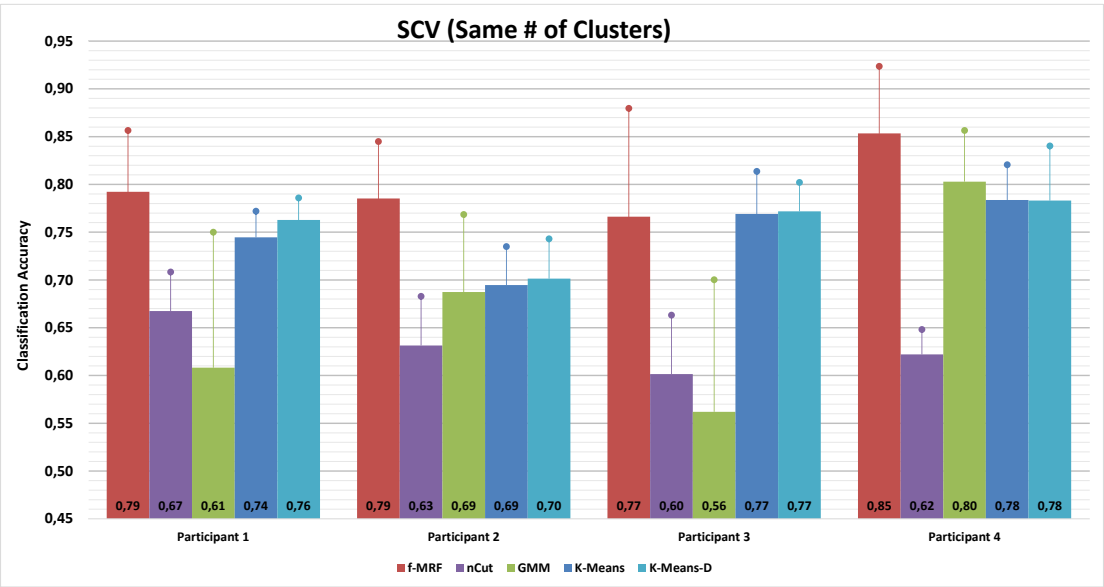
4.4.1 Classification Performance

In this section, by using the **SCV**, **SKL** and **AVG** approaches, we compute the classification accuracy for *f-MRF*, SCSC (nCut), GMM and K-Means segmentation algorithms. Moreover, in order to evaluate how *unary features* affect the performance, K-Means algorithm is applied on the training feature matrix, $\boldsymbol{F}_{tr}$, which is constructed by using all voxels in the brain (referred as K-Means-D).

As it is explained in section (4.2), representative power of the segmentation algorithms are evaluated by using the classification performance. For each participant, segmentation results of the four algorithms are obtained for all possible initializations of the parameters. For the algorithms SCSC (nCut), GMM, K-Means and K-Means-D, the number of clusters are is initialized with $\beta_c \in \{30, 50, 100, 200, 300, 500, 700\}$. For the *f-MRF*, in addition to the β_c, weight of the functional energy takes the values $\beta_f \in \{0.05, 0.3, 0.5, 1, 2, 2.5, 3, 3.5, 4, 5\}$. Therefore, we have 70 different partitioning results for the *f-MRF*, while number of the results is 7 for the compared algorithms. Under three approaches, namely **SCV**, **SKL** and **AVG**, classification accuracy for each partitioning result is computed. In Figs. 4.14a, 4.15a and 4.16a classifica-



(a)

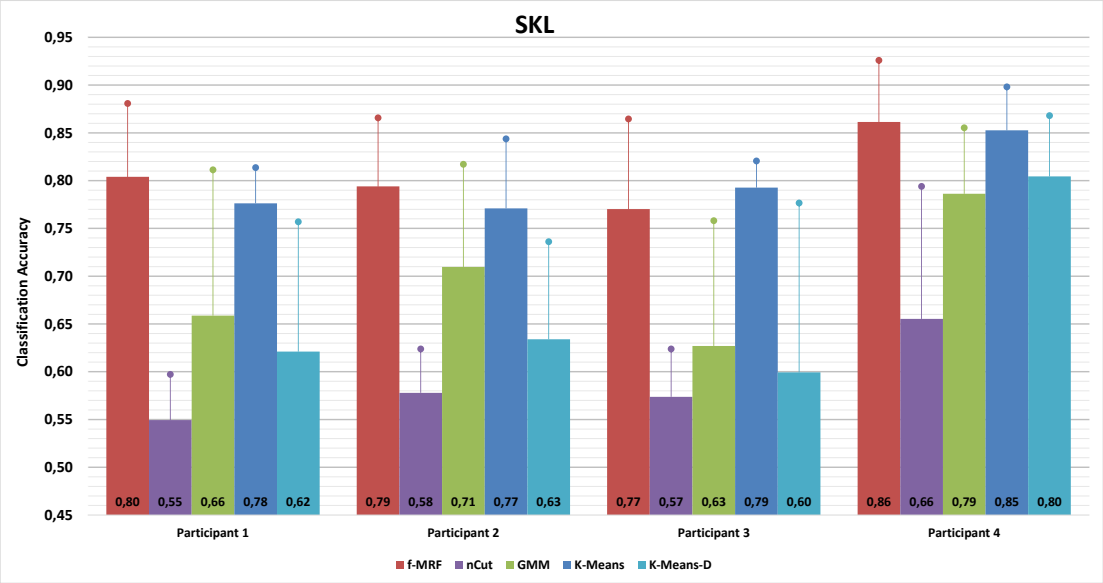


(b)

