Functional Neuroimaging: A Sparse Modelling Approach

Shulin Yan

Submitted in part fulfillment of the requirements for the degree of Doctor of Philosophy in Computing of Imperial College London and the Diploma of Imperial College London, September 2014
Abstract

Developments in technology have enabled scientists to study brain function in an unprecedented way. Functional neuroimaging is the use of neuroimaging technologies to capture information about the state of a brain, with the goal of studying the relationship between mental functions and brain activity. One such technology is functional magnetic resonance imaging (fMRI), which produces a signal that can be used to create cross-sectional images of the brain. These images can be used to measure brain activity in different sections of the brain.

In fMRI recording there is a tradeoff between spatial and temporal resolution. My first contribution in this thesis is to present a novel algorithm for generating cross-sectional images. This is a signal processing problem with high dimensionality, but few measurements. My algorithm uses ideas from sparse modelling because variations in functional MR images are sparse over time in the wavelet domain. It will enable high resolution images to be generated using fewer measurements.

Sequences of functional MR images are recorded while subjects perform different tasks. The second contribution of this thesis is a machine learning technique to predict different tasks from the captured fMRI sequences. Existing methods perform poorly at this prediction task due to the curse of dimensionality. I overcome this problem by designing a novel sparse modelling method based on the assumption that the active brain region in response to a target task is sparse in the whole brain area.

The final contribution to this thesis is the design of different assessment criteria for selecting the most relevant voxels to interpret the neural activity. The conventional selection method uses the assessment of predictive performance, resulting in many false positive selections due to the small number of samples. To overcome this problem, I introduce the concept of stability. My method selects the relevant voxels using the assessments of both predictive performance and stability, which significantly reduces the selection error while maintaining the predictive performance.
Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.
Acknowledgements

I would like to take this opportunity to express my thanks to all those who have assisted and supported me from various aspects.

First, I would like to thanks my supervisor, Professor. Yike Guo, for his support and professional guidance. Without his continuous encouragement and great patience, this thesis would not be possible.

I would also like to thank all my lovely friends and colleagues in Imperial College London for their constance help in the programming aspect of my research as well as their friendship that brought much happiness to my life at Imperial.

Finally, I would like to thank my parents for their constant support, love and encouragement.
Dedication

To my dearest Mama and Baba.
‘A person who never made a mistake never tried anything new.’

‘The important thing is not to stop questioning. Curiosity has its own reason for existing.’

‘If you can’t explain simple, you don’t understand it well enough.’

‘Look deep into nature, and then we will understand everything better.’

‘We cannot solve our problems with the same thinking we used when we created them.’

Albert Einstein (1879-1955)
Contents

Abstract .......................................................... 1

Copyright Declaration ............................................ 3

Acknowledgements .................................................. 5

1 Introduction ....................................................... 19

1.1 Motivations and Objectives .................................. 19

1.2 Contributions .................................................. 22

1.3 Thesis Organisation .......................................... 24

1.4 Statement of Originality ...................................... 24

1.5 Publications ................................................... 24

2 Functional Neuroimaging ..................................... 28

2.1 Overview of Techniques ..................................... 29

2.1.1 Electrophysiology based Techniques .................. 30

2.1.2 Hemodynamic Response based Techniques .......... 32

2.1.3 Electrophysiology vs. Hemodynamic Response ....... 34
2.1.4 Multimodal Techniques ........................................ 37

2.2 Functional Magnetic Resonance Imaging .......................... 38
  2.2.1 Imaging Technique ............................................. 40
  2.2.2 Data Analysis .................................................. 47

3 Sparse Modelling .................................................. 57
  3.1 Compressive Sensing Reconstruction Methods .................. 61
    3.1.1 $\ell_p$ Optimisation Algorithms ............................ 63
    3.1.2 Greedy Algorithms .......................................... 67
    3.1.3 Sparse Bayesian Learning .................................... 68
  3.2 Stable Reconstruction Methods ................................ 73
    3.2.1 Ensemble Based Methods .................................... 78
    3.2.2 Group Selection Methods .................................... 80
  3.3 Measurement Design ............................................ 83
    3.3.1 Random Methods ............................................. 83
    3.3.2 Informative Methods ......................................... 85

4 Linear Dynamic Sparse Modelling for Functional MR Imaging 90
  4.1 Introduction ................................................... 90
  4.2 Problem Formulation .......................................... 94
    4.2.1 Sparsity of Variations ...................................... 94
    4.2.2 Linear Dynamic Sparse Model ............................... 97
## CONTENTS

### 4.3 Linear Dynamic Sparse Modelling

- 4.3.1 Hierarchical Bayesian Kalman Filter
- 4.3.2 Informative Measurement Design

### 4.4 Experimental Results

- 4.4.1 Implementation Details of the fMRI Sequence Reconstruction
- 4.4.2 Comparing Different Reconstruction Algorithms
- 4.4.3 Comparing Different Measurement Algorithms

### 4.5 Discussion and Conclusion

### 5 Robust Linear Sparse Modelling for fMRI Data Analysis

#### 5.1 Introduction

#### 5.2 Linear Sparse Modelling in Multivariate Pattern Analysis

#### 5.3 Random Subspace Sparse Bayesian Learning

#### 5.4 Materials and Experiments

- 5.4.1 Dataset
- 5.4.2 Experimental Protocol
- 5.4.3 Results

#### 5.5 Conclusions

### 6 Balancing Stability and Predictive Performance of fMRI Models

#### 6.1 Introduction

#### 6.2 Methods
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>AUC Values for Four Different Classifiers.</td>
<td>138</td>
</tr>
<tr>
<td>6.1</td>
<td>Optimal Results Achieved with Highest Scores for Simulation Dataset 1</td>
<td>158</td>
</tr>
<tr>
<td>6.2</td>
<td>Comparison of Selection Accuracy. Optimal Results Achieved with Highest Scores of Three Different Methods (Simulation Datasets 2 and 3).</td>
<td>159</td>
</tr>
<tr>
<td>6.3</td>
<td>Comparison of Predictive Performance. Optimal Results Achieved with Highest Scores of Three Different Methods (Simulation Datasets 2 and 3).</td>
<td>160</td>
</tr>
<tr>
<td>6.4</td>
<td>Optimal Results Achieved with Highest Scores of Three Different Methods (Subject 1 and 2).</td>
<td>163</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Example of Different Weighted MR Images. This figure is generated with the fMRI data provided by [BBH+13]. .......................... 42

