POPULATION BASED SPATIO-TEMPORAL PROBABILISTIC MODELLING OF FMRI DATA

by

NAHED ALOWADI

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
The University of Birmingham
May 2018

To my parents, my son and my daughter

ABSTRACT

High-dimensional functional magnetic resonance imaging (fMRI) data is characterized by complex spatial and temporal patterns related to neural activation. Mixture based Bayesian spatio-temporal modelling is able to extract spatiotemporal components representing distinct haemodyamic response and activation patterns.

A recent development of such approach to fMRI data analysis is so-called spatially regularized mixture model of hidden process models (SMM-HPM). SMM-HPM can be used to reduce the four-dimensional fMRI data of a pre-determined region of interest (ROI) to a small number of spatio-temporal prototypes, sufficiently representing the spatio-temporal features of the underlying neural activation. Summary statistics derived from these features can be interpreted as quantification of (1) the spatial extent of sub-ROI activation patterns, (2) how fast the brain respond to external stimuli; and (3) the heterogeneity in single ROIs.

This thesis aims to extend the single-subject SMM-HPM to a multi-subject SMM-HPM so that such features can be extracted at group-level, which would enable more robust conclusion to be drawn.

To pave the way for such extension of SMM-HPM, we proposed a normalized form of the haemodynamics response function (HRF), so as to de-couple the haemodynamics response magnitude from the HRF shape. Numerical experiments have been conducted to demonstrate the benefit of this normalization.

To extend the single-subject SMM-HPM, we formulate a hierarchy of multi-subject SMM-HPM models, ranging from the most constrained model to the most flexible one, so as to find the optimal common model for extracting informative features that can be

used in comparing different populations.

The multi-subject SMM-HPM has been verified through extensive numerical experiments using both synthetic and real fMRI data. The results of the synthetic experiments show how a robust and accurate multi-subject model can be learned from the data by the optimization method we have developed. The results of the experiments with real data show how the multi-subject SMM-HPM is able to extract spatio-temporal patterns within individual ROIs from different populations, which enables us to discriminate them.

ACKNOWLEDGEMENTS

I would like to express my heartfelt appreciation and gratitude to my supervisor Prof. Peter Tino and my second supervisor Dr. Yuan Shen for their careful supervision and thoughtful guidance. The experience that I gained from them will definitely influence the rest of my life.

Special thanks to Prof. Zoe Kourtzi who allowed us to benefit from her knowledge in the area of cognitive neuroscience. I am also grateful to Dr. Rui Wang for providing us with the real fMRI data.

I must also acknowledge my thesis group members, Prof. John Barnden, Prof. Jeremy Wyatt and Dr. Eike Ritter, for their valuable advice.

All thanks and love to my family (my parents, my precious son Rayan, my beautiful daughter Rahaf, my husband, my brothers, and my sister) without their unconditional and continuous love, care, encouragement and understanding ; it would be impossible for me to complete this thesis.

Many thanks go to my sponsor King Abdulaziz University, and to the University of Birmingham.

This thesis was partly copy-edited for the conventions of language, grammar and spelling by Janet's Proofreading Service.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AFNI Analysis of Functional NeuroImages. 31

AR Auto-Regression model. 36

BOLD Blood Oxygenation Level Dependent signal. 17

CG Cingulate Gyrus. 99

CRP Chinese Restaurant Process. 48

DCM Dynamic Causal Modeling. 48

DP Dirichlet Process. 37

DPP Dirichlet Process Prior. 46

EEG Electro Encephalography. 1

EM Expectation-Maximization. 41

fMRI Functional Magnetic Resonance Imaging. 1

FSL FMRIB Software Library. 31

FWHM Full Width at Half Maximum. 23

GLM General Linear Model. 4, 5

GMRF Gaussian Markov Random Field. 35

HDPP Hierarchical Dirichlet Process Prior. 46

HRF Haemodynamic Response Function. 17

IOG Inferior Occipital Gyrus. 99

ISI Inter-Stimulus Interval. 60

LDC Linear Discriminant Classifier. 30

LiG Limbic Gyrus. 99

MAP Maximum A Posteriori. 41

MCMC Markov Chain Monte Carlo. 36

MEG Magneto Encephalography. 1

MFG Middle Frontal Gyrus. 99

ML Maximum Likelihood. 41

MNI Montreal Neurological Institute. 23

MoAR Mixture of Auto-Regressions. 45

MOG Medial Occipital Gyrus. 99

MPL Maximum Pseudo-Likelihood. 45

MRF Markov Random Field. 36

MRI Magnetic Resonance Imaging. 1, 15

MVPA Multi-Variate Pattern Analysis. 27

n-HRF normalized Haemodynamic Response Function. 64

OI Optical Imaging. 1

PET Positron Emission Tomography. 1

PSC Percent signal change. 98

Pu Putamen. 99

SFG Superior Frontal Gyrus. 99

SGLMM Spatial Generalized Linear Mixed Model. 39

SMM Spatial Mixture Model. 41

SMM-HPM Spatially regularized Mixture Model of Hidden Process Models. 4

SNR Signal-to-Noise Ratio. 16

SPM Statistical Parametric Mapping. 31

SSBFs Sparse Spatial Basis Functions. 36

u-HRF unnormalized Haemodynamic Response Function. 64

VB Variational Bayes. 36

LIST OF NOTATIONS

A ICA mixing matrix. 27

A^{-1} inverse of ICA mixing matrix. 28

C hidden spatial or temporal components. 27

C^k consensus matrix of prototype k. 77

D clustering distance. 66

L number of iteration. 77

$M^{(l)}$ connectivity matrix. 77

N free normalization parameter. 54

P number of cognitive processes. 55

S number of stimuli. 55

T time points (volumes). 18

TR repetition time. 18

U number of subjects. 76

V number of voxels. 53

W peak width. 67

X design matrix. 26

Y observed data. 26

a response magnitude. 55

b level shift in the fMRI signal in null prototype. 56

d^* voxel space dimensionality. 84

df degree of freedom. 84

g HRF shape function (gamma function). 55

h haemodynamic response. 55

r_v the location of the voxel v. 54

A_S spatial prior estimation accuracy. 61

$A_{\tilde{g}}$ shape function estimation accuracy. 62

A_a response magnitude estimation accuracy. 62

T^p time-to-peak. 67

Ψ Inverse Wishart scale matrix. 84

Σ covariance matrix (prototype shape). 54

Θ^{HRF} haemodynamic response function parameters. 71

Θ^{NIS} Noise parameters. 71

Θ^{NRL} response magnitude parameters. 71

Θ^{STM} spatio-temporal parameters. 53

Θ^{S} spatial parameters. 54

Θ^{T} temporal parameters. 54

β regression coefficients. 26

$\mathcal{L}$ negative log model likelihood. 57

ϵ noise component. 55

η fixed relative difference. 61

κ shape parameter of gamma function. 56

λ tuning parameter. 66

μ^{κ} shape of gamma function mean. 80

μ^{θ} scale of gamma function mean. 80

μ^{s} prototype position mean. 86

μ mean vector (prototype position). 54

ν a fraction of number of samples. 89

π mixing weights. 40

ρ tuning parameter. 89

σ^2 variance (prototype noise). 55

$\sigma^{2^{\kappa}}$ shape of gamma function variance. 80

$\sigma^{2^{\theta}}$ scale of gamma function variance. 80

$\sigma^{2^{s}}$ prototype position variance. 86

θ^{init} parameter initialization value. 61

θ^{tr} parameter ground-truth value. 61

θ scale parameter of gamma function. 56

x signal component. 55

CHAPTER 1

INTRODUCTION

With the growing interest in studying human brains, several techniques have been developed to enable researchers to study brain activities. They include Positron Emission Tomography (PET), Electro Encephalography (EEG), Magneto Encephalography (MEG), Optical Imaging (OI) and Functional Magnetic Resonance Imaging (fMRI). Each of these techniques has its own importance and application area. Due to its high spatial resolution, fMRI is particularly popular.

fMRI measures the metabolic changes (the increase of the oxygenated blood volume and flow) that are a consequence of the neural activities in the brain using a scanner with strong magnetic fields (Magnetic Resonance Imaging (MRI) scanner). Over the past two decades, fMRI has been the main tool to investigate human brains non-invasively. It mainly aims to localize activation regions and determine brain connectivity in response to specific external stimuli. Due to the high dimensionality as well as the complex spatial and temporal correlation of the fMRI data, advanced data modelling techniques need to be applied in order to infer the relationship between the external stimuli and the neuronal response (activation).

In this work we propose a method for fusing information obtained by behavioural modelling (fast and slow learners) with probabilistic modelling of fMRI data gathered at different stages of training the subjects. Traditionally, whole brain analysis of fMRI signals is used. However, there may be subtle differences between cortical activation patterns in fast and slow learners at the level of individual ROIs. Whole brain analysis is not appropriate for this setting. We develop a hierarchy of population models based on the previous single subject model - a spatially regularized mixture model of hidden process models (SMM-HPM) [1]. In this way we can answer targeted questions regarding differences in cortical activation structures in the two populations (slow and fast learners) in a model based way. SMM-HPM reduces the high-dimensional fMRI data of a pre-determined region of interest (ROI) to a small number of spatio-temporal prototypes. This prototype-based modelling method enables us to extract three novel cortical activation signatures (features) from each prototype. The first feature characterizes the spatial pattern of neural activation within single-ROI (spatial feature); the second one characterizes the haemodynamics response shape (temporal feature); and the third one characterizes the heterogeneity in single-ROI (spatio-temporal feature). We study whether there are significant differences in the three features between the populations of fast and slow learners. This may provide a basis for further more focused study of the neural correlates of learning in cognitive science and in brain disorders.

In the literature, spatial structure in fMRI data has not been explicitly modelled by those earlier fMRI data modelling methods. Instead, the structure was indirectly incorporated through smoothing the fMRI data over neighbouring voxels [2, 3, 4, 5, 6, 7, 8, 9]. As a result, the spatial correlation in fMRI data is treated in a separate, preprocessing phase. This is disadvantageous because the whole image is smoothed equally while in reality the spatial correlation varies across different activation regions. To deal with this issue, the spatial behaviour of the fMRI data should be considered as a part of an encompassing model that accounts for both spatial and temporal correlations in the fMRI data. The Bayesian framework is the optimal approach to naturally describe and model both the spatial and temporal behaviours of fMRI data because any neuroscientific knowledge about spatial and temporal correlation in the fMRI data can be formulated as prior probability distributions.

Depending on whether they adopt an implicit or explicit approach to modelling spatial coherence of the neural activation, most Bayesian spatio-temporal fMRI models can be categorized into two groups. In those models based on the implicit modelling approach, smoothness constraints are imposed on all temporal parameters that were inferred from fMRI time series on individual voxels. This can ensure that each of these parameter varies smoothly across the voxels but note that such parameters are estimated for every voxel. The smoothness constraints could be formulated as a Markov random field model [5], Gaussian kernels [2], spatial wavelet shrinkage [3, 4], anisotropic averaging spatial filtering [6, 7], adaptive spatial filtering (spatial basis filters) [8], or surface-based filtering (spatially informed basis functions) [10, 11]. We refer to such models as spatially regularized Bayesian spatio-temporal models. In those methods based on the explicit modelling approach, prior knowledge about the spatial coherence needs to be incorporated explicitly by modelling the spatial pattern of neuron activation via a parametric model. An example of such model is Gaussian mixture model. The means and covariance matrices of these Gaussian distributions represent the location and spread of the neural activations. We refer to such models as mixture-based Bayesian spatio-temporal models. It is worth noting that for the second approach, fMRI data is actually modelled at the cluster level rather than at the voxel level. In this setting, all clusters have distinct temporal patterns and spatial extent while fMRI signal at individual voxels is modelled as a mixture of several temporal patterns corresponding to those clusters. The activation is determined by assigning the voxels to the most likely components. It also enables inference of the shape and the location of the activation response.

These two Bayesian modelling approaches have been widely adopted for single-subject analysis: [12, 13, 14, 15, 16, 2, 17, 18, 19, 20, 21, 22, 21, 23, 24] have applied spatially regularized Bayesian spatio-temporal modelling, whears [25, 26, 27, 28, 29, 30, 31, 32, 33, 34] have applied mixture based Bayesian spatio-temporal modelling.

Modelling the fMRI data by the mixture model approach is more efficient than the spatially regularized Bayesian approach because it is no longer necessary to estimate temporal parameters for all voxels. Also, it explicitly models the activation shape and location providing a more interpretable model in which each component corresponds to an underlying neuron activation source. A recent development of the mixture model approach is so-called Spatially regularized Mixture Model of Hidden Process Models (SMM-HPM)[1]. This model is used to identify the spatio-temporal patterns within single ROIs. It advances the previous spatio-temporal mixture models in the following aspects:

- SMM-HPM adopts a hidden process model (HPM) as a localized temporal prototype. HPM assumes that there is a series of overlapping hidden cognitive processes that probabilistically generate the fMRI time series, which enables the inference of the contribution of each individual cognitive process (e.g., visual analysis process, perceptual judgement process, and motor response process) to the observed fMRI time series. In the literature, General Linear Model (GLM) is the conventional model for the temporal aspects of the fMRI data in which single cognitive process describes the haemodynamic response.

- SMM-HPM employs a parametric form of the HPM, which enables imposition of biological constraints on the HRF and therefore the shape of the HRF can be vary according to the cognitive process.

- SMM-HPM can detect the neuronal activation naturally in one step. Previous studies use statistical maps which treat the temporal and the spatial aspects separately, and result in splitting the analysis into two steps.

- SMM-HPM can infer the response magnitude and the response shape from the data.

- SMM-HPM utilizes a small number of free parameters since it is a prototype based model[1].

- SMM-HPM examined the heterogeneity within a specific ROI by using HPM as a localized temporal prototype and allowing more than one prototype (component) to be estimated in each ROI.

[1] In a broad sense, our prototypes can been seen as a dictionary elements, however, in this case, our dictionary elements are model-based live in space of voxels not in space of measurements

Multi-subject data analysis is a natural extension of single-subject analysis. This allows for a principled and integrated test on statistical significance for any neuroscientific finding derived from fMRI data analysis. Recently, there has been a clear trend showing that Bayesian modelling approaches have been increasingly adopted to multi-subject fMRI data modelling. Several approaches based on spatial regularization have been adopted to model multi-subject fMRI data [35, 36, 18, 37, 38, 39]. Similarly, mixture based Bayesian spatio-temporal modelling has also been developed for modelling of multi-subject fMRI [40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52].

For group-level modelling of fMRI data, employing a mixture model approach is sufficient. It helps in modelling the fMRI data at a higher level of features, such as activation location and intensity, which provides more subtle information about the neuronal activities and their variations within and between subjects, and makes it less sensitive to the misregistration problem in modelling group data.

In this thesis, the main goal is to extend the single-subject SMM-HPM to a multi-subject SMM-HPM, so that we can extract group-level features of those spatio-temporal prototypes that can be inferred from the fMRI data using the SMM-HPM model. The proposed multi-subject model resembles the Gaussian mixture model of [44] in that the activation pattern is modelled by a mixture of Gaussian distributions over the voxel locations. However, it has many unique features compared to [44] and many other previous studies:

- The temporal aspects of the fMRI data have been modelled by HPM instead of the commonly used model GLM, which helps in considering the underlying cognitive processes.

- Entire fMRI time series has been modelled while majority of the group-level analysis methods model statistical maps. This makes our model more realistic because modelling the entire fMRI time series consider the evolution in the response magnitudes over time. The drawback of modelling entire fMRI time series is that it is time consuming (the computational time is large).

- The number of components has been determined automatically from the data based on consensus clustering method, which is computationally more efficient compared to the commonly used approach based on a Dirichlet process prior (DPP).

- This model estimates not only the activation intensity, location, and shape; but also the shape of the haemodynamic response and the time series of response magnitudes.

SMM-HPM has three distinct sets of model parameters: spatial parameters, HRF shape parameters and response magnitude parameters. We thus ask what is the optimal multi-subject SMM-HPM that is suitable for all three sets of the model parameters. To investigate this research question, we formulated the multi-subject SMM-HPM as a hierarchy of model formations, from the most constrained model, where the parameters are fixed across subjects except for the haemodynamic response magnitudes, to the most flexible one, where the parameters are to be inferred for individual subjects from their corresponding data sets while those individual parameters are controlled by the group-level priors that are inferred from the data set pooled together. We can determine the optimal common model by computing the out-of-sample negative log likelihood of each model in the hierarchy. The optimal common model is the one that has the lowest negative log likelihood. From the optimal common model, we can extract informative features (spatial feature, temporal feature, and spatio-temporal feature) that can be used in comparing the fMRI data of different populations.

1.1 Motivation

In spite of the fact that considerable effort has been devoted to the problem of modelling fMRI data, it is still attracting the interest of researchers for new development, which could make further contributions to fMRI data analysis. The major contribution of [1] is to reduce a four-dimensional [1] fMRI data set to a small number of spatio-temporal prototypes. Each prototype consists of three sets of fMRI features. The first set characterizes the spatial pattern of neural activation within single ROIs (spatial feature); the second one characterizes the haemodynamics response shape (temporal feature); and the third set characterizes the cross-correlation between the time series of haemodynamics response magnitudes of the two prototypes (spatio-temporal feature). More importantly, they all have direct interpretability. The resulting summary statistics from those features can be interpreted as quantification of (1) spatial extent of sub-ROI (prototype) activation patterns, (2) how fast the brain responds to external stimuli; and (3) heterogeneity within single ROIs[2].

In this thesis, we further develop this framework so that such features can be extracted at group-level, which enables more robust conclusions to be drawn. We further hypothesize that the group-level effects can be captured by a hierarchy of group-level models representing a decreasing degree of group model specificity (degree of model constraints).

1.2 Contribution

The primary contributions of this thesis:

- **Extend a single-subject fMRI data model (that is, single-subject SMM-HPM) to a population-based one in a principled way (a hierarchy of model formations with increasing complexity in each level of the hierarchy)**

[1] Three spatial dimensions and one temporal dimension.

[2] Negative cross-correlation between the two time series of haemodynamic responses magnitudes from the two (subROI) prototypes are the most significant cause for heterogeneity within the ROI

The main contribution of this thesis is to develop a conceptual common model (multi-subjects SMM-HPM model) that can examine the heterogeneity within specific active regions (ROIs) across different groups of subjects, and at the same time can discriminate between fMRI data from different groups of subjects. There are many challenges one has to meet so as to achieve this goal, for example, appropriate group-level models for variations in haemodynamic response and for variations in spatial extent of HPM prototypes among the subjects. Therefore, modelling the multi-subjects SMM-HPM requires three hierarchical levels of model complexity:

First level: Model L1G-SMM-HPM: this level is the most constrained one. The assumption of this model is that within a single ROI, the multi-subject fMRI time series share most of their properties, namely, the shape of the haemodynamic response, the location and the shape of the neuronal response sources (we call them prototypes), and the number of the neuronal response sources. Only the haemodynamic response magnitudes are considered subject dependent. We start with this assumption because the heamodynamic response magnitudes depend on the stimuli and in our experiment different subjects see different stimuli sequences, which mean that the heamodynamic response magnitudes should be subject-dependent and we have to estimate them for each subject.

Second level: Model L2G-SMM-HPM: because our model is ROI-based model and the size of the ROIs are small, which means there is no big variabilities in the location and the shape of the neuronal response sources within the ROIs, in the second level we weaken the constraints by assuming that the haemodynamic response shapes are different across subjects in addition to the magnitudes. The location and the shape of the neuronal response sources, and the number of the neuronal response sources remain shared across subjects.

Third level: Model L3G-SMM-HPM: this model is the least constrained one. To allow further variabilities, we assume that the remaining property, which is the location and the shape of the neuronal response sources, is subject-dependent. In this level, different subjects only share the number of the neuronal response sources. All of the other properties of the fMRI time series; the magnitude and the shape of the haemodynamic response, as well as the location and shape of the neuronal response sources can vary across subjects.

- **Such a hierarchical formulation of the population based fMRI data model enables finding the optimal common model, and extracting novel informative features that can be used in contrasting different populations**. The optimal common model can be detected based on computing the out-of-sample (both spatially and temporally) negative log likelihood of each model in the hierarchy. The optimal model is the one that has the lowest negative log likelihood. To discriminate between different groups of subjects, three novel features can be extracted from the optimal model: a spatial feature, which is described by the prototypes volume (extent of prototypes) (left panel of Fig. 1.1); a temporal feature, which is described by the haemodynamic response time to peak (how fast is the response) (middle panel of Fig. 1.1); and a spatio-temporal feature; which is described by the zero lag cross-correlation between the haemodynamic response magnitudes time series of the prototypes within a specific ROI (high cross-correlation means that the ROI is homogeneous and one prototype is enough, low cross-correlation means that the ROI is heterogeneous and there is a need for more than one prototype) (right panel of Fig. 1.1).

1.3 Research questions

1. **How can the idea of the population-based fMRI data model be formulated?**
   To answer this question, modelling the multi-subject version of the single-subject SMM-HPM is performed at three hierarchical levels with different degrees of model constraints at each hierarchical level. Starting from the most constrained model,

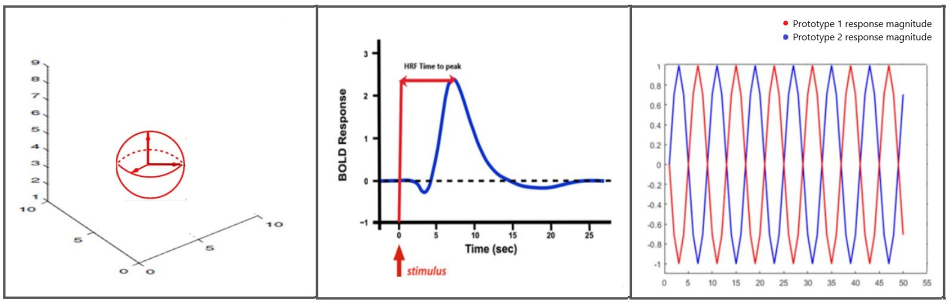

Figure 1.1: Features extracted from the optimal model to discriminate between the fMRI data of different population: spatial feature - prototypes volume (left panel), temporal feature - haemodynamic response time to peak (middle panel), and spatio-temporal feature - zero lag cross-correlation between the haemodynamic response magnitudes time series of the prototypes within the ROI (right panel).

where the population shares the same fMRI data characteristics but with subject specific haemodynamic response magnitudes, to the most relaxed model, where different subjects have different fMRI data characteristics controlled by appropriate group-level priors.

2. **What is the most constrained model that still can describe the population based fMRI data and what can be learnt from it?**

   To answer this question, out of sample negative log likelihood has been computed for each model in the hierarchy in order to find the optimal model that can describe the population. From the optimal model, three different features can be identified: a spatial feature (prototypes volume), a temporal feature (haemodynamic response time to peak) , and a spatio-temporal feature (zero lag cross-correlation between the haemodynamic response magnitude time series of different prototypes within the ROI). These features can be used in analysing within ROI cortical activation, and in contrasting different populations (e.g., fast vs. slow learners with respect to a cognitive task)

1.4 Thesis outline

The remainder of this thesis is organized into the following chapters:

Chapter two gives an overview of functional magnetic resonance imaging (fMRI). It begins by a brief description of neuroscience and neuroimaging, followed by a detailed background about fMRI and its analysis methods.

Chapter three explains the spatio-temporal modelling of fMRI data. Specifically, the two main Bayesian-based model-driven approaches; which are spatially regularized Bayesian spatio-temporal modelling, and mixture-based Bayesian spatio-temporal modelling. This chapter also reviews related works of each approach both for single-subject fMRI data modelling and multi-subject fMRI data modelling.

Chapter four provides a detailed description of the single-subject SMM-HPM. It also proposes a modification which is normalizing the haemodynamic response function (HRF). This modification is essential to extend the single-subject SMM-HPM to a multi-subject fMRI data model (multi-subject SMM-HPM).

Chapter five presents the main contribution of this thesis which is to extend the single-subject SMM-HPM to multi-subject SMM-HPM. It explains the methodology that has been adopted for this extension through a hierarchy of model formations that represent different degrees of group specificity at each level. Furthermore, this chapter describes and discusses the extensive numerical experiments that have been developed to validate and examine the performance of the proposed multi-subject models using synthetic and real data.

Chapter six presents the summary of the presented work and the plan of the future work.

1.5 Publication

Alowadi, N., Shen, Y. and Tino, P., 2016. Prototype-Based Spatio-Temporal Probabilistic Modelling of fMRI Data. In Advances in Self-Organizing Maps and Learning Vector Quantization (pp. 193-203). Springer, Cham.

CHAPTER 2

BACKGROUND

This chapter gives an overview of functional magnetic resonance imaging (fMRI). Section 2.1 gives a brief description of cognitive neuroscience and neuroimaging. Section 2.2 gives detailed background information about fMRI and its analysis methods.

2.1 Cognitive neuroscience and neuroimaging

Cognitive neuroscience studies the neural basis of the brain's cognitive performance. It relates human cognition (perception, thoughts, beliefs, memory, decision making, attention, language understanding, and problem solving) to the neurons' activities in the brain. Studies have shown that different brain regions have different functions. Cognitive neuroscience arose in the late twentieth century with the emergence of brain imaging techniques (neuroimaging: fMRI, PET, EEG, OI and MEG) as tools for analysing brain cognition [53].

Neuroimaging is considered today as one of the most successful research fields. It has a number of different technologies to image the brain directly or indirectly, which help in understanding the brain and its cognition. These technologies provide different information. Structural neuroimaging provides information about the structure of the brain, which helps in diagnosing intracranial diseases, stroke, and tumours. Functional neuroimaging provides information about the relationship between the brain's neuronal activity in specific areas and specific cognition function, which helps in diagnosing metabolic diseases. It is mainly used in cognitive neuroscience because it provides a way to image the brain's activities while subjects perform specific cognitive tasks [54].

Many functional neuroimaging techniques are available. They are varied in what they measure, and in the resolution of their temporal and spatial results.

Electroencephalography (EEG) measures the electrical activity of the brain by measuring voltage variations in brain areas from the electrical currents. It is one of the first utilized functional neuroimaging techniques, from back in 1920. It helps in diagnosing sleep problems and brain tumours, distinguishing between seizures types, confirming brain death and examining head injuries. It has many advantages. It is safe, non-invasive and cheap compared to other techniques. Its temporal resolution is high, but on the other hand, its spatial localization resolution is uncertain. This is due to the fact that EEG electrodes are separated from neuronal sources in the brain by cerebrospinal fluid (CSF), the skull, and the scalp [55].

Magnetoencephalography (MEG) originates back to 1960. It is similar to the EEG, but it measures the magnetic fields resulting from the electrical currents. Therefore, it is more accurate[1], particularly in identifying the location of the brain's activities. MEG is very useful in diagnosing brain tumours, and defects in motor areas and primary auditory; and in identifying the sources of epileptic seizures. The MEG is mostly used in combination with fMRI. As with EEG, the integration of MEG and fMRI works under the hypothesis that the regions with the greater fMRI BOLD responses have larger possibility of being electrically active over the time period of interest [56, 57].

Positron emission tomography (PET) measures metabolic changes (blood flow, oxygen use and metabolic activity) at the cellular level by injecting a small dose of radiotracer into the blood, which can be harmful, and then scans the subject with a PET scan. It provides a 3D image of how organs and tissues work, which is used in cancer and cognitive problems' detection, such as Alzheimer's; and in diagnosing brain tumours and seizures. However, its use is limited because it is expensive [58].

Optical Imaging (OI) is the most recent method for brain investigation. It measures blood and tissue oxygenation changes in the brain using near-infrared (NIR) light [59]. It can provide images of brain metabolism or intrinsic activity. However, it does not provide a full coverage of the brain volume.

Functional magnetic resonance imaging (fMRI) dates back to 1990. It measures the metabolic changes (the increase of the oxygenated blood volume and flow) that are a consequence of the neural activities in the brain using a scanner with strong magnetic fields (MRI scanner). The fMRI is friendlier compared to the other techniques. There is no need to inject the subjects with a radio-tracer as with PET or to place electrodes on their heads as with EEG. It is safe to all individuals including children and it can be used repeatedly. It is accessible to many more researchers than PET had been. This is because fMRI could be performed on many standard MRI scanners, and by the 1990s

[1] MEG is more accurate than EEG in terms of identifying the location of the brain activities. This is because MEG measures local magnetic fields inside the brain while surface EEG measures a mixture of electric signals from the whole brain.

MRI systems had proliferated such that nearly every medical center had at least one scanner and often several [54]. Moreover, although fMRI has a lower temporal resolution compared to EEG[1], its superior spatial resolution (fMRI provide images with high spatial resolution) makes it preferable. There are also several drawbacks of fMRI, fMRI includes confined space in which participants must be placed, which can induce claustrophobia in susceptible participants; subjects in the scanner are required to lie absolutely still since any movement can induce changes in the Signal-to-Noise Ratio (SNR); and subjects should protect their ears with ear plugs because of the acoustic noise required to obtain scans [57, 60, 54]. The cost also can be a disadvantage for fMRI if there is not a readily available instrument to acquire the images. Relative to EEG-based techniques, fMRI is expensive (MRI scanners cost millions of dollars, and their maintenance can be expensive as well). Relative to PET and MEG-based techniques, fMRI have similar costs for implementation [57].

Although all of these neuroimaging techniques are important in different application areas, functional magnetic resonance imaging (fMRI) has become the predominant technique in the cognitive neuroscience studies in the last two decades [54], and will be the focus of this thesis.

2.2 Functional Magnetic Resonance Imaging (fMRI)

2.2.1 fMRI and the BOLD signal and the HRF

The neural activation cannot be measured directly by fMRI. However, fMRI exploits the fact that the neuronal activation is associated with metabolic changes: increases in oxygenated blood volume and flow in the brain's activated areas. The most common method of fMRI depends on measuring these changes in blood oxygenation as an indirect

[1] fMRI has low temporal resolution because temporal resolution depends on the time between acquisitions of successive brain volumes, which is in second, and because the BOLD response peaks approximately 5 seconds after neuronal firing begins in an area. This means that it is hard to distinguish BOLD responses to different events which occur within a short time window [57].

measurement of neural activities in the brain [61, 62, 54, 60]. These changes in oxygenation are called the Blood Oxygenation Level Dependent signal (BOLD) signal by Ogawa et al. [63]. The BOLD signal arises from the interplay of blood flow, blood volume, and blood oxygenation in response to changes in neuronal activity. In short, under an active state, the local concentration of oxygenated haemoglobin increases, which increases homogeneity of magnetic susceptibility, resulting in an increase in T2*-weighted MRI signal. This BOLD signal is recorded during the fMRI scan. Hence, the fMRI signal is the BOLD signal [60].