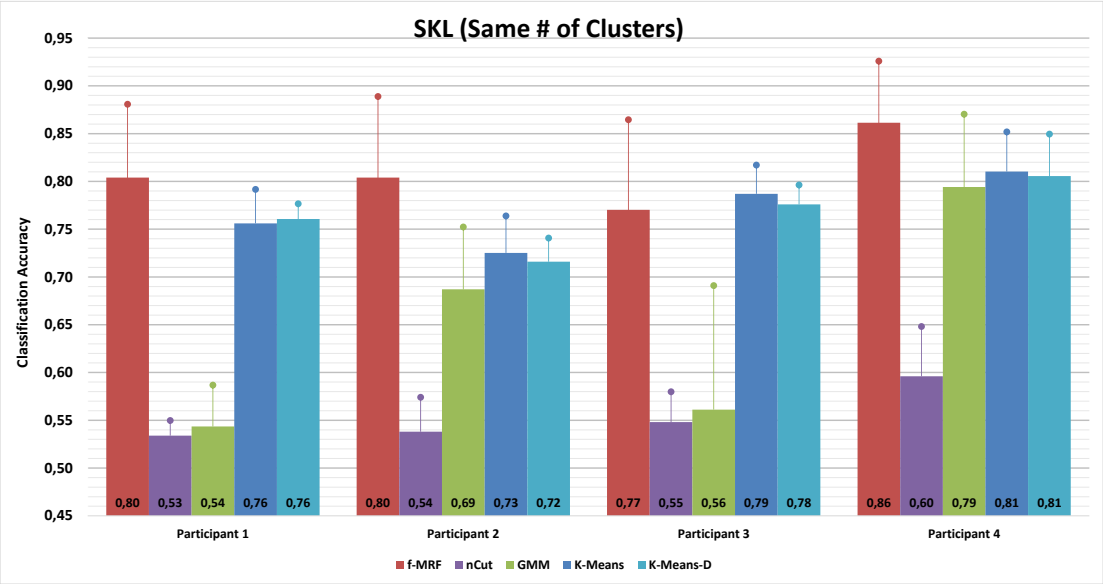
Figure 4.14: Classification accuracy for the *f-MRF*, SCSC (nCut), GMM, K-Means and K-Means-D clustering algorithms computed under **SCV** approach.

tion results are plotted. The bar plots show average accuracy of the all partitioning results for each clustering algorithm. The dots on each bar indicate the maximum performance we get with the corresponding clustering method.

As it an be seen from the plots, the classification performance of *f-MRF* outperforms



(a)

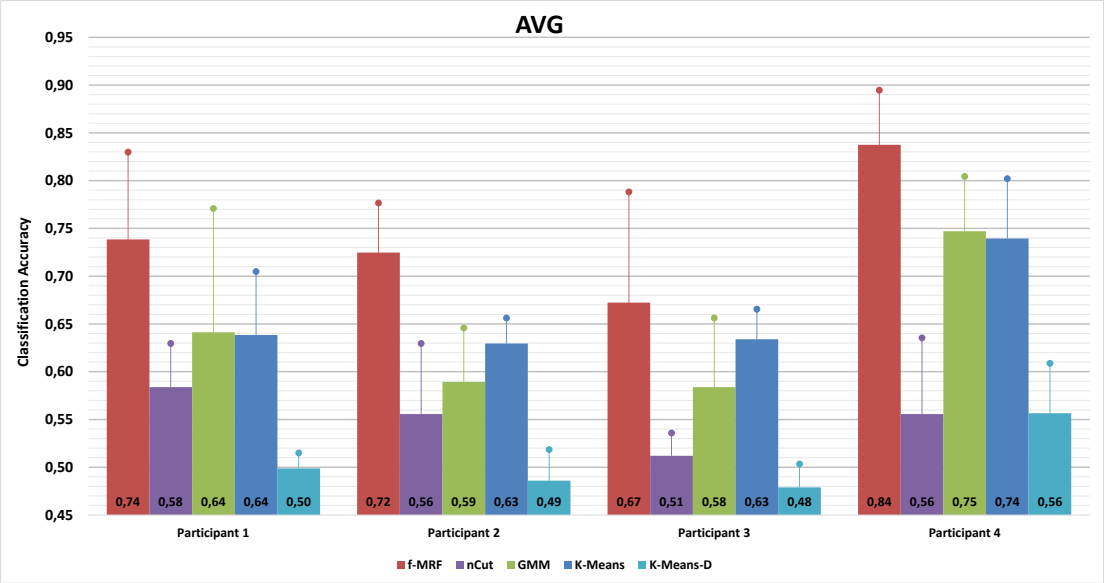


(b)