3.1 Best Approximation of A Point in $\mathbb{R}^2$ by A One-dimensional Support Using $\ell_p$ norm, where $p = 0, 1, 2$. Reproduce from [DDEK11]. ............... 64

3.2 Best Approximation of A Point in $\mathbb{R}^2$ by A One-dimensional Support Using Lasso. .......................................................... 65

3.3 The Effect of Threshold on Sensitivity and Specificity. The dark red bell refers to the distribution of positive group, and the light red bell refers to the distribution of negative group, where the distributions are determined by classifiers. Reproduced from [Sr09]. ......................... 75

3.4 Examples of ROC Curves for Different Predictive Models. Reproduced from [Sr09]. ............................................................. 76

3.5 Illustration of Elastic Net: (a) Spheres of Lasso, Ridge and Elastic Net, (b) Best Approximation of a Point in $\mathbb{R}^2$ by A One-dimensional Support Using Elastic Net. Reproduced from [ZF14]. ............................. 80

3.6 Illustrations of Group Lasso and Sparse Group Lasso for Two Groups with $w_1 \in \mathbb{R}^1$ and $w_2 \in \mathbb{R}^2$. Reproduced from [YL06, FHT10]. ............ 82
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.7</td>
<td>Framework of the Nonadaptive Measurement Design Methods (Open-Loop)</td>
<td>88</td>
</tr>
<tr>
<td>3.8</td>
<td>Framework of the Adaptive Measurement Design Methods (Closed-Loop)</td>
<td>88</td>
</tr>
<tr>
<td>4.1</td>
<td>Discrete Wavelet Transform (DWT) of a One Dimensional Signal</td>
<td>94</td>
</tr>
<tr>
<td>4.2</td>
<td>One-level Daubechies-4 DWT for an Functional MR Image</td>
<td>95</td>
</tr>
<tr>
<td>4.3</td>
<td>Two-level Daubechies-4 DWT for an Functional MR Image</td>
<td>96</td>
</tr>
<tr>
<td>4.4</td>
<td>Example of Sparse Variations. The variations are calculated with 96% energy</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>supports of DWT coefficients of images</td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>Linear Dynamic Model</td>
<td>97</td>
</tr>
<tr>
<td>4.6</td>
<td>Framework of fMRI Sequence Reconstruction</td>
<td>99</td>
</tr>
<tr>
<td>4.7</td>
<td>Comparison of Graphic Models for Kalman Filter and HB-Kalman</td>
<td>102</td>
</tr>
<tr>
<td>4.8</td>
<td>The Relationship between Student’s t Distribution and Gaussian Distribution</td>
<td>106</td>
</tr>
<tr>
<td>4.9</td>
<td>Implementation of the Functional MR Image Reconstruction. Note that</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>the red point in the bottom-left image indicates the measurements.</td>
<td></td>
</tr>
<tr>
<td>4.10</td>
<td>Selected Points in (k)-space with Three Different Measurement Methods.</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>Note that the white points indicate the selected ones.</td>
<td></td>
</tr>
<tr>
<td>4.11</td>
<td>Reconstruction Errors (HB-Kalman vs. SBL)</td>
<td>113</td>
</tr>
<tr>
<td>4.12</td>
<td>Sparsity of Image (w_t). (</td>
<td>supp(w_t)</td>
</tr>
<tr>
<td></td>
<td>of DWT coefficients of image at time (t).</td>
<td></td>
</tr>
<tr>
<td>4.13</td>
<td>Reconstruction Errors (Random Sampling vs. VD Sampling)</td>
<td>116</td>
</tr>
<tr>
<td>4.14</td>
<td>Reconstruction Errors (IMD vs. VD Sampling and Random Sampling)</td>
<td>117</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

4.15 Reconstructions of Functional MR Images (4nd, 8th, 12th, 16th, 20th frames). .................................................. 118

5.1 Framework of RS-SBL .................................................. 129

5.2 Surfaces for Usability and Coverage with $L = 1 \sim 100$ and $M = 1 \sim p$,
(a)-(b) $S=1000$, $n=228000$ and (c)-(d) $S=50$, $n=228000$. .................. 135

5.3 Classification Accuracies Across 10 Subjects with Four Different Classifiers: SVM-Lin, SBC-Lin, SBC and RS-SBC. ................. 136

5.4 ROC Curves for Four Different Classifiers. ......................... 138

5.5 An Example of ‘Individual’ Predictive Map vs. Active Map for Relation
vs. Rest Task. ................................................................. 140

5.6 An Example of ‘Individual’ Predictive Map vs. Active Map for Match vs.
Rest Task. ................................................................. 141

5.7 An Example of ‘Individual’ Predictive Map vs. Active Map for Relation
vs. Rest Task. ................................................................. 142

5.8 An Example of ‘Individual’ Predictive Map vs. Active Map for Relation
vs. Match Task. ................................................................. 143

6.1 Framework for RS-SBL Multivariate Voxel Selection (MVS) .......... 148

6.2 Framework of Proposed Method ........................................ 152

6.3 Examples of Scores of Three Different Methods under Different Parameter Settings (Simulation Dataset 1). ............................... 157