The underlying haemodynamic response (blood flow increase) evoked due to the neuronal activation is called the Haemodynamic Response Function (HRF) by Friston [64]. It can be described as the ideal, noiseless response to an infinitesimally brief stimulus [54].

The HRF underlies the basic features of the BOLD signal. It can be considered as a generalized approximation of the BOLD signal curve. Hence, the BOLD signal can be modelled by the HRF, and the shape of the HRF can vary between subjects and between the brain regions of one subject. The haemodynamic response is very slow compared to the neuronal activity. Just after the neuronal activity, there is a slight undershoot for 1 to 2 seconds. Then, the haemodynamic response takes about 5 to 7 seconds to reach its peak. After that there is a long undershoot lasting between 15 to 20 seconds before the haemodynamic response returns back to its baseline. Based on these features of the haemodynamic response, the shape of the HRF underlying the BOLD signal is sketched in Fig.(2.1) [60, 54, 65].

Typically, in the fMRI analysis, they assumed that the response to a stimulus is well modelled by linear convolution of the stimulus with the HRF, and the nonlinear effects, such as nonlinearities in the vascular response, are largely ignored due possibly to several reasons. First, the assumption that BOLD responses were approximately linear over a range of stimulus durations promised to greatly simplify analysis. Second, nonlinearities are believed to be relatively small compared to the overall BOLD effects for events spaced more widely than 2. Third, most work on development of expected hemodynamic

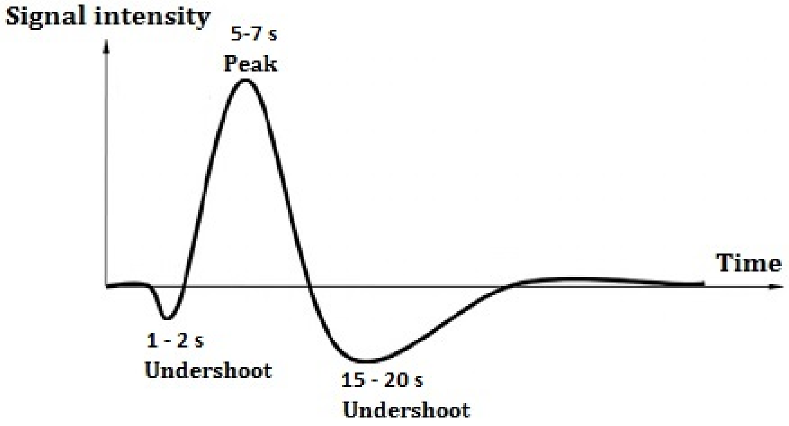

Figure 2.1: Shape of haemodynamic response function (HRF) underlying BOLD signal.

Source: Figure modified from [66]

responses has focused on determining canonical responses to single stimuli rather than exploring interactions among them. Finally, existing nonlinear models require fitting a large number of parameters, which may not be practical for many multicondition fMRI experiments due to overfitting and loss of power. In addition, the interpretation of parameter estimates with such models becomes more problematic [67]. However, modelling the nonlinearities in the BOLD response have been considered by some studies [68, 69, 70, 71].

2.2.2 fMRI time series

The acquired fMRI data consists of a sequence of brain volumes (magnetic resonance images) acquired repeatedly at T separate time points (T varies between 100 to 2000 time points) with repetition time (TR) equal typically to 3 seconds. Each brain volume consists of multiple uniformly spaced elements, called voxels. This means that one volume is a three-dimensional matrix of voxels (3D activation map). Hence, fMRI data is four-dimensional; three dimensions represent the spatial features of the fMRI data and one dimension represents the temporal features of the fMRI data. The fMRI time series in any voxel is the temporal evolution of the brain activation at that location.

In one fMRI experiment, typical brain volumes have $(64 \times 64 \times 30)$ voxels (i.e. 122,880 voxels) sampled for T time points. This produces 122,880 time series of length equal to T. The experiment is often repeated for the same subject or for multiple subjects (around 10 to 40 subjects) several times [1]. Consequently, fMRI data is massive and comprises hundreds of thousands of fMRI time series [62, 65, 54].

The fMRI data mostly suffer from distortion because of head motion; physiological oscillations, such as breathing and heartbeats; and variations in the image acquisition time, and in the magnetic static field. Consequently, the fMRI time series consists of the BOLD signals (the component of interest) and noise [65].

2.2.3 fMRI experimental objectives

There are three common objectives for the fMRI experiments: localize the activation regions for each type of stimuli; determine brain connectivity; and predict the psychological and physiological state of the brain [65, 58].

Localizing the activation regions for specific type of stimuli is the most common objective of the fMRI experiment. In the experiment, the subject's brain is scanned many times while the subject performs specific cognitive tasks.

Determining brain connectivity has received increased interest recently. It aims to reveal brain networks by finding how different brain regions interact, or understanding the transmission of information between different brain regions. It is most often performed as resting state fMRI without a specific task, in which the brain pseudo-randomly activates under little or no guiding external influence. Since no task performance is required on the part of the subject, the resting state implementation has the advantage of being a passive method of interrogating functional brain networks and their functional connectivity [57]. Brain connectivity can be structural connectivity, functional connectivity, or effective connectivity. In the structural connectivity, the connectivity network is determined based on the anatomical interaction that connect different brain regions. In functional

[1] The data acquisition process continues for 5 to 20 minutes, called a run and is repeated for a number of sessions.

connectivity, the connectivity network is determined based on the statistical dependence between signals from different regions. In the effective connectivity, the connectivity network is determined based on the causal dependence between signals from different regions (activation signal in one region causes activation signal in another region) [72, 61, 62].

2.2.4 fMRI experimental design

The two main designs used in fMRI experiments are block experimental design and rapid event-related experimental design, respectively.

The block experimental design, Fig. (2.2), presents a stimulus continuously for a long period (20 to 30 s would be typical durations), followed by absence of stimuli or by a comparison stimulus for a long period. The block experimental design provides high statistical detection power to detect brain activated regions (high spatial resolution); however, it provides poor information about the onset and the width of the haemodynamic response (low temporal resolution). It also suffers from the effects of fatigue, anticipation, boredom, and habituation, particularly in the case of a large block length.

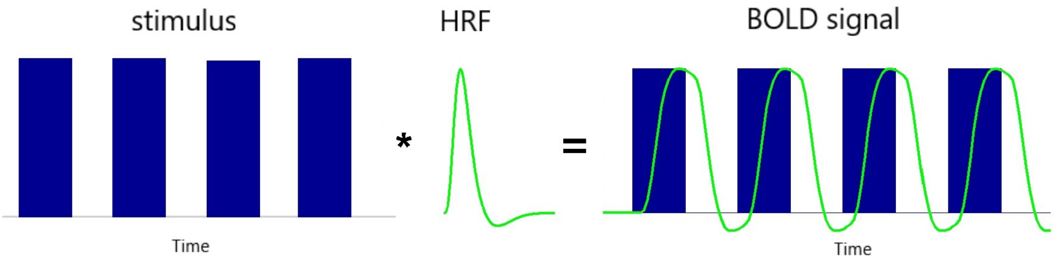

Figure 2.2: Typical modeling of the BOLD signal at a given voxel for block design experimental. The BOLD signal is modeled as the convolution of the experimental stimulus and the hemodynamic response function (HRF).

Source: Figure obtained from [65]

The rapid event-related experimental design, Fig. (2.3), presents many types of stimuli for short durations (about 2s). Rapid event-related design experiment provides fMRI data with high temporal resolution. It also can avoid the effects of fatigue, anticipation, boredom, and habituation. However, its statistical power to detect brain activated regions is low (low spatial resolution).

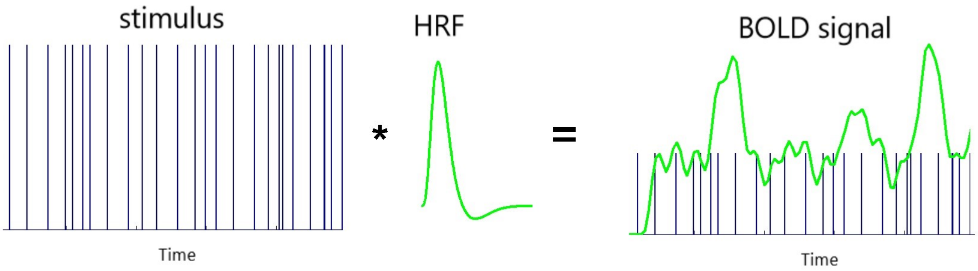

Figure 2.3: Typical modeling of the BOLD signal at a given voxel for rapid event-related design experimental. The BOLD signal is modeled as the convolution of the experimental stimulus and the hemodynamic response function (HRF).

Source: Figure obtained from [65]

Which experimental design is optimal depends on the goal of the experiment, the nature of the cognitive tasks, the ability of the resulting signal to track changes over time resulting from the task, and the statistical analysis that will be used in the experiment.

2.2.5 fMRI data preprocessing

Preprocessing the fMRI data is essential before performing statistical analysis due to the low signal-to-noise ratio, the distortions that mostly occur during data acquisition, and the large variability in the fMRI data. Some preprocessing steps aim to detect and repair distortions in the data caused by the scanners (variations in the image acquisition time and in the magnetic field), or by the subjects (head motion and physiological oscillations). Other steps aim to reduce the variabilities in the data (standardize brain regions within and across subjects) in order to increase the sensitivity and validity, particularly in the case of group-based analysis. These pre-processing steps are not fixed. A particular step is used based on the aim of the statistical analysis and the fMRI data itself [73]. fMRI preprocessing steps include:

Slice timing correction: during the fMRI experiment, the brain is scanned sequentially at different time points; therefore, the same time series in different slices (layers of the brain) are sampled at different time points, and hence temporally shifted and appear different (left panel in Fig. 2.4). These differences depend on the repetition time (TR). The acquisition time of one slice equals to (TR/Number of slices), hence the last slice is acquired almost TR seconds later than the first slice.

To correct the slice timing, the time series of all voxels in the different slices are shifted so they appear as if they are measured simultaneously (right panel in Fig. 2.4). The most popular slice timing correction is based on using a reference slice and temporally interpolating the time series of the other slices to match the timing of the reference slice. Slice timing correction is effective when the TR is short (low variability). In the case of long TR, it is better if the slice timing correction is skipped because it could introduce errors.

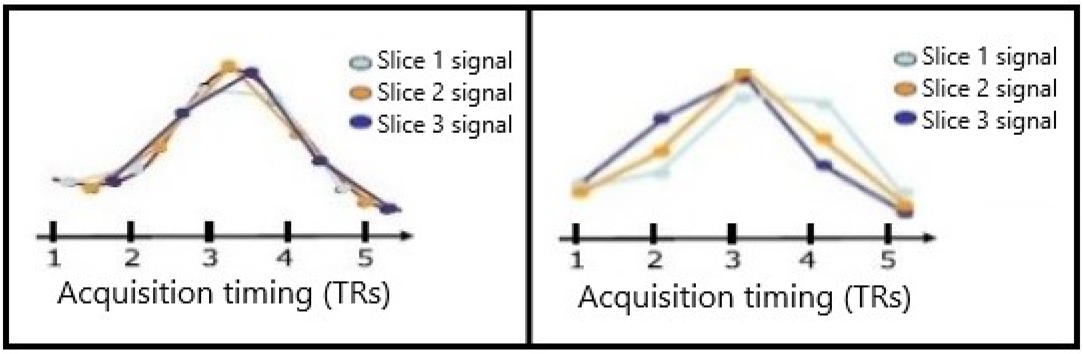

Figure 2.4: Illustration of slice timing correction. Assume three brain slices, exhibiting a similar time course, are sampled sequentially during each TR. Since the voxels are sampled at different time points relative to one another, their respective time courses will appear shifted (left panel). Slice timing correction shifts the time series so they can be considered to have been measured simultaneously (right panel).

Source: Figure obtained from [65]

Motion correction: motion occurs frequently during the fMRI data acquisition because of the subject's head movement and the physiological oscillations (breathing and heartbeat). Motions cause a mismatch in the locations of the time series in subsequent volumes. Even a small number of motion events could cause large distortion in the time series.

Head motion can be corrected by aligning all the volumes to a reference volume (the first volume or the mean volume) with rigid-body transformations: three rotation parameters (around the x, y, and z axes), and three translation parameters (up-down, left-right, and forward-backward). Assessment of the similarity between any volume and the reference volume is performed by optimizing a cost function (mutual information or sum of squared differences) in order to find the optimal parameter values.

The physiological motion can be addressed by monitoring and recording the heartbeat time and the breathing time, and correct for their effects on the data. A severe amount of motion results in excluding the subject from the study completely.

Co-registration and Normalization: co-registration or intra-subject registration is the process of aligning the functional image to the structural image of the same subject.

In the case of group-based analysis, because of the high variabilities in the shapes and features of the brains of different subjects,it is necessary to transform each subject's anatomical image into a standard atlas space, such as the Montreal Neurological Institute (MNI), and Talairach template brain. That means that a specific voxel in all subjects should represent the same brain location. This transformation is known as normalization or inter-subject registration. By normalizing the data, activations' locations become more interpretable and the results can be generalized and compared across different subjects and studies. However, normalization reduces the spatial resolution of the data and introduces errors. Currently, there are several approaches to deal with the variabilities of fMRI data in group-based analysis. One approach is smoothing, and another is identifying regions of interest (ROIs) and restricting the analysis to these regions only.

Smoothing: smoothing means blurring the fMRI images by convolving them with a Gaussian kernel. The distribution of the Gaussian kernel is described by the Full Width at Half Maximum (FWHM) of its height. Broader (Wider) FWHM produces smoother images. In the smoothed image, the number and the shape of the voxels remains the same, but the resolution of the image is reduced. Smoothing is essential in the group-based analysis to increase the overlap of activated regions between subjects by averaging the signal over a large area. This increases the results' significance and increases the analysis' validity since realistic neighbouring voxels are spatially correlated. Smoothing also increases the signal-to-noise ratio; because in the smoothed image the signal of each voxel not only originates from the voxel itself but from the neighbouring voxels as well; which reduces the effects of the random noise. The drawback of smoothing is that it could mask important variabilities between subjects. In a recent fMRI analysis approach: spatio-temporal analysis, there is no need for smoothing, as the spatio-temporal model itself deals with this issue.

2.2.6 fMRI data modelling methods

Due to the high dimensionality and the complex spatial and temporal correlation of fMRI data, it should be analysed and modelled in order to infer the relationship between the stimuli (cognitive task) and the neuronal response (temporal and spatial resolutions of the neuronal activities). fMRI data modelling methods can be categorized into two types: model-driven methods and data-driven methods.

Model-driven methods

Model-driven methods assume that there exists a model generating the observed fMRI data. These methods model the relationship between the experimental stimulus and the BOLD response. They also model the underlying HRF and noise, and try to fit the model to the observed fMRI data.

Modelling the BOLD signal and the HRF (effect of interest)

Due to the supposed linear time invariant relationship between the BOLD response and the stimulus, the BOLD response can be modelled by convolving the stimulus function with the appropriate HRF model. The shape of the signal resulting from this convolution closely represents the BOLD response. However, this assumption is poor in certain situations, such as in the case when nonlinearities are predominant when there are short separations (less than 3 s) between stimuli. Such nonlinearities are predicted by nonlinear biophysical models, for example, the balloon model [70] Another approach that can be used in the GLM setting is to extend the idea of convolution to include second-order nonlinear terms using Volterra kernels [68, 71].

In the literature, there are many methods adopted to model the HRF. One of the most popular methods is the parametrized HRF; where an analytical function (e.g.Gamma HRF) with a small number of free parameters learned from the data is used to model the HRF. Another popular approach is to use basis functions (canonical HRF, canonical HRF and its derivative, and constrained basis set) [54, 71].

Modelling the noise (effect of no interest)

fMRI data consists of the BOLD signal and noise. Noise is variance in the signal due to uncontrolled or unpreventable events (e.g. head motion, breathing, heartbeats, scanner instability). To improve the fit of the model, the noise should be modelled. The fMRI data has two types of noise: white noise and coloured noise. White noise is unstructured random noise and cannot be modelled. Coloured noise (scanner instability noise , head motion noise, and breathing and heartbeat noise) is a structured noise resulting from consistent sources of variabilities. This type of noise should be modelled [54]:

- High frequency noise due to the temporal correlation of the fMRI time series in one voxel. This type of noise can be corrected by convolving the time series with a smoothing function, such as a Gaussian curve, or by calculating and removing the correlation between the neighbouring time scans using a first-order auto-regression model (Auto-Regression model (AR)).

- Low frequency noise due to the scanner instability. It is one of the most obvious coloured noises. A high pass filter, which means a high frequency signal may pass, and pre-whitening or pre-colouring, which means removing the temporal autocorrelation by estimating it and construct pre-whitening temporal filter to undo it, are used as two steps to remove this noise.

- Movement noise due to head motion. This type of noise is removed by calculating the degree of the movement (realignment of the movement) and then transforming the images accordingly.

- Physiological noise due to the breathing and heartbeat. This type of noise can be removed by measuring it during scanning and then removing it from the signal in the pre-processing phase, or adding it as a covariate of no interest into the model design matrix. Nevertheless, some studies left it un-modelled.

Model-driven conventional method

The General Linear Model (GLM) of Eq. (2.1) is the conventional and most adopted model-driven method.

$$Y = X\beta + \varepsilon, \tag{2.1}$$

where Y: observed data, X: design matrix, β: regression coefficients, i.e. weight of each regressor, and ε: remaining noise (White noise).

Each column (regressor) in the design matrix X is the result of convolving the stimulus function and HRF to represent the BOLD signal. In addition to the regressors that represent the factor of interest (BOLD signal), other regressors can be added; such as a constant regressor which represents the intercept to model a baseline signal (the signal during rest periods), and an uncorrected coloured noise regressor if there is a need to model it.

The goal of the GLM[1] is to estimate the regression coefficients β (signal intensity at each voxel) that describe the observed data Y correctly by minimizing (optimizing) the residual error $\varepsilon = Y - X\beta$ (usually, by minimizing the sum of the square error). If the design matrix X is non-singular, the minimization of the sum of the square error is equivalent to $\beta = (X^T X)^{-1} X^T Y$. Thresholding is used to infer the existence of the activation by comparing each voxel's intensity to a significant threshold value. The null hypothesis of no effect is rejected if the voxel intensity is larger than the threshold value.

[1] In the case of fMRI data analysis, solving GLM is under-determined (for individual voxel, number of parameters is more than number of observed fMRI data). Pseudo inverse can be used to solve it.

In the literature, most of the studies that applied GLM on modelling fMRI data produce a statistic image (t-, F-, or Z- map) with signal intensity at each voxel. Signal intensity measures the evidence of the activation in the corresponding voxel. High intensity means there is an effect and the voxel is active. Low or zero intensity means there is no effect and the voxel is non-active.

The drawbacks of the GLM is that it is a mass-univariate method [1] that assumes that the voxels are independent; while in reality, neighbouring voxels are spatially coherent.

Data-driven methods

In this method, there is no underlying model. The goal is to find a structure (meaningful temporal or spatial pattern) within the data for the brain activation based on the assumption that the task-related activation leads to a distinctive structure in the data. A number of data-driven methods are available: Independent Component Analysis (ICA), clustering, parcellation, and Multi-Variate Pattern Analysis (MVPA).

Independent Component Analysis (ICA)

ICA decomposes the observed fMRI data Y into a set of underlying sources (hidden components) based on the assumption that the observed data Y is a linear combination of hidden components C:

$$Y = A \times C, \tag{2.2}$$

where Y: observed fMRI data, A: mixing matrix of mixing coefficients, which define the weight (amplitude) of each hidden component C, and C: hidden spatial or temporal components.

The hidden components C are statistically independent (the value of one component does not provide any information about the value of any other component), and have non-Gaussian distribution. In Eq. (2.2), mixing matrix A and hidden components C are

[1] mass-univariate method models the fMRI data in each voxel (voxel-wise inference) and assumes that noise covariance is diagonal (independent noise over the voxel space).

unknown. To estimate C:

$$C = Y \times A^{-1}, \tag{2.3}$$

where A^{-1}: is the inverse of the mixing matrix A.

This means A should be estimated first in order to calculate its inverse A^{-1} and find C. Based on the assumption that the components C are independent and have non-Gaussian distribution, the mixing matrix A can be estimated to a good approximation of it by maximizing the non-Gaussianity (maximizing the non-Gaussianity is equivalent to minimizing the mutual information). After estimating A, the original hidden components C can be recovered by multiplying the observed signals Y with the inverse of the mixing matrix A^{-1} simply by Eq. (2.3). Here it is assumed that the mixing matrix is square. If the number of basis vectors is greater than the dimensionality of the observed vectors, the task is over-complete but is still solvable with the pseudo inverse.

ICA was introduced the first time for analysing fMRI data by McKeown [74]. Following his successful study , many studies applied ICA in modelling fMRI data. There are two ICA approaches: temporal ICA and spatial ICA. Temporal ICA detects temporal independent components by assuming that each voxel signal is a mixture of independent time courses (points). Spatial ICA detects spatial independent components within the fMRI data by assuming that the fMRI volumes (images) are a mixture of independent spatial components. The choice of which of these two approaches is better to model fMRI data is controversial. In the literature, spatial ICA is adopted more than temporal ICA [62, 75, 76, 77]. Some studies such as Stone et al. [78] apply both spatial and temporal ICA together to model fMRI data. At first, spatial ICA is applied to reduce the dimensionality. Then, temporal ICA is applied to estimate the temporal response (haemodynamic response).

The drawbacks of ICA are that it is non-deterministic (different run of the ICA on the same data provides different components and different numbers of components), and its interpretability is low (no statistical framework to assess the results).

Clustering

K-means clustering is the most popular clustering method that has been used in modelling fMRI data [79, 80, 81, 82]. In the literature, Many of the studies applied K-means directly on the fMRI time series, but this often produces unsatisfactory and inadequate results, due to the fact that K-means is sensitive to noise and fMRI data has a high noise level. To deal with this issue, K-means has been applied on the cross-correlation between the fMRI time series of the voxels [83, 84]. Applying the K-means on the cross-correlation helps in reducing the noise and improving the performance of the K-means.

Other clustering methods that have been adopted to model fMRI data: hierarchical clustering [85, 86], and support vector clustering which provides high quality results [87, 88].

Parcellation

Parcellation means grouping voxels into small anatomically or functionally homogeneous areas called parcels. It was proposed by Thirion et al. [89]. Brain parcellations can be performed in an anatomical context based on prior anatomy and connectivity knowledge known from an existing brain atlas, such as, the Talairairach atlas and automatic anatomical labelling. It also can be performed in functional context [89, 90, 91, 62, 47, 92, 93, 94, 95]. Parcellation can serve as a basis for any further analysis of fMRI data [96].

Mostly, brain parcellation is developed by applying clustering algorithms on brain images. The most popular clustering techniques that are used for parcellation are K-mean clustering, hierarchical clustering (e.g. Ward's algorithm), spectral clustering and clustering based on mixture models. Independent component analysis (ICA) and principle component analysis (PCA) can be used as well to develop a parcellation.

The problem with brain parcellation is the lack of reproducibility. Parcellation results from a specific context may not fit a slightly different context, particularly, for multi-subjects' parcellation. One solution to enhance the reproducibility of parcellation is by random parcellation [97].

Multi-Variate Pattern Analysis (MVPA)

In the last decade, MVPA has been increasingly used to model fMRI data. This is mainly because MVPA overcomes the drawbacks of the mass-univariate model-driven methods (e.g GLM), where the correlation among the neighbouring voxels is ignored and there is a need for a threshold, which may be affected by the experimental conditions; and the drawbacks of the exploratory non-parametric data-driven methods (e.g. ICA), in which there is no statistical framework to assess the results.

MVPA was introduced in 2001 by Haxby et al. [98]. It models the neural response as a pattern of activity [99]. While standard fMRI analysis maps the experimental conditions (cognitive tasks stimulus) to a brain activated region, MVPA conversely maps the activated pattern to cognitive tasks (brain reading).

In the literature, MVPA mostly is used as a supervised classification method to find the relationship between the spatial pattern and the experimental conditions (brain state stimulus) of the fMRI activity, i.e. for each pattern determine the experimental condition to which it belongs. A number of classifiers have been used in modelling fMRI data including: Gaussian naive Bayes, support vector machine (SVM), Linear Discriminant Classifier (LDC), and neural network.

SVM is the most popular classifier used for fMRI data modelling, due to its flexibility in dealing with high dimensional data in a reasonable time, and modelling data from diverse sources. SVM has two phases: a training phase to find the statistical properties of an activated pattern in the fMRI training data in order to discriminate between the cognitive tasks; and a test phase to predict and classify the cognitive tasks of test data [100, 101, 102, 103, 104, 100].

The limitation of the MVPA are it is complex to implement. However, there are some libraries that implement MVPA, such as: SVM-light[1],LIBSVM[2],and PyMVPA[3];The application of a classifier is not as straightforward as the statistical and exploratory method; Different experimental designs and data samples require different classification approaches

[1] www.cs.cornell.edu/people/tj/svm_light

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvm

[3] www.pymvpa.org

and parameter tuning methods to avoid overfitting and to keep the results reliable; the high dimensionality and limited number of samples could easily bias the analyses; Successful application of a classifier to fMRI data relies on tight cooperation between neuroscientists and experts in machine learning techniques [62].

2.2.7 fMRI software packages

Currently, there are a number of fMRI data modelling software packages:

Statistical Parametric Mapping (SPM)[1] is the most popular. It was developed in the mid 1990s by Karl Friston and his colleagues at University College London using MATLAB, which makes it widely accessible and easy to use. SPM provides model-driven (mass-univariate) fMRI data analyses based on GLM. SPM is open source, and mostly used for data pre-processing and read and write data files, even if it is not used for the analysis of the data. It has unique connectivity modelling tools (dynamic causal modelling and psycho-physiological interaction); but it has limited visualization capabilities [54].

BrainVoyager[2] has been developed by Rainer Goebel and his colleagues. It is a commercial solution available for all platforms. BrainVoyager provides model-driven (univariate) and data-driven (multivariate) fMRI data analyses. It is easy to use, and has a user- friendly interface [54].

Analysis of Functional NeuroImages (AFNI)[3] developed in the early days of fMRI by Robert Cox and his colleagues at the Medical College of Wisconsin (now, AFNI is maintained by the National Institute of Mental Health). AFNI is sort of open source C programs for UNIX. It provides high visualization abilities, but on the other hand, it provides less sophisticated statistical modelling compared to SPM and FSL [54].

FMRIB Software Library (FSL)[4] developed by Stephen Smith and his colleagues at Oxford University in 2000. It is open source with an extensive library of statistical

[1] `http://www.fil.ion.ucl.ac.uk/spm/software/download/`

[2] `http://www.brainvoyager.com/downloads/downloads.html`

[3] `https://afni.nimh.nih.gov/download`

[4] `https://fsl.fmrib.ox.ac.uk/fsldownloads_registration`

and analysis tools for fMRI data. In recent years, FSL has been leading the statistical modelling of fMRI data. It has a robust toolbox for ICA, and for the analysis of diffusion tensor imaging data. It provides high visualization ability and rapid analysis for large data sets (FSL supports grid computing)[54].

Which of these different fMRI software packages is more appropriate depends on the analysis aspects and requirements. Based on the modelling methods: in ICA modelling, FSL is preferable; and in the case of dynamic causal modelling, SPM is preferable. Based on computing platforms: for UNIX, almost all the packages are appropriate; but for Windows, SPM is better [54]. Based on the size of the data set: for a large dataset, FSL is the superior software package [54]. FSL has become the most common package and it has been used by a number of researchers recently [54]. There is also the possibility of using more than one package, such as using SPM for data pre-processing and using FSL for data modelling and analysis [105, 54]

2.3 Summary

In this chapter, we have provided a brief overview of cognitive neuroscience and neuroimaging. Because in this work we use fMRI data, we have explained functional magnetic resonance imaging (fMRI) in detail: what is fMRI, fMRI experimental objectives, the experimental design of fMRI experiments, pre-processing steps of fMRI data, the methods used in modelling fMRI data, and the most popular fMRI software packages. The following chapter explains the spatio-temporal methods of modelling fMRI data, and reviews related works.

CHAPTER 3

SPATIO-TEMPORAL MODELLING AND LITERATURE REVIEW

This chapter explains the spatio-temporal modelling of fMRI data and its approaches; specifically, the Bayesian-based model-driven approaches. It also reviews the previous studies in each approach for single-subject and multi-subject fMRI data modelling. Section 3.1 introduces the spatio-temporal modelling of fMRI data. The two main Bayesian based model-driven approaches for the spatio-temporal modelling of fMRI data, spatially regularized Bayesian spatio-temporal modelling; and mixture based Bayesian spatio-temporal modelling, are described in sections 3.1.1 and 3.1.2, respectively.

3.1 Spatio-temporal modelling of fMRI data

Spatio-temporal modelling of fMRI data models both the spatial and the temporal behaviours of fMRI time series (i.e. BOLD signal dispersion in both space and time). The temporal behaviour of this dispersion is characterized in time by the haemodynamic response function (HRF). The spatial behaviour of this dispersion is characterized by assuming that each voxel's effect is constrained by its neighbouring voxel's response.