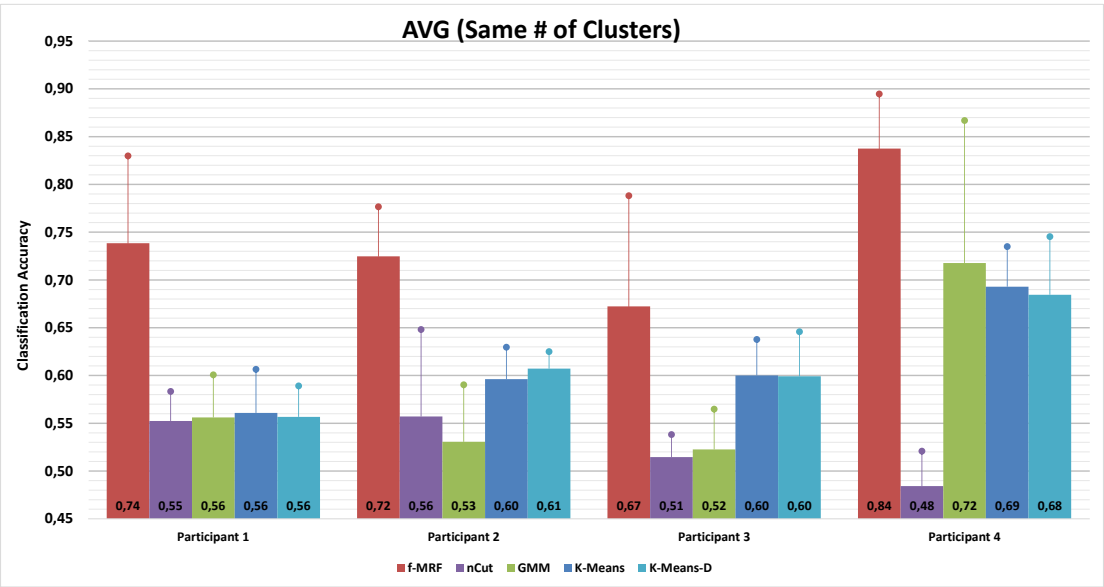
Figure 4.15: Classification accuracy for the *f-MRF*, SCSC (nCut), GMM, K-Means and K-Means-D clustering algorithms computed under **SKL** approach.

the other algorithms for all participants when the maximum performance is considered. On the average accuracies, only for participant 3, the K-Means algorithm gives competitive results.

Recall that *f-MRF* finds less number of clusters than the initialized. Hence, in order



(a)



(b)

Figure 4.16: Classification accuracy for the *f-MRF*, SCSC (nCut), GMM, K-Means and K-Means-D clustering algorithms computed under **AVG** approach.

to make a comparison under same conditions, SCSC (nCut), GMM and K-Means algorithms are initialized with same cluster numbers with *f-MRF* results (see Table 4.1 for the initial number of clusters). In Figs. 4.14b, 4.15b and 4.16b, the classification performances computed in this setting are plotted. On the data of participants 1,2

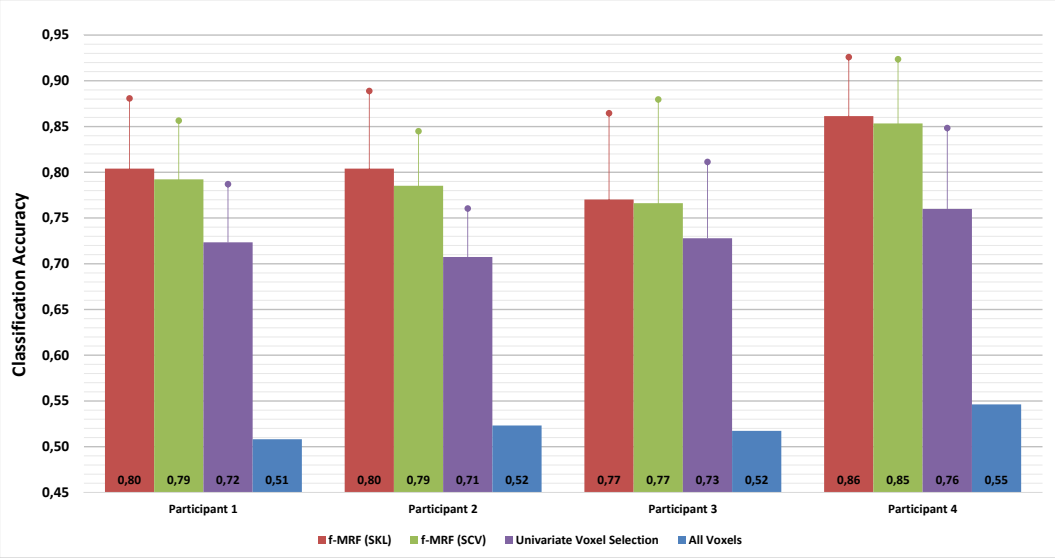


Figure 4.17: Classification accuracy of various voxel selection approaches. Accuracy is computed after selecting the voxels by using **SKL** and **SCV** on *f-MRF* clustering results, univariate voxel selection, and using all voxels.

and 4, performance of the compared algorithms decreases. However, again only for participant 3, K-Means yield competitive performance in **SCV**, **SKL** approaches.

For all participants, when the clusters are represented with the average time series of the constituent voxels, there are approximately 5% decrease in the performance of *f-MRF* (see the plots in Figs. 4.16a and 4.16b). However, when it is compared to other algorithms, accuracy of the *f-MRF* is significantly higher. This result indicates that clusters of the *f-MRF* provide better representation for the fMRI time series.