6.4 ROC Curves with AUC scores of Three Different Methods with Optimal
Parameter Settings for Simulation Dataset 1. ................................ 158

6.5 ROC Curves of Three Different Methods with Optimal Parameter Settings
for Simulation Datasets 2 and 3. ........................................ 160
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6</td>
<td>Examples of Scores of Three Different Methods under Different Parameter Settings (Simulation Datasets 2 and 3)</td>
<td>162</td>
</tr>
<tr>
<td>6.7</td>
<td>ROC Curves of Three Different Methods with Optimal Parameter Settings for Real Datasets 1 and 2.</td>
<td>164</td>
</tr>
<tr>
<td>6.8</td>
<td>Examples of Scores of Three Different Methods under Different Parameter Settings (Subject 1)</td>
<td>165</td>
</tr>
<tr>
<td>6.9</td>
<td>Examples of Scores of Three Different Methods under Different Parameter Settings (Subject 2)</td>
<td>166</td>
</tr>
<tr>
<td>6.10</td>
<td>Brain Maps Generated by Three Different Methods (Subject 1)</td>
<td>167</td>
</tr>
<tr>
<td>6.11</td>
<td>Brain Maps Generated by Three Different Methods (Subject 2)</td>
<td>168</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivations and Objectives

For centuries, understanding human brain functions has been a topic of study for scientists and philosophers. However, the research was extremely slow due to limited techniques available. The rise of functional neuroimaging technology has resulted in the brain finally beginning to relinquish its secrets. Functional neuroimaging technology [PF07] allows the brain to be viewed without invasive neurosurgery. It has led to the capture of large quantities of digital information about the state of a brain, and provides the ability to relate complex brain functions to patterns of neural activity. Functional neuroimaging measures the brain activity when a subject is performing a task, and brain function is investigated through the analysis of this neuroimaging data. It consists of various techniques which give us images of the function of the brain. The focus of my work is the functional resonance imaging (fMRI) technique, which is an important noninvasive functional neuroimaging technique that has been developing rapidly. The fMRI technique measures signals coming directly from functionally induced changes with little known risk, and provides brain images with excellent spatial resolution leading to good delineation of the
spatial extent on an activated area. Functional MRI techniques provide a powerful tool to help understand brain functions. However, our current ability to study brain functions is still severely limited by the fMRI image quality as well as the analysis technologies. For this reason, the objective of this thesis is to improve both of these aspects of the fMRI technique. To do this we have to address the following challenges.

**Image Quality** Many researchers have been working on improving the image quality of MRI images by increasing their spatial resolution. The simplest way to do so is to increase the number of measurements used for constructing a single image. However, due to physical constraints, this will increase the scanning duration, which in turn increases scanning cost as well as discomfort for the subject, and even reduces the temporal resolution of fMRI images. Therefore, one big challenge of this area is how to reconstruct high quality images (i.e. images with high spatial resolution) with a limited number of measurements. Some researchers have managed to overcome this challenge by developing novel reconstruction algorithms, among which, compressive sensing reconstruction methods [LDP07, GBK08, WLD+06, LV09] distinct from others [NNM91, CZLZ09]. They enable accurate reconstructions given fewer measurements than are traditionally required by the Nyquist sampling theory. Other researchers [LDP07, LLLZ12, RB11, SNPS10] focused on designing efficient measurement strategies, which aim to select an optimal number of measurements so as to further improve reconstruction accuracy. Both methods have been tailored to address the reconstruction problem of the MRI images, but do not take into account the additional properties of the fMRI images which can be exploited to design more efficient methods. Additionally, both techniques were often pursued independently, and the use of two separately designed methods does not always give an optimal solution.
1.1. Motivations and Objectives

**Data Analysis** Functional MRI researchers have been showing increased interest in using multivariate pattern analysis (MVPA) since its first appearance in 2001 [HGF+01]. Due to the fact that MVPA forms simultaneous analysis of multivariate voxels, it is more sensitive to the patterns within brain activities and more robust to noises compared to the conventional univariate analysis. The early MVPA users focused on the implementation of brain reading studies, which aim to provide predictive model to predict specific stimuli with input brain images. However, predictive accuracy is not the only objective of fMRI research, discovery of new neuroscientific knowledge is also important. Therefore, the recent MVPA methods are also investigated to implement the brain mapping studies, which aim to construct brain maps highlighting the activated brain regions in response to specific stimuli. Among the existing MVPA methods, linear predictive models (e.g. linear support vector machine) are normally preferred. One one hand, they can accomplish both objectives with a single model. One they other hand, they are not easy to overfit the high-dimensional fMRI data compared to the nonlinear predictive models (e.g. nonlinear support vector machine). However, as the brain image is high-dimensional, when a whole brain image is explored, the MVPA methods face the curse of dimensionally problem. The conventional methods [DVS+08, MEKT11, FDV08] address this problem by introducing a feature selection process beforehand or limiting analysis to a small region of the brain, referred to as regions of interest (ROI). Some advanced methods [KGB06, HSR+07, KHPH11] use a “searchlight” method that applies MVPA analysis at each location in the brain. These solutions can overcome the overfitting problem, but the input voxel reduction may result in loss of significant information that defeat the purpose of MVPA. [YSY+08, CCR+09, RSAM10, VGT12] overcome the problem by employing sparse modelling methods. These methods can incorporate feature selection and predictive modelling into one single process and analysis the whole brain voxels simultaneously. Some of these methods [CCR+09, RSAM10] focus on the predictive performance (i.e. brain reading), while some others [VGT12, YSY+08] focus more on the performance of constructing brain maps (i.e. brain mapping). The standard sparse modelling algorithms
manage to find the minimum number of predictors which yields the best predictive power. As high correlations exists among voxels due to the spatial correlation, only a fraction of the correlated relevant voxels can be correctly selected with the sparse modelling methods. [CCR+09, RSAM10] have solved this problem by involving additional penalties to control the spatial correlations, though the values of the corresponding parameters have a significant effect on both the predictive and selection accuracy. As there is no ground truth for the relevant voxels (i.e. voxels of activated brain regions), they optimise the values of the parameters based on predictive accuracy of the model. This can guarantee the predictive performance. However, because of the small number samples, the selection is very sensitive to the sampling variations, so many false positive selections are introduce into the brain mapping process.