In the standard model-driven approach (GLM), the spatial behaviour of fMRI data is not explicitly modelled. In the literature, they deal with the spatial extent of neuroal response indirectly by smoothing (averaging the signal over neighbouring voxels) the fMRI data. The most used smoothing approach is FWHM fixed-width Gaussian kernels [2], which defines the activation size (the number of voxels in the neighbourhood). Other smoothing approaches include spatial wavelet shrinkage, which provides relatively little smoothing compared to Gaussians [3, 4]; Markov random field filtering [5]; anisotropic averaging spatial filtering [6, 7]; adaptive spatial filtering (spatial basis filters) [8]; or surface-based filtering (spatially informed basis functions) [10, 11]. The limitation of these approaches is that they mainly consider the spatial behaviour of fMRI data in the pre-processing phase, before the analysis of fMRI data. Consequently, the amount of smoothing is determined independently from the data. To deal with this issue, the spatial behaviour of the fMRI data should be considered as a part of the model by incorporating the spatial and temporal modelling into one encompassing model. The Bayesian framework is the optimal way to naturally describe and model both the spatial and temporal behaviours of fMRI data. It is a statistical inference method that employs Bayes' theorem to probabilistically infer the model's parameters as a joint posterior distribution on the parameters of interest.

$$p(\Theta|Y) = \frac{p(Y|\Theta)p(\Theta)}{p(Y)}, \tag{3.1}$$

where Y the observed data, Θ the model's parameters, $p(\Theta|Y)$ the posterior distribution (the probability of the model's parameters given the observed data), $p(Y|\Theta)$ the likelihood of the observed data given model's parameters, which can be considered as the generative model, $p(\Theta)$ the prior probability of the parameters, and $p(Y)$ the probability of the observed data.

Nowadays, Bayesian inference has become increasingly common in modelling and analysing fMRI data. Prior probability helps in incorporating valuable information about the model and its parameters in a principled manner.

Bayesian spatio-temporal modelling of fMRI data can be categorized as either spatially regularized Bayesian spatio-temporal modelling or mixture based Bayesian spatio-temporal modelling.

3.1.1 Spatially regularized Bayesian spatio-temporal modelling

In this approach of the Bayesian spatio-temporal modelling of fMRI data, a spatial prior is adopted to spatially constrain the mass univariate method (modelling the fMRI data in each voxel individually). This spatial prior implements adaptive spatial regularization on the posterior probability to reflect prior knowledge that the neuronal responses are spatially coherent. Recently, there have been several spatially regularized Bayesian methods in the literature for modelling fMRI data. These methods usually begin with a GLM model. Activation localization results from these methods have shown that the inferences produced by these methods have higher sensitivity compared to the standard modelling of fMRI data based on image smoothing.

Single-subject spatially regularized Bayesian spatio-temporal modelling

The earliest work that applied this approach to fMRI data modelling is Gossl et al. [13]. Gossl and his colleagues proposed a Bayesian spatio-temporal framework based on a GLM mass-univariate model to model fMRI data. They have utilized a Gaussian Markov Random Field (GMRF) prior as the spatial prior on the regression coefficients to characterize the spatial dependencies of fMRI data. To infer from the posterior, Markov Chain Monte Carlo (MCMC) has been used to draw samples for the parameters from the posterior distributions, which is time consuming and computationally expensive. Fharmier et al. [14] replaced the GMRF prior with Markov Random Field (MRF) prior. The MRF improves the performance by overcoming the problem of over-smoothing areas of high spatial curvature, such as the border between the high and low activation areas. Woolrich et al. [15] modelled the fMRI data using a Bayesian spatio-temporal framework but with a MRF prior on the Auto-Regression model (AR) noise parameters. As in the previous studies, they utilized MCMC to perform the posterior inference using Gibbs sampling. For the purpose of enhancing the efficiency of their model, in [106], they adopted Variational Bayes (VB) to approximate the posterior densities by factorizing over voxels space , which is computationally more efficient compared to the full Bayesian method (MCMC). Penney et al. [16] adopted a Laplacian spatial prior on the regression coefficients of a GLM, which helps in penalizing the differences between adjacent voxels. They adopted VB to infer from the posterior. In [2], Flandin and Penny have improved their previous work [16] by replacing the Laplacian prior with sparse prior. Their main goal is to decompose the fMRI data into spatial sets to easily separate the noise from the signal. They applied wavelet transform on the resulting regression coefficients' image to decompose it, and then applied Sparse Spatial Basis Functions (SSBFs) prior on the wavelet coefficients. They have used VB to approximate the posterior distributions. Compared to the previous Laplacian prior, this SSBFs prior is more robust to noise and computationally more efficient, it allows for spatially variant smoothing. In [17], Harrison et al. have proposed a Bayesian schema to analyse fMRI data with a spatial Gaussian process prior based on a diffusion kernel, which helps in modelling spatial non-stationarities. Bowman et al. [18] have modelled the spatial correlation between the voxels in a Bayesian framework with a parcellation-based Gaussian prior (anatomically informed spatial prior). Groves et al. [19] have developed a Bayesian method to combine adaptive spatial prior and fixed informative shrinkage prior. The fixed informative shrinkage prior encodes the permitted values for the signal parameters. For the inference from the posterior, they applied VB. Quir'os et al. [20] applied Bayesian spatio-temporal model with a GMRF prior on the location and magnitude of the activation in each voxel. In this model, MCMC has been utilized to sample from the posterior in order to infer the model's parameters.

Mostly, the previous works imposed the spatial prior on the activation magnitudes (regression coefficients). Therefore, they also derive the posterior probability of activation magnitudes. In order to derive the posterior probability of the activation itself, the Bayesian variable selection approaches in the following works have imposed the spatial prior on the activation indicator variables rather than imposing it on the activation magnitudes. This method incorporates a binary indicator for each voxel to determine the activation by identifying the non-zero indicator variables. Smith et al. [21] proposed a Bayesian variable selection approach to detect the activation. They fit GLM voxel-wise; then in order to represent whether the voxel is active or not, they detect whether the corresponding GLM regression coefficient is non-zero based on the value of the corresponding latent indicator variable. To handle the spatial interaction between voxels, they applied a spatial Ising prior on the indicator variables. The Ising prior is a special type of the Markov Random Field prior (binary MRF), it clusters the variables that have similar binary values together. This approach produces reliable inference, particularly in the case of low SNR. However, it ignores the temporal correlation of the time series. Lee et al. [22] have improved the spatio-temporal Bayesian variable selection approach of Smith et al. [21] by capturing the temporal correlation of each voxel using a first-order autoregressive model (AR). In [23], Zhang et al. have also provided a Bayesian variable selection approach for modelling both brain activation patterns and brain connectivity. The BOLD response has been modelled voxel-wise with a linear regression model and then a spike-and-slab prior has been applied on the regression coefficients to detect activated regions. They adopted a MRF prior on the indicator variables for capturing the spatial connectivity. To account for the temporal correlation they employ wavelet transforms. Li et al. [24] have proposed a joint Ising and Dirichlet Process (DP) prior in a Bayesian variable selection framework. They have applied the Ising prior on the indicator variable to identify the spatial dependence between voxels; and the DP prior on the regression coefficients to group the regression coefficients of the voxels that have similar intensity effects together.

Multi-subjects spatially regularized Bayesian spatio-temporal modelling

Much of the multi-subject modelling of fMRI data has adopted a hierarchical two-stage spatio-temporal approach. This approach separates the group-level inference from the subject-level inference, in which summary statistics of the parameters estimation obtained from the inference in the subject-level (first stage) are passed to the group-level inference (second stage). Compared to the all-in-one approach [107] , this two-stage summary statistics approach reduces the computational burden of analysing fMRI data.

The first stage includes voxel-specific and subject-specific modelling of fMRI data; in which a temporal model, such as GLM, is fitted voxel-wise for each subject and then the resulting summary statistics (regression parameters and their variance) are passed to the second stage. The second stage relates the summary statistics to the group-level parameters (e.g. activation level mean) in order to estimate the group-level parameters.

Woolrich et al. [35] is one of the first leading studies that applied the two-stage spatio-temporal approach in a Bayesian framework. They applied GLM to the lower-level of the hierarchy and inferred the group-level activation in the top-level in a Bayesian reference analysis framework using a reference prior, i.e, a non-informative prior. They employed both MCMC and a posterior approximation approach for the inference at the group-level. In Bowman et al. [18], the GLM has also been fitted voxel-wise in the first stage for each subject but in ROIs based analysis. In the second stage, the spatial correlations in the BOLD signal between voxels within the ROIs have been calculated in Markovian assumptions. These spatial correlations are then utilized to detect the group level activation. The two-stage spatio-temporal approach of Derado et al. [37] can consider both the spatial and temporal correlations at the group level. In the first stage, the GLM has been fitted voxel-wise to each subject's fMRI data. In the second stage, they have constructed an autoregressive model to model simultaneously the spatial and temporal correlations. To infer the parameters, maximum likelihood (ML) has been employed. In [38], Sanyal et al. have generalized the single-subject spatio-temporal model of Flandin and Penny [2] to a multi-subject spatio-temporal model. In this generalization, they assumed that the sparse spatial priors of the wavelet coefficients at the same locations are common across the subjects. Zhang and his colleagues in [36] have developed a group-level spatio-temporal model based on their single-subject model in [23]. For capturing the spatial correlation within and between subjects, hierarchical Dirichlet process priors have been applied on both the subject-level and group-level. They have also compared the results of the single-subject model with the results of the multi-subject model and found that a multi-subject modeling strategy leads to a more accurate detection of the activated areas (more accurate activation maps)[36]. Musgrove et al.[39] have also extended the single-subject Bayesian variable selection approaches proposed in [21] and [22] to model group-level fMRI data. For modelling the spatial dependency in the regression coefficients according to the values of the corresponding latent indicator variables, they adopted a Spatial Generalized Linear Mixed Model (SGLMM) prior in conjunction with parcellation; which is computationally more efficient compared to the Ising model. To account for the temporal correlation, they have employed a second-order autoregressive AR(2), which provides a trade-off between the complexity of the higher-order autoregressive model and the simplicity of the first-order autoregressive model.

3.1.2 Mixture based Bayesian spatio-temporal modelling

A mixture model is a probabilistic model assuming that the observed data is generated from a finite mixture of component models. Such model components could be given as Gaussian probability distributions and it is referred as to Gaussian mixture model (GMM) (for an example of GMM, see Fig. (3.1)).

Mathematically, given an observed data item y, the likelihood of a mixture model is

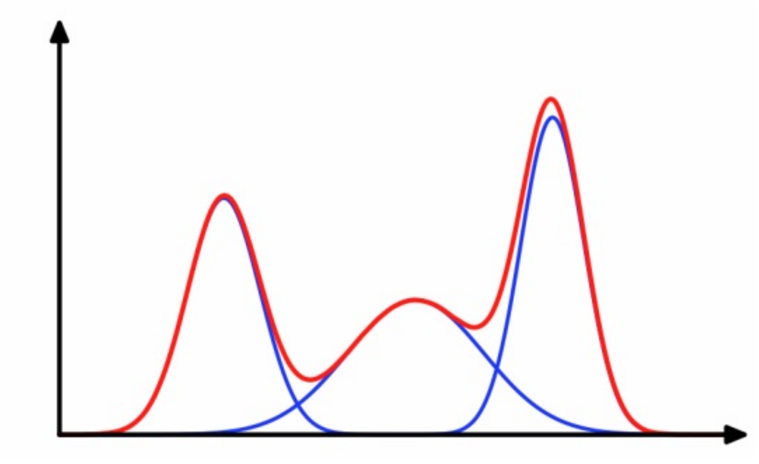

Figure 3.1: Gaussian mixture model (GMM) (red) of three Gaussian components distributions (Blue)

Source: Figure obtained from https://dirichletprocess.weebly.com/clustering.html

the weighted sum of (K) component densities:

$$p(y) = \sum_{k=1}^{K} \pi_k p(y|k, \theta_k) \quad \text{with} \quad \sum_{k=1}^{K} \pi_k = 1 \tag{3.2}$$

where y represents data, k the component index, K the number of components, π_k the k-th mixture weight, and θ_k the parameters of the k-th component model.

The drawback of this mixture model in modelling spatial data is that the mixture weights are assumed to be constant across all observations. This assumption could be invalid for spatial data. This is the case because such mixture models are used to explain the data at individual locations and the distribution of mixture weights could be location-dependent. To adapt the mixture approach to spatial modelling, so-called spatial mixture model (SMM) has been formulated to account for this location-dependence. SMM is mathematically defined as follows:

$$p(y(r)) = \sum_{k=1}^{K} \pi_{(k|r)} p(y(r)|k, \theta_k), \tag{3.3}$$

where $y(r)$ represents the observed data at location r, K the number of components,

r the location, and $\pi_{(k|r)}$[1] the k-th mixture weight for location r. SMM can be seen as an instance of mixture-of-experts and the term $\pi_{(k|r)}$ can be interpreted as so-called gate function. In a Bayesian setting, it can be interpreted as a spatial prior, that is, $p(k|r)$ is proportional to the probability of the data at r being generating by the k-th component. More importantly, spatial correlations in the data can be modelled through the spatial prior. For example, $p(k|r)$ can be defined using a smooth parametrized function of r so as to take into account the smoothness in changes of the weights between the adjacent locations. In the fMRI data modelling literature, the activation map of individual activation sources is often modelled as a Gaussian-shaped surface. This is because: (i) It has a reasonable shape, being high in and around the activation centre and decreasing from this centre; (ii) It is a simple, parametrized model; (iii) It has sufficient flexibility, serving as the base shape for modelling a surface of complicated shape. By Bayes' rule, such a Gaussian spatial prior (gate function) is defined as follows:

$$p(k|r) = \frac{N(r, \mu_k, \Sigma_k)p(k)}{\sum_{k=1}^{K} N(r, \mu_k, \Sigma_k)p(k)}, \tag{3.4}$$

where $N(r, \mu_k, \Sigma_k)$ denote a three-dimensional Gaussian distribution over r with mean μ_k and covariance matrix Σ_k. It is also worth noting that for the component representing background activity, its spatial prior is usually assumed to be uniform rather than Gaussian.

The point estimates of Spatial Mixture Model (SMM) parameters could be learnt by Maximum Likelihood (ML) or Maximum A Posteriori (MAP). In the cases where latent variables are included in SMM, Expectation-Maximization (EM) can be employed. To avoid local maxima, stochastic optimization algorithms such as simulated annealing can be employed. For a fully Bayesian approach, the posterior distribution over SMM parameters is to be inferred. To compute such a posterior distribution exactly in a statistical sense, the workhorse algorithm is MCMC. For an approximate inference in SMM, VB or other

[1] Note that $\pi(k)$ represents location-independent prior over component index k and $\pi_{(k|r)}$ the location dependent one.

variational approximation methods could be employed.

The spatial mixture model is a suitable spatial model for fMRI data modelling. This is based on two assumptions about the fMRI data: (1) the activated regions are spatially extended. (2) the activation pattern is smooth. Moreover, there are multiple activation sources.

A spatial mixture model in fMRI data assumes that signals are generated from a mixture of components distributions, and the activation is estimated by assigning the voxels to the most likely components (the nearest component). Modelling fMRI data by the spatial mixture model approach is more efficient the than spatially regularized Bayesian approach. The spatial mixture model approach explicitly models the activation shape and location, providing a more interpretable model in which each component corresponds to an underlying neural activation source [27, 48]. It is computationally more efficient with a small number of parameters. For group-level modelling of fMRI data, SMM approaches model the fMRI data at a higher level (feature level, such as activation location and intensity) than the spatially regularized Bayesian approaches, which model the data voxel-wise. This makes it less sensitive to the mis-registration problem in modelling group data and makes it more adequate for group modelling.

Single-subject Mixture based Bayesian spatio-temporal modelling

A number of studies in the literature have utilized a mixture model to model fMRI data, following the successful implementation of the mixture model on fMRI data by Everitt and Bullmore in 1999 [25].

Everitt and Bullmore [25] have developed a two-component mixture model to represent the surface of a statistic which is derived from fMRI data by using a mass-univariate GLM model. One component uses a non-central χ^2 distribution to model the probability distribution which generates the statistics on those activated voxels and another one uses a χ^2 distribution for modelling the statistics on the non-active voxels. This accounts for the pre-assumed fact that every voxel is either activated or not. Further, the maximum likelihood algorithm has been employed to learn the model parameters. Following this, the posterior probabilities of being activated can be computed for all voxels, which effectively results in a brain activation map. More importantly, the resulting activation map is equivalent to a map of p-values derived from hypothesis testing on test statistics [25].

Hartvig and Jensen[26] extended the mixture model of Everitt and Bullmore [25] by taking into account the spatial correlation between the statistics of the neighbouring voxels, that is, a voxel has a higher chance to be active if its neighbouring voxels are all active and vice versa. This is implemented by using both the statistic of a voxel and those of its neighbouring voxels to compute the (marginal) posterior probability of that voxel being activated. Also, note that they used a Gamma distribution as the probability density for the activated voxels and a Gaussian distribution for the inactivated voxels.

In contrast to Hartvig and Jensen[26], Woolrich et al. [28] have developed a principled way to incorporate the smoothness prior (MRF) into the fMRI models by introducing a class label for each voxel (class of activated voxels versus that of inactivated ones) and enforcing a smooth change of those class labels across the voxels. They modelled the active components as Gamma distributions, and the inactive component as Gaussian distribution. In [29], Woolrich and Behrens have improved their previous method by adopting computationally more efficient inferential techniques, that is, VB instead of MCMC.

Ggorgolewski et al. [108] have developed a mixture model that is similar to the one in [28] but the mixture weights are defined at cluster level as in [26] .

Penny and Friston [27] have further developed the above methodologies. First, they have integrated the mixture-based approach to spatial modelling of fMRI activation pattern with a GLM-based approach to temporal modelling of fMRI time series, Second, more than two components are allowed so as to account for the existence of multiple activation sources. Third, they proposed a very flexible model for the mixture weights, a multinomial distribution. At the same time, for each of the activation sources, the spatial variation of their corresponding mixture weights is modelled by a Gaussian distribution function (that is a smooth function). This represents another principled approach to incorporate the smoothness prior. It is worth noting that the number of the mixture components are fixed a priory, and they modelled the temporal model by GLM on the component level (for each cluster) rather than the voxel level, which reduces the number of the parameters in the model because single time series represents the voxels within one cluster.

Oikonomou and Blekas [30] have analysed fMRI data by a spatial mixture of linear regression with a sparse prior over the linear regression coefficients for model order selection, and a spatial MRF prior over the mixing coefficients for the spatial correlation between voxels.

In order to identify a procedure for brain lesion-segmentation, Ozenne and Subtil [31] developed a spatial mixture of Gaussian and Gamma distributions that includes adaptive large-range spatial prior. They employed the Potts model as a specification for the spatial prior. However, Potts model is only able to consider the short-range spatial dependencies (adjacent voxels belong to same spatial structure). Therefore they extended the Potts model using multi-order regional potential to be able to consider the large-range spatial dependencies. They have found that large-range regional regularization significantly improves the accuracy of the lesion segmentation in the case of the white matter disease compared to the short-range regional regularization.

In the interest of estimating the HRF and the within and between single-trial variability (the variability in the brain's response corresponding to different stimuli over specific period of time), Brigne et al. [32] have modelled the fMRI data based on Gaussian mixture model and applied maximum likelihood to infer the model.

Llera et al. [33] have developed methods to model fMRI data by a mixture of Gaussian and inverse-Gamma components, or a mixture of Gaussian and Gamma components learned by Variational Bayes (VB), in order to compare the performance of these methods with the classical methods to model fMRI data by a mixture of Gaussian and Gamma components or a mixture of Gaussian and inverse-Gamma components learned by Maximum Likelihood (ML). They found that the mixture of Gaussian and inverse-Gamma components learned by Variational Bayes (VB) is the most robust and computationally efficient method.

Nguyen et al. [34] have proposed a two-stage mixture method for time-series data modelling. In stage one, they adopted a Mixture of Auto-Regressions (MoAR) model to perform temporal clustering. In stage two, they have fitted a MRF model to smooth stage one's clustering outcomes. The results of this approach show that the addition of the second stage increases the performance accuracy. In [109], Nguyen and his colleagues explained the importance of adopting a Maximum Pseudo-Likelihood (MPL) estimation approach in fitting the model as an alternative to the Maximum Likelihood (ML) approach, which converges to zero in the case of long time series.

Multi-subjects Mixture based Bayesian spatio-temporal modelling

To the best of our knowledge, the first mixture-based methodology for modelling the spatial activation patterns across multiple fMRI data sets was developed by Kim et al. [40]. These data sets were obtained from a single subject during different visits and/or at different imaging facilities. However, the methodologies developed for such so-called multi-site fMRI data are applicable to multisubject fMRI data.

To model the variabilities of neural activation exhibited in these multi-site fMRI data, they have developed a Gaussian mixture model similar to that of Penny and Friston [27]. The difference between these two works are as follows: the model in [40] is a spatial model of the β-map produced by individually inferring GLM from fMRI time series across all voxels whereas the model in [27] is a spatio-temporal model inferred from the four-dimensional fMRI data.

In [40], the model parameters were estimated in a fully Bayesian manner using Gibbs sampling except that the number of mixture components (say K) is fixed a priori as in [27]. In [41], Kim et al. have further developed their model by allowing for a full Bayesian approach to infer K from the data. To achieve this, an infinite mixture model is adopted for mixture-based spatial modelling while a Dirichlet process is used as a prior on those infinitely many mixture weights. The Dirichlet Process Prior (DPP) penalises large K values and the posterior distribution over K is inferred from the data.

To infer a common activation pattern that consider the variations between different multi-site fMRI data, in [42], Kim and his colleagues have developed a spatial mixture model (SMM) based on a Hierarchical Dirichlet Process Prior (HDPP) with random effects (RE) on the components' shape parameters. HDPP allows the automatic inference of the number of components from the data and the sharing of the mixture model across a number of images. Random effects allow variations in the components of the mixture model (activation intensity and location) across the images.

Thirion et al [43] have also performed group-level modelling for fMRI data based on the Dirichlet process mixture model. Their spatial mixture model is dissimilar to [42] in that they applied the Dirichlet process mixture model at the subject-level to extract the spatial model activation pattern, and then matched the activation pattern across subjects in a Bayesian framework.

The Bayesian hierarchical mixture model that has been developed by Xu and his colleagues in [45] to model the fMRI data of groups of subjects differs from the previous works in that Xu and his colleagues have modelled both the subject-level and group-level variabilities by applying Gaussian mixture models on each level. This approach represents the underlying structure perfectly and allows estimating the proportion of subjects who have activation on a specific location. However, it is a complex method (large number of parameters and slow estimation method, MCMC).

The Bayesian hierarchical mixture model that has been developed by Xu and his colleagues in [45] to model the fMRI data of groups of subjects differs from the previous works in that Xu and his colleagues have modelled both the subject-level and group-level variabilities by applying Gaussian mixture models on each level. For this reason, this model can represent the underlying structure (subject-level and group-level) perfectly [45], can make inference on the activation patterns at all levels: the group-level, the subject-level and the voxel-level, and can estimate the proportion of subjects who have activation on a specific location.

For the purpose of clustering the activation based on the haemodynamic features, Fouque et al. [46] have applied a spatial mixture model on the haemodynamic features (HRF shape, such as time to peak and width). They estimated the haemodynamic response function voxel-wise and then applied a multivariate spatial Gaussian mixture model on the extracted haemodynamic features.

Unlike the previous works, which have been interested in localizing brain activation, Jbabdi and his colleagues in [47] applied a hierarchical infinite mixture of Gaussians with DPP on the fMRI data of a group of subjects to study the group-level brain connectivity.

Gershman et al. [48] suppose that the fMRI data is generated by a superposition (linear combination) of latent sources, such as Gaussian radial basis functions. This superposition is covariate-dependent, which means that it relates the latent sources to the covariate variables through the mixing weight to show how much each component is activated responding to different covariates. This approach is different from the previous approaches in that it is a mixture in signal space (mixing); while other approaches are a mixture in model space. Mixing allows multiple components to contribute to the voxels, whereas the mixture assigns a single component to each voxel. To model fMRI data of a group of subjects, Gershman et al. applied a hierarchical model such that the latent sources of each subject are considered as a spatial transformation of the group-level latent sources (template).

Lashkari et al.[49] have proposed a hierarchical Bayesian mixture model with DPP with the aim of defining the patterns of the functional specificity, which means that different areas in the brain are specific for different functions, that appear consistently across multi-subject fMRI data. In each subject, they model the response in each voxel to each stimulus as a binary activation variable (activation profile). To identify the functional specificity systems (group of voxels that become active responding to specific stimulus), voxels with similar activation profiles are clustered together based on the assumption that the activation profiles of the voxels are generated by a mixture model. To model the variabilities across groups of subjects, they applied DPP on the functional systems of the group of subjects. The functional system resembles the active component on the previous studies but it is for a specific stimulus type.

Roge and his colleagues [50] have localized the activation in multi-subject fMRI data by a fully unsupervised method (no information is available about the task or the stimuli, such as resting state fMRI) based on the assumption that active regions are consistent across subjects. For this purpose, they developed non-parametric GMM with GPP to smooth the activation, and a Chinese Restaurant Process (CRP) prior to determine the clusters. MCMC with an enhanced split-merge procedure has been utilized for inferring the model, which reduces the computation times significantly.

In [51], Churchill and his colleagues have improved the conventional Gaussian mixture model, which is typically used in the clustering of the fMRI data, for the purpose of investigating group-level functional connectivity. This improvement includes simultaneously estimating the active regions and the functional connectivity between them using an expectation-maximization (EM) method.

Raman et al. [52] a proposed unified model to simultaneously infer the effective connectivity for each subject by Dynamic Causal Modeling (DCM), and define the population-based connectivity clusters using finite Gaussian mixture model. Parameter inference has been accomplished by MCMC.

3.2 Summary

It has been shown from this review that Bayesian framework is the optimal way to naturally describe and model both the spatial and temporal behaviour of fMRI data. Within the Bayesian framework, the spatial behaviour of the fMRI data is considered as a part of the model by incorporating the spatial and temporal modelling into one encompassing model. Moreover , Prior probability helps in incorporating valuable information about the model and its parameters in a principled manner. The literature identified two main methods for the Bayesian based spatio-temporal modelling of fMRI data: spatially regularized Bayesian spatio-temporal modelling; and mixture-based Bayesian spatio-temporal modelling.

In the spatially regularized Bayesian spatio-temporal modelling, an adaptive spatial prior regularizes the posterior probability to reflect prior knowledge that the neuronal responses are spatially coherent. Different spatial priors have been used in the literature: Gaussian Markov random field (GMRF) prior [13, 20]; Markov random field (MRF) prior [14, 15, 23, 36]; Laplacian spatial prior [16]; sparse spatial basis functions (SSBFs) prior [2, 38]; Gaussian process priors (GPPs) [19]; Gaussian process priors based on diffusion kernel [17]; parcellation-based Gaussian prior [18]; and Ising prior [21, 22, 24]. Most of the studies applied the spatial prior on the regression coefficients of the GLM; excluding [15], who applied it on the autoregressive (AR) noise parameters, and [21, 22, 23, 36, 24] who provided a Bayesian variable selection approach by applying the spatial prior on the activation indicator variables.

In the mixture based Bayesian spatio-temporal modelling, the spatial characteristics of fMRI data are modelled explicitly in addition to the temporal characteristics. As it appears from this review, modelling fMRI data by a mixture model approach is more efficient than the spatially regularized Bayesian approach. In the literature, there are a number of studies that have adopted mixture models to model multi-subject fMRI data. Beside the variations between these methods in their objectives and data used, the technical differences in their analysis methods can be discussed under the following aspects:

- **Specification of the number of the mixture components**. There are three different approaches to deal with this issue. In the first approach, the number of mixture components is usually set to a fixed number [46]. For example, a model with two components is formulated to account for two distinct states of brain activation, namely active versus non-active states. In the second approach, an infinite mixture approach has become very popular. In theory, the possible number of mixture components ranges from 1 to infinity but a prior is employed to penalize larger numbers. Examples of the prior employed are the Dirichlet process prior (DPP) [40, 41, 42, 43, 45, 47, 49] and the Chinese restaurant process (CRP) prior [50]. In the third approach, which could be considered as a trade-off between the first two approaches, instead of a fixed number of components or infinitely many components, the optimal number of components is inferred from the data via a model selection procedure [51, 52].

- **The spatial distributions of the components**. Almost all the studies have adopted Gaussian mixture models to represent the spatial extent of the active and non-active components [40, 41, 42, 45, 46, 47, 48, 49, 50, 51, 52]. However, [43] have used Gamma distribution for active component and Gaussian distribution for non-active component.

- **One versus two stage approaches**. In almost all of these studies[40, 41, 42, 43, 45, 47, 49, 50, 51], a two-stage approach was employed. At stage 1, a statistical map (i.e., t-, F-, or Z-map) is first inferred. At stage 2, this map is modelled as a Gaussian mixture. The exceptions are [46] and [27], where both the temporal parameters and the parameters in the spatial mixture model are learned jointly from the data.

- **The computational methods for Bayesian inference**. In addition to Maximum Likelihood and Maximum A Posterior approaches, there are three major computational tools that have employed Bayesian spatio-temporal fMRI data analysis: Markov Chain Monte Carlo (MCMC) [40, 41, 42, 45, 46, 47, 48, 50, 52], Variational Bayes (VB) [49], and Expectation Maximization (EM) [43, 51].

- **Temporal model**. All these studies model the temporal aspects of fMRI data by a GLM.

In this thesis, the main goal is to build a group-level fMRI data model. Therefore, a mixture-based Bayesian spatio-temporal modelling approach has been adopted to build a group-level fMRI data model. Shen et al. [1] have developed a regularized spatial mixture model of hidden process models (SMM-HPM) to identify spatio-temporal patterns within single ROI, while adopting a parametric approach to model the HRF. The aim of this research is to extend the single-subject SMM-HPM to apply it in a population-based fMRI data modelling.

In the next chapter, there is a detailed explanation of the single-subject SMM-HPM and the modification that has been applied to it for the purpose of group-level modelling. The extension of this model to group-level modelling is in Chapter (4).

CHAPTER 4

SINGLE-SUBJECT SMM-HPM WITH NORMALIZED HRF

This chapter first describes the SMM-HPM model [1] and then proposes its modification based on a normalized HRF which is essential to extend the SMM-HPM model for group fMRI data modelling. Section 4.1 provides a detailed explanation for the single-subject SMM-HPM model. Section 4.2 describes the normalized HRF proposed for the model. Learning the modified model is presented in section 4.3. Experiments to validate this modification are given in sections 4.4 and 4.5. This Chapter's contents corresponds to the paper "Prototype-Based Spatio-Temporal Probabilistic Modelling of fMRI Data" published in the international conference "11th Workshop on Self-Organizing Maps 2016" (WSOM 2016), held at Rice University in Houston, Texas, 6-8 January 2016.