Finally, performance of the *f-MRF* method is compared with the univariate voxel selection approach. Recall that activation score of a voxel is estimated as maximum value of the *unary features*. Starting from the 100 voxels with highest activation score, classification accuracy is computed using the selected voxels until 3500 voxels ($\sim 10\%$ of all voxels) are chosen in increments of 100. In other words, we employ a univariate voxel selection approach based on the voxel activation score. In Fig. (4.17), classification accuracy results are plotted. We also compute the accuracy by using all voxels. The results clearly indicates that *f-MRF* outperforms the univariate technique we employed. Moreover, classification performance of our univariate voxel selection approach is higher than performance of the nCut and GMM algorithms in

some cases, which shows that set of *unary features* is itself informative. There are dramatic decreases on the accuracy when all of the voxels are employed. The performances drop to just above chance level. This means that the non-activated voxels excessively affect the classifier.

4.5 Summary

This chapter covers analysis of the proposed method and comparative evaluation results obtained on a real fMRI data of visual object recognition. First, we analyze the effect of different initialization of the parameters on energy function of the *f-MRF*. It is observed that our iterative solution algorithm is robust to parameter initializations. In other words, the convergence point of the energy minimization is approximately same for all parameter configurations. Moreover, initialization of the energy weights $\{\beta_d, \beta_p, \beta_f\}$ has little effect on the initial and final energy compared to the parameter β_c, initial number of clusters. On the clustering results, we see that larger values of the smoothing weights, i.e., β_p and β_f, make the model less tolerant to different label assignments for the neighboring voxels. Resulting in less number of clusters, the model converges earlier.

Second, we compare quality of the *f-MRF* clustering results with well-accepted segmentation algorithms, namely K-Means, spatially constrained spectral clustering (nCut) and Gaussian Mixture Models (GMM). In order to see, how partitioning increase representative power of the data, we evaluate performance of the segmentation algorithms under brain decoding tasks, by using classification accuracy. In the comparative results, it is observed that *f-MRF* is able to isolate non-activated voxels from activated voxels by gathering them into a large cluster. Considering the fact that vast amount of the voxels in a standard fMRI experiment are non-activated, we can say that *f-MRF* successfully capture structure of the data. Moreover, compared to the K-Means and GMM algorithms where no spatial regularization is employed, *f-MRF*, due to its pairwise energy terms, yields less scattered clusters that consist of several spatially continuous segments. In the brain decoding experiments, by using the cluster results, voxel selection by means of cluster selection (**SCV** and **SKL**), and voxel agglomeration (**AVG**) tasks are performed. In all three settings, *f-MRF* outperforms

the compared segmentation algorithms. Lastly, classification accuracy is computed using all voxels and voxels selected by a univariate selection approach. Here again, *f-MRF* gives higher classification accuracy than the compared techniques. Moreover, near chance level performance of all voxels emphasizes that prior analysis should be useful before the brain decoding tasks.

CHAPTER 5

CONCLUSION AND SUGGESTIONS TO FUTURE WORK

In this study, a segmentation method that is specially tailored for the fMRI data is presented. The proposed method, *f-MRF*, estimates latent cluster labels of the voxels by using *maximum a posterior* (MAP) estimation, where it is equivalently reformulated as energy minimization under Markov Random Fields (MRF) framework. Without making any assumptions about the size and shape of the clusters, *f-MRF* is able to fit structure of the fMRI data. More specifically, in an fMRI experiment, large amount of the voxels are noisy, redundant or non-activated, hence uninformative. *f-MRF* isolates the uninformative voxels by gathering them into a few large clusters. And, the activated voxels are collected into much smaller and functionally homogeneous clusters.

By simultaneously employing two different feature sets, namely *unary* and *pairwise features*, *f-MRF* differs from the existing clustering techniques applied on the fMRI data, where *unary* and *pairwise features* are complementary to each other under the MRF framework. For crafting these features, *f-MRF* exploits basic assumptions and well-accepted analysis techniques in the fMRI literature.

Set of *unary features* corresponds to the activation statistics of an individual voxel. As it is in the univariate approach, time courses of a voxel is statistically compared with theoretical BOLD signals that are generated under various experimental conditions. We also try to capture the delayed voxel behavior by applying time shifts on the BOLD signals. A higher similarity value between time series of a voxel and a BOLD signal indicates that voxel is highly activated under the corresponding condition. *Pairwise features*, on the other hand, is a measure of statistical similarity

between two neighboring voxels, which is also referred as functional connectivity in the literature. By incorporating the pairwise similarity into the model, we aim to ensure both functional homogeneity and spatial continuity of the clusters. *Pairwise features* basically reveal functional relations of a voxel in its neighborhood. Thus, in the context of MRFs, they can be regarded as functional textures in the fMRI data.

Total energy of the *f-MRF* can be decomposed into three components namely, unary, Potts and functional energy terms. While unary energy is modeling the data, Potts energy is responsible for spatial smoothness of the labels. And, the functional energy is a link between the observations and the labels. In other words, it regulates spatial coherence of the labels by considering the local interactions of voxels.