In this thesis, I aim to investigate and provide efficient methodologies and algorithms to improve the fMRI image quality as well as boost multivariate pattern analysis. For fMRI imaging, I use the special properties of fMRI sequences in the algorithm design. I investigate a systematic methodology that integrates efficient reconstruction and measurement design algorithm so as to improve the image reconstruction accuracy. For the fMRI analysis, my objective is to develop novel multivariate analysis methods in order to obtain more powerful prediction as well as provide accurate relevant voxel selection so that the neural activity can be accurately interpreted.

1.2 Contributions

Sparse modelling is an important component in many state of the art signal processing and machine learning tasks. It is based on the sparsity assumption, which states that the useful information in a high dimensional data is often sparse. Sparse modelling methods have been widely explored and proposed to efficiently solve some classical signal reconstruction and predictive modelling problems. In this thesis, I investigate application of
sparse modelling in fMRI and make the following three contributions:

- **A linear dynamic sparse modelling method to address the image reconstruction problem.**

  My proposed linear dynamic sparse modelling method consists of both measurement design and reconstruction processes. The integration of these processes boosts image quality and accomplishes image reconstruction with an underdetermined number of measurements. The algorithms for both processes are developed based on a key assumption that variations of functional MR images are sparse over time in the wavelet domain. They process cooperatively by following a linear dynamic model, which are introduced to formulate an fMRI sequence in an efficient way.

- **A novel linear sparse modelling method to implement the multivariate pattern analysis in fMRI study.**

  I develop the method based on a key assumption that active brain regions in response to a target task are sparse in the whole brain area. My proposed method works directly on the whole brain fMRI images and addresses the overfitting problem. It integrates the predictive modelling and relevant voxel selection into one process so that both the powerful predictive model and meaningful interpretation can be obtained simultaneously.

- **A novel method to boost the relevant voxel selection accuracy.**

  I introduce the concept of stability into the selection process. Instead of selecting voxels based on their predictive power alone, my proposed method selects relevant voxels by assessing both their predictive power and the level of stability. I use simulated and real fMRI data to demonstrate that my method reduces the numbers of false positive and false negative selections while simultaneously maintaining the predictive performance.
1.3 Thesis Organisation

This thesis is organised as follows. Chapter 2, is a review of the current state-of-the-art functional neuroimaging techniques, and presents a background survey of the fMRI technique, including its signal generation, signal sampling, image reconstruction and data analysis methods. In Chapter 3, I introduce what is sparse modelling and its corresponding techniques that were designed to solve signal processing and predictive modelling problems. In Chapter 4, I introduce my contribution on functional MR imaging technique, presenting my linear dynamic sparse modelling method for improving the image quality. In Chapter 5 and 6, I detail my contributions on fMRI analysis. I propose a novel linear sparse modelling method for solving the MVPA problem of fMRI analysis in Chapter 5, and a method for improving the selection accuracy of MVPA in Chapter 6. Finally, in Chapter 7, I conclude my work and present ideas for further work.

1.4 Statement of Originality

I declare that the content of thesis is composed by myself, and the work it presents is my own. All use of the previously published work of others has been listed in the bibliography.

1.5 Publications

The following publications have been written during the course of this PhD study. I mark the publications where I am the first author using * and the paper which won the best student paper award using #.

- An Approximation Approach to Measurement Design in the Reconstruction of

This paper presents a novel measurement design method for addressing the measurement design problem of fMRI sequence reconstruction. This method selects feasible measurements such that the mutual information between the unknown image and measurements is maximised with a given budget (i.e. the number of measurements). It calculates the mutual information by utilising the correlations of adjacent functional MR images: the variations of functional MR images are sparse over time in the wavelet domain. The experimental results demonstrated that the proposed method succeeded in reconstruction functional MR images with greater accuracy than random sampling.


This paper presents a novel method, called Hierarchical Bayesian Kalman (HB-Kalman) filter, for reconstructing dynamic sparse signals. This method is derived from the principles behind the Kalman filter and Sparse Bayesian Learning (SBL), and therefore succeeds in promoting sparsity and accurately tracks time varying sparse signals. Two case studies using real-world data show how the proposed method outperforms the traditional Kalman filter as well as the compressive sensing method when tracking dynamic sparse signals.


This paper presents a novel linear sparse modelling method for MVPA in fMRI studies. It models the MVPA problem using a linear sparse model based on an as-
sumption that the brain regions responding to a brain state is sparse compared to the whole brain region. The linear sparse model is built using linear Sparse Bayesian learning integrated with random subspace method. The experimental results from a real fMRI dataset demonstrate that our method has distinct predictive performance in comparison to other popular MVPA methods, and the detected relevant voxels are located in informative brain areas.

• **Balancing the Stability and Predictive Performance for Multivariate Voxel Selection in fMRI Study**. In: *The 2014 International Conference on Brain Informatics and Health*. Warsaw, Poland, 2014 [YYW+14].

This paper presents a novel method for improving the voxel selection accuracy of MVPA methods for fMRI studies. Rather than selecting the voxels based on their predictive power which is a common assessment used by classical MVPA methods, our method selects voxels which not only provides the best predictive performance but also are stable (i.e. consistently selected) when analysing using different sets of samples. The experiment results of both simulation and real fMRI datasets demonstrated that our method can simultaneously reduce the numbers of false positive and false negative selections while maintaining the predictive performance.