63

CHAPTER 5

MULTI-SUBJECT SMM-HPM

The issues that need to be considered when defining a group level SMM-HPM include variations in haemodynamic response and spatial extent of HPM prototypes among the subjects. We will formulate the group level SMM-HPM as a hierarchy of model formations, from the most constrained (except for response magnitudes, all subjects share the same model parameters) to the most flexible (subjects can have different individual SMM-HPM parameters, however, they are constrained by appropriate common group-level priors). In particular we will consider three levels in the hierarchy. The first level model (L1G-SMM-HPM) is the most constrained formulation in which different subjects share the same prototypes, both in the spatial and temporal sense, i.e., the same spatial priors and the same haemodynamic response shape. They only differ in haemodynamic response magnitudes. In the second level model (L2G-SMM-HPM), we allow subjects to have different haemodynamic response shapes. However, parameters of the normalized HRFs of individual subjects are assumed to come from the same group-level prior. The same spatial priors are shared by all subjects. In the third level model (L3G-SMM-HPM), the constraint on spatial priors is relaxed. Now, besides individual response magnitudes and HRF shapes, each subject can also have different spatial priors. As before, the spatial prior shape parameters of individual subjects are constrained by stipulating that they come from the same group-level prior.

Sections 5.1, 5.2, and 5.3 describe L1G-SMMHPM, L2G-SMM-HPM, and L3G-SMM-

HPM, respectively. Results and discussion of both synthetic and real data experiments are in section 5.4.

5.1 First level multi-subject SMM-HPM: L1G-SMM-HPM

Given the noise-free signal, we assume that the observation on individual subject, voxel, and volume are independent from each other (this is a conditional independence). The L1G-SMM-HPM model likelihood is

$$p(\mathbf{Y}) = \prod_{u=0}^{U} \prod_{v=0}^{V} \prod_{t=0}^{T} p\left(y_{uvt}; \Theta^{STM}\right), \tag{5.1}$$

where Θ^{STM} collects all group-level model parameters, u, v and t are the subject, voxel and volume indices, respectively. $p\left(y_{uvt}; \Theta^{STM}\right)$ is modelled as a spatial mixture model:

$$\begin{aligned} p\left(y_{uvt}; \Theta^{STM}\right) &= \sum_{k=0}^{K} p(k|v; \Theta^{S}) \cdot p\left(y_{uvt}|k; \Theta^{T}_{u}\right) \\ = \quad & \sum_{k=0}^{K} p(k|v; \Theta^{S}) \cdot p(y_{uvt}|k; \Theta^{NRL}_{u}, \Theta^{HRF}, \Theta^{NIS}) \end{aligned} \tag{5.2}$$

Spatial parameters $\Theta^{S} = \{\mu_k, \Sigma_k\}$ contain the prototype locations and shapes. Temporal parameters $\Theta^{T}_{u} = \{\Theta^{NRL}_{u}, \Theta^{HRF}, \Theta^{NIS}\}$, contain haemodynamic response shape parameters for each process p and prototype k, $\Theta^{HRF}= \{\kappa_{k,p}, \theta_{k,p}\}$; and noise parameters for each prototype k, $\Theta^{NIS}= \{\sigma^2_k\}$. All these spatial and temporal parameters are shared across subjects. On the other hand, the haemodynamic response magnitude parameters ${\Theta^{NRL}}_{u} = \{a_{u,k,p,s}\}$ are subject specific, per subject u, prototype k, process p and stimulus s. That means that for each subject u in the active prototypes $k = 1, 2, ..., K$, the haemodynamic response of each hidden cognitive process p, for each stimulus s at time t

is

$$h_{u,k,p,s}(t) = a_{u,k,p,s} \cdot \delta(t - (t_{p,s} + \tau_{p,s})) \bigotimes g_{k,p}(t) \tag{5.3}$$

5.1.1 Learning of the L1G-SMM-HPM

The model parameters are fitted in the MAP estimation framework, maximising the posterior

$$p(\Theta_{STM}|\mathbf{Y}) = p(\mathbf{Y}|\Theta^{STM}) \cdot p(\Theta^{STM}). \tag{5.4}$$

The likelihood ($\mathcal{L}$) can be expressed as:

$$\begin{aligned} & p\left(\mathbf{Y}|\Theta^{STM}\right) \qquad (5.5) \\ = & \prod_{u=1}^{U}\prod_{v=1}^{V}\prod_{t=1}^{T} \frac{1}{\sum_{k=0}^{K} p(v|k;\Theta_k^S)} \sum_{k=0}^{K} p(v|k;\Theta_k^S) \cdot p(y_{uvt}|k;\Theta_{u,k}^T) \\ = & \prod_{u=1}^{U}\prod_{v=1}^{V}\prod_{t=1}^{T} \frac{1}{\sum_{k=1}^{K} \mathcal{N}\left(v|\mu_k,\Sigma_k\right) + \frac{1}{N}} \left\{ \sum_{k=1}^{K} \mathcal{N}(v|\mu_k,\Sigma_k) \cdot \right. \\ & \left. \mathcal{N}\left(y_{uvt}; x(t; a_{u,k,p,s}, \kappa_{k,p}, \theta_{k,p}), \sigma_k^2\right) + \frac{1}{N} \cdot \mathcal{N}\left(y_{uvt}; b, \sigma_0^2\right) \right\}. \end{aligned}$$

As in the single subject case, the prior is modelled as

$$\begin{aligned} & p(\Theta^{STM}) \qquad (5.6) \\ = & p(N) \cdot p(b) \cdot \prod_{k=1}^{K} p(\mu_k) \cdot \prod_{k=1}^{K} p(\Sigma_k) \cdot \prod_{u=1}^{U}\prod_{k=1}^{K}\prod_{p=1}^{P}\prod_{s=1}^{S} p(a_{u,k,p,s}) \\ & \cdot \prod_{k=1}^{K}\prod_{p=1}^{P} p(\kappa_{k,p}, \theta_{k,p}) \cdot \prod_{k=0}^{K} p(\sigma_k^2). \end{aligned}$$

We maximized the posterior by minimizing the negative log posterior using scaled conjugate-gradient optimization. Below we list the relevant gradients.

For each prototype k, we need to calculate

$$\nabla_{\Theta_k^{STM}} \left\{ -\log \left(p(y_{uvt}|\Theta^{STM}) \right) - \log \left(p(\Theta^{STM}) \right) \right\}. \tag{5.7}$$

In what follows we present the derivatives of the (negative log) likelihood $\mathcal{L} = -\log \left(p(y_{uvt}|\Theta^{STM}) \right)$.

The derivative of $\mathcal{L}$ (Eq. 5.5) with respect to the temporal parameters Θ^T in the $k-th$ prototype:

$$\begin{aligned} \nabla_{\Theta_k^T} \{\mathcal{L}\} &= \nabla_{\Theta_k^T} \left\{ -\log \left(\sum_{k=0}^{K} p(k|v;\Theta_k^S) \cdot p(y_{uvt}|k;\Theta_k^T) \right) \right\} \\ &= -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} \frac{1}{\sum_k p(k|v).p(y_{uvt}|k)} . \nabla_{\Theta_k^T} \left\{ \frac{\sum_k p(v|k).p(y_{uvt}|k))}{\sum_k p(v|k)} \right\} \\ &= -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} \frac{p(k|v)}{\sum_k p(k|v)p(y_{uvt}|k)} . \nabla_{\Theta_k^T} p(y_{uvt}|k) \end{aligned} \tag{5.8}$$

The derivative of $p(y_{uvt}|k)$ with respect to the temporal parameters Θ^T in the $k-th$ prototype:

$$\begin{aligned} \nabla_{\Theta_k^T} \Big\{ p(y_{uvt}|k) \Big\} &= \nabla_{\Theta_k^T} \left\{ \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp \left(\frac{-(y_{uvt} - x_t^k(\theta))^2}{2\sigma_k^2} \right) \right\} \\ &= p(y_{uvt}|k) . \frac{y_{uvt} - x_t^k}{\sigma_k^2} . \nabla_{\Theta_k^T} x_t^k \end{aligned} \tag{5.9}$$

Substitute the value of $\nabla_{\Theta_k^T} \{p(y_{uvt}|k)\}$ in Eq.(5.8)

$$\nabla_{\Theta_k^T} \{\mathcal{L}\} = -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(k|v, y_{uvt}) \cdot \frac{y_{uvt} - x_t^k}{\sigma_k^2} \cdot \nabla_{\Theta_k^T} x_t^k, \tag{5.10}$$

where x_t^k is the haemodynamic response at each time step t for prototype k. The derivative of (Eq. 5.10) with respect to each of the temporal parameters is as follow:

Response magnitudes:

$$\frac{\mathrm{d}\,\mathcal{L}}{\mathrm{d}\,a_{u,k,p,s}} = -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(k|v, y_{uvt}) . \frac{y_{uvt} - x_t^k}{\sigma_k^2} . \frac{x_t^k}{a_{u,k,p,s}} \tag{5.11}$$

HRF scale parameter with constraint $\theta_{k,p} > 0$:

We suppose that $\alpha_{k,p} = \log(\theta_{k,p})$

$$\begin{aligned} \frac{\mathrm{d}\ \mathcal{L}}{\mathrm{d}\ \theta_{k,p}} &= \frac{\mathrm{d}\ \mathcal{L}}{\mathrm{d}\ \alpha_{k,p}} \cdot \frac{\mathrm{d}\ \alpha_{k,p}}{\mathrm{d}\ \theta_{k,p}} \\ &= -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(k|v, y_{uvt}) \cdot \frac{y_{uvt} - x_t^k}{\sigma_k^2} \cdot x_t^k \cdot \left[\frac{t}{\theta_{k,p}} - \kappa_{k,p} + 1\right], \end{aligned} \tag{5.12}$$

HRF shape parameter with constraint $\kappa_{k,p} > 1$:

We suppose that $\beta_{k,p} = \log(\kappa_{k,p} - 1)$

$$\begin{aligned} \frac{\mathrm{d}\ \mathcal{L}}{\mathrm{d}\ \kappa_{k,p}} &= \frac{\mathrm{d}\ \mathcal{L}}{\mathrm{d}\ \beta_{k,p}} \cdot \frac{\mathrm{d}\ \beta_{k,p}}{\mathrm{d}\ \kappa_{k,p}} \\ &= -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(k|v, y_{uvt}). \frac{y_{uvt} - x_t^k}{\sigma_k^2} \cdot x_t^k \cdot \log \frac{t}{(\kappa_{k,p} - 1)\theta_{k,p}} \end{aligned} \tag{5.13}$$

Noise with constraint $\sigma_k^2 > 0$:

We suppose that $r_k = \log(\sigma_k^2)$

$$\begin{aligned} \frac{\mathrm{d}\ \mathcal{L}}{\mathrm{d}\ \sigma_k^2} &= \frac{\mathrm{d}\ \mathcal{L}}{\mathrm{d}\ r_k} \cdot \frac{\mathrm{d}\ r_k}{\mathrm{d}\ \sigma_k^2} \\ &= \sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} \frac{p(k|v, y_{uvt})}{2} \cdot \left[1 - \frac{(y_{uvt} - x_t^k)^2}{\sigma_k^2}\right] \end{aligned} \tag{5.14}$$

The derivative of $\mathcal{L}$ (Eq. 5.5) with respect to the spatial prior parameters Θ^S in the

k-th prototype:

$$\begin{aligned}
\nabla_{\Theta_k^S}\left\{\mathcal{L}\right\} &= \nabla_{\Theta_k^S}\left\{-\log\left(\sum_{k=0}^{K} p(k|v;\Theta_k^S)\cdot p(y_{uvt}|k;\Theta_k^T)\right)\right\} \\
&= -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T}\frac{1}{\sum_k p(k|v).p(y_{uvt}|k)}.\nabla_{\Theta_k^S}\left\{\frac{\sum_k p(v|k).p(y_{uvt}|k))}{\sum_k p(v|k)}\right\} \\
&= -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T}\frac{1}{\sum_k p(k|v).p(y_{uvt}|k)} \\
&\quad\cdot\left\{\frac{p(v|k)\cdot p(y_{uvt}|k) - p(v|k)\sum_k p(k|v)p(y_{uvt}|k)}{\sum_k p(v|k)}\right\}\cdot\nabla_{\Theta_k^S}p(v|k) \\
&= -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T}\nabla_{\Theta_k^S}p(v|k)\cdot\{p(k|v) - p(k|v,y_{uvt})\} \qquad (5.15)
\end{aligned}$$

The derivative of (Eq. 5.15) with respect to the prototype location:

$$\nabla_{\Theta_k^\mu}\mathcal{L} = \sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(v|k)\cdot\Sigma_k^{-1}\cdot(r_v-\mu_k)\cdot\{p(k|v)-p(k|v,y_{uvt})\}\,. \qquad (5.16)$$

The derivative of (Eq. 5.15) with respect to the prototype shape and with constraint Σ_k is positive definite. We optimize I_k instead of Σ_k in which $\Sigma_k = L_k L_k^T$ (cholesky decomposition) and $I_k = L_k^{-1}$:

$$\begin{aligned}
\nabla_{\Theta_k^I}\mathcal{L} &= \sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(v|k)\cdot(I_k^{-1})^T\cdot I_k\cdot(r_v-\mu_k)\cdot(r_v-\mu_k)^T \\
&\quad\cdot\{p(k|v)-p(k|v,y_{uvt})\} \qquad (5.17)
\end{aligned}$$

We now show the derivatives of the (negative log) prior $\mathcal{P} = -\log p(\Theta^{STM})$:

Derivative of the HRF prior $p(\kappa_{k,p},\theta_{k,p}) \propto \exp^{\log((T^p-T^p_{min})(T^p_{max}-T^p))+\log((W-W_{min})(W_{max}-W))}$, where $W = 2\sqrt{2\ln 2} = 2.35$ and $\frac{W}{2} = 1.175$:

$$\begin{aligned}
\frac{\mathrm{d}\,\mathcal{P}}{\mathrm{d}\,\theta_{k,p}} &= -2\left[\frac{\frac{1}{2}(T^p_{max}+T^p_{min})-T^p}{(T^p-T^p_{min})(T^p_{max}-T^p)}\cdot\kappa_{k,p} - 1\right. \\
&\quad\left.+\frac{\frac{1}{2}(W_{max}+W_{min})-W}{(W-W_{min})(W_{max}-W)}\cdot 2.35\sqrt{\kappa_{k,p}}\right], \qquad (5.18)
\end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}\,\mathcal{P}}{\mathrm{d}\,\kappa_{k,p}} &= -2\left[\frac{\frac{1}{2}(T^p_{max}+T^p_{min})-T^p}{(T^p-T^p_{min})(T^p_{max}-T^p)}\cdot\theta_{k,p}\right.\\ &\left.+\frac{\frac{1}{2}(W_{max}+W_{min})-W}{(W-W_{min})(W_{max}-W)}\cdot\frac{1.175\theta_{k,p}}{\sqrt{\kappa_{k,p}}}\right] \qquad (5.19)\end{aligned}$$

Derivative of the spatial prior $p(\Sigma_k)=\frac{1}{|\Sigma_k|^2}$ with constraint Σ_k is positive definite. We optimize I_k instead of Σ_k in which $\Sigma_k=L_kL_k^T$ (cholesky decomposition) and $I_k=L_k^{-1}$:

$$\nabla_{\Theta_k^I}\mathcal{P}=4\frac{1}{I_k^T} \qquad (5.20)$$

Derivative of the noise parameters prior $p(\sigma_k^2)=\frac{1}{(\sigma_k^2)^2}$ with constraint $\sigma_k^2>0$: We suppose that $r_k=\log(\sigma_k^2)$

$$\frac{\mathrm{d}\,\mathcal{P}}{\mathrm{d}\,\sigma_k^2}=\frac{\mathrm{d}\,\mathcal{P}}{\mathrm{d}\,r_k}\cdot\frac{\mathrm{d}\,r_k}{\mathrm{d}\,\sigma_k^2}=2\frac{1}{\sigma_k^3} \qquad (5.21)$$

5.1.2 Initialization of the L1G-SMM-HPM

We have adopted a data-driven approach to initialize L1G-SMM-HPM. To initialize the number of the prototypes (that is, K) for a given ROI, we employ ''Consensus Clustering[1]" and proceed as follows:

1. Group-level functional clustering of fMRI time series using K-means methods for a given K. The clustering distance D between voxels v_1 and v_2 for a group of U subjects is defined as

$$D(v_1,v_2)=d(v_1,v_2)-\lambda\cdot\left(\frac{1}{U}\sum_{u=1}^{U}C_0^2\big(y_{u,v_1},y_{u,v_2}\big)\right). \qquad (5.22)$$

[1] Consensus clustering is a method to represent the consensus across multiple runs of a clustering algorithm (with random restart) by integrating the resulted clustering solutions. This method improve the stability and the robustness of the clustering algorithms.

where $d(\cdot,\cdot)$ denotes the Euclidean distance in the voxel space, $C_0^2\big(y_{u,v_1}, y_{u,v_2}\big)$ denotes zero-lag cross-correlation between the fMRI time series on voxels v_1 and v_2 for subject u, and $\lambda = 0.1$ is a tuning parameter. This results in a voxel-cluster configuration (with K clusters).

2. Use the resulting voxel-cluster configuration to construct the connectivity matrix $M^{(l)}$. The entries of $M^{(l)}$ are specified as:

$$M^{(l)}(i,j) = \begin{cases} 1 & \text{if items } i \text{ and } j \text{ belong to the same cluster,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.23}$$

3. Repeat this group-level functional clustering algorithm L times, each with an independent random restart (resampling). This results in L connectivity matrices, say $\{M^{(l)} : l = 1, 2, ...,\text{L}\}$.

4. Use these L connectivity matrices to construct a consensus matrix C storing for each pair of data items (in our case voxels' fMRI time series) the proportion of times in which these items are clustered together, that is,

$$C(i,j) = \frac{\sum_l M^{(l)}(i,j)}{L}, \tag{5.24}$$

As this consensus matrix is constructed with the number of clusters fixed to k, we denote it by C^k.

5. We reconstruct ten such consensus matrices with their k-values ranging from $k = 1$ to $k = 10$. The optimal number of prototypes k^{opt} is determined such that $C^{k^{\text{opt}}}$ is a perfect consensus matrix with entries equal to one or zero only [112] .

To initialize the spatial prior parameters $\Theta_S = \{\mu_k, \Sigma_k\}$ for the active prototypes within a given ROI, we performed re-clustering of the voxels in that ROI using an agglomerative hierarchical clustering algorithm. The similarity measure (with average linkage) used in this algorithm is $1 - C^{k^{\text{opt}}}$ where $C^{(k^{\text{opt}})}$ is the consensus matrix with the optimal number of the clusters, that is, k^{opt}. The agglomerative hierarchical clustering algorithm stops when the number of branches equals to K. The resulting sub-trees determine the cluster members. For each cluster, we fit a three-dimensional Gaussian distribution to the location of all voxels in this cluster and use its μ and Σ to initialize the spatial prior parameters of the corresponding prototype.

To initialize the HRF shape parameters $\Theta^{HRF} = \{\kappa_{k,p}, \theta_{k,p}\}$, the haemodynamic response magnitudes $\Theta_u^{NRL} = \{a_{u,k,p,s}\}$, and the noise parameter $\Theta^{NIS} = \{\sigma_k^2\}$ in active prototypes, we determine the most representative voxels for each active prototype k by ranking all voxels by $p(v|k)$. We take the first n voxels by rank with $\sum_{i=1}^{n} p(v_i|k) = 20\%$. Following this, fMRI data on these voxels Y^m are used to initialize the corresponding HPM model. We construct a grid of all permissible combinations of the values of HRF shape parameters (θ, κ), as seen in Fig. (5.1). HRF shape parameters (θ, κ) are permissible

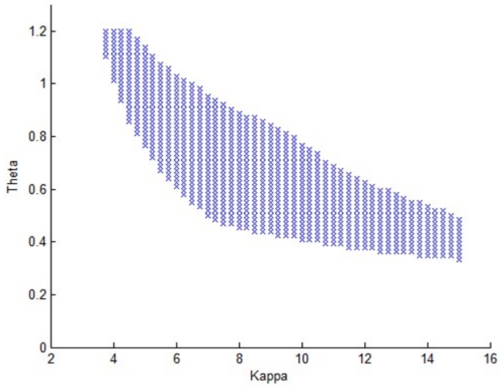

Figure 5.1: HRF shape parameters (θ, κ) permissible range grid

if the corresponding time-to-peak ($T = (\kappa - 1)\theta$) and peak width $\left(W = 2\sqrt{2\ln 2} \cdot \sqrt{\kappa}\theta\right)$ are both within their permissible ranges. The permissible range[1] are given by $[W_{min} = 3s, W_{max} = 6s]$ and $[T_{min} = 3s, T_{max} = 7s]$, respectively.

For each combination of HRF shape parameters (θ, κ) in the grid and using the fMRI data of the most representative voxels, we proceed as follows

[1] This permissible range has been artificially cut. We removed the narrow corners of the grid (increase the lower bound and decrease the upper bound) to make a finer grid with number of points that is sufficient for our experiment.

1. For each subject, the haemodynamic response magnitudes is computed by applying GLM. We define a regressor in the design matrix X for each pair of stimulus s and process p using the values of the HRF shape parameters. The resulting X is a matrix of size T $\times$ $P \cdot S$. The regression coefficient vector β_u contains all haemodynamic response magnitude parameters for subject u. A (least-squares) estimate of β_u is given by $\hat{\beta}_u = (X^T X)^{-1} X^T Y_u^m$ where Y_u^m is the fMRI data in the selected voxels of subject u.

2. For each subject, we computed the variance of the difference between the fMRI data Y_u^m of the most representative voxels and the estimated signal $\widehat{Y_u} = \beta \times X$ from the GLM method that was applied to the computation of the haemodynamic response magnitude. The noise is the mean of these variances.

3. Using the HRF shape parameters value that we have from the grid, the haemodynamic response magnitude parameter value that we have from step (1), and the noise parameter value that we have from step (2) , we optimize the HRF shape parameters, the haemodynamic response magnitude parameter, and the noise parameter iteratively by minimizing $\mathcal{L}$ in the same way as for the full model but using the HPM model.

Because the model that we fit here to the data is not a mixture model but a HPM, this makes the initializing and learning much simpler (when compared to the full model), which in turn allows us to use a large number of combination of HRF shape parameters (θ, κ) in the grid to initialize the HRF shape parameters and computing the corresponding haemodynamic response magnitude parameters and noise. The best solutions (initializations) for the HRF shape parameters, the haemodynamic response magnitudes parameters, and the noise parameter are the ones with the least $\mathcal{L}$.

For the null prototype, we initialize its parameters N, b, σ_0^2 as follow:

- N is initialized by the number of voxels within the ROI.

- To initialize b, for each subject, we compute the mean of the fMRI data of the least representative voxels Y_u^l. These voxels are again ranked by $p(v|k)$. This time we take the last n voxels in the rank with $\sum_{i=1}^{n} p(v_i|k) = 20\%$. b is initialized as the average of the means of the subject level fMRI data of the least representative voxels.

- To initialize the noise $\{\sigma_0^2\}$, for each subject, we compute the variance of the fMRI data of the least representative voxels Y_u^l. The noise $\{\sigma_0^2\}$ is initialized as the mean of the variance of the subject level fMRI data of the least representative voxels.

5.2 Second level multi-subject SMM-HPM: L2G-SMM-HPM

Compared to L1G-SMM-HPM, L2G-SMM-HPM allows different subjects to have different haemodynamic response shapes for each process and prototype. Variations in the individual HRF shape parameters $\Theta_u^{HRF} = \{\kappa_{u,k,p}, \theta_{u,k,p}\}$ are controlled by two factors: (1) a group level local spherical Gaussian $\mathcal{N}(\kappa_{u,k,p}, \theta_{u,k,p}|\mu_{k,p}^{\kappa}, \sigma_{k,p}^{2^{\kappa}}, \mu_{k,p}^{\theta}, \sigma_{k,p}^{2^{\theta}})$ and (2) the distribution $p(\kappa_{u,k,p}, \theta_{u,k,p})$ (see Eq. (4.15) in chapter (4)) controlling the admissible range of HRF shape parameters:

$$p(\Theta_u^{HRF}|\Theta^{HRF}) \propto \mathcal{N}(\kappa_{u,k,p}, \theta_{u,k,p}|\mu_{k,p}^{\kappa}, \sigma_{k,p}^{2^{\kappa}}, \mu_{k,p}^{\theta}, \sigma_{k,p}^{2^{\theta}}) \cdot p(\kappa_{u,k,p}, \theta_{u,k,p}). \tag{5.25}$$

The group-level HRF shape parameters Θ^{HRF} include

$$\Theta^{HRF} = \{\mu_{k,p}^{\kappa}, \sigma_{k,p}^{2^{\kappa}}, \mu_{k,p}^{\theta}, \sigma_{k,p}^{2^{\theta}}\},$$

where ${\mu^{\kappa}}_{k,p}$ and ${\sigma^{2^{\kappa}}}_{k,p}$ are the mean and the variance of the subject-level HRF shape parameters $\kappa_{u,k,p}$, respectively; and ${\mu^{\theta}}_{k,p}$ and ${\sigma^{2^{\theta}}}_{k,p}$ are the mean and the variance of the subject-level HRF scale parameters $\theta_{u,k,p}$, respectively.

Assuming that the observations are independent over subjects, voxels and volumes, the model likelihood of L2G-SMM-HPM given fMRI time series ($\mathbf{Y}$) of a group of subjects reads:

$$p(y) = \prod_{u=0}^{U}\prod_{v=0}^{V}\prod_{t=0}^{T} p\left(y_{uvt}; \Theta^{STM}\right) \tag{5.26}$$

$p\left(y_{uvt}; \Theta^{STM}\right)$ is modelled by a spatial mixture model:

$$\begin{aligned} & p\left(y_{uvt}; \Theta^{STM}\right) \\ = & \sum_{k=0}^{K} p(k|v; \Theta^{S}) \cdot p\left(y_{uvt}|k; \Theta^{T}\right) \\ = & \sum_{k=0}^{K} p(k|v; \Theta^{S}) \cdot p(y_{uvt}|k; \Theta_u^{NRL}, \Theta_u^{HRF}, \Theta^{NIS}) \end{aligned} \tag{5.27}$$

All the parameters $\{\Theta^S, \Theta_u^{NRL}, \Theta^{NIS}\}$ are the same as the parameters of L1G-SMM-HPM except for the HRF shape parameters. For active prototypes $k = 1, 2, ..., K$, the haemodynamic response shape function (HRF) (normalized gamma function) of each hidden cognitive process p, for each stimulus s is subject specific as:

$$\tilde{g}_{u,k,p}(t) = \left(\frac{t}{t_{max}}\right)^{\kappa_{u,k,p}-1} \exp\left(-\frac{t - t_{max}}{\theta_{u,k,p}}\right), \tag{5.28}$$

where $t_{max} = (\kappa_{u,k,p} - 1)\theta_{u,k,p}$.

5.2.1 Learning of the L2G-SMM-HPM

As in L1G-SMM-HPM, we learn the L2G-SMM-HPM parameters Θ^{STM} in a Bayesian manner (MAP estimation), by maximizing the posterior $p(\Theta^{STM}|\mathbf{Y})$. The difference is that in L2G-SMM-HPM the posterior is also maximized with respect to the haemodynamic response shape of each subject Θ_u^{HRF}. The model posterior

$$p(\Theta^{STM}|\mathbf{Y}) = p(\mathbf{Y}|\Theta^{STM}) \cdot p(\Theta^{STM}) \tag{5.29}$$

is calculated using the likelihood $\mathcal{L}$

$$\begin{aligned} & p\left(\mathbf{Y}|\Theta^{STM}\right) && (5.30)\\ = & \prod_{u=1}^{U}\prod_{v=1}^{V}\prod_{t=1}^{T}\frac{1}{\sum_{k=0}^{K}p(v|k;\Theta_{k}^{S})}\sum_{k=0}^{K}p(v|k;\Theta_{k}^{S})\cdot p(y_{uvt}|k;\Theta_{u,k}^{T})\\ = & \prod_{u=1}^{U}\prod_{v=1}^{V}\prod_{t=1}^{T}\frac{1}{\sum_{k=1}^{K}\mathcal{N}\left(v|\mu_{k},\Sigma_{k}\right)+\frac{1}{N}}\left\{\sum_{k=1}^{K}\mathcal{N}(v|\mu_{k},\Sigma_{k})\cdot\right.\\ & \left.\mathcal{N}\left(y_{uvt};x(t;a_{u,k,p,s},\kappa_{u,k,p},\theta_{u,k,p}),\sigma_{k}^{2}\right)+\frac{1}{N}\cdot\mathcal{N}\left(y_{uvt};b,\sigma_{0}^{2}\right)\right\}. \end{aligned}$$

The prior is factorized as:

$$\begin{aligned} & p(\Theta^{STM}) && (5.31)\\ = & p(N)\cdot p(b)\cdot\prod_{k=1}^{K}p(\mu_{k})\cdot\prod_{k=1}^{K}p(\Sigma_{k})\cdot\prod_{u=1}^{U}\prod_{k=1}^{K}\prod_{p=1}^{P}\prod_{s=1}^{S}p(a_{u,k,p,s})\\ & \cdot\prod_{u=1}^{U}\prod_{k=1}^{K}\prod_{p=1}^{P}p(\kappa_{u,k,p},\theta_{u,k,p})\cdot\prod_{u=1}^{U}\prod_{k=1}^{K}\prod_{p=1}^{P}p(\kappa_{u,k,p},\theta_{u,k,p}|\mu_{k,p}^{\kappa},\sigma_{k,p}^{2^{\kappa}},\mu_{k,p}^{\theta},\sigma_{k,p}^{2^{\theta}})\\ & \cdot\prod_{k=0}^{K}p(\sigma_{k}^{2}), \end{aligned}$$

where $p(\kappa_{u,k,p},\theta_{u,k,p}|\mu_{k,p}^{\kappa},\sigma_{k,p}^{2^{\kappa}},\mu_{k,p}^{\theta},\sigma_{k,p}^{2^{\theta}})=\mathcal{N}(\kappa_{u,k,p},\theta_{u,k,p}|\mu_{k,p}^{\kappa},\sigma_{k,p}^{2^{\kappa}},\mu_{k,p}^{\theta},\sigma_{k,p}^{2^{\theta}})$. All the other priors are the same as in L1G-SMM-HPM.