Unary energy is defined as negative log-likelihood of the voxels given the cluster labels. The voxel space is represented by a Gaussian Mixture Model (GMM) over $d-$dimensional *unary features*, where each component of the mixture corresponds to a cluster. The unary energy term basically determines characteristics of the clusters. More specifically, we can claim that a cluster is activated or non-activated by inspecting parameters (mean and variance) of the model. Potts energy term, on the other hand, is used for the purpose of spatial regularization. The unary energy term, itself, results in highly scattered clusters (recall test results of the compared clustering algorithms). In order to enforce spatial continuity, Potts energy term applies a penalty when neighboring voxels are assigned to different clusters. Although the functional energy term has the same structure with Potts energy, amount of the penalty is determined by *pairwise features*. Accordingly, the more functionally similar neighboring voxels, the larger it costs to set different cluster labels. Hence, functional energy term ensures not only spatial continuity but also functional homogeneity of the clusters, which also motivates the model to preserve the functional textures.

In order to find the a configuration that minimizes the total energy, we employ a two-step iterative procedure. First, given the GMM parameters, cluster labels are computed by sub-optimally minimizing the energy via α-*expansion* algorithm. In turn, model parameters are re-estimated given the labeling. Starting from a random configuration, yielding a better fit to the voxel space every iteration, the algorithm continues until there is no change on the labels.

The tests that we have conducted on the real fMRI data -a visual recognition experiment- indicate that *f-MRF* reveals non-activated voxels as a single large cluster and collects the activated voxels into much smaller clusters. Having such a partition on the voxel space, i.e., feature space, where the data is expected to be represented better, we propose using the *f-MRF* for brain decoding tasks. Accordingly, voxel selection or voxel agglomeration, i.e, feature extraction, steps can be applied on the clustering result which already provides homogeneous groups of similarly activated voxels. Hence, we evaluate the quality of partitioning of the feature space by means of classification accuracy.

In the experiments, voxel selection is performed by means of selecting the most informative clusters. For this purpose, two heuristics namely **SCV** and **SKL** are proposed. In both criteria, greedily evaluating the clusters on the training data, a feature matrix is constructed by simply concatenating voxels of the selected clusters. As a feature agglomeration practice, on the other hand, each cluster is represented with the average time series of its constituent voxels (**AVG**), which is a common practice in cluster-based analysis. In the case of **AVG**, the feature matrix is constructed by concatenation of the representative signals.

As it can be seen from test results, classification performance of *f-MRF* under all three tasks outperforms the comparative methods. Our findings on this dataset indicate that methods that are based on the univariate voxel activation (K-Means and GMM) generally represents the data better than the methods using pairwise similarity of the voxels (nCUT). However, *f-MRF*, as a hybrid of both approaches, yields much better performance. This can be explained by *f-MRF*'s reasonably preserving the spatial coherence and functional textures in the data.

5.1 Future Work

One of the major disadvantage of the *f-MRF* is that it introduces more hyper-parameters which are already hard to optimize. In this study, the weight parameters $\{\beta_d, \beta_p, \beta_f\}$ of the energy terms are not optimized. Instead, by keeping the β_d and β_p constant, a grid search is applied on a set of empirically determined values for β_f. As we can

see, the performance of the *f-MRF* under **SKL** heuristic is very promising. Hence, we plan to construct an optimization function based on **SKL** criterion in order to estimate weight parameters. If we optimize the final clustering by using the **SKL** criterion which also determines the selected clusters, we expect to boost classification performance. More specifically, the clusters will be specially tailored for the classification task.

Moreover, in the cluster selection tests (**SKL** and **SCV**), number of selected clusters is generally more than one. This fact can be used to construct a hierarchy, so that at higher levels of the hierarchy, a subset of voxels that yields maximum classification performance can be discovered.

REFERENCES

[1] Kâmil Uludag, David J Dubowitz, and Richard B Buxton. Basic principles of functional mri. *Clinical MRI. Elsevier, San Diego*, pages 249–287, 2005.

[2] Xiaoping Hu and Essa Yacoub. The story of the initial dip in fmri. *Neuroimage*, 62(2):1103–1108, 2012.

[3] Andreas Meyer-Lindenberg. From maps to mechanisms through neuroimaging of schizophrenia. *Nature*, 468(7321):194–202, 2010.

[4] R Henson and K Friston. Convolution models for fmri. *Statistical parametric mapping: The analysis of functional brain images*, pages 178–192, 2011.

[5] Simon JD Prince. *Computer vision: models, learning, and inference*. Cambridge University Press, 2012.

[6] Onal Itir, Ozay Mete, and Yarman Vural Fatos T. Modeling voxel connectivity for brain decoding. In *Pattern Recognition in Neuroimaging, 2015 International Workshop on*. IEEE, 2015.

[7] Tom M Mitchell, Rebecca Hutchinson, Radu S Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1-2):145–175, 2004.

[8] Kenneth a. Norman, Sean M. Polyn, Greg J. Detre, and James V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430, 2006.

[9] Ruth Heller, Damian Stanley, Daniel Yekutieli, Nava Rubin, and Yoav Benjamini. Cluster-based analysis of FMRI data. *NeuroImage*, 33(2):599–608, 2006.

[10] Bertrand Thirion, Gaël Varoquaux, Elvis Dohmatob, and Jean Baptiste Poline. Which fMRI clustering gives good brain parcellations? *Frontiers in Neuroscience*, 8(8 JUL):1–13, 2014.

[11] Stephen M Smith. Overview of fmri analysis. *The British Journal of Radiology*, 2014.

[12] Choong Wan Woo, Anjali Krishnan, and Tor D. Wager. Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, 91:412–419, 2014.