• **Linear Dynamic Sparse Modelling for Functional MR Imaging**. In: *Brain Informatics: Brain Data Computing and Health Studies* [YNWG14].

This paper is the extended journal version of our paper [YNWG13]. It presents a linear dynamic sparse modelling method which is composed of measurement design and reconstruction processes to improve the image quality for functional MR imaging. This method models an fMRI sequence as a linear dynamic sparse model which is based on a key assumption that variations of functional MR images are sparse over time in the wavelet domain. Novel measurement design and reconstruction algorithms following the model are investigated and proposed to implement the measurement design and reconstruction processes respectively. The experimental
results demonstrated that our proposed method succeeded in boosting the quality of functional MR images with a limited number of measurements.
Chapter 2

Functional Neuroimaging

A human brain has approximately 86 billion nerve cells or neurons that constantly interact with each other through tens of thousands of synapses [ACG\textsuperscript{+09}]. Networks of these neurons form a complicated and complex information processing system. This system enables human to obtain information from environment, process the information, and output appropriate responses so as to accomplish complex human brain functions, such as perception, language, memory, reasoning, emotion, decision-making and etc. Information that is input to and outputted from the system is in the form of active neuron patterns, which are located in different regions of the brain. For example, tactile sensations on skin activate neurons in the primary somatosensory cortex of the brain. In response, activation of different groups of neurons in the primary motor cortex lead to the movements of different muscle groups in the body. In other words, the intermediate processes of the brain (i.e. brain functions) are presumed to operate with patterns of neural activity.

The emergence of the functional neuroimaging techniques has made it possible to visualise activities of the brain; previously brain structure has been inaccessible without access to neurosurgery. Functional neuroimaging techniques, such as EEG, MEG, PET and fMRI, provide a non-invasive way to directly or indirectly measure the neural activity,
and allow scientists to understand human brain functions in terms of neural activity. They have dramatically improved our understanding of brain functions. For instance, cognitive scientists and psychologists have used the fMRI images to compare the mechanisms of healthy and unhealthy brains. By studying differences between healthy and unhealthy brains, they enable diagnosis of diseases (e.g. Parkinson, Alzheimer’s) and close monitoring of treatment efficacy to be made possible through non-invasive scans. Brain-computer is another significant application of functional neuroimaging, it enables the brain to control computers or machines. It can help patients with motor disabilities and provided a breakthrough in prosthetics control technologies. A recent study [FHW+14] employed electroencephalogram (EEG) to correlate the EEG signals and arm movements, in order to establish a connection between brain impulses and robotic prosthetic to accomplish simple movements such as reaching and grasping. The brain-computer can also help patients with perception disabilities, such as [Dob00] helped blind patients to look by using devices (e.g. electrodes) to stimulate the visual cortex.

Section 2.1 provides an overview of current functional neuroimaging techniques. Section 2.2, focuses on a specific technique called functional magnetic resonance imaging technique, by detailing its technique principles as well as the state-of-the-art analysis methods.

2.1 Overview of Techniques

In 1929, Berger [Ber29] invented the first successful functional neuroimaging technique, electroencephalogram (EEG) that records the neuron signal from the human scalp surface. Then, in 1968, another similar technique, magnetoencephalogram (MEG) [Coh68] was reported. After that, the technique that uses radio-active tracer to report the Cerebral Blood Flow (CBF) activity [OL79] appeared in 1979, and the technique (i.e. functional Magnetic Resonance Imaging (fMRI)) [OL90] that measures the neural activity based on the principle of Blood Oxygenation Level Dependent (BOLD) appeared in 1990.
Current state-of-the-art functional neuroimaging techniques mainly fall into two groups based on their principles [Shi08, LP09, CFM10]: electrophysiological and hemodynamic response. The former group includes EEG and MEG techniques, and latter includes Positron Emission Tomography (PET), Single Photon Emission computed Tomography (SPET) and fMRI techniques.

### 2.1.1 Electrophysiology based Techniques

EEG and MEG are non-invasive electrophysiological techniques. They directly measure neural activity by capturing the electrical signals generated during communications among neurons. The neural communication process is completed by signal transmission from a synaptic of the source neuron to a dendrite of the target neuron. A neuron has multiple dendrites. When a dendrite actives in response to an excitatory synaptic input, it results in a current flowing from other dendrites to the active one.

**Electroencephalography (EEG)**

EEG measures a summation of the electrical currents, which are recorded along the scalp surface of the brain. These electrical currents have to pass through various layers, such as skull and skin, before being detected from the scalp surface. The currents are directed into different orientations by the tissues they go through. Currents directed by the gyri are radial currents whereas currents directed by the sulci lead to tangential currents, where gyri and sulci are the ridges and grooves in the cerebral cortex respectively. EEG can capture both oriented current components, while it is more sensitive to the radial components. Because there could be an infinite number of possible neural generators (e.g. regions along cortical surfaces) generating the consistent EEG signal and the spatial resolution of EEG signals is low, the source localisation which localises the neural generators of measured EEG signals is very complicated. This normally requires knowledge of the
2.1. Overview of Techniques

tissues’ conductivities as they have different influences on the currents. The skull itself has low conductivity that reduces the magnitude of the signal, therefore signal attenuation is a necessary component to introduce in EEG techniques. The EEG device is relatively simple and portable; its cost is considerably less than other functional neuroimaging techniques. EEG equipment consists of electrodes with conductive media, amplifiers with filter, analogue to digital (A/D) converters, and is connected to a computer for recording and signal processing. To reduce the contact impedance at the electrode to skin interface so as to obtain appropriate conductance, skin preparation and a conductive medium is usually needed. However, this causes little discomfort to participators.