As with L1G-SMM-HPM, scaled conjugate-gradient optimization algorithms are applied to optimize L2G-SMM-HPM parameters iteratively. The gradients of model L2G-SMM-HPM with respect to the haemodynamic response magnitude, noise and spatial prior parameters are the same as in L1G-SMM-HPM. The difference is in the gradient of L2G-SMM-HPM with respect to the subject specific haemodynamic response shape parameters.

The derivatives of the (negative log) likelihood $\mathcal{L}=-\log\left(p(y_{uvt}|\Theta^{STM})\right)$ (Eq. 5.30) with respect to HRF scale parameter and with constraint $\theta_{u,k,p}>0$:
We suppose that $\alpha_{u,k,p}=\log(\theta_{u,k,p})$

$$\begin{aligned}\frac{\mathrm{d}\,\mathcal{L}}{\mathrm{d}\,\theta_{u,k,p}} &= \frac{\mathrm{d}\,\mathcal{L}}{\mathrm{d}\,\alpha_{u,k,p}} \cdot \frac{\mathrm{d}\,\alpha_{u,k,p}}{\mathrm{d}\,\theta_{u,k,p}} \\ &= -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(k|v, y_{uvt}) \cdot \frac{y_{uvt} - x_t^k}{\sigma_k^2} \cdot x_t^k \cdot \left[\frac{t}{\theta_{u,k,p}} - \kappa_{u,k,p} + 1\right],\end{aligned} \tag{5.32}$$

The derivatives of the (negative log) likelihood $\mathcal{L} = -\log\big(p(y_{uvt}|\Theta^{STM})\big)$ (Eq. 5.30) with respect to HRF shape parameter and with constraint $\kappa_{u,k,p} > 1$:
We suppose that $\beta_{u,k,p} = \log(\kappa_{u,k,p} - 1)$

$$\begin{aligned}\frac{\mathrm{d}\,\mathcal{L}}{\mathrm{d}\,\kappa_{u,k,p}} &= \frac{\mathrm{d}\,\mathcal{L}}{\mathrm{d}\,\beta_{u,k,p}} \cdot \frac{\mathrm{d}\,\beta_{u,k,p}}{\mathrm{d}\,\kappa_{u,k,p}} \\ &= -\sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(k|v, y_{uvt}) . \frac{y_{uvt} - x_t^k}{\sigma_k^2} \cdot x_t^k \cdot \log\frac{t}{(\kappa_{u,k,p} - 1)\theta_{u,k,p}}\end{aligned} \tag{5.33}$$

The model L2G-SMM-HPM has extra terms

$$\mathcal{P}_{u,k,p} = -\log\mathcal{N}(\kappa_{u,k,p}, \theta_{u,k,p}|\mu^{\kappa}_{k,p}, \sigma^{2^{\kappa}}_{k,p}, \mu^{\theta}_{k,p}, \sigma^{2^{\theta}}_{k,p})$$

in the HRF shape prior. The derivatives read:

$$\frac{\mathrm{d}\,\mathcal{P}_{u,k,p}}{\mathrm{d}\,\kappa_{u,k,p}} = \frac{\kappa_{u,k,p} - \mu^{\kappa}_{k,p}}{\sigma^{2^{\kappa}}_{k,p}} \cdot (\kappa_{u,k,p} - 1), \tag{5.34}$$

$$\frac{\mathrm{d}\,\mathcal{P}_{u,k,p}}{\mathrm{d}\,\theta_{u,k,p}} = \frac{\theta_{u,k,p} - \mu^{\theta}_{k,p}}{\sigma^{2^{\theta}}_{k,p}} \cdot (\theta_{u,k,p}) \tag{5.35}$$

5.2.2 Initialization of the L2G-SMM-HPM

In L2G-SMM-HPM, all the parameters are initialized as in L1G-SMM-HPM except the subject specific HRF shape parameters of the active prototypes. HRF shape parameters of individual subjects are initialized separately by applying the procedure described in Section 5.1.2 for initializing L1G-SMM-HPM. Group level HRF shape parameters $\Theta^{HRF} = \mu^{\kappa}_{k,p}, \sigma^{2^{\kappa}}_{k,p}, \mu^{\theta}_{k,p}, \sigma^{2^{\theta}}_{k,p}$ are then initialized as the mean and variance of those subject specific initial HRF shape parameters.

5.3 Third level multi-subject SMM-HPM: L3G-SMM-HPM

Compared to L2G-SMM-HPM, L3G-SMM-HPM further allows different subjects to have different spatial priors (prototype location $\mu_{u,k}$ and shape $\Sigma_{u,k}$). Variations in the haemodynamic response shapes and haemodynamic response magnitudes are modelled as in L2G-SMM-HPM. Variations in the prototype locations $\mu_{u,k}$ among the subjects are controlled by a group level Gaussian prior

$$p(\mu_{u,k}) = \mathcal{N}\big({\mu^{s}}_{k}, \sigma^{2^{s}}_{k}\big). \tag{5.36}$$

As for the prototype shape $\Sigma_{u,k}$, there are two factors contributing to the prior:

1. A group level Inverse Wishart distribution $\mathcal{IW}\big(\Psi_k, df_k\big)$ to prevent shrinking shapes to small regions ($\mathcal{IW}$ is a multivariate counterpart of inverse-gamma ($\mathcal{IG}$) distribution. For $\mathcal{IW}(x)$, $\mathcal{IW}$ penalizes both small and larger x.) Here, $\Psi_k = \Sigma_k \cdot (df_k - d^* - 1)$ is the scale matrix, where Σ_k is the mean prototype shape of $\Sigma_{u,k}$, df_k is the degree of freedom (in our case the number of subjects), and $d^*= 3$ is the voxel space dimensionality.

2. Subject specific Jeffrey's priors $\frac{1}{|\Sigma^2_{u,k}|}$ to prevent extending shape to large regions.

Hence, the prior for the prototype shape has the form

$$p(\Sigma_{u,k}) \propto \mathcal{IW}\big(\Psi_k, df_k\big) \cdot \frac{1}{|\Sigma^2_{u,k}|} \tag{5.37}$$

To summarize, the group level spatial parameters Θ^S are:

$$\Theta^S = \{\mu^s_k, \sigma^{2^s}_k, \Psi_k\}.$$

Assuming independent observations over subjects, voxels and volumes, the formula of model L3G-SMM-HPM for modelling the fMRI time series ($\mathbf{Y}$) of group of subjects:

$$p(Y) = \prod_{u=0}^{U}\prod_{v=0}^{V}\prod_{t=0}^{T} p\left(y_{uvt}; \Theta^{STM}\right), \tag{5.38}$$

where $p\left(y_{uvt}; \Theta^{STM}\right)$ is modelled by a spatial mixture model:

$$p\left(y_{uvt}; \Theta^{STM}\right) = \sum_{k=0}^{K} p(k|v; \Theta^S_u) \cdot p\left(y_{uvt}|k; \Theta^T_u\right) \tag{5.39}$$

$$= \sum_{k=0}^{K} p(k|v; \Theta^S_u) \cdot p(y_{uvt}|k; \Theta^{NRL}_u, \Theta^{HRF}_u, \Theta^{NIS})$$

L3G-SMM-HPM has the same parameters of L2G-SMM-HPM, except the spatial prior parameters, which are subject specific $\Theta^S_u = \{\mu_{u,k}, \Sigma_{u,k}\}$ (location and shape of prototype k at subject u). That means that $p(k|v; \Theta^S_u)$ denotes the probability that the k-th prototype generates the fMRI time series $\mathbf{Y}$ in voxel v for subject u.

$$p(k|v; \Theta^S_u) = \frac{p(v|k; \Theta^S_{u,k})}{\sum_{k=0}^{K} p(v|k; \Theta^S_{u,k})}, \tag{5.40}$$

For active prototypes $k = 1, 2, ..., K$, $p(v|k)$ modelled as a multivariate Gaussian:

$$p(v|k) = \mathcal{N}(r_v|\mu_{u,k}, \Sigma_{u,k}), \tag{5.41}$$

5.3.1 Learning of the L3G-SMM-HPM

As in L1G-SMM-HPM and L2G-SMM-HPM, we learn the L3G-SMM-HPM parameters Θ^{STM} in a Bayesian manner (MAP estimation) by maximizing the posterior $p(\Theta^{STM}|\mathbf{Y})$. The difference is that in L3G-SMM-HPM the posterior is also maximized with respect to the spatial prior of each subject Θ^S_u. The model posterior

$$p(\Theta^{STM}|\mathbf{Y}) = p(\mathbf{Y}|\Theta^{STM}) \cdot p(\Theta^{STM}), \tag{5.42}$$

where the likeliood and the prior are specified as follow:

The likelihood $\mathcal{L}$:

$$\begin{aligned} & p\left(\mathbf{Y}|\Theta^{STM}\right) \\ = & \prod_{u=1}^{U}\prod_{v=1}^{V}\prod_{t=1}^{T}\frac{1}{\sum_{k=0}^{K} p(v|k;\Theta^S_{u,k})}\sum_{k=0}^{K} p(v|k;\Theta^S_{u,k}) \cdot p(y_{uvt}|k;\Theta^T_{u,k}) \\ = & \prod_{u=1}^{U}\prod_{v=1}^{V}\prod_{t=1}^{T}\frac{1}{\sum_{k=1}^{K}\mathcal{N}\left(v|\mu_{u,k},\Sigma_{u,k}\right) + \frac{1}{N}}\left\{\sum_{k=1}^{K}\mathcal{N}(v|\mu_{u,k},\Sigma_{u,k})\cdot\right. \\ & \left.\mathcal{N}\left(y_{uvt};x(t;a_{u,k,p,s},\kappa_{u,k,p},\theta_{u,k,p}),\sigma_k^2\right) + \frac{1}{N}\cdot\mathcal{N}\left(y_{uvt};b,\sigma_0^2\right)\right\}. \end{aligned} \tag{5.43}$$

The prior is factorized as:

$$\begin{aligned} & p(\Theta^{STM}) \\ = & p(N)\cdot p(b)\cdot\prod_{u=1}^{U}\prod_{k=1}^{K} p(\mu_{u,k}|{\mu^s}_k,\sigma_k^{2^s})\cdot\prod_{u=1}^{U}\prod_{k=1}^{K} p(\Sigma_{u,k}) \\ & \cdot\prod_{u=1}^{U}\prod_{k=1}^{K} p(\Sigma_{u,k}|\Psi_k,df_k)\cdot\prod_{u=1}^{U}\prod_{k=1}^{K}\prod_{p=1}^{P}\prod_{s=1}^{S} p(a_{u,k,p,s}) \\ & \cdot\prod_{u=1}^{U}\prod_{k=1}^{K}\prod_{p=1}^{P} p(\kappa_{u,k,p},\theta_{u,k,p})\cdot\prod_{u=1}^{U}\prod_{k=1}^{K}\prod_{p=1}^{P} p(\kappa_{u,k,p},\theta_{u,k,p}|\mu^\kappa_{k,p},\sigma^{2^\kappa}_{k,p},\mu^\theta_{k,p},\sigma^{2^\theta}_{k,p}) \\ & \cdot\prod_{k=0}^{K} p(\sigma_k^2), \end{aligned} \tag{5.44}$$

where $p(\mu_{u,k}|\ {\mu^s}_k,\ {\sigma^{2^s}}_k) = \mathcal{N}\big(\mu_{u,k}|{\mu^s}_k,\sigma_k^{2^s}\big)$ and $p(\Sigma_{u,k}|\Psi_k,df_k) = \mathcal{IW}\big(\Sigma_{u,k}|\Psi_k,df_k\big)$. All the other priors are the same as in L2G-SMM-HPM.

Scaled conjugate-gradient optimization algorithms are employed to optimize L3G-SMM-HPM parameters iteratively. The gradients of L3G-SMM-HPM with respect to the haemodynamic response magnitudes, HRF shape, and noise parameters are the same as in L2G-SMM-HPM. The difference is in the gradient of L3G-SMM-HPM with respect to the subject specific spatial prior parameters.

The derivatives of the (negative log) likelihood $\mathcal{L} = -\log p(y_{uvt}|\Theta^{STM})$ (Eq. 5.43) with respect to spatial prior location:

$$\nabla_{\Theta^{\mu}_{u,k}} \mathcal{L} = \sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(v|k) \cdot \Sigma_{u,k}^{-1} \cdot (r_v - \mu_{u,k}) \cdot \{p(k|v) - p(k|v, y_{uvt})\}\,, \tag{5.45}$$

The derivatives of the (negative log) likelihood $\mathcal{L} = -\log p(y_{uvt}|\Theta^{STM})$ (Eq. 5.43) with respect to spatial prior shape and with constraint $\Sigma_{u,k}$ is positive definite. We optimize $I_{u,k}$ instead of $\Sigma_{u,k}$ in which $\Sigma_{u,k} = L_{u,k}L_{u,k}^{T}$ (cholesky decomposition) and $I_{u,k} = L_{u,k}^{-1}$:

$$\begin{aligned}\nabla_{\Theta^{I}_{u,k}} \mathcal{L} \;&=\; \sum_{u=1}^{U}\sum_{v=1}^{V}\sum_{t=1}^{T} p(v|k) \cdot (I_{u,k}^{-1})^{T} \cdot I_{u,k} \cdot (r_v - \mu_{u,k}) \cdot (r_v - \mu_{u,k})^{T} \\ &\quad \cdot \{p(k|v) - p(k|v, y_{uvt})\}\end{aligned} \tag{5.46}$$

Compare to L2G-SMM-HPM, L3G-SMM-HPM has extra terms in the spatial prior: $\mathcal{P}1_{u,k} = -\log \mathcal{N}\big(\mu_{u,k}|{\mu^{s}}_{k}, \sigma_{k}^{2^{s}}\big)$ and $\mathcal{P}2_{u,k} = -\log \mathcal{IW}\big(\Sigma_{u,k}|\Psi_{k}, df_{k}\big)$.

The derivative of $\mathcal{P}1_{u,k} = -\log \mathcal{N}\big(\mu_{u,k}|{\mu^{s}}_{k}, \sigma_{k}^{2^{s}}\big)$ with respect to $\mu_{u,k}$:

$$\nabla_{\Theta^{\mu}_{u,k}} \mathcal{P}1_{u,k} \;=\; \frac{\mu_{u,k} - \mu_{k}^{s}}{\sigma_{k}^{2^{s}}}. \tag{5.47}$$

The derivative of $\mathcal{P}2_{u,k} = -\log \mathcal{IW}\big(\Sigma_{u,k}|\Psi_{k}, df_{k}\big)$ with respect to $\Sigma_{u,k}$ and with constraint $\Sigma_{u,k}$ is positive definite. We optimize $I_{u,k}$ instead of $\Sigma_{u,k}$ in which $\Sigma_{u,k} = L_{u,k}L_{u,k}^{T}$

(cholesky decomposition) and $I_{u,k} = L_{u,k}^{-1}$:

$$\nabla_{\Theta^I_{u,k}} \mathcal{P}2_{u,k} = -(df_k + d^* + 1) \cdot (I_{u,k}^{-1})^T - (\Psi_k) \cdot I_{u,k}, \qquad (5.48)$$

5.3.2 Initialization of the L3G-SMM-HPM

For Model L3G-SMM-HPM, all of model parameters are initialized as in L2-SMM-HPM, except for the spatial parameters. Spatial parameters of individual subjects are first initialized separately by applying the procedure employed to initialize L1-SMM-HPM (described in Section 5.1.2). Based on these individual estimates we initialise the group-level prior for the spatial parameters.

Individual prototype locations are considered as random samples from a group level prior on the location vectors. The location prior is given as a spherical Gaussian distribution specified by $\{\mu_k^s, \sigma_k^2\}$ for prototype k. The group-level hyperparameters μ_k^s and σ_k^2 are computed as the empirical mean and variance of the individually estimated location vectors.

Similarly, individual prototype covariance matrices are considered random samples from a group level prior given as an Inverse Wishart distribution specified by $\{\Psi_k, df_k\}$ for prototype k. Its hyperparameters $\Theta^S = \{\Psi_k, df_k\}$ are initialized as follow: df is the degree of freedom resembling the degree of freedom of a student's t test distribution that sets the certainty of the prior. The initialization value of df is equal to the number of subjects. Ψ is the scale matrix describing the position of the Inverse Wishart distribution in the parameter space in which the average of the subjects-specific covariance matrices $\Sigma_{u,k}$. is equal to $\frac{\Psi}{df-d^*-1}$, and hence $\Psi = \Sigma_k \cdot (df - d^* - 1)$. To initialize the scale matrix we first need to average the subjects-specific covariance matrices $\Sigma_{u,k}$ and then multiply the average Σ_k by $(df - d^* - 1)$. Since the covariance matrices live on the Riemannian manifold of (semi-)definite matrices, the mean of $\Sigma_{u,k}$ has to be computed e.g. using the method of Smake and Kawanabe [113] that robustly estimates the mean of covariance matrices by minimizing the Beta divergence iteratively by

$$\Sigma_k^{(j+1)} = \frac{\sum_{u=1}^{U} \psi_\rho(\Sigma_{u,k}; \Sigma_k^j, \nu)\Sigma_{u,k}}{\nu \sum_{u=1}^{U} \psi_\rho(\Sigma_{u,k}; \Sigma_k^j, \nu) - \gamma|\Sigma_k^j|^{\frac{(\nu-d^*-1)\rho}{2}}} \tag{5.49}$$

$\Sigma_k^{(j+1)}$: prototype shape matrices mean at $(j+1)$ for prototype k.

Σ_k^j: prototype shape matrices mean at j. We initialize it by the group level prototype shape that we computed using our clustering method described in section (5.1.2).

$\Sigma_{u,k}$: prototype shape of subject u in prototype k .

ν: a fraction (1/20) of the number of samples $(\Sigma_{u,k})$.

ρ: tuning parameters takes a value from $\{0, 2^{-20}, ..., 2^1, 1\}$.

d^*: matrix dimension.

$\psi_\rho(\Sigma_{u,k}; \Sigma_k^j, \nu)$ defined as:

$$\psi_\rho(\Sigma_{u,k}; \Sigma_k, \nu) = |\Sigma_{u,k}|^{\frac{(\nu-d^*-1)\rho}{2}} exp\left\{-tr\left(\frac{\rho}{2}\Sigma_k^{-1}\Sigma_{u,k}\right)\right\} \tag{5.50}$$

γ defined as:

$$\begin{aligned}\gamma = &\frac{U\rho(d^*+1)}{2^{\frac{\nu d^*}{2}}\Gamma_{d^*}(\frac{\nu}{2})(\rho+1)}\left(\frac{2}{\rho+1}\right)^{\frac{\nu d^*(\rho+1)}{2}-\frac{d^*(d^*+1)\rho}{2}} \\ &\times\Gamma_{d^*}\left(\frac{\nu(\rho+1)}{2}-\frac{(d^*+1)\rho}{2}\right),\end{aligned} \tag{5.51}$$

where Γ_{d^*} is multivariate Gamma function defined as:

$$\Gamma_{d^*}(X) = \pi^{\frac{d^*(d^*-1)}{4}} \prod_{j=1}^{d^*} \Gamma\left[X + \frac{(1-j)}{2}\right] \tag{5.52}$$

In each iteration, we compute the difference between $\Sigma_k^{(j+1)}$ and Σ_k^j:

$$D_k^{d^*} = \frac{\sum_{l=1}^{d^*}\sum_{m=1}^{d^*}|\Sigma_{k,l,m}^{(j+1)} - \Sigma_{k,l,m}^j|}{{d^*}^2} \tag{5.53}$$

If the difference $D_k^{d^*}$ is sufficiently small, we stop the method and $\Sigma_k^{(j+1)}$ is the resulting mean of subject-level prototype shape $\Sigma_{u,k}$.

5.4 Results and discussion

To validate the three group-level spatio-temporal fMRI models from the hierarchy developed in the previous section, we have conducted extensive numerical experiments using both synthetic and real fMRI data. The real fMRI data were generated by a joint behavioural and fMRI experiment studying how humans learn probabilistic sequential structures encoded in visual stimulus sequences. Based on the analysis of those behavioural data, a cohort of participants can be categorized into two subgroups, namely groups of fast and slow learners.

5.4.1 Synthetic data

In this validation experiment, we generated synthetic fMRI data sets of 18 virtual subjects. Each data set emulates fMRI data from a single ROI, because the methodology developed here is tailored for ROI-based analysis rather than whole-brain analysis. Such virtual ROIs consists of 1000 voxels arranged on a three-dimensional regular grid (i.e., 10 × 10 × 10). This size is comparable with that of a large ROI. The synthetic fMRI time series on individual voxels were generated using a two-prototype SMM-HPM model as follow:

1. For each of the two prototypes, generate the fMRI signal $x_k(t)$, $k \in \{1, 2\}$, for the k-th prototype with the corresponding HPM.

2. For each voxel v, compute the corresponding weight distribution $p(k|v)$.

3. Generate synthetic fMRI time series for subject u, voxel v and time t as

$$y_{uvt} = x_{k^*}(t) + \epsilon(t),$$

where prototype index k^* is the k-value of the prototype with the highest $p(k|v)$ value.

For L1G-SMM-HPM, we assume two prototypes with distinct HPMs representing two separate neural activations. The spatial prior of these two prototypes are computed using Eq. (4.3) and Eq.(4.4) with $\mu_1 = (3, 5, 5)$, $\mu_2 = (7, 5, 5)$, and $\Sigma_1 = \Sigma_2 = 1.5 \cdot \mathbf{I}_3$, where $\mathbf{I}$ denotes an identity matrix. The two HPM models are set up as follows:

- The haemodynamic responses were evoked by a sequence of 50 stimuli with inter-stimulus interval (ISI) equal to 3.0 time units.

- Each of these stimuli triggers two virtual cognitive processes, which are separated in time by 1.5 time units.

- These two processes evoke haemodynamic responses with distinct HRFs. The HRF shape parameters $\{\kappa_{k,p}, \theta_{k,p}\}$ for the two processes in the two prototypes are parametrized as $\kappa_{1,1} = \kappa_{2,1} = 4.7348, \theta_{1,1} = \theta_{2,1} = 1.0431, \kappa_{1,2} = \kappa_{2,2} = 18.6742, \theta_{1,2} = \theta_{2,2} = 0.3409$. These values of the κ and θ give two quite different HRFs for the two prototypes.

- The haemodynamic response of process p evoked by stimulus s is modelled as the product of the HRF shape function g_p and the response magnitude $a_{u,k,p,s}$ using Eq.(4.9).

- The data was generated by regularly measuring the fMRI signal at a frequency of two volumes per time unit. This yields 300 fMRI volumes.

We assume that the shape function is constant in time. Thus, the HRF shape parameters are the same for all stimuli. The variation in the haemodynamic response across stimuli come from the variation in the response magnitude.

We generated different haemodynamic response magnitudes for individual subjects while keeping everything else fixed across the subjects. The haemodynamic response magnitude a for subjects $u = 1, 2, ..., 18$ as function of stimulus $s = 1, ..., 50$ for process $p = 1, 2$ in prototype $k = 1, 2$ is defined as:

$$a_{u,k,p,s} = h_p * f_{u,k}\left(\frac{2\pi}{8} \cdot s \cdot \text{ISI} + \delta_u\right) * i_k, \tag{5.54}$$

where (1) h denotes the maximum response magnitude. Its value is set to 1 for process 1 and 0.8 for process 2, that is, $h_1 = 1$ and $h_2 = 0.8$. For each of the two processes, the h value remains unchanged across subjects, prototypes and stimuli; (2) f specifies how the response magnitudes evolve over time, in the form of a sine function or a unit square function ('square'); Further, we divide 18 subjects into three subgroups, namely $\{u$: $f_{u,1} = f_{u,2} =$ 'sine'$\}$, $\{u$: $f_{u,1} = f_{u,2} =$ 'square'$\}$ and $\{u$: $f_{u,1} =$ 'sine', $f_{u,2} =$ 'square'$\}$. (3) $\delta \in [0, \pi)$ and $i \in \{1, -1\}$ together specify the phase shift of f. Note that δ varies randomly across the subjects, while i differs between the two prototypes.

Note that for L1G-SMM-HPM, we keep both HRF shape parameters and SMM location and spread of the two prototypes fixed across the 18 subjects.

The synthetic data of L2G-SMM-HPM has been generated in the same way as the generation of the data of L1G-SMM-HPM, except that HRF shape varies from subject to subject. Thus, we sample the HRF shape parameters for each subject from a group-level prior on these parameters given as $\mathcal{N}(\kappa_{u,k,p}, \theta_{u,k,p} | \mu^{\kappa}_{k,p}, \sigma^{2^{\kappa}}_{k,p}, \mu^{\theta}_{k,p}, \sigma^{2^{\theta}}_{k,p})$ using Eq. (5.55).

$$\begin{aligned} \kappa_{u,k,p} &= \mu^{\kappa}_{k,p} + \sigma^{2^{\kappa}}_{k,p} \cdot \chi, \\ \theta_{u,k,p} &= \mu^{\theta}_{k,p} + \sigma^{2^{\theta}}_{k,p} \cdot \chi, \end{aligned} \tag{5.55}$$

where χ is a sample from the standard normal distribution (zero mean and unit variance normal distribution); $\mu^{\kappa}_{1,1} = \mu^{\kappa}_{2,1} = 4.7348, \mu^{\theta}_{1,1} = \mu^{\theta}_{2,1} = 1.0431, \mu^{\kappa}_{1,2} = \mu^{\kappa}_{2,2} = 18.6742, \mu^{\theta}_{1,2} = \mu^{\theta}_{2,2} = 0.3409$ (these values give two quite different HRFs for the two prototypes); and $\sigma^{\kappa}_{1,1} = \sigma^{\kappa}_{1,2} = \sigma^{\kappa}_{2,1} = \sigma^{\kappa}_{2,1} = 0.25, \sigma^{\theta}_{1,1} = \sigma^{\theta}_{1,2} = \sigma^{\theta}_{2,1} = \sigma^{\theta}_{2,1} = 0.025$. We choose σ^{κ} bigger than σ^{θ} because the range of κ is bigger than the range of θ.

The synthetic data of L3G-SMM-HPM has been generated in the same way of the generation of the data of L2G-SMM-HPM except that the spatial prior varies across the subjects. This means that for each of the two prototypes, location and spread (that is mean vector and covariance matrix) differs between the subjects. We sampled the prototype location of each subject from the group level prototype location prior given as $\mathcal{N}\big(\mu_{u,k}|{\mu^s}_k, \sigma_k^{2^s}\big)$ with $\mu_1^s = (3,5,5)$, $\mu_2^s = (7,5,5)$, and $\sigma_1^{2^s} = \sigma_2^{2^s} = 0.01$. Similarly, we sampled individual prototype covariance matrices from their group level prior given as an Inverse Wishart distribution prior $\mathcal{IW}\big(\Sigma_{u,k}|\Psi_k, df_k\big)$ with $\Psi_k = 1.5 \cdot \mathbf{I}_3 \cdot (df - d^* - 1)$.

For each level of model, the generated data have been divided into a training set and a test set: training set to learn model parameters Θ^{STM}, and test set to validate the models.

Synthetic data experiments results

To examine how accurate these three levels multi-subject fMRI models can be learned from the data, we conducted an extensive numerical experiment using synthetic data. Fig.(5.2) shows the synthetic data experiment. The design of this validation experiment is given as follows:

1. Primarily, we infer each of the three models from the data generated by the same model, for example, we learn L1G-SMM-HPM model parameters from the synthetic data generated by L1G-SMM-HPM. But we also examine how well each of the three models can be learned from the data when there exists discrepancy between the inferential and data-generating model (so-called model misfit).

2. To quantify how well the inferred model fits the data, we use out-of-sample negative log likelihood [1]. To this end, we split each synthetic data set into a training set and a testing set. The training set is used to learn model parameters Θ^{STM}, while the testing is used to calculate the negative log likelihood. The original data set is split according to both voxels and volumes.

[1] We used the natural log

3. To account for the uncertainty arising from the optimization process, we repeat the experiment[1] with ten independent random run (each run with different split of the data into training-set and test-set) and obtain ten measurements of the out-of-sample negative log likelihood. The mean and standard deviation are subsequently computed.

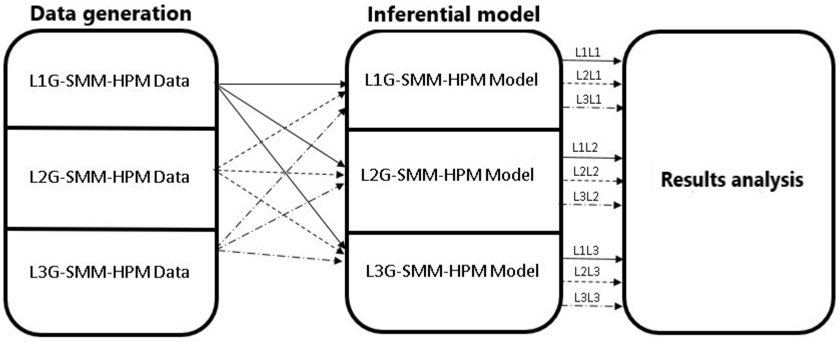

Figure 5.2: Controlled experiments using synthetic fMRI data.

The results for out-of-sample negative log likelihood (mean ± standard deviation) is summarized in table 5.1. It shows that for any of the three data sets, the lowest out-of-sample negative log likelihood is always observed in the case where we do not have the model misfit (the diagonal in the Table 5.1). The same is true for any of the three inference models (see each of the three columns in the table). Moreover, in term of both the bias and the variance, the negative log likelihood of all the models with all the datasets is low, which show how robust our learning method is.

Table 5.1: Out of sample negative log Likelihood. The best results are marked with bold font.

	L1G-SMM-HPM	L2G-SMM-HPM	L3G-SMM-HPM
L1G-SMM-HPM data	**1.1934** ± [0.0398]	1.3275 ±[0.0555]	1.3494 ± [0.0464]
L2G-SMM-HPM data	1.3357 ± [0.0401]	**1.2302** ±[0.0416]	1.3546 ± [0.0429]
L3G-SMM-HPM data	1.3997 ± [0.0312]	1.3761 ±[0.0466]	**1.2905** ± [0.0518]

In addition to out-of-sample negative log likelihood, we also used other quantities to

[1] One experiment takes on average two days to produce the results.

test the learning performance. Such measures are based on the absolute difference between the ground truth and the estimated model parameters In the following, we describe the definition of these measures in detail.