[13] Yukiyasu Kamitani and Yasuhito Sawahata. Spatial smoothing hurts localization but not information: pitfalls for brain mappers. *Neuroimage*, 49(3):1949–1952, 2010.

[14] T. Vincent, R. Risser, and P. Ciuciu. Spatially Adaptive Mixture Modeling for Analysis of fMRI Time Series. *Medical Imaging, IEEE Transactions on*, 29(4):1059–1074, 2010.

[15] R. Cameron Craddock, G. Andrew James, Paul E. Holtzheimer, Xiaoping P. Hu, and Helen S. Mayberg. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8):1914–1928, 2012.

[16] Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, Christine Keribin, and Bertrand Thirion. A supervised clustering approach for fMRI-based inference of brain states. *Pattern Recognition*, 45(6):2041–2049, 2012.

[17] Mark W. Woolrich, Timothy E J Behrens, and Stephen M. Smith. Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage*, 21(4):1748–1761, 2004.

[18] Srikanth Ryali, Tianwen Chen, Kaustubh Supekar, and Vinod Menon. A parcellation scheme based on von Mises-Fisher distributions and Markov random fields for segmenting brain regions using resting-state fMRI. *NeuroImage*, 65:83–96, 2013.

[19] A fully Bayesian approach to the parcel-based detection-estimation of brain activity in fMRI. *NeuroImage*, 41(3):941–969, 2008.

[20] Bernard Ng, Rafeef Abugharbieh, and Ghassan Hamarneh. Group MRF for fMRI activation detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2887–2894, 2010.

[21] Abi Berger. How does it work?: Magnetic resonance imaging. *BMJ: British Medical Journal*, 324(7328):35, 2002.

[22] Donald W McRobbie, Elizabeth A Moore, Martin J Graves, and Martin R Prince. *MRI from Picture to Proton*. Cambridge university press, 2006.

[23] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.

[24] Scott H Faro and Feroze B Mohamed. *BOLD fMRI: A guide to functional imaging for neuroscientists*. Springer Science & Business Media, 2010.

[25] Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

[26] Seiji Ogawa and Tso-Ming Lee. Magnetic resonance imaging of blood vessels at high fields: in vivo and in vitro measurements and image simulation. *Magnetic Resonance in Medicine*, 16(1):9–18, 1990.

[27] Seiji Ogawa, Tso-Ming Lee, Asha S Nayak, and Paul Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic resonance in medicine*, 14(1):68–78, 1990.

[28] Seiji Ogawa, David W Tank, Ravi Menon, Jutta M Ellermann, Seong G Kim, Helmut Merkle, and Kamil Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

[29] Charles Smart Roy and Charles S Sherrington. On the regulation of the blood-supply of the brain. *The Journal of physiology*, 11(1-2):85–158, 1890.

[30] Linus Pauling and Charles D Coryell. The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, 22(4):210, 1936.

[31] Richard B Buxton, Kâmil Uludağ, David J Dubowitz, and Thomas T Liu. Modeling the hemodynamic response to brain activation. *Neuroimage*, 23:S220–S233, 2004.

[32] Melissa Kristin Carroll. *FMRI" mind Readers": Sparsity, Spatial Structure, and Reliability*. Citeseer, 2011.

[33] Marc G Berman, John Jonides, and Derek Evan Nee. Studying mind and brain with fmri. *Social cognitive and affective neuroscience*, 1(2):158–161, 2006.

[34] S.A. Bunge and I. Kahn. Cognition: An overview of neuroimaging techniques. In Larry R. Squire, editor, *Encyclopedia of Neuroscience*, pages 1063 – 1067. Oxford, 2009.

[35] Reidar P Lystad and Henry Pollard. Functional neuroimaging: a brief overview and feasibility for use in chiropractic research. *The Journal of the Canadian Chiropractic Association*, 53(1):59, 2009.

[36] Karl J Friston, Andrew P Holmes, Keith J Worsley, JP Poline, Chris D Frith, Richard SJ Frackowiak, et al. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

[37] Solveig Badillo, Thomas Vincent, and Philippe Ciuciu. Group-level impacts of within- and between-subject hemodynamic variability in fMRI. *NeuroImage*, 82:433–448, 2013.

[38] Barry Horwitz. The elusive concept of brain connectivity. *Neuroimage*, 19(2):466–470, 2003.

[39] Karl J Friston, Peter Jezzard, and Robert Turner. Analysis of functional mri time-series. *Human brain mapping*, 1(2):153–171, 1994.

[40] Sepideh Sadaghiani, Guido Hesselmann, Karl J Friston, and Andreas Kleinschmidt. The relation of ongoing brain activity, evoked neural responses, and cognition. *Frontiers in systems neuroscience*, 4, 2010.

[41] Jesse Rissman, Adam Gazzaley, and Mark D'Esposito. Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, 23(2):752–763, 2004.

[42] KJ Friston, CD Frith, PF Liddle, and RSJ Frackowiak. Functional connectivity: the principal-component analysis of large (pet) data sets. *Journal of cerebral blood flow and metabolism*, 13:5–5, 1993.

[43] Guillaume Flandin, Ferath Kherif, Xavier Pennec, Grégoire Malandain, Nicholas Ayache, and Jean-Baptiste Poline. Improved detection sensitivity in functional mri data using a brain parcelling technique. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002*, pages 467–474. Springer, 2002.