Magnetoencephalography (MEG)

MEG measures the external magnetic field, which is formed by the current flows rather than the currents themselves. As with EEG, MEG captures the signal along the scalp surfaces. However unlike EEG, it detects the neuromagnetic fields that are generated by the tangential current sources. As magnetic fields are not influenced by tissues conductivity, it is not influenced by the signal attenuation caused by the low electrical conductivity of the skull. The spatial resolution of MEG is higher in comparison to EEG, and therefore its source localisation process is much easier resulting in sources which can be localised with millimetre precision. On the other hand, the MEG device is more complex and expensive than EEG. Magnetic shielding is necessary to protect the target signal from external noises because the magnitude of the target neuromagnetic field is much weaker than the one of ambient magnetic (e.g. earth magnetic field). Different shielding methods have been used, such as superconducting shields and shielded mu-metal rooms.
2.1.2 Hemodynamic Response based Techniques

Hemodynamic response (HR) is a process that adjusts blood flow to deliver nutrients which are necessary for supporting neural tissues in order that the tissues perform their functions. Hemodynamic based techniques measure neural activity by measuring changes in the composition (e.g. oxygen) of blood near a neural event. They are based on a phenomenon that the increase in neural activity can lead to the increase in regional cerebral blood flow (CBF) to the active region, where the hemodynamic response lasts longer and covers relatively larger areas in space than the electrophysiological activity [Shi08].

Positron Emission Tomography (PET)

PET is the first scanning method to give functional information about the brain based on hemodynamic response. Patients are first injected with radioactive water, containing a radioisotope tracer. The radioisotope undergoes a positron emission decay, during which a positron is emitted. The emitted positron travels in the brain tissues and collides with an electron, resulting in a pair of photons moving towards the opposite directions with the same energy. PET scanners are built to detect and measure radiation in the form of photon pairs. To measure the CBF signal, radioactive water is often labelled with the oxygen-15 radioisotope. The 2 minutes short half-life of this radioisotope contributes to constraints such as scans to be performed within 2 mins, where the half-life of a radioisotope is the time taken for half its atoms to decay. A particle cyclotron that is required to produce oxygen-15 needs to be in close proximity to PET scanner as well as the subjects ready to be injected because of the short half-life of the radioisotope. The cost of a PET device is more expensive than other devices because of the need for a particle cyclotron as well as expensive radioactive water. On the other hand, the temporal resolution of the PET images is very poor compared to the electrophysiology neuroimages. This is due to the limitations of the PET device as well as the metabolism of the radioisotope. On the other hand,
its spatial resolution is good. PET is capable of localising neural activity to an area of
5mm/voxels within the brain, the resolution of which is determined by the quality of the
PET camera and the number of the detector rings.

**Single-Photon Emission Computerised Tomography (SPECT)**

As with PET, SPECT technique measures the regional CBF using radioisotopes. The ra-
dioisotopes which are commonly used for blood flow detection, are technetium 99 and
iodine 123. These two radioisotopes have much longer half-life (several hours) than
oxygen-15! used by PET. This enables the investigation of long-lasting tasks or events
such as walking and swimming. The spatial resolution of the SPECT images are max-
imised by using multiple detector rings, while it is still worse than the spatial resolution
of the PET images. As SPECT does not require a cyclotron and its detector rings are
much cheaper, its cost is much lower in comparison to PET. Moreover, since the par-
ticipants are exposed to radiation, the repeated scanning is limited with the PET and the
SPECT techniques. In consequence, a task is difficult to be repeated within an experiment
session.

**Functional Magnetic Resonance Imaging (fMRI)**

Functional MRI is a non-invasive method that requires no radioactive materials but uses
a more complicated physiological mechanism: when neural activity in a brain region in-
creases, the surrounding blood oxygenation (BOLD) will increase; and this leads to a
change in the magnetic properties of the blood flow. The fMRI technique measures the
blood flow by the change of the magnetic resonance (detailed in Section 2.2). The spa-
tial resolution of the fMRI images can be improved by using scanners which provide a
stronger magnetic field. This is because stronger magnets increases the signal-to-noise ra-
tio of the MRI signal. The magnets in use today in MRI are in range between $0.5T$ to $3T$,
and a powerful scanner can resolve down to 1mm/voxels. Functional MRI is considered to provide the best spatial resolution images amongst the functional neuroimaging techniques. However, the fMRI technique requires the strict fixation of subjects’ heads, which is not necessary for other techniques. Tiny movements of the subjects can blur and ruin the fMRI data, while they do not affect the quality of other images (e.g. PET). The studies of fMRI not only focus on the positive BOLD activity (the signal increase with the task), some researches are turning to study the negative BOLD activity (e.g. [SWP04, BSW07]). However, the correlation between the negative BOLD activity and the electrophysiological activity remains to be investigated. The cost of fMRI is more reasonable in comparison to PET and SPECT, in addition temporal resolution is higher enabling brain activity to be measured in seconds. Furthermore, the tasks within an experiment session are repeatable because fMRI does not require the presence of radioisotope.

### 2.1.3 Electrophysiology vs. Hemodynamic Response

Techniques based on electrophysiology or hemodynamic response have different features in terms of temporal resolution and spatial resolution. Electrophysiology based techniques can only pick up the signals that are produced by large numbers of neurons, which activate synchronously and are aligned tangentially or radially to the scalp surface. They only detect a global signal of the neural activity, therefore the spatial resolution is very low. As a result, it is impossible to localise where the activity is happening inside of the brain, so the active cortical sources have to be estimated in order to localise the brain activity. Two methods are common used to implement the estimation [Shi08]. One method searches for a single source in the brain so that field signals calculated from that source are consistent with the measured values. The other method searches for a series of sources distributed along the cortical surface, such as dynamic statistical parametric mapping methods. As these techniques directly measure neural activity, they can provide
an excellent temporal resolution so that brain activity can be captured within the span of milliseconds. For this reason, they are excellent for measuring the time-course of neural events. Hemodynamic response based techniques, by contrast, measure neural signal within the brain rather than from the scalp surface. As a result, higher spatial resolution neuroimages are generated. However, as the hemodynamic response is much longer than the electrophysiological activity (6s ∼ 8s), their temporal resolutions are not as good as electrophysiology images.