The accuracy of spatial prior parameters (μ_k,Σ_k) for prototype k was measured through the symmetrized Kullback–Leibler divergence between the ground-truth spatial prior multivariate Gaussian distributions $\mathcal{N}_k^g(\mu_k^g, \Sigma_k^g)$ and the estimated one $\mathcal{N}_k^e(\mu_k^e, \Sigma_k^e)$ using Eq. (4.24). The accuracy of the subject-level spatial prior parameters ($\mu_{u,k}$,$\Sigma_{u,k}$) is measured through:

$$\begin{aligned} A_{S_{u,k}} &= \operatorname{Tr}\Big(\frac{\big(\Sigma_{u,k}^g\big)^{-1}\Sigma_{u,k}^e + \big(\Sigma_{u,k}^e\big)^{-1}\Sigma_{u,k}^g}{2}\Big) \\ &+ (\mu_{u,k}^e - \mu_{u,k}^g)^\intercal \frac{\big(\Sigma_{u,k}^e\big)^{-1} + \big(\Sigma_{u,k}^g\big)^{-1}}{2}(\mu_{u,k}^e - \mu_{u,k}^g) - 3, \end{aligned} \tag{5.56}$$

where the average error of the spatial prior estimation: $A_S = \frac{1}{U \cdot K}\sum_u \sum_k A_{S_{u,k}}$.

The accuracy of the haemodynamic response shape parameters $\kappa_{k,p}$ and $\theta_{k,p}$ for prototype k and process p was measured through the L_1 distance between the ground truth HRF $g_{k,p}^g$ and the estimated HRF $g_{k,p}^e$ using Eq. (4.25). Subject-level haemodynamic response shape parameters $\kappa_{u,k,p}$ and $\theta_{u,k,p}$ accuracy is measured through:

$$A_{g_{u,k,p}} = \frac{1}{n}\sum_{i=1}^{n}\Big|g_{u,k,p}^g(i\Delta t) - g_{u,k,p}^e(i\Delta t)\Big|, \tag{5.57}$$

where n is the number of sample points ($n = 2000$) and $\Delta t = 0.01$, and the overall error of the haemodynamic response shape estimation: $A_g = \frac{1}{U \cdot K \cdot P}\sum_u \sum_k \sum_p A_{g_{u,k,p}}$.

The accuracy of the subject level haemodynamic response magnitude ($a_{u,k,p,s}$) estimation was measured through two summary statistics:

(i) L_1 difference between the ground truth and the estimated response magnitudes:

$$A_{a_{u,k,p}} = \frac{1}{S}\sum_{s=1}^{S}\Big|a_{u,k,p,s}^g - a_{u,k,p,s}^e\Big| \tag{5.58}$$

Where S is the number of stimuli. The average error of response magnitude estimation is given by $A_a = \frac{1}{U \cdot K \cdot P} \sum_u \sum_k \sum_p A_{a_{u,k,p}}$;

(ii) Zero-lag cross correlation between the estimated time series of haemodynamic response magnitudes of the two prototypes for specific process, denoted by eC_0^p. Due to the way the synthetic data is generated, the ground truth value of gC_0^p is -1:

$$eC_0^p = \frac{1}{u} \sum_{u=1}^{U} eC_0^{u,p}\big(a_{u,k_1,p}, a_{u,k_2,p}\big) \tag{5.59}$$

where $eC_0^{u,p}(\cdot,\cdot)$ denotes the estimated value of zero-lag cross-correlation between the haemodynamic response magnitude a of subject u for particular process p in the the two prototypes. The average value of zero-lag cross-correlation $eC_0 = \frac{1}{P} \sum_p eC_0^p$.

Table 5.2 display the performance of parameters estimation in the inferential model L1G-SMM-HPM from the data generated by L1G-SMM-HPM, L2G-SMM-HPM, and L3G-SMM-HPM (Row 2 to Row 4). Tables 5.3 and 5.4 show the same results but for the inferential model L2G-SMM-HPM and L3G-SMM-HPM, respectively. For all the models, both the bias and the variance are low, which show how accurate and robust our model optimization is. As expected, in each table the highest accuracy of the parameter learning is observed when there is a match between the dataset and the inferential model. However, if we look carefully at the results in these tables 5.2, 5.3 and 5.4, we can see that there is no big difference between the results when there is a match between the dataset and the inferential model (Bold entries) and the results when there is a miss-match between the dataset and the inferential model, but the consistency of that the highest accuracy is observed when there is a matching is reassuring (given that our model is a complex latent variable model operating on quite limited noisy observations). Furthermore, it is also notable to mention that although the standard deviations are small, they are sometimes larger than the means. This is maybe because we repeated the experiments only ten times and maybe there are outliers but with this small number of measurements (10 measurements), it is hard to claim that these are real outliers.

Table 5.2: L1G-SMM-HPM parameters estimation performance. The best results are marked with bold font.

Data	A_S	A_g	eC_0	A_a
L1-Data	**0.0044**±[0.0121]	**0.0254**±[0.0381]	**-0.9185**±[0.0223]	**0.0406**±[0.0415]
L2-Data	0.0063±[0.0315]	0.0267±[0.0367]	-0.8751±[0.0559]	0.0458±[0.0758]
L3-Data	0.0078±[0.0271]	0.0319±[0.0588]	-0.8721±[0.0451]	0.0461±[0.0594]

Table 5.3: L2G-SMM-HPM parameters estimation performance. The best results are marked with bold font.

Data	A_S	A_g	eC_0	A_a
L1-Data	0.0053±[0.0640]	0.0326±[0.0546]	-0.8785±[0.0479]	0.0489±[0.0778]
L2-Data	**0.0049**±[0.0417]	**0.0224**±[0.0319]	**-0.8976**±[0.0571]	**0.0434**±[0.0645]
L3-Data	0.0079±[0.0981]	0.0373±[0.0736]	-0.8724±[0.0321]	0.0467±[0.0743]

Table 5.4: L3G-SMM-HPM parameters estimation performance. The best results are marked with bold font.

Data	A_S	A_g	eC_0	A_a
L1-Data	0.0072±[0.0784]	0.0331±[0.0325]	-0.8752±[0.0442]	0.0491±[0.0567]
L2-Data	0.0066±[0.0465]	0.0346±[0.0341]	-0.8732±[0.0329]	0.0475±[0.0141]
L3-Data	**0.0060**±[0.0648]	**0.0318**±[0.0554]	**-0.8769**±[0.0541]	**0.0451**±[0.0546]

5.4.2 Real data

Two-session fMRI data of 21 participants (mean age = 21 years) were used in this work. These data are taken from a fMRI study investigating how humans learn probabilistic sequential structures [114].

To investigate the humans' sequence learning, two types of probabilistic sequences of different complexity level were generated (labeled as Level 0 or Level 1 sequences). The process underlying the Level 0 sequences is an i.i.d. process and the probability distribution used to specify this i.i.d. process is a multinomial distribution over symbols from an alphabet ({A, B, C, D} in this study). Note that it is a memory-less process. In contrast, the process underlying Level 1 sequences is a first-order Markov process. Each symbol in this process is a random sample from a multinomial distribution conditional on its previous symbol. Therefore, a memory structure of length 1 is introduced into those Level 1 sequences.

The first session fMRI data set was acquired before any training, whereas the second one was acquired after the participants had been trained with both Level 0 and Level 1 sequences. They are referred as pre- and post training sessions, respectively. Each fMRI session comprised nine runs. Each run included ten blocks with two trials per block, and two fixation blocks at the beginning and the end of each run. In each trial, a sequence of 10 symbols was presented to the participants in the screen center one at a time. Each symbol is shown for 250ms followed by a white fixation dot for 250ms. At the end of each trial, a response cue appeared on the screen before a test comprising 4 stimuli appeared for 1.5s. Participants were asked to predict which symbol they expected to appear next. After the response of the participants by pressing the key corresponding to the symbol location on the screen, a white fixation dot appeared for 5.5s before the next trial. All trials except fixation trials involve three processes: (1) a visual analysis process (2) a perceptual judgement process and (3) a motor response process. The fMRI data sets were acquired at the Birmingham University Imaging Centre with a 3-T Philips Achieva MRI scanner. In each scanning session, Echo Planar Imaging (EPI) data were acquired from 32 slices (whole brain coverage, TR: 2000 ms, TE: 35 ms, $2.5 \times 2.5 \times 4$ mm resolution).

Each participant from the cohort involved in this study can be categorized either as fast or slow learners based on their behavioural performance[1]. This results in two subgroups of the cohort: a fast learner and a slow learner group. Alongside the fMRI data, we also obtained a group of identified ROIs with statistically significant three-way interactions between session (pre versus post training), structure (random guess versus probabilistically structured sequences), and learning (fast versus slow learners). However, not all activation patterns shown by these ROIs are related to learning. Therefore, they are further divided into two subgroups of ROIs: one with a statistically significant shift of Percent signal change (PSC) from pre to post session and the one without it. For this work, we have choose four ROIs (MFG, SFG, CG, and Pu) from the first group and three ROIs (MOG, IOG, and LiG) from the second one. The last three ROIs are used as controls for a sanity check.

[1] Rui Wang, Yuan Shen, Peter Tino, A. Welchman, Z. Kourtzi, Learning predictive statistics: dynamics and strategies, Journal of Vision, Accepted for publication.

The first ROI group consists of Middle Frontal Gyrus (MFG) of 480 voxels, Superior Frontal Gyrus (SFG) of 349 voxels, Cingulate Gyrus (CG) of 134 voxels, and Putamen (Pu) of 44 voxels. The second group consists of Medial Occipital Gyrus (MOG) of 175 voxels, Inferior Occipital Gyrus (IOG) of 448 voxels, and Limbic Gyrus (LiG) of 303 voxels.

Real data experiments results

We applied our hierarchical multi-subject SMM-HPM model on the fMRI data of each group separately, in order to examine the ability of our model in (1) jointly describing multiple fMRI data sets from a precisely defined group of subjects and (2) in discriminating between different groups of subjects based on their fMRI data. For each ROI, we applied our models (L1G-SMM-HPM, L2G-SMM-HPM, and L3G-SMM-HPM) to the fMRI data of two different groups, fast learners and slow learners. We repeated the experiment[1] ten times, with random independent initialization in each repetition (the results shown below are the mean along with the standard deviation ± across the ten repetitions of the experiment for each model).

In order to find the optimal model, for each model in the hierarchy, we compute the out-of-sample negative log Likelihood (both spatially and temporally). Fig.(5.3) shows the real data experiment.

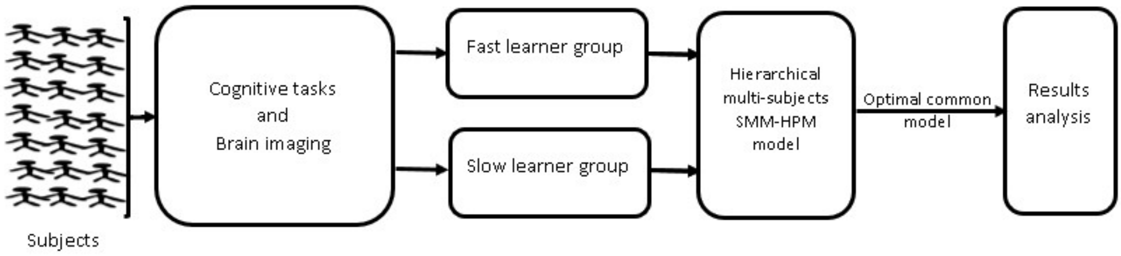

Figure 5.3: Real data experiment using real fMRI data of one session and one ROI .

The results show that there is no different between these three level models in the

[1] One experiment takes on average two days to produce the results.

hierarchy but the second-level model have the lowest out-of-sample negative log likelihood. In our experiment, we use the second-level multi-subject SMM-HPM: L2G-SMM-HPM for discriminating between the fast learners group GrF and slow learners group GrS within each ROI using three different features: spatial feature, temporal feature, and spatio-temporal feature (actually, the results of the extracted features are consistent over the three levels). We derive a summary statistic from those features. For the spatial feature and the temporal feature, for each of four session-prototype pairs, we plot this statistic (mean ± std). Recall that 'session' could be either pre- or post training, while 'prototype' is indexed either by 1 or 2. For the spatio-temporal features, for each session (either pre- or post learning), we plot the statistics (mean ± std). To examine the statistical significance of the result of each feature, we use a t-test[1] and a rank test (Wilcoxon rank test) [2]. To examine the effect of the learning, we compute the relative percent reduction (RPR)[3] in the p-value of the pre-learning session and post-learning session.

$$RPR = \frac{P_{pre} - P_{post}}{P_{pre}} \cdot 100, \qquad (5.60)$$

where P_{pre} is the p-value of pre-learning session, and $P_{post}$is the p-value of post-learning session. It is worth mentioning that there is no limit on the value of the RPR results (it can be very large or very small). In general, positive results means that there is an effect for the learning (learning increase the separation between the two groups), and negative results means that there is no effect for the learning (learning does not increase the separation between the two groups).

Spatial feature To perform the outlined analysis based on the spatial features, we use the so-called prototype volume as a summary statistic of the spatial prior for those prototypes inferred individually for different subjects, groups, and/or ROIs. For

[1] Student's-t test is used to compare the mean of two normally distributed samples, preferably of equal size and variance.

[2] The rank test (Wilcoxon) is used to compare the median of tow samples without any assumption about the samples distribution (it is a nonparametric test which is based solely on the order in which the observations from the two samples fall)

[3] We computed the relative difference (relative percent reduction (RPR)) instead of the absolute difference because the absolute difference is less informative in the case of small p-values.

each of these prototypes, its volume is computed as a product of the eigenvalues of the corresponding shape matrix (Covariance matrix of the multivariate Gaussian distribution that describe the spatial prior).

The computed prototype volumes are displayed in Fig.(5.4) for MFG, SFG, CG, Pu and in Fig.(5.4) for IOG, MOG, LiG. In Table(5.5), for the frontal ROIs (MFG and SFG), we found that there existed a statistically significant difference in prototype volume between fast and slow learners for the two prototypes with larger spatial extent (shaded cells in Table(5.5)). RPR values for the large prototypes (Bold entries in Table(5.5)) show that there is an effect for the learning in increasing the separation between the two groups in term of the volume of the prototypes. Moreover, from Fig.(5.4), we see that for the larger prototypes (prototype 1 for MFG and prototype 2 for SFG) in both frontal ROIs, the fast learners prototype volumes are on average larger than those of slow learners across the sessions. For all prototypes in the small ROIs (CG and Pu) as well as for the small prototype in MFG (prototype 2) and small prototype in SFG (prototype 1), such difference is insignificant. The above observation suggests that we may explain away the observed difference by large size of those prototypes and/or ROI. To test this suggestion, we performed the same analysis for the three control ROIs. Note that the neural activation of these ROIs is not related to the learning. Even though these ROIs are large and the prototypes in them also have large spatial extent, there is no statistically significant difference in volume between fast and slow learns (see Fig.(5.5) and Table(5.6)). As a result, we may now claim that fast learners have larger homogeneous sub-ROIs (prototypes) than slow learners and this difference is related to learning.

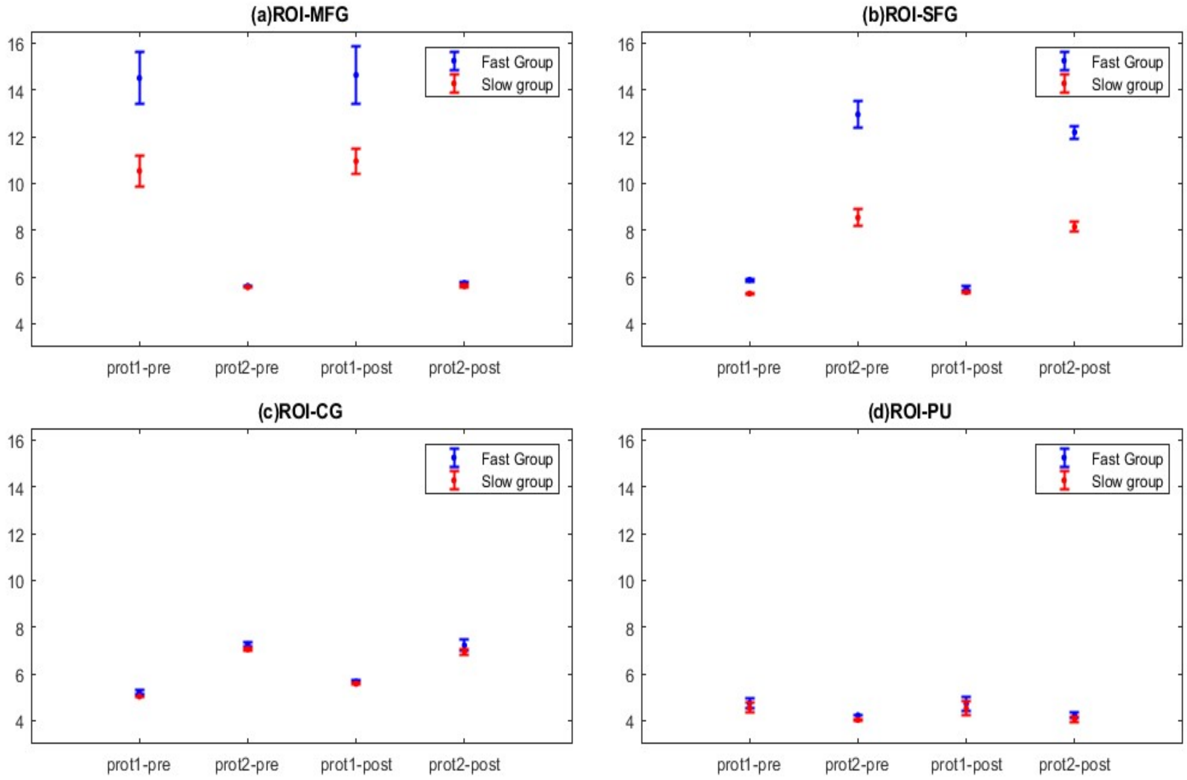

Figure 5.4: The volume of the two prototypes (prot1 and prot2) on the interesting ROIs ((a) ROI-MFG, (b) ROI-SFG, (c) ROI-CG, and (d) ROI-Pu) for both the fast learners (blue) and slow learners (red) groups in the pre-learning (pre) and post-learning (post) sessions.

Table 5.5: Prototypes volume statistics for interesting ROIs (ROI-MFG, ROI-SFG, ROI-CG, and ROI-Pu) for the two prototypes (prot1 and prot2) in the pre-learning (pre) and post-learning (post) sessions

	Pre-learning p-value		Post-learning p-value		RPR	
	Prot-1	Prot-2	Prot-1	Prot-2	Prot-1	Prot-2
ROI-MFG t-test	0.0085	0.1976	0.0069	0.2908	**19**%	-47%
ROI-MFG rank test	0.0207	0.3147	0.003	0.2241	**86**%	29%
ROI-SFG t-test	0.0587	2.30E-06	0.0613	1.10E-09	-4%	**99**%
ROI-SFG rank test	0.1712	0.0003	0.1841	0.0002	-7%	**33**%
ROI-CG t-test	0.51	0.7693	0.6743	0.4911	-32%	36%
ROI-CG rank test	0.9231	1	0.7362	0.951	20%	5%
ROI-Pu t-test	0.3252	0.0891	0.3503	0.5175	-8%	-481%
ROI-Pu rank test	0.1447	0.0778	0.4376	1	-202%	-1e+03%

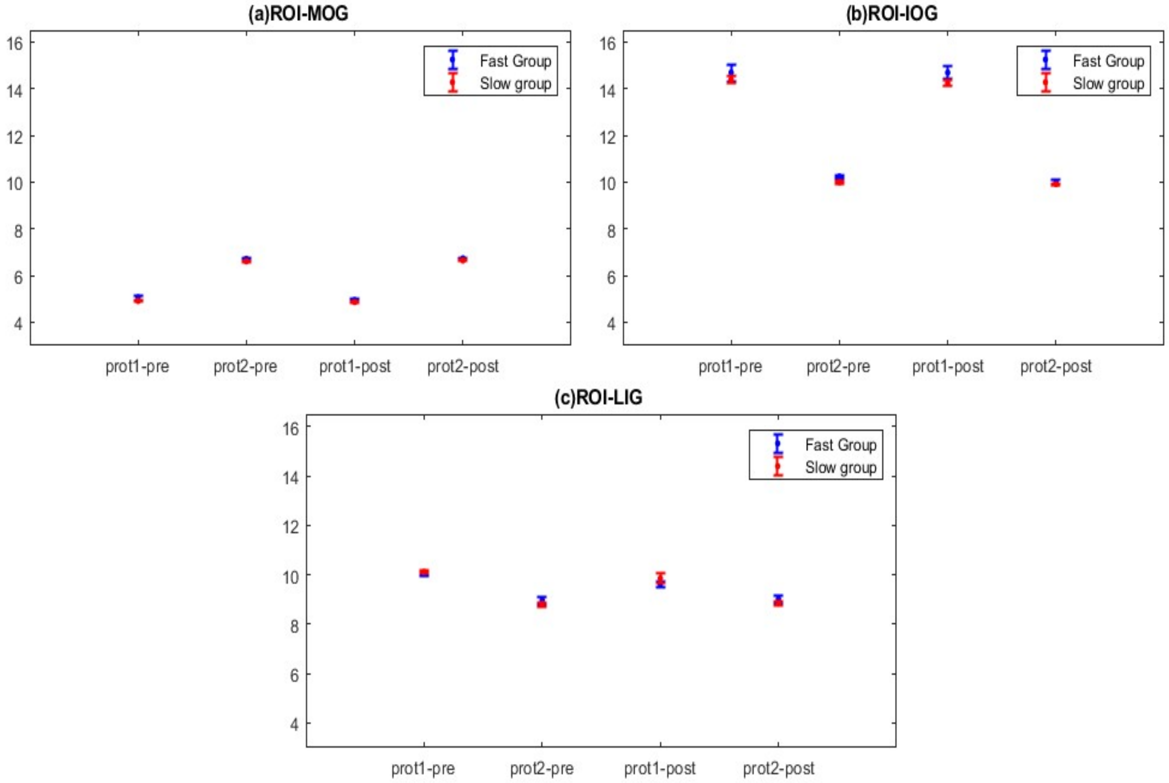

Figure 5.5: The volume of the two prototypes (prot1 and prot2) on the control ROIs ((a) ROI-MOG, (b) ROI-IOG, and (c) ROI-LiG) for both the fast learners (blue) and slow learners (red) groups in the pre-learning (pre) and post-learning (post) sessions.

Table 5.6: Prototype volume statistics for control ROIs (ROI-MOG, ROI-IOG, and ROI-LiG) for the two prototypes (prot1 and prot2) in the pre-learning (pre) and post-learning (post) sessions

	Pre-learning p-value		Post-learning p-value		RPR	
	Prot-1	Prot-2	Prot-1	Prot-2	Prot-1	Prot-2
ROI-MOG t-test	0.3106	0.1571	0.1746	0.5512	44%	-251%
ROI-MOG rank test	0.3747	0.1323	0.1031	0.1257	72%	5%
ROI-IOG t-test	0.2438	0.2746	0.2782	0.8378	-14%	-205%
ROI-IOG rank test	0.4363	0.2224	0.1963	0.9314	55%	-319%
ROI-LiG t-test	0.0927	0.6829	0.7664	0.5793	-727%	15%
ROI-LiG rank test	0.1304	0.5743	1	0.6126	-667%	-7%

Temporal feature From the estimated HRF parameters, we reconstructed the haemodynamic response time to peak $T^p = (\kappa_{k,p_2} - 1) \cdot \theta_{k,p_2}$ to quantify how fast the response is. This statistic was computed only for the perceptual judgement process (i.e. Process 2 (p_2) in the model) because it is the process of most interest. Fig. (5.6) shows that for all interesting ROIs, fast learners have earlier time to peak response than slow learners after the training session. This is statistically significant (shaded cells in Table (5.7)). Moreover, RPR values show that in general there is an effect on the learning in increasing the separation between the fast and slow learners in terms of their haemodynamic response time to peak (Bold entries in Table (5.7)). As in the spatial features, to test this suggestion, we performed the same analysis for the three control ROIs. The results (Fig(5.7) and shaded cells and Bold entries in Table (5.8)) show that the same significant results appeared in the control ROIs. As a consequence, we may claim that fast learners have earlier time to peak response than slow learners but this difference is not related to learning.

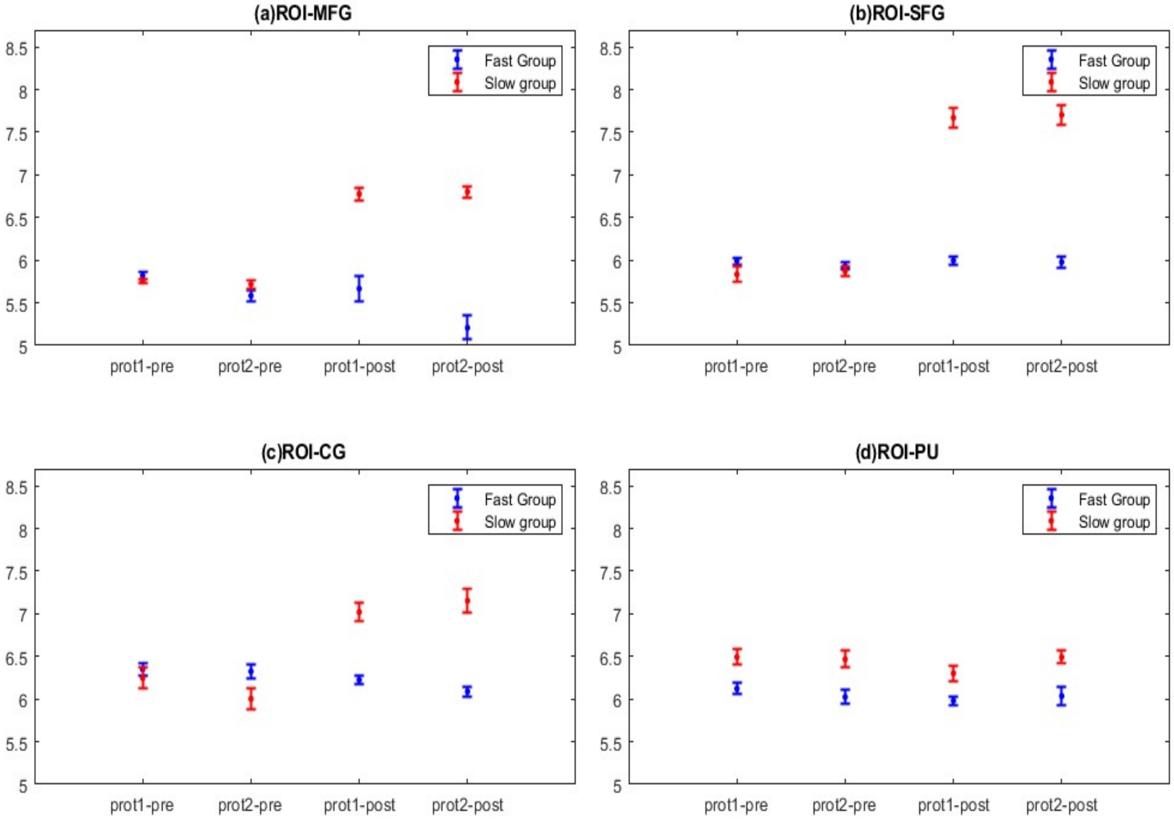

Figure 5.6: Haemodynamic response time to peak of the perceptual judgement process for the two prototypes (prot1) and (prot2) on the interesting ROIs ((a) ROI-MFG, (b) ROI-SFG, (c) ROI-CG, and (d) ROI-Pu) for both the fast learners (blue) and slow learners (red) groups in the pre-learning (pre) session and post-learning (post) session.

Table 5.7: HRF time to peak statistics for interesting ROIs (ROI-MFG, ROI-SFG, ROI-CG, and ROI-Pu) for the two prototypes (prot1 and prot2) in the pre-learning (pre) and post-learning (post) sessions

	Pre-learning p-value		Post-learning p-value		RPR	
	Prot-1	Prot-2	Prot-1	Prot-2	Prot-1	Prot-2
ROI-MFG t-test	0.2189	0.0006	7.80E-11	2.81E-20	**100%**	**100%**
ROI-MFG rank test	0.14	0.0114	1.39E-08	2.04E-12	**100%**	**100%**
ROI-SFG t-test	0.1311	0.1445	3.80E-36	1.12E-31	**100%**	**100%**
ROI-SFG rank test	0.1352	0.1211	2.30E-22	2.90E-20	**100%**	**100%**
ROI-CG t-test	0.4754	0.0045	0.0003	1.20E-06	**99%**	**99%**
ROI-CG rank test	0.7273	0.0052	0.0008	1.50E-05	**99%**	**99%**
ROI-Pu t-test	0.0363	0.0141	0.0465	0.0017	-28%	**88%**
ROI-Pu rank test	0.0075	0.0184	0.0255	0.0032	-240%	**83%**

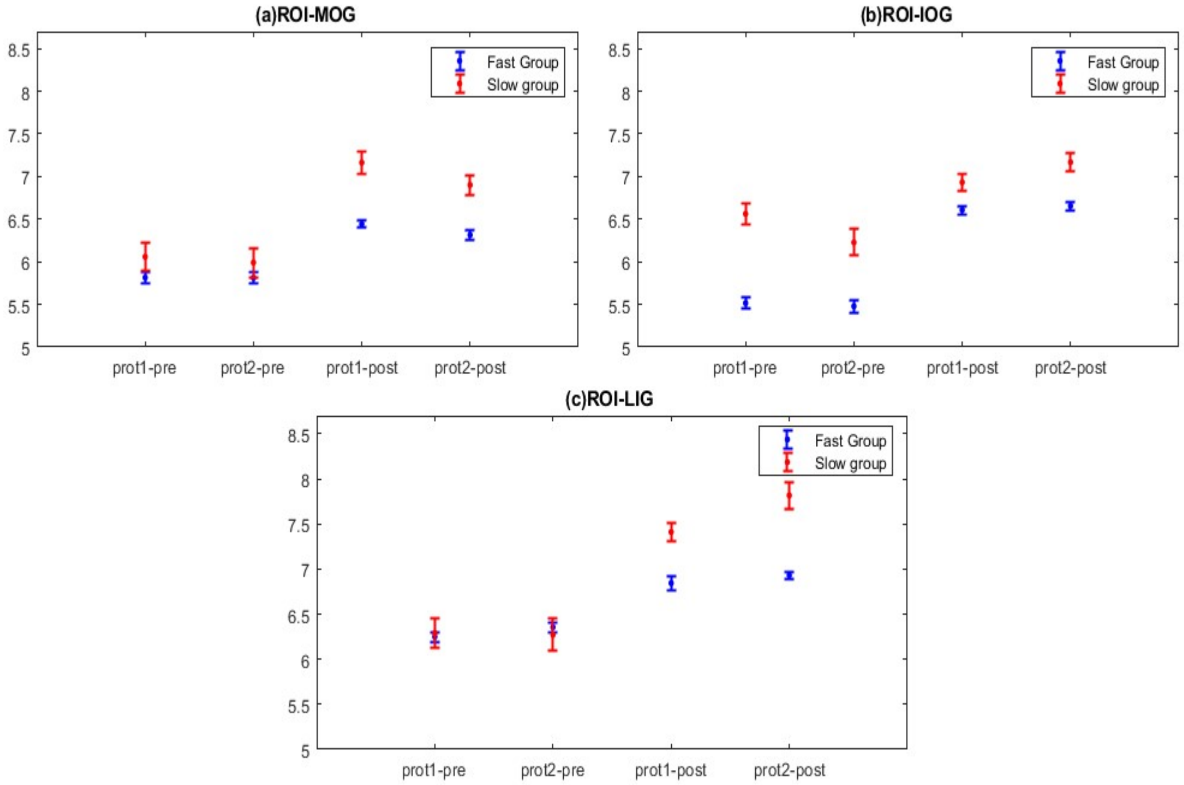

Figure 5.7: Haemodynamic response time to peak of the perceptual judgement process for the two prototypes (prot1) and (prot2) on the control ROIs ((a) ROI-MOG, (b) ROI-IOG, and (c) ROI-LiG) for both the fast learners (blue) and slow learners (red) groups in the pre-learning (pre) session and post-learning (post) session.