[44] Danial Lashkari, Ed Vul, Nancy Kanwisher, and Polina Golland. Discovering structure in the space of fMRI selectivity profiles. *NeuroImage*, 50(3):1085–1098, 2010.

[45] Xilin Shen, Xenophon Papademetris, and R Todd Constable. Graph-theory based parcellation of functional subunits in the brain from resting-state fmri data. *Neuroimage*, 50(3):1027–1035, 2010.

[46] Thomas Blumensath, Saad Jbabdi, Matthew F. Glasser, David C. Van Essen, Kamil Ugurbil, Timothy E J Behrens, and Stephen M. Smith. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. *NeuroImage*, 76:313–324, 2013.

[47] Archana Venkataraman, Koene RA Van Dijk, Randy L Buckner, and Polina Golland. Exploring functional connectivity in fmri via clustering. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 441–444. IEEE, 2009.

[48] Dietmar Cordes, Vic Haughton, John D Carew, Konstantinos Arfanakis, and Ken Maravilla. Hierarchical clustering to measure connectivity in fmri resting-state data. *Magnetic resonance imaging*, 20(4):305–317, 2002.

[49] Marotesa Voultsidou, Silke Dodel, and J. Michael Herrmann. Neural networks approach to clustering of activity in fMRI data. *IEEE Transactions on Medical Imaging*, 24(8):987–996, 2005.

[50] Gang Chen, B Douglas Ward, Chunming Xie, Wenjun Li, Guangyu Chen, Joseph S Goveas, Piero G Antuono, and Shi-Jiang Li. A clustering-based method to detect functional connectivity differences. *NeuroImage*, 61(1):56–61, 2012.

[51] Polina Golland, Yulia Golland, and Rafael Malach. Detection of spatial activation patterns as unsupervised segmentation of fmri data. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*, pages 110–118. Springer, 2007.

[52] Richard Baumgartner, Gordon Scarth, Claudia Teichtmeister, Ray Somorjai, and Ewald Moser. Fuzzy clustering of gradient-echo functional mri in the human visual cortex. part i: Reproducibility. *Journal of Magnetic Resonance Imaging*, 7(6):1094–1101, 1997.

[53] Ewald Moser, Markus Diemling, and Richard Baumgartner. Fuzzy clustering of gradient-echo functional mri in the human visual cortex. part ii: Quantification. *Journal of Magnetic Resonance Imaging*, 7(6):1102–1108, 1997.

[54] Martin J McKeown, Scott Makeig, Greg G Brown, Tzyy-Ping Jung, Sandra S Kindermann, Anthony J Bell, and Terrence J Sejnowski. Analysis of fmri data by blind separation into independent spatial components. Technical report, DTIC Document, 1997.

[55] VD Calhoun, T Adali, VB McGinty, JJ Pekar, TD Watson, and GD Pearlson. fmri activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis. *NeuroImage*, 14(5):1080–1088, 2001.

[56] VD Calhoun, T Adali, GD Pearlson, and JJ Pekar. A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

[57] Cyril Goutte, Peter Toft, Egill Rostrup, Finn Å Nielsen, and Lars Kai Hansen. On clustering fmri time series. *NeuroImage*, 9(3):298–310, 1999.

[58] Cyril Goutte, Lars Kai Hansen, Matthew G Liptrot, and Egill Rostrup. Feature-space clustering for fmri meta-analysis. *Human brain mapping*, 13(3):165–183, 2001.

[59] Bertrand Thirion and Olivier Faugeras. Feature detection in fmri data: the information bottleneck approach. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003*, pages 83–91. Springer, 2003.

[60] Yulia Golland, Polina Golland, Shlomo Bentin, and Rafael Malach. Data-driven clustering reveals a fundamental subdivision of the human cortex into two global systems. *Neuropsychologia*, 46(2):540–553, 2008.

[61] Benjamin Thyreau, Bertrand Thirion, Guillaume Flandin, and Jean-Baptiste Poline. Anatomo-functional description of the brain: a probabilistic approach. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE, 2006.

[62] Danial Lashkari, Ramesh Sridharan, Edward Vul, Po-Jang Hsieh, Nancy Kanwisher, and Polina Golland. Search for patterns of functional specificity in the brain: a nonparametric hierarchical bayesian model for group fmri data. *Neuroimage*, 59(2):1348–1368, 2012.

[63] Lili He and Ian R Greenshields. An mrf spatial fuzzy clustering method for fmri spms. *Biomedical Signal Processing and Control*, 3(4):327–333, 2008.

[64] Martijn Van Den Heuvel, Rene Mandl, and Hilleke Hulshoff Pol. Normalized cut group clustering of resting-state fmri data. *PloS one*, 3(4):e2001, 2008.

[65] Alexander L Cohen, Damien A Fair, Nico UF Dosenbach, Francis M Miezin, Donna Dierker, David C Van Essen, Bradley L Schlaggar, and Steven E Petersen. Defining functional areas in individual human brains using resting functional connectivity mri. *Neuroimage*, 41(1):45–57, 2008.

[66] Vincent Michel, Alexandre Gramfort, Gael Varoquaux, and Bertrand Thirion. Total variation regularization enhances regression-based brain activity prediction. In *Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), 2010 First Workshop on*, pages 9–12. IEEE, 2010.

[67] X Descombes, F Kruggel, and D Y von Cramon. Spatio-temporal fMRI analysis using Markov random fields. *IEEE transactions on medical imaging*, 17(6):1028–1039, 1998.