As discussed above, the imaging techniques have advantages in some aspects and disadvantages in others. Consequently, different imaging techniques are used for different applications. Because of the excellent temporal resolution of the electrophysiology based techniques, they have been widely used to investigate temporal patterns of brain activity and are helpful in clinical research. Specifically, they are good for investigating brain diseases, which have remarkable frequency or Event-Relevant Potential/Field (ERP/ERF) components. ERP and ERF are the measured brain responses, which are calculated by means of EEG and MEG respectively. For instance, Laganaro and his colleagues [LMS+09] used EEG to diagnose lexical semantic impairments by comparing ERP components of patients to health subjects, and they found early ERP abnormalities in the patient group. Accurate diagnosing epilepsy and predicting epileptic seizures [Smi05] have been realised by analysing the EEG frequencies, and MEG [WWS+05] has been combined to EEG to further boost the diagnostic accuracy. In addition to the clinical field, EEG has been used in public applications. It has been employed to implement the EEG-based brain computer interface, which establishes a direct communication pathway between the brain and a computing system. [GWPB12] used this interface to realise an emotion-based music recommendation system and [MAP+10] employed it to implement cursor control. Others designed the interface to help disable users to control external devices through computers, such as visual stimulation enabling to write [KSC+06], or a wheelchair [TMW05]. The common analysis method for EEG and MEG is Independent
Components Analysis (ICA) and ERP.

Hemodynamic response based techniques study brain functions by investigating the spatial pattern of brain activity rather than the temporal pattern. They allow the brain to be explored in more details because of their high spatial resolution. They are used to determine precisely which part of the brain is handling brain functions such as memory, language, and reasoning. They are used regularly to investigate brain diseases such as Parkinson and Alzheimer’s. With PET or fMRI, researchers aim to differentiate active brain regions between patient and healthy subjects when a task is performed. For example, [WH05] found in performing automated movements, Parkinsons disease patients had greater activity in the cerebellum, premotor area, parietal cortex, precuneus and prefrontal cortex. [GSR+05] used memory tasks to study the Alzheimer's disease, they indicated a dissociation in Alzheimers disease based on differently impaired brain regions. This type of study is mainly completed via univariate analysis (e.g. generic linear model). Some other researchers [NPDH06, VDE+11] investigate relationships between brain functions and diseases in order to build predictive models. This type of study is often implemented by decoding the brain functions with the responses of multiple voxels in the brain, and it is usually implemented via multivariate pattern analysis (MVPA) which can capture more information than univariate analysis from the neuroimages.

Brain connectivity is another study of the functional neuroimaging techniques, and it has attracted recent attentions of researchers. It can give us a picture of how interacting neural populations give rise to brain functions. Coherence and correlation analysis are two simple methods to measure the similarity of the signals in two brain regions. The coherence analysis works on the frequency domain; it measures the linear dependence of the frequency components of two signals, while the correlation analysis measures the linear relationship between two signals in their original domain (i.e. time domain). Both the electrophysiology and hemodynamic response based techniques can be utilised to implement the connectivity study. The coherence analysis of the EEG signal and correlation
analysis of the fMRI signal are two commonly used approaches and have demonstrated some significant findings. For instance, Wheaton [WNB+05], which used the EEG technique, demonstrated that there was an increasing coherence between the left parietal region and the left frontal region of a subject when he/she was gesturing. [OOK+04] worked on finding the coupling brain regions that were involved in the verbal working memory. By utilising the fMRI technique, they found the anterior cingulate cortex and the left prefrontal cortex in the subjects with good memory performance have a higher correlation than those in the subjects with worse memory performance.

2.1.4 Multimodal Techniques

More recently, researchers began to explore multimodal functional neuroimaging techniques, which are implemented by combining the electromagnetic and the hemodynamic response based techniques so as to overcome the shortcomings of each other. EEG and fMRI are two modalities which are often combined. Many researchers have used these two techniques in the same experiment to acquire different types of neuroimages simultaneously. Combined modalities take the advantages of the two techniques and result in a signal with high resolution in both time and space domain. Epilepsy is a popular study of using this multimodal (i.e. EEG-fMRI) technique [SHDH+06, AABG+03]. Because the signal of epilepsy changes too fast in time domain to be captured by the hemodynamic response based techniques, most previous epilepsy studies used the EEG technique. However, the localisation of epileptiform activity to a specific brain region is limited because of the low spatial resolution of EEG images. Simultaneously recording the EEG and fMRI signals provides an opportunity to investigate the epileptiform activity in terms of both its spatial and temporal mechanisms. However, the use of EEG within the fMRI device introduces prominent artefacts. To obtain accurate neuroimages, these artefacts need to be eliminated from the EEG signals in real time and corresponding methods have been
explored and developed.

EEG-PET is another multimodal method [ONH+00]. Even though PET images have lower spatial resolution than the fMRI images, the PET signal is less disturbed by the EEG devices. In other words, it introduces much less artefacts than the EEG-fMRI technique. [IJF05] proposed a multimodal technique by combing MEG and fMRI. Unlike the EEG-fMRI technique, MEG and fMRI images have to be recorded separately because of some technical issues. The functional neuroimaging techniques have been also cooperated with other devices to improve the accuracy of the study. For example, an eye tracker is used to track the eye movements so that the attention that paid by the subjects are registered. It has been used to remove eye movement artefacts in neuroimages [POK12]. A camera, which records the subjects motion during an fMRI experiment, can help to remove the motion noise [MW10].