Table 5.8: HRF time to peak statistics for control ROIs (ROI-MOG, ROI-IOG, and ROI-LiG) for the two prototypes (prot1 and prot2) in the pre-learning (pre) and post-learning (post) sessions

	Pre-learning p-value		Post-learning p-value		RPR	
	Prot-1	Prot-2	Prot-1	Prot-2	Prot-1	Prot-2
ROI-MOG t-test	0.1326	0.0654	2.64E-10	3.00E-06	**100%**	**99%**
ROI-MOG rank test	0.1734	0.1124	4.31E-07	5.91E-06	**99%**	**99%**
ROI-IOG t-test	9.56E-08	0.0034	1.67E-02	1.30E-03	-1.7e+07%	**62%**
ROI-IOG rank test	2.22E-07	0.048	1.27E-02	3.00E-04	-5.7e+06%	**99%**
ROI-LiG t-test	0.6269	0.6990	1.40E-03	1.66E-07	**99%**	**100%**
ROI-LiG rank test	0.967	0.5412	4.38E-06	2.92E-10	**99%**	**100%**

Spatio-temporal feature which is described by zero lag cross-correlation of temporal evolution of response magnitudes of the two prototypes k_1 and k_2 for a particular process (perceptual judgement process p_2, which we are interested in) within a specific ROI:

$$C_0^2 = \frac{1}{u} \sum_{u=1}^{U} C_0^{2_u}(a_{u,k_1,p_2}, a_{u,k_2,p_2}) \tag{5.61}$$

High cross-correlation indicates that the ROI is homogeneous and one prototype (along with the null prototype) is sufficient for characterising that ROI. Low cross-correlation means that the ROI is heterogeneous and there is a need for more than one prototype. The computed correlation coefficients are displayed with error bars (mean ± std). The results show that in the all interesting ROIs (Fig. (5.8)), the two prototypical patterns of response magnitudes are more positively correlated for the slow learners group than for the fast learners group in the after training session. This observation is statistically significant for all ROIs except the smallest ROI-Pu (shaded cells in Table (5.9)). Moreover, RPR values show that there is an effect of the learning in increasing the separation between the two groups in term of the zero lag cross-correlation of temporal evolution of response magnitudes of the two prototypes in all ROIs except the smallest one (Bold entries in Table (5.9)). As in the previous features, to test this suggestion, we performed the same analysis for the three control ROIs. The results (Fig(5.9) and Table (5.10)) show that there is no statistically significant difference in the term of the cross-correlation of temporal evolution of response magnitudes between fast and slow learns. Moreover, negative results in the RPR means that there is no effect for the learning and learning does not increase the separation between the two groups (this is what we expect for the control ROIs). As a consequence, we may claim that fast learners have more heterogeneous ROIs than slow learners and this difference is related to learning.

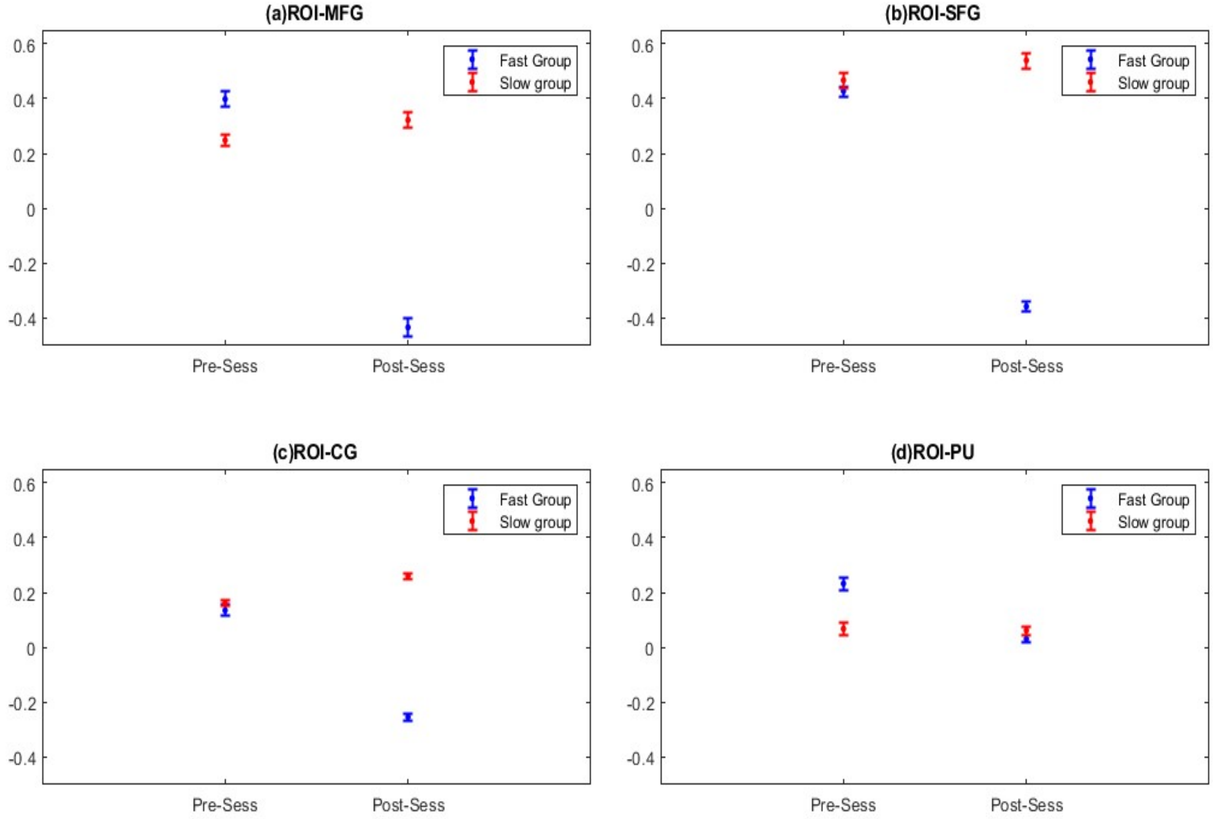

Figure 5.8: Zero-lag cross correlation between the estimated haemodynamic response magnitudes time series of the two prototypes on the interesting ROIs ((a) ROI-MFG, (b) ROI-SFG, (c) ROI-CG, and (d) ROI-Pu)) for both the fast learners (blue) and slow learners (red) groups in the pre-learning session (Pre-Sess) and post-learning (Post-Sess) session.

Table 5.9: Statistics of the zero-lag cross correlation between the estimated haemodynamic response magnitudes time series of the two prototypes for interesting ROIs (ROI-MFG, ROI-SFG, ROI-CG, and ROI-Pu)

	Pre-learning p-value	Post-learning p-value	RPR
ROI-MFG t-test	0.0079	1.38E-05	**99**%
ROI-MFG rank test	0.0056	2.11E-06	**99**%
ROI-SFG t-test	0.2774	2.74E-09	**100**%
ROI-SFG rank test	0.1136	3.18E-09	**100**%
ROI-CG t-test	0.4161	1.05E-03	**99**%
ROI-CG rank test	0.5812	3.47E-03	**99**%
ROI-Pu t-test	0.0019	0.205	-1e+04%
ROI-Pu rank test	0.001	0.1734	-17240%

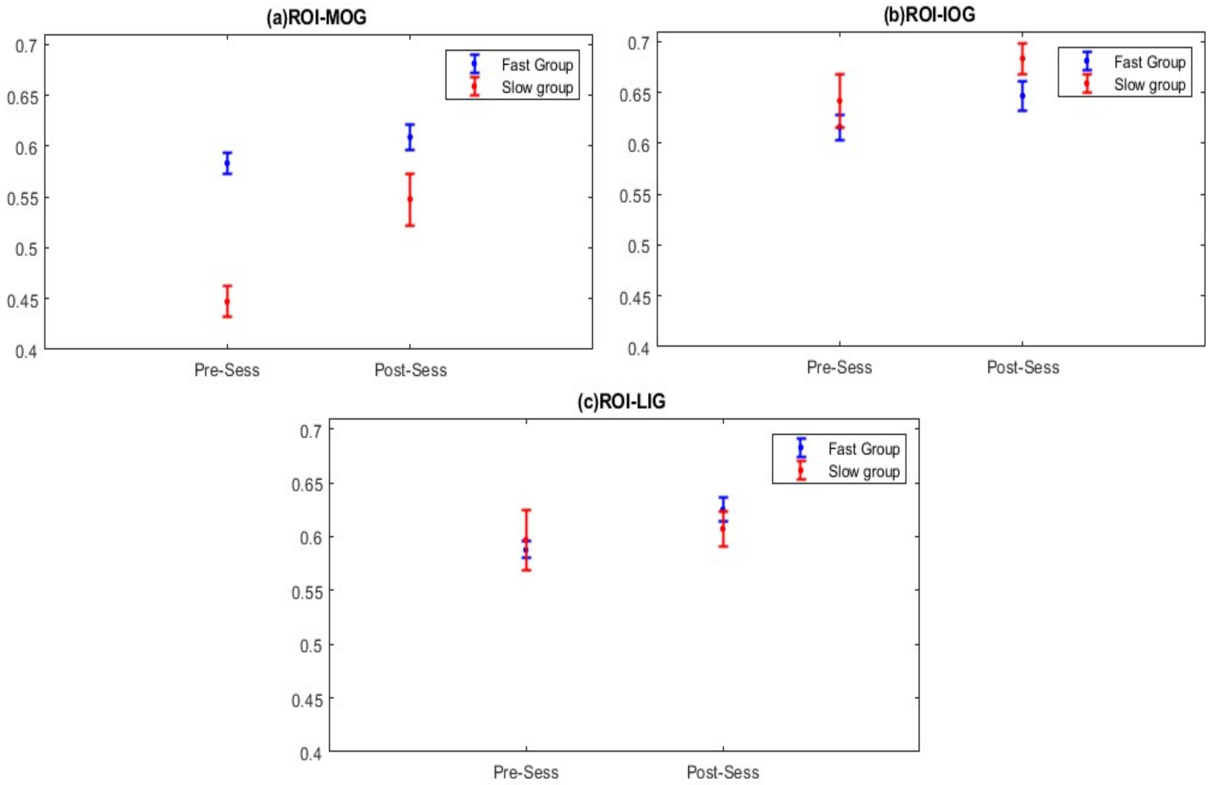

Figure 5.9: Zero-lag cross correlation between the estimated haemodynamic response magnitudes time series of the two prototypes on the control ROIs ((a) ROI-MOG, (b) ROI-IOG, and (c) ROI-LiG) for both the fast learners (blue) and slow learners (red) groups pre-learning session (Pre-Sess) and post-learning (Post-Sess) session.

Table 5.10: Statistics of the zero-lag cross correlation between the estimated haemodynamic response magnitudes time series of the two prototypes for control ROIs (ROI-MOG, ROI-IOG, and ROI-LiG)

	Pre-learning p-value	Post-learning p-value	RPR
ROI-MOG t-test	0.0151	0.3765	-2e+03%
ROI-MOG rank test	0.0352	0.7601	-2e+03%
ROI-IOG t-test	0.5241	0.5612	-7%
ROI-IOG rank test	0.2118	0.4878	-130%
ROI-LiG t-test	0.7077	0.4526	36%
ROI-LiG rank test	0.269	0.3828	-42%

5.5 Summary

In this chapter, we developed the multi-subject version of the SMM-HPM as a hierarchical model formations (L1G-SMM-HPM, L2G-SMM-HPM, and L3G-SMM-HPM). Such hierarchical formations enabled us to identify the optimal common model (the one that has the lowest negative log likelihood) that can describe any population and can discriminate between different populations.

In the synthetic data experiments, both the out-of-sample negative log likelihood, and the absolute difference between the ground truth and the estimated model parameters show how robust and accurate our model is.

Our multi-subject SMM-HPM is a prototype based spatio-temporal model. This fact enabled us to extract three novel features (a spatial feature, a temporal feature, and a spatio-temporal feature) and use them to discriminate between different groups of subjects. In the real data experiments, the results of extracting these features for each groups show that the temporal features, and the spatio-temporal features can be used to discriminate between different populations. However, the spatial features can be used only in the case of large ROIs.

CHAPTER 6

CONCLUSION

This chapter presents the general conclusions of the work presented in this thesis and suggests several directions for future work.

6.1 Thesis Reflections

This thesis was a result of our quest to find the optimal multi-subject fMRI data model that can describe the fMRI data of any population. The scientific method consists of asking questions, and trying to systematically work towards the answer. In this thesis we introduced the following research questions:

1. How can the idea of the population-based fMRI data model be formulated?

2. What is the most constrained model that still can describe the population based fMRI data and what can be learnt from it?

To answer these questions, we formulated (in chapter 5) the multi-subject fMRI data model as a hierarchy of model formations, from the most constrained model to the most flexible one. The models in this hierarchy differ in which degree the spatio-temporal features of fMRI data are allowed to vary between subjects. Such hierarchical formations enabled us to identify the optimal common model (the one that has the lowest out-ofsample negative log likelihood). From the optimal common model, we can extract informative features and use them to discriminate between the fMRI data of different populations.

6.2 Work summary

The main contribution of this thesis is to extend the single-subject SMM-HPM model to a population-based one in a principled way (a hierarchy of model formations with increasing complexity in each level of the hierarchy), and to define the optimal common model that can discriminate between fMRI data of different populations.

The first step of extending single-subject SMM-HPM to multi-subject is to normalize the HRF model (that is, the gamma function with two shape parameters κ and θ). This modification is essential for building a population based model because the variations of κ and θ not only lead to variations in the haemodynamic response shape but also to variations in its peak height. Normalization of the HRF shape to one with unit peak height can make sure that all variations in the intensity of haemodynamic response are solely captured by the haemodynamic response magnitude parameters. In Chapter (4), we demonstrate through numerical experiments that such a modification not only constitutes a more natural model formulation, but also makes the parameter estimation more robust.

Then, we proceeded to the multi-subject extension of the single-subject SMM-HPM and have developed a hierarchy of multi-subject SMM-HPM models ranging from the most constrained model, where the subjects share all the model parameters except the heamodynamic response magnitudes (L1G-SMM-HPM), to the most flexible one, where the subjects have different individual parameters controlled by group-level priors (L3G-SMM-HPM). The intermediate level of multi-subject SMM-HPM (that is, L2G-SMM-HPM) allows for variation of the HRF shape but keeps the spatial prior fixed. Such hierarchical formations enabled us to identify the optimal common model using the out-of-sample negative log likelihood, and to examine the impact of the model flexibility on identification of spatio-temporal patterns for a given population.

To validate these three multi-subject models in the hierarchical framework, we conducted an extensive numerical experiment using synthetic data (section 5.4.1). We first used these models as data-generating models to generate three corresponding synthetic datasets with known ground-truth model setting. Then, each of these three models are used as inferential models to fit all of the three synthetic datasets individually. The quality of model fitting is tested by (1) the out-of-sample negative log likelihood and (2) the absolute difference between the ground truth and the estimated model parameters. For (2), we used (i) a symmetrized Kullback-Leibler divergence between two Gaussian distributions (that is, the ground-truth spatial prior and the estimated one), (ii) a L1 distance between the ground truth HRF and the estimated HRF, (iii) a L1 difference between the ground truth and the estimated response magnitudes, and (iv) a zero-lag cross correlation between the estimated time series of haemodynamic response magnitudes of the two prototypes. To quantify the uncertainty arising from the parameter estimation, we repeated the experiment with ten independent, random initialisation and summarize all results in mean ± standard deviation, These experiments demonstrates how robust and accurate our model is.

To assess the performance of these three multi-subject models of the hierarchical framework in describing multiple fMRI data and in discriminating between different groups of subjects based on their fMRI data, we applied them to real data comprises the fMRI data of two different groups of learners (fast learners and slow learners) from two different sessions (pre-learning session and post-learning session) on specific ROIs (section 5.4.2). As in the synthetic data experiments, we repeated the real data experiment ten times with new initialization each time. The results (mean ± standard deviation) of computing the out-of-sample negative log likelihood enabled us to identify the optimal common model (the model with the lowest out-of-sample negative log likelihood). The fact that the proposed multi-subject model is a prototype based spatio-temporal model enabled us to extract three informative novel features from the optimal model: a spatial feature (prototypes volume); a temporal feature (haemodynamic response time to peak); and a spatio-temporal feature (zero lag cross-correlation between the haemodynamic response magnitudes time series of the two prototypes). In general, the results of extracting these features for each group show that both the temporal feature and spatio-temporal feature can be used to discriminate between different populations. However, the spatial features can be used only in the case of large ROIs. Moreover, for the temporal and spatio-temporal features, learning increase the separation between these two groups regardless of the ROIs size, but for the spatial features, learning increases the separation only in the case of the large ROIs.

6.3 Future work

The work presented in this thesis opens up several new directions for further work. We have started to explore some of them.

- Adopt the proposed multi-subject SMM-HPM to predict the group membership of new subjects.

- Extend the multi-subject SMM-HPM to model the interaction between the cognitive processes (visual analysis process, perceptual judgement process and motor response process) and examine if there is an overlapping between the cognitive processes trigged by the same stimulus. This overlapping can be detected by determining the appropriate number of the cognitive processes, which can be one, two or three, using model selection approach.

- Improve the optimization method that was employed to learn the parameters in the second level multi-subject SMM-HPM (L2G-SMM-HPM) and the third level multi-subject SMM-HPM (L3G-SMM-HPM) by optimizing the HRF shape parameters with marginalization over all possible values, and optimizing the spatial prior parameters with marginalization over the neighbouring voxels, respectively. We have already started to work on this improvement (see Appendix A).

APPENDIX A

MARGINALIZATION METHOD

This appendix explains the optimization of the second level multi-subject SMM-HPM (L2G-SMM-HPM) with a marginalization method.

A.0.1 Second level model with marginalization: MargL2G-SMM-HPM

As in L2G-SMM-HPM, this model allows different subjects to have different haemodynamic response shapes beside different haemodynamic response magnitudes. The spatial prior and the number of the prototypes are fixed across subjects. However, we learn the group-level HRF shape parameters instead of the subject-level HRF shape parameters by optimizing them with marginalization over the possible values h. We know the permissible range of the HRF response shape parameters (κ, θ). (θ, κ) is permissible if the corresponding time-to-peak $T^p = (\kappa - 1)\theta$ and peak width $W = 2\sqrt{2\ln 2} \cdot \sqrt{\kappa}\theta$ are both within their permissible ranges. The permissible ranges are given by $[W_{min} = 3s, W_{max} = 6s]$ and $[T^p_{min} = 3s, T^p_{max} = 7s]$, respectively. We built a grid of all permissible combination of the values of HRF response shape parameters (κ, θ), as seen in Fig.(A.1). We marginalize over this grid's values.

Assuming that the observations are independent over subjects, voxels and volumes, the model likelihood of MargL2G-SMM-HPM given fMRI time series $(\mathbf{Y})$ of a group of

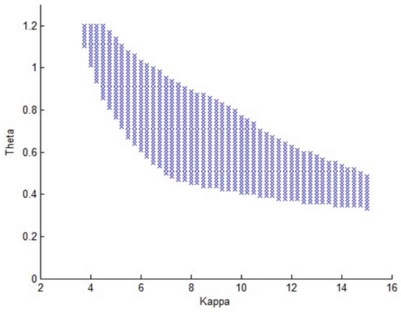

Figure A.1: HRF response shape parameters (κ, θ) permissible rang grid

subjects reads:

$$p(y) = \prod_{u=0}^{U} \prod_{v=0}^{V} \prod_{t=0}^{T} p\left(y_{uvt}; \Theta^{STM}\right) \tag{A.1}$$

With marginalization over the permissible values of the HRF shape parameters (κ, θ), $p\left(y_{uvt}; \Theta^{STM}\right)$ is modelled as:

$$\begin{aligned} p\left(y_{uvt}; \Theta^{STM}\right) &= \sum_{h \in G}^{H} p\big(y_{uvt}; \Theta^{STM}\big) \cdot p(h|\Theta^{HRF}) \\ = \sum_{h \in G}^{H} &\left\{ p\big(y_{uvt}|\Theta^{S}, h, \Theta_u^{NRL}, \Theta^{NIS}\big) \cdot \prod_{k=1}^{K} \prod_{p=1}^{P} p(h_{k,p}|\Theta^{HRF}) \right\} \\ = \sum_{h \in G}^{H} &\left\{ \sum_{k=0}^{K} p(k|v; \Theta^{S}) \cdot p(y_{uvt}|k; h, \Theta_u^{NRL}, \Theta^{NIS}) \cdot \prod_{k=1}^{K} \prod_{p=1}^{P} p(h_{k,p}|\Theta^{HRF}) \right\}, \end{aligned}$$

where $h = h_{k,p} = (\kappa_{k,p}, \theta_{k,p})$, $H =$ all possible $h \in G$, G is a grid of permissible values of κ and θ, $p(k|v; \Theta^{S})$ denotes the prior probability for the k-th prototype generating fMRI time series at voxel v. $p\big(y_{uvt}|k; h, \Theta_u^{NRL}, \Theta^{NIS}\big)$ is the likelihood of y_{uvt} being generated by the k-th prototype. The probability $p(h_{k,p}|\Theta^{HRF})$ is given as normalized Gaussian:

$$\begin{aligned} p(h_{k,p}|\Theta^{HRF}) &= \frac{\mathcal{N}(h_{k,p}|\mu_k^{\kappa}, \sigma_k^{2^{\kappa}}, \mu_k^{\theta}, \sigma_k^{2^{\theta}})}{\sum_{\tilde{h}_{k,p} \in G} \mathcal{N}(\tilde{h}_{k,p}|\mu_k^{\kappa}, \sigma_k^{2^{\kappa}}, \mu_k^{\theta}, \sigma_k^{2^{\theta}})} \\ = &\frac{\mathcal{N}(\kappa_{k,p}; \mu_{k,p}^{\kappa}, \sigma_{k,p}^{2^{\kappa}}) \cdot \mathcal{N}(\theta_{k,p}; \mu_{k,p}^{\theta}, \sigma_{k,p}^{2^{\theta}})}{\sum_{(\tilde{\kappa}_{k,p}, \tilde{\theta}_{k,p}) \in G} \mathcal{N}(\tilde{\kappa}_{k,p}; \mu_{k,p}^{\kappa}, \sigma_{k,p}^{2^{\kappa}}) \cdot \mathcal{N}(\tilde{\theta}_{k,p}; \mu_{k,p}^{\theta}, \sigma_{k,p}^{2^{\theta}})}, \end{aligned} \tag{A.2}$$

where $\mu_{k,p}^{\kappa}$ and $\sigma_{k,p}^{2^{\kappa}}$ are the mean and the variance of the subject-level HRF shape parameters, respectively; and $\mu^{\theta}_{k,p}$ and $\sigma^{2^{\theta}}_{k,p}$ are the mean and the variance of the subject-level HRF scale parameters, respectively.

Learning the Model MargL2G-SMM-HPM

We learn MargL2G-SMM-HPM parameters Θ^{STM} in the usual Bayesian manner (MAP estimation), posterior:

$$p(\Theta^{STM}|\mathbf{Y}) = p(\mathbf{Y}|\Theta^{STM}) \cdot p(\Theta^{STM}).$$

Model likelihood:

$$\begin{aligned} & p(\mathbf{Y}|\Theta^{STM}) = \prod_u \prod_v \prod_t p(y_{uvt}|\Theta^{STM}) \\ = & \prod_u \prod_v \prod_t \sum_h \{p(y_{uvt}|\Theta^{S}, h, \Theta^{NRL}_u, \Theta^{NIS}) \cdot p(h|\Theta^{HRF})\} \end{aligned}$$

The prior is factorized as:

$$\begin{aligned} & p(\Theta^{STM}) \\ = & p(N) \cdot p(b) \cdot \prod_{k=1}^{K} p(\mu_k) \cdot \prod_{k=1}^{K} p(\Sigma_k) \cdot \prod_{u=1}^{U} \prod_{k=1}^{K} \prod_{p=1}^{P} \prod_{s=1}^{S} p(a_{u,k,p,s}) \\ & \cdot \prod_{k=1}^{K} \prod_{p=1}^{P} p(\mu^{\kappa}_{k,p}, \mu^{\theta}_{k,p}) \cdot \prod_{k=1}^{K} \prod_{p=1}^{P} p(\sigma^{2^{\kappa}}_{k,p}) \cdot \prod_{k=1}^{K} \prod_{p=1}^{P} p(\sigma^{2^{\theta}}_{k,p}) \\ & \cdot \prod_{k=0}^{K} p(\sigma^2_k), \end{aligned} \tag{A.3}$$

where $p(\mu^{\kappa}_{k,p}, \mu^{\theta}_{k,p}) \propto \exp^{+\log((T^p - T^p_{min})(T^p_{max} - T^p)) + \log((W - W_{min})(W_{max} - W))}$, $p(\sigma^{2^{\kappa}}_{k,p}) = \frac{1}{(\sigma^{2^{\kappa}}_{k,p})^2}$, and $p(\sigma^{2^{\theta}}_{k,p}) = \frac{1}{(\sigma^{2^{\theta}}_{k,p})^2}$. All the other priors are the same as in L2G-SMM-HPM.

Scaled conjugate-gradient optimization algorithms are used to optimize these parameters iteratively. The gradient of the negative log likelihood $-\log p(\mathbf{Y}|\Theta^{STM})$ with respect to the MargL2G-SMM-HPM parametersis Θ^{STM}:

$$\begin{aligned}
&\nabla_{\Theta^{STM}}\left\{-\log p(Y|\Theta^{STM})\right\}\\
=\;&\nabla_{\Theta^{STM}}\left\{-\log\left(\prod_u\prod_v\prod_t\sum_h p(y_{uvt}|\Theta^{STM})\cdot p(h|\Theta^{HRF})\right)\right\}\\
=\;&\nabla_{\Theta^{STM}}\left\{\sum_u\sum_v\sum_t -\log\left(\sum_h p(y_{uvt}|\Theta^{STM})\cdot p(h|\Theta^{HRF})\right)\right\} \qquad \text{(A.4)}
\end{aligned}$$

In this marginalization optimization method, we sum over the all possible haemodynamic response shape values, which produce a very small probability and the logarithm of small number approaches infinity. To solve this problem we employ Jensen's inequality and optimize the lower bound:

$$\begin{aligned}
\leqslant\;&\nabla_{\Theta^{STM}}\left\{\sum_u\sum_v\sum_t\sum_h -\log\left(p(y_{uvt}|\Theta^{STM})\cdot p(h|\Theta^{HRF})\right)\right\}\\
\leqslant\;&\nabla_{\Theta^{STM}}\left\{\sum_h\sum_u\sum_v\sum_t -\log\left(p(y_{uvt}|\Theta^{STM})\right)-\sum_u\sum_h\log\left(p(h|\Theta^{HRF})\right)\right\}\\
\leqslant\;&\sum_h\nabla_{\Theta^{STM}}\left\{-\log\left(p(Y|\Theta^{STM})\right)\right\}-N\sum_h\nabla_{\Theta^{STM}}\left\{\log\left(p(h|\Theta^{HRF})\right)\right\} \qquad \text{(A.5)}
\end{aligned}$$

As apparent in Eq. (A.5), we marginalize over all the possible values of $h=(\kappa,\theta)\in G$. This is very time consuming, particularly because of the summation over subjects, voxels, volumes and prototypes. In order to reduce the computational time, we built a small grid for the permissible range of the HRF shape parameters. To build this grid, we use functional k-means clustering based on the L2-Distance between the HRF signals of the original grid points, see Fig. (A.2).

In Eq. (A.5), $\{-\log p(Y|\Theta^{STM})\}$ is the likelihood of Model L1G-SMM-HPM. Its derivatives with respect to $(\Theta^{SMM},\Theta_u^{NRL},\Theta^{NIS})$ has not changed. Its derivatives with respect to (Θ^{HRF}) is zero, because h here is a constant value from the grid of the permissible values.