[68] Wei Liu, Suyash P Awate, and P Thomas Fletcher. Group analysis of resting-state fMRI by hierarchical Markov random fields. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 15(Pt 3):189–96, 2012.

[69] Mark W Woolrich, Timothy EJ Behrens, Christian F Beckmann, and Stephen M Smith. Mixture models with adaptive spatial regularization for segmentation with an application to fmri data. *Medical Imaging, IEEE Transactions on*, 24(1):1–11, 2005.

[70] Andrew Delong, Anton Osokin, Hossam N Isack, and Yuri Boykov. Fast approximate energy minimization with label costs. *International journal of computer vision*, 96(1):1–27, 2012.

[71] Lili He and Ian R. Greenshields. An MRF spatial fuzzy clustering method for fMRI SPMs. *Biomedical Signal Processing and Control*, 3(4):327–333, 2008.

[72] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[73] Stan Z Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[74] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

[75] Nikos Komodakis and Nikos Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pages 2985–2992, 2009.

[76] Pushmeet Kohli and Carsten Rother. Higher-Order Models in Computer Vision. *Image Processing and Analysis with Graphs*, pages 1–28, 2012.

[77] Alexander Fix, Aritanan Gruber, Endre Boros, and Ramin Zabih. A graph cut algorithm for higher-order markov random fields. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1020–1027. IEEE, 2011.

[78] Pushmeet Kohli, M Pawan Kumar, and Philip HS Torr. P^3 & beyond: Move making algorithms for solving higher order functions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(9):1645–1656, 2009.

[79] Oliver Woodford, Philip Torr, Ian Reid, and Andrew Fitzgibbon. Global stereo reconstruction under second-order smoothness priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2115–2128, 2009.

[80] Pushmeet Kohli and M Pawan Kumar. Energy minimization for linear envelope mrfs. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1863–1870. IEEE, 2010.

[81] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 259–302, 1986.

[82] DM Greig, BT Porteous, and Allan H Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 271–279, 1989.

[83] Yuri Boykov, Olga Veksler, and Ramin Zabih. Efficient Approximate Energy Minimization via Graph Cuts. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 20(12):1222–1239, 2001.

[84] Vladimir Kolmogorov and Ramin Zabih. What Energy Functions Can Be Minimized via Graph Cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

[85] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[86] Yuanqing Li, Praneeth Namburi, Zhuliang Yu, Cuntai Guan, Jianfeng Feng, and Zhenghui Gu. Voxel selection in fmri data analysis based on sparse representation. *Biomedical Engineering, IEEE Transactions on*, 56(10):2439–2451, 2009.

[87] Martin A. Lindquist. The Statistical Analysis of fMRI Data, 2008.

[88] Abdelhak Mahmoudi, Sylvain Takerkart, Fakhita Regragui, Driss Boussaoud, and Andrea Brovelli. Multivoxel pattern analysis for fMRI data: A review. *Computational and Mathematical Methods in Medicine*, 2012, 2012.

[89] Daniel a. Handwerker, John M. Ollinger, and Mark D'Esposito. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21(4):1639–1651, 2004.

[90] Peter A Bandettini. Functional mri limitations and aspirations. In *Neural Correlates of Thinking*, pages 15–38. Springer, 2009.

[91] Geoffrey M Boynton, Stephen A Engel, Gary H Glover, and David J Heeger. Linear systems analysis of functional magnetic resonance imaging in human v1. *The journal of neuroscience*, 16(13):4207–4221, 1996.

[92] Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. Modeling the hemodynamic response function in fmri: efficiency, bias and mismodeling. *Neuroimage*, 45(1):S187–S198, 2009.

[93] Karl J Friston, JOHN Ashburner, J Heather, et al. Statistical parametric mapping. *Neuroscience Databases: A Practical Guide*, page 237, 2003.

[94] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images: the analysis of functional brain images*. Academic press, 2011.

[95] Michelle Hampson, Bradley S Peterson, Pawel Skudlarski, James C Gatenby, and John C Gore. Detection of functional connectivity using temporal correlations in mr images. *Human brain mapping*, 15(4):247–262, 2002.

[96] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–3868, 2006.

[97] Orhan Fırat, Mete Özay, Itır Önal, Ilke Öztekin, and Fatoş T Yarman Vural. Enhancing local linear models using functional connectivity for brain state decoding. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(3):46–57, 2013.

[98] Itir Onal, Emre Aksan, Burak Velioglu, Orhan Firat, Mete Ozay, Ilke Oztekin, and FT Yarman Vural. Modeling the brain connectivity for pattern analysis. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 3339–3344. IEEE, 2014.

[99] Martin M Monti. Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Frontiers in human neuroscience*, 5:28, 2011.

[100] Francis M Miezin, L Maccotta, JM Ollinger, SE Petersen, and RL Buckner. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage*, 11(6):735–759, 2000.

[101] Negar Bazargani and Aria Nosratinia. Joint maximum likelihood estimation of activation and hemodynamic response function for fmri. *Medical image analysis*, 18(5):711–724, 2014.

[102] R. Zabih and V. Kolmogorov. Spatially Coherent Clustering Uusing Graph Cuts. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:437–444, 2004.

[103] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86, 1951.

[104] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1 Suppl):S199–S209, 2009.