The derivatives of $\mathcal{L}^h=\log p(h|\Theta^{HRF})=\log\left\{\frac{\mathcal{N}(\kappa_{k,p};\mu^{\kappa}_{k,p},\sigma^{2^{\kappa}}_{k,p})\cdot\mathcal{N}(\theta_{k,p};\mu^{\theta}_{k,p},\sigma^{2^{\theta}}_{k,p})}{\sum_{(\tilde{\kappa}_{k,p},\tilde{\theta}_{k,p})\in G}\mathcal{N}(\tilde{\kappa}_{k,p};\mu^{\kappa}_{k,p},\sigma^{2^{\kappa}}_{k,p})\cdot\mathcal{N}(\tilde{\theta}_{k,p};\mu^{\theta}_{k,p},\sigma^{2^{\theta}}_{k,p})}\right\}$ with respect to $(\Theta^{SMM},\Theta_u^{NRL},\Theta^{NIS})$ equal to zero, and with respect to $(\Theta^{HRF}=\mu^{\kappa}_{k,p},\sigma^{2^{\kappa}}_{k,p},\mu^{\theta}_{k,p},\sigma^{2^{\theta}}_{k,p})$

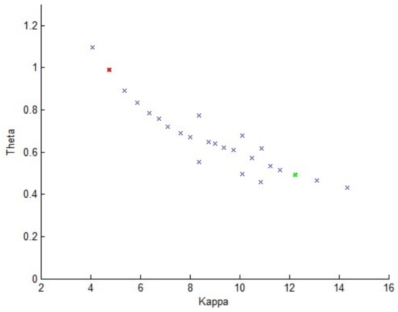

Figure A.2: Small permissible range grid based on L2-Distance between the HRF signals of the original grid points (κ, θ)

is the following:

$$\frac{\mathrm{d}\ \mathcal{L}^h}{\mathrm{d}\ \mu^\kappa_{k,p}} = \frac{\kappa-\mu^\kappa_{k,p}}{\sigma^{2^\kappa}_{k,p}} - \frac{\sum_{(\tilde\kappa_{k,p},\tilde\theta_{k,p})} \mathcal{N}(\tilde\kappa_{k,p},\tilde\theta_{k,p};\mu^\kappa_{k,p},\sigma^{2^\kappa}_{k,p},\mu^\theta_{k,p},\sigma^{2^\theta}_{k,p})\cdot \frac{\tilde\kappa-\mu^\kappa_{k,p}}{\sigma^{2^\kappa}_{k,p}}}{\sum_{(\tilde\kappa_{k,p},\tilde\theta_{k,p})} \mathcal{N}(\tilde\kappa_{k,p},\tilde\theta_{k,p};\mu^\kappa_{k,p},\sigma^{2^\kappa}_{k,p},\mu^\theta_{k,p},\sigma^{2^\theta}_{k,p})} \tag{A.6}$$

$$\frac{\mathrm{d}\ \mathcal{L}^h}{\mathrm{d}\ \mu^\theta_{k,p}} = \frac{\theta-\mu^\theta_{k,p}}{\sigma^{2^\theta}_{k,p}} - \frac{\sum_{(\tilde\kappa_{k,p},\tilde\theta_{k,p})} \mathcal{N}(\tilde\kappa_{k,p},\tilde\theta_{k,p};\mu^\kappa_{k,p},\sigma^{2^\kappa}_{k,p},\mu^\theta_{k,p},\sigma^{2^\theta}_{k,p})\cdot \frac{\tilde\theta-\mu^\theta_{k,p}}{\sigma^{2^\theta}_{k,p}}}{\sum_{(\tilde\kappa_{k,p},\tilde\theta_{k,p})} \mathcal{N}(\tilde\kappa_{k,p},\tilde\theta_{k,p};\mu^\kappa_{k,p},\sigma^{2^\kappa}_{k,p},\mu^\theta_{k,p},\sigma^{2^\theta}_{k,p})} \tag{A.7}$$

$$\begin{aligned}\frac{\mathrm{d}\ \mathcal{L}^h}{\mathrm{d}\ \sigma^{2^\kappa}_{k,p}} &= \frac{(\kappa-\mu^\kappa_{k,p})^2-\sigma^{2^\kappa}_{k,p}}{2(\sigma^{2^\kappa}_{k,p})^2}\\ &-\frac{\sum_{(\tilde\kappa_{k,p},\tilde\theta_{k,p})} \mathcal{N}(\tilde\kappa_{k,p},\tilde\theta_{k,p};\mu^\kappa_{k,p},\sigma^{2^\kappa}_{k,p},\mu^\theta_{k,p},\sigma^{2^\theta}_{k,p})\cdot \frac{(\tilde\kappa-\mu^\kappa_{k,p})^2-\sigma^{2^\kappa}_{k,p}}{2(\sigma^{2^\kappa}_{k,p})^2}}{\sum_{(\tilde\kappa_{k,p},\tilde\theta_{k,p})} \mathcal{N}(\tilde\kappa_{k,p},\tilde\theta_{k,p};\mu^\kappa_{k,p},\sigma^{2^\kappa}_{k,p},\mu^\theta_{k,p},\sigma^{2^\theta}_{k,p})}\end{aligned} \tag{A.8}$$

$$\begin{aligned}\frac{\mathrm{d}\ \mathcal{L}^h}{\mathrm{d}\ \sigma^{2^\theta}_{k,p}} &= \frac{(\theta-\mu^\theta_{k,p})^2-\sigma^{2^\theta}_{k,p}}{2(\sigma^{2^\theta}_{k,p})^2}\\ &-\frac{\sum_{(\tilde\kappa_{k,p},\tilde\theta_{k,p})} \mathcal{N}(\tilde\kappa_{k,p},\tilde\theta_{k,p};\mu^\kappa_{k,p},\sigma^{2^\kappa}_{k,p},\mu^\theta_{k,p},\sigma^{2^\theta}_{k,p})\cdot \frac{(\tilde\theta-\mu^\theta_{k,p})^2-\sigma^{2^\theta}_{k,p}}{2(\sigma^{2^\theta}_{k,p})^2}}{\sum_{(\tilde\kappa_{k,p},\tilde\theta_{k,p})} \mathcal{N}(\tilde\kappa_{k,p},\tilde\theta_{k,p};\mu^\kappa_{k,p},\sigma^{2^\kappa}_{k,p},\mu^\theta_{k,p},\sigma^{2^\theta}_{k,p})}\end{aligned} \tag{A.9}$$

We now show the derivatives of the (negative log) prior $\mathcal{P} = -\log p(\Theta^{STM})$:

For $p(\mu^{\kappa}_{k,p}, \mu^{\theta}_{k,p}) \propto \exp^{+\log((T^p - T^p_{min})(T^p_{max} - T^p)) + \log((W - W_{min})(W_{max} - W))}$:

$$\begin{aligned}\frac{\mathrm{d}\,\mathcal{P}}{\mathrm{d}\,\mu^{\theta}_{k,p}} &= -2\left[\frac{\frac{1}{2}(T^p_{max} + T^p_{min}) - T^p}{(T^p - T^p_{min})(T^p_{max} - T^p)} \cdot \mu^{\kappa}_{k,p} - 1\right. \\ &\left. + \frac{\frac{1}{2}(W_{max} + W_{min}) - W}{(W - W_{min})(W_{max} - W)} \cdot 2.35\sqrt{\mu^{\kappa}_{k,p}}\right], \qquad (A.10)\end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}\,\mathcal{P}}{\mathrm{d}\,\mu^{\kappa}_{k,p}} &= -2\left[\frac{\frac{1}{2}(T^p_{max} + T^p_{min}) - T^p}{(T^p - T^p_{min})(T^p_{max} - T^p)} \cdot \mu^{\theta}_{k,p}\right. \\ &\left. + \frac{\frac{1}{2}(W_{max} + W_{min}) - W}{(W - W_{min})(W_{max} - W)} \cdot \frac{1.175\mu^{\theta}_{k,p}}{\sqrt{\mu^{\kappa}_{k,p}}}\right] \qquad (A.11)\end{aligned}$$

For $p(\sigma^{2^{\kappa}}_{k,p}) = \frac{1}{(\sigma^{2^{\kappa}}_{k,p})^2}$:

$$\frac{\mathrm{d}\,\mathcal{P}}{\mathrm{d}\,\sigma^{2^{\kappa}}_{k,p}} = 2\frac{1}{\sigma^{2^{\kappa}}_{k,p}} \qquad (A.12)$$

For $p(\sigma^{2^{\theta}}_{k,p}) = \frac{1}{(\sigma^{2^{\theta}}_{k,p})^2}$:

$$\frac{\mathrm{d}\,\mathcal{P}}{\mathrm{d}\,\sigma^{2^{\theta}}_{k,p}} = 2\frac{1}{\sigma^{2^{\theta}}_{k,p}} \qquad (A.13)$$

All the other priors have the same derivative as in L2G-SMM-HPM.

Initialization of the MargL2G-SMM-HPM

Initialization of the MargL2G-SMM-HPM is the same as the initialization of the L2G-SMM-HPM.

Results

We applied the same synthetic experiments that have been used for the L2G-SMM-HPM in section (5.4.1), and using the same synthetic data (only L2-data) and the same statistics. Compared to the results of optimizing the L2G-SMM-HPM, MargL2G-SMM-HPM provides better results:

Table A.1: Out-of-sample negative log likelihood

	MargL2G-SMM-HPM
L2-Data	1.015±[0.0211]

Table A.2: MargL2G-SMM-HPM parameters estimation performance

Data	A_S	A_g	eC_0	A_a
L2-Data	0.0007 ±[0.0325]	0.01124±[0.0387]	-0.9511±[0.0455]	0.0036±[0.0298]

APPENDIX

LIST OF REFERENCES

[1] Yuan Shen, Stephen D Mayhew, Zoe Kourtzi, and Peter Tiňo. Spatial–temporal modelling of fMRI data through spatially regularized mixture of hidden process models. *NeuroImage*, 84:657–671, 2014.

[2] Guillaume Flandin and William D Penny. Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage*, 34(3):1108–1125, 2007.

[3] Alle Meije Wink and Jos BTM Roerdink. Denoising functional mr images: a comparison of wavelet denoising and gaussian smoothing. *IEEE transactions on medical imaging*, 23(3):374–387, 2004.

[4] Michael Hilton, Todd Ogden, David Hattery, Guinevere Eden, and Bjorn Jawerth. Wavelet denoising of functional mri data. *Wavelets in Medicine and Biology*, pages 93–114, 1996.

[5] Xavier Descombes, Frithjof Kruggel, and D Yves von Cramon. fMRI signal restoration using a spatio-temporal markov random field preserving transitions. *NeuroImage*, 8(4):340–349, 1998.

[6] Andres Fco. Sole, Shing-Chung Ngan, Guillermo Sapiro, Xiaoping Hu, and Antonio Lopez. Anisotropic 2-d and 3-d averaging of fMRI signals. *IEEE transactions on medical imaging*, 20(2):86–93, 2001.

[7] Hae Yong Kim, Javier Giacomantone, and Zang Hee Cho. Robust anisotropic diffusion to produce enhanced statistical parametric map from noisy fMRI. *Computer Vision and Image Understanding*, 99(3):435–452, 2005.

[8] Ola Friman, Magnus Borga, Peter Lundberg, and Hans Knutsson. Adaptive analysis of fMRI data. *NeuroImage*, 19(3):837–845, 2003.

[9] Jean-Baptiste Poline and BM Mazoyer. Analysis of individual brain activation maps using hierarchical description and multiscale detection. *IEEE transactions on medical imaging*, 13(4):702–710, 1994.

[10] Stefan J Kiebel, Rainer Goebel, and Karl J Friston. Anatomically informed basis functions. *NeuroImage*, 11(6):656–667, 2000.

[11] Alexandre Andrade, Ferath Kherif, Jean-François Mangin, Keith J Worsley, Anne-Lise Paradis, Olivier Simon, Stanislas Dehaene, Denis Le Bihan, and Jean-Baptiste Poline. Detection of fMRI activation using cortical surface mapping. *Human brain mapping*, 12(2):79–93, 2001.

[12] Christopher R Genovese. A bayesian time-course model for functional magnetic resonance imaging data. *Journal of the American Statistical Association*, 95(451): 691–703, 2000.

[13] Christoff Gössl, Dorothee P Auer, and Ludwig Fahrmeir. Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57(2):554–562, 2001.

[14] L Fahrmeir, C Gössl, and A Hennerfeind. Spatial smoothing with robust priors in functional mri. In *Exploratory Data Analysis in Empirical Research*, pages 50–57. Springer, 2003.

[15] Mark William Woolrich, Mark Jenkinson, J Michael Brady, and Stephen M Smith. Fully bayesian spatio-temporal modeling of fMRI data. *IEEE transactions on medical imaging*, 23(2):213–231, 2004.

[16] William D Penny, Nelson J Trujillo-Barreto, and Karl J Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, 2005.

[17] Lee Michael Harrison, W Penny, J Ashburner, N Trujillo-Barreto, and KJ Friston. Diffusion-based spatial priors for imaging. *NeuroImage*, 38(4):677–695, 2007.

[18] F DuBois Bowman, Brian Caffo, Susan Spear Bassett, and Clinton Kilts. A bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage*, 39(1):146–156, 2008.

[19] Adrian R Groves, Michael A Chappell, and Mark W Woolrich. Combined spatial and non-spatial prior for inference on mri time-series. *NeuroImage*, 45(3):795–809, 2009.

[20] Alicia Quirós, Raquel Montes Diez, and Dani Gamerman. Bayesian spatiotemporal model of fMRI data. *NeuroImage*, 49(1):442–456, 2010.

[21] Michael Smith and Ludwig Fahrmeir. Spatial bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 102(478):417–431, 2007.

[22] Kuo-Jung Lee, Galin L Jones, Brian S Caffo, and Susan Spear Bassett. Spatial bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Analysis (Online)*, 9(3):699, 2014.

[23] Linlin Zhang, Michele Guindani, Francesco Versace, and Marina Vannucci. A spatio-temporal nonparametric bayesian variable selection model of fMRI data for clustering correlated time courses. *NeuroImage*, 95:162–175, 2014.

[24] Fan Li, Tingting Zhang, Quanli Wang, Marlen Z Gonzalez, Erin L Maresh, James A Coan, et al. Spatial bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *The Annals of Applied Statistics*, 9(2):687–713, 2015.

[25] Brian S Everitt and Edward T Bullmore. Mixture model mapping of brain activation in functional magnetic resonance images. *Human brain mapping*, 7(1):1–14, 1999.

[26] Niels Vaever Hartvig and Jens Ledet Jensen. Spatial mixture modeling of fMRI data. *Human brain mapping*, 11(4):233–248, 2000.

[27] Will Penny and Karl Friston. Mixtures of general linear models for functional neuroimaging. *IEEE transactions on medical imaging*, 22(4):504–514, 2003.

[28] Mark William Woolrich, Timothy Edward John Behrens, Christian F Beckmann, and Stephen M Smith. Mixture models with adaptive spatial regularization for segmentation with an application to fMRI data. *IEEE transactions on medical imaging*, 24(1):1–11, 2005.

[29] Mark William Woolrich and Timothy E Behrens. Variational bayes inference of spatial mixture models for segmentation. *IEEE Transactions on Medical Imaging*, 25(10):1380–1391, 2006.

[30] Vangelis P Oikonomou and Konstantinos Blekas. An adaptive regression mixture model for fMRI cluster analysis. *IEEE transactions on medical imaging*, 32(4): 649–659, 2013.

[31] Brice Ozenne, Fabien Subtil, Leif Østergaard, and Delphine Maucort-Boulch. Spatially regularized mixture model for lesion segmentation with application to stroke patients. *Biostatistics*, 16(3):580–595, 2015.

[32] Christopher J Brignell, William J Browne, Ian L Dryden, and Susan T Francis. Mixed effect modelling of single trial variability in ultra-high field fMRI. *arXiv preprint arXiv:1501.05763*, 2015.

[33] A Llera, D Vidaurre, RHR Pruim, and CF Beckmann. Variational mixture models with gamma or inverse-gamma components. *arXiv preprint arXiv:1607.07573*, 2016.

[34] Hien D Nguyen, Geoffrey J McLachlan, Jeremy FP Ullmann, and Andrew L Janke. Spatial clustering of time series via mixture of autoregressions models and markov random fields. *Statistica Neerlandica*, 70(4):414–439, 2016.

[35] Mark W Woolrich, Timothy EJ Behrens, Christian F Beckmann, Mark Jenkinson, and Stephen M Smith. Multilevel linear modelling for fMRI group analysis using bayesian inference. *Neuroimage*, 21(4):1732–1747, 2004.

[36] Linlin Zhang, Michele Guindani, Francesco Versace, Jeffrey M Engelmann, Marina Vannucci, et al. A spatiotemporal nonparametric bayesian model of multi-subject fMRI data. *The Annals of Applied Statistics*, 10(2):638–666, 2016.

[37] Gordana Derado, F DuBois Bowman, and Clinton D Kilts. Modeling the spatial and temporal dependence in fMRI data. *Biometrics*, 66(3):949–957, 2010.

[38] Nilotpal Sanyal and Marco AR Ferreira. Bayesian hierarchical multi-subject multiscale analysis of functional mri data. *NeuroImage*, 63(3):1519–1531, 2012.

[39] Donald R Musgrove, John Hughes, and Lynn E Eberly. Fast, fully bayesian spatiotemporal inference for fMRI data. *Biostatistics*, 17(2):291–303, 2016.

[40] Seyoung Kim, Padhraic Smyth, Hal Stern, and Jessica Turner. Parametric response surface models for analysis of multi-site fMRI data. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, pages 352–359, 2005.

[41] Seyoung Kim, Padhraic Smyth, and Hal Stern. A nonparametric bayesian approach to detecting spatial activation patterns in fMRI data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 217–224. Springer, 2006.

[42] Seyoung Kim, Padhraic Smyth, and Hal Stern. A bayesian mixture approach to modeling spatial activation patterns in multisite fMRI data. *IEEE transactions on medical imaging*, 29(6):1260–1274, 2010.

[43] Bertrand Thirion, Alan Tucholka, Merlin Keller, Philippe Pinel, Alexis Roche, Jean-François Mangin, and Jean-Baptiste Poline. High level group analysis of fMRI data based on dirichlet process mixture models. In *Information Processing in Medical Imaging*, pages 482–494. Springer, 2007.

[44] Seyoung Kim and Padhraic Smyth. Hierarchical dirichlet processes with random effects. In *NIPS*, pages 697–704, 2006.

[45] Lei Xu, Timothy D Johnson, Thomas E Nichols, and Derek E Nee. Modeling inter-subject variability in fMRI activation location: A bayesian hierarchical spatial model. *Biometrics*, 65(4):1041–1051, 2009.

[46] Anne-Laure Fouque, Philippe Ciuciu, and Laurent Risser. Multivariate spatial gaussian mixture modeling for statistical clustering of hemodynamic parameters in functional mri. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 445–448. IEEE, 2009.

[47] Saad Jbabdi, Mark William Woolrich, and Timothy Edward John Behrens. Multiple-subjects connectivity-based parcellation using hierarchical dirichlet process mixture models. *NeuroImage*, 44(2):373–384, 2009.

[48] Samuel J Gershman, David M Blei, Francisco Pereira, and Kenneth A Norman. A topographic latent source model for fMRI data. *NeuroImage*, 57(1):89–100, 2011.

[49] 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.

[50] Rasmus E Røge, Kristoffer H Madsen, MikkelN Schmidt, and Morten Mørup. Unsupervised segmentation of task activated regions in fMRI. In *Machine Learning*

for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on, pages 1–6. IEEE, 2015.

[51] Nathan W Churchill, Kristoffer Madsen, and Morten Mørup. The functional segregation and integration model: Mixture model representations of consistent and variable group-level connectivity in fMRI. *Neural computation*, 2016.

[52] Sudhir Raman, Lorenz Deserno, Florian Schlagenhauf, and Klaas Enno Stephan. A hierarchical model for integrating unsupervised generative embedding and empirical bayes. *Journal of neuroscience methods*, 269:6–20, 2016.

[53] Matthew A.L. Ralph James L. McClelland. Cognitive neuroscience. *International Encyclopedia of the Social and Behavioral Sciences*, 2:95–102, 2015.

[54] Russell A Poldrack, Jeanette A Mumford, and Thomas E Nichols. *Handbook of functional MRI data analysis.* Cambridge University Press, 2011.

[55] Ramesh Srinivasan. Methods to improve the spatial resolution of eeg. *International Journal of Bioelectromagnetism*, 1(1):102–111, 1999.

[56] Zhongming Liu, Lei Ding, and Bin He. Integration of eeg/meg with mri and fMRI. *IEEE engineering in medicine and biology magazine*, 25(4):46–53, 2006.

[57] Bruce Crosson, Anastasia Ford, Keith M McGregor, Marcus Meinzer, Sergey Cheshkov, Xiufeng Li, Delaina Walker-Batson, and Richard W Briggs. Functional imaging and related techniques: An introduction for rehabilitation researchers. *Journal of rehabilitation research and development*, 47(2):vii, 2010.

[58] Mathews Paul Jezzard, Peter and Stephen Smith, editors. *Functional mri an introduction to methods.* Oxford Medical Publication, 2001.

[59] Elizabeth MC Hillman. Optical brain imaging in vivo: techniques and applications from animal to man. *Journal of biomedical optics*, 12(5):051402, 2007.

[60] Matthijs Vink, M Raemaekers, A van der Schaaf, R Mandl, and N Ramsey. Preprocessing and analysis of functional mri data. 2007.

[61] Linlin Zhang. *Bayesian nonparametric models for functional magnetic resonance imaging (fMRI) data.* PhD thesis, Rice University, 2015.

[62] Yongnan Ji. *Data-driven fMRI data analysis based on parcellation.* PhD thesis, University of Nottingham, 2001.

[63] 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.

[64] Karl J Friston, Chris D Frith, Robert Turner, and Richard SJ Frackowiak. Characterizing evoked hemodynamics with fMRI. *Neuroimage*, 2(2):157–165, 1995.

[65] Martin A Lindquist. The statistical analysis of fMRI data. *Statistical Science*, pages 439–464, 2008.

[66] Jeroen CW Siero, Alex Bhogal, and J Martijn Jansma. Blood oxygenation level–dependent/functional magnetic resonance imaging. *PET clinics*, 8(3):329–344, 2013.

[67] Tor D Wager, Alberto Vazquez, Luis Hernandez, and Douglas C Noll. Accounting for nonlinear bold effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. *NeuroImage*, 25(1):206–218, 2005.

[68] Karl J Friston, Oliver Josephs, Geraint Rees, and Robert Turner. Nonlinear event-related responses in fMRI. *Magnetic resonance in medicine*, 39(1):41–52, 1998.

[69] Karl J Friston, Andrea Mechelli, Robert Turner, and Cathy J Price. Nonlinear responses in fMRI: the balloon model, volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477, 2000.

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

[71] Massimo Filippi, Roberta Messina, and Maria A Rocca. fMRI of the sensorimotor system. *fMRI Techniques and Protocols*, pages 523–543, 2016.

[72] Karl J Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

[73] Massimo Filippi. *fMRI techniques and protocols*. Springer, 2009.

[74] 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.

[75] Vince D Calhoun, Jingyu Liu, and Tülay Adalı. A review of group ica for fMRI data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45 (1):S163–S172, 2009.

[76] Erik Barry Erhardt, Srinivas Rachakonda, Edward J Bedrick, Elena A Allen, Tülay Adali, and Vince D Calhoun. Comparison of multi-subject ica methods for analysis of fMRI data. *Human brain mapping*, 32(12):2075–2095, 2011.

[77] O Coulon, J-F Mangin, J-B Poline, M Zilbovicius, D Roumenov, Y Samson, V Frouin, and I Bloch. Structural group analysis of functional activation maps. *NeuroImage*, 11(6):767–782, 2000.

[78] JV Stone, J Porrill, C Buchel, and K Friston. Spatial, temporal, and spatiotemporal independent component analysis of fMRI data. In *Proc. Leeds Statistical Research Workshop*, pages 7–9. Citeseer, 1999.

[79] 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.

[80] MC McIntyre, A Wennerberg, R Somorjai, and G Scarth. Activation and deactivation in functional brain images. *Neuroimage*, 3(3):S82, 1996.

[81] 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.

[82] G Scarth, A Wennerberg, R Somorjai, T Hindmarsh, and M McIntyre. The utility of fuzzy clustering in identifying diverse activations in fMRI. *NeuroImage*, 3(3):S89, 1996.

[83] Xavier Golay, Spyros Kollias, Dieter Meier, Anton Valavanis, and Peter Boesiger. Fuzzy membership vs. probability in cross correlation based fuzzy clustering of fMRI data. *NeuroImage*, 5(4 PART I), 1997.

[84] Peter Toft, Lars Kai Hansen, Finn Årup Nielsen, LK Hansen, Nick Lange, Niels Mørch, FA Nielsen, Olaf B Paulson, Robert Savoy, Bruce Rosen, et al. On clustering of fMRI time series. 1997.

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

[86] Peter Filzmoser, Richard Baumgartner, and Ewald Moser. A hierarchical clustering method for analyzing functional mr images. *Magnetic resonance imaging*, 17(6): 817–826, 1999.

[87] Defeng Wang, Lin Shi, Daniel Yeung, Pheng-Ann Heng, Tien-Tsin Wong, and Eric Tsang. Support vector clustering for brain activation detection. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, pages 572–579, 2005.

[88] Defeng Wang, Lin Shi, Daniel S Yeung, Eric CC Tsang, and Pheng Ann Heng. Ellipsoidal support vector clustering for functional mri analysis. *Pattern Recognition*, 40(10):2685–2695, 2007.

[89] Bertrand Thirion, Guillaume Flandin, Philippe Pinel, Alexis Roche, Philippe Ciuciu, and Jean-Baptiste Poline. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Human brain mapping*, 27(8): 678–693, 2006.

[90] Bertrand Thirion, Alan Tucholka, and Jean-Baptiste Poline. Parcellation schemes and statistical tests to detect active regions on the cortical surface. In *Proceedings of COMPSTAT'2010*, pages 565–572. Springer, 2010.

[91] Xilin Shen, F Tokoglu, Xenios Papademetris, and R Todd Constable. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage*, 82:403–415, 2013.

[92] Salima Makni, Jérôme Idier, Thomas Vincent, Bertrand Thirion, Ghislaine Dehaene-Lambertz, and Philippe Ciuciu. A fully bayesian approach to the parcel-based detection-estimation of brain activity in fMRI. *Neuroimage*, 41(3):941–969, 2008.

[93] Lotfi Chaari, Florence Forbes, Thomas Vincent, and Philippe Ciuciu. Hemodynamic-informed parcellation of fMRI data in a joint detection estimation framework. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 180–188, 2012.

[94] Mohanad Albughdadi, Lotfi Chaari, Jean-Yves Tourneret, Florence Forbes, and Philippe Ciuciu. Hemodynamic brain parcellation using a non-parametric bayesian approach. 2016.

[95] Mohanad Albughdadi, Lotfi Chaari, Florence Forbes, Jean-Yves Tourneret, and Philippe Ciuciu. Multi-subject joint parcellation detection estimation in functional mri. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 74–77. IEEE, 2016.

[96] Bertrand Thirion, Gaël Varoquaux, Elvis Dohmatob, and Jean-Baptiste Poline. Which fMRI clustering gives good brain parcellations? *Frontiers in neuroscience*, 8:167, 2014.

[97] Benoit Da Mota, Virgile Fritsch, Gaël Varoquaux, Tobias Banaschewski, Gareth J Barker, Arun LW Bokde, Uli Bromberg, Patricia Conrod, Jürgen Gallinat, Hugh Garavan, et al. Randomized parcellation based inference. *NeuroImage*, 89:203–215, 2014.

[98] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[99] James V Haxby. Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage*, 62(2):852–855, 2012.

[100] Janaina Mourão-Miranda, Arun LW Bokde, Christine Born, Harald Hampel, and Martin Stetter. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional mri data. *NeuroImage*, 28(4): 980–995, 2005.

[101] Ze Wang, Anna R Childress, Jiongjiong Wang, and John A Detre. Support vector machine learning-based fMRI data group analysis. *NeuroImage*, 36(4):1139–1151, 2007.

[102] Ze Wang. A hybrid svm–glm approach for fMRI data analysis. *Neuroimage*, 46(3): 608–615, 2009.

[103] Xiaomu Song and Alice M Wyrwicz. Unsupervised spatiotemporal fMRI data analysis using support vector machines. *NeuroImage*, 47(1):204–212, 2009.

[104] Stephen LaConte, Stephen Strother, Vladimir Cherkassky, Jon Anderson, and Xiaoping Hu. Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, 26(2):317–329, 2005.

[105] Mehdi Behroozi and Mohammad Reza Daliri. Software tools for the analysis of functional magnetic resonance imaging. *Basic and Clinical Neuroscience*, 3(5):71–83, 2012.

[106] Mark W Woolrich, Timothy EJ Behrens, and Stephen M Smith. Constrained linear basis sets for hrf modelling using variational bayes. *NeuroImage*, 21(4):1748–1761, 2004.

[107] Karl J Friston, William Penny, Christophe Phillips, S Kiebel, G Hinton, and John Ashburner. Classical and bayesian inference in neuroimaging: theory. *NeuroImage*, 16(2):465–483, 2002.

[108] Krzysztof J Gorgolewski, Amos J Storkey, Mark E Bastin, and Cyril R Pernet. Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in human neuroscience*, 6, 2012.

[109] Hien D Nguyen, Geoffrey J McLachlan, Pierre Orban, Pierre Bellec, and Andrew L Janke. Maximum pseudolikelihood estimation for model-based clustering of time series data. *Neural computation*, 29(4):990–1020, 2017.

[110] Rebecca A Hutchinson, Tom M Mitchell, and Indrayana Rustandi. Hidden process models. In *Proceedings of the 23rd international conference on Machine learning*, pages 433–440. ACM, 2006.

[111] Nahed Alowadi, Yuan Shen, and Peter Tiňo. Prototype-based spatio-temporal probabilistic modelling of fMRI data. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 193–203. Springer, 2016.

[112] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1):91–118, 2003.

[113] Wojciech Samek and Motoaki Kawanabe. Robust common spatial patterns by minimum divergence covariance estimator. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2040–2043. IEEE, 2014.

[114] Rui Wang, Yuan Shen, Peter Tino, Andrew E Welchman, and Zoe Kourtzi. Learning predictive statistics: strategies and brain mechanisms. *Journal of Neuroscience*, 37 (35):8412–8427, 2017.