2.2 Functional Magnetic Resonance Imaging

In this thesis, I focus on the studies of the functional magnetic resonance imaging (fMRI) technique which provides good spatial and relatively good temporal resolution brain images in a non-invasive way. As discussed above, it has been one of the most successful techniques that are used for investigating brain functions. Over the last decade, it has been employed to investigate how human processes thought, visual perception, memory, affection, and used in clinical and commercial researches.

The functional MRI technique relies on the fact that the changes of cerebral blood flow are closely linked to neural activity of the brain. The primary fMRI methods are based on Blood Oxygenation Level Dependent (BOLD) effect. Those methods map neural activity by imaging the change in blood flow which is related to energy consumed by brain cells. As neural cells do not serve energy itself, the energy required for performing neu-
ral activity is supplied by chemical reaction of oxygen. When neural activity increases, the demand for oxygen is increased. Oxygen is carried by the haemoglobin molecule in red blood cells, it is released by the transformation of haemoglobin from oxygenated to deoxygenated. As the amount of blood flow sent to the activity region for providing oxygen is always more than needed, relative surplus in local blood oxygen is produced. In 1936, Pauling and Coryell [PC36] found that the deoxygenated haemoglobin (dHb) is more magnetic than oxygenated haemoglobin (Hb). This different magnetic properties lead to different strengths of blood’s MR signals. As the degree of oxygenation varies according to the levels of neural activity, the difference in MR signals is used to measure neural activity.

Experiment design is a significant component in the functional neuroimaging study. Correct design of the experiments is critical to the understanding of brain functions. A good designed experiment should be able to correctly formulate the questions to be explored. It requires the designer to determine the inference one wish to draw from the experiment. Besides, the designer need to control the extraneous variable and sample size, minimise the potential experiment bias and etc. In fMRI, the experimental paradigms mainly fall into two categories: blocked design and event-related design. In a block design, the task conditions alternate in blocks, and each block/trial last a certain duration. This design offers a considerable statistical power when the number of blocks is relatively large. On the contrary, the duration of each trial in an event-related design is not fixed, which allows more real-world testing than the blocked design. It can capture the transient changes in brain signal, while the statistical power is scarified.

Imaging technique and data analysis are another two key components of fMRI studies, which have a significant impact on the understanding of brain functions. In this chapter, I focus on the researches of these two components. I first introduce the imaging technique in Section 2.2.1. In Section 2.2.2, I will present the overview of the existing popular functional MRI analysis methods.
2.2.1 Imaging Technique

The physics and hardware of Functional Magnetic Resonance Imaging (MRI) technique are complicated. However, from the perspective of signal processing, it can be divided into three processes: signal generation, signal sampling and image reconstruction. The signal generation process generates collected encoded signal of blood flow. The signal sampling process samples the encoded signal by following specific strategies, and the image reconstruction process reconstructs brain images from the sampled encoded signal. The concepts are based on [Bux09, LL00].

Signal Generation

**Nuclear Magnetic Moments**  A nucleus is formed by protons and neutrons which are combined in pairs and with oppositely oriented spin. As all protons and neutrons have the same and stable magnitude of angular momentum, only the nucleus (e.g. $^1$H and $^{13}$C) which consists of odd numbers of protons and neutrons has a spin. A nucleus with non-zero spin creates a magnetic field (called magnetic moment) around. Given the spin angular momentum $\vec{J}$ which is a measure of the amount of rotation a nucleus has, the magnetic moment represented by a vector $\vec{\mu}$ is:

$$\vec{\mu} = \gamma \vec{J},$$

(2.1)

where $\gamma$ is a constant which is called the gyromagnetic ratio, the value of which is different for different nuclei. MR can only detect magnetic field around a macroscopic object other than a microscopic one (e.g. nucleus). In the absence of other effects, the direction of magnetic moment is completely random because of the thermal random motion. Therefore, there is no macroscopic magnetisation which can be measured by MRI devices.
2.2. Functional Magnetic Resonance Imaging

**Net Magnetization** To activate the macroscopic magnetisation, a static magnetic field \( B_0 \) along \( z \)-axis is placed. This makes nuclei precess around the field axis, where frequency of the precession (i.e. Larmor frequency) is given by:

\[
\nu_0 = \gamma B_0.
\]  
(2.2)

With the added static magnetic, a macroscopic magnetisation \( M \) summarising all microscopic magnetic moments \( \mu \) is generated; where \( M_\ast \) indicates the magnitude of macroscopic magnetisation along the corresponding direction. In this situation, the macroscopic magnetisation is represented by \( M_0 \) which is a weak equilibrium magnetisation aligned with the field \( B_0 \).

**Radiofrequency Pulse Excitation** Even though the added magnetic field \( B_0 \) produces a macroscopic magnetisation, this magnetisation is still not be able to be observed. This is because this macroscopic magnetisation induces a non-oscillating voltage (i.e. DC voltage) which cannot be measured by receiver coils of the MRI devices. In order to generate a measurable signal, an oscillating Radiofrequency (RF) pulse which tips the macroscopic magnetisation away from the equilibrium is applied. When the RF frequency matches the precession frequency \( \nu_0 \) of nuclei, a net magnetic field \( B_1 \) wobbling back and forth is produced. This magnetic field makes the macroscopic magnetisation precess around it. Since the receiver coil can only detect oscillating signals, it measures the transverse magnetisation \( M_{xy} \) but not the magnetisation along the \( z \)-axis (i.e. \( M_z \)). Because the magnitude of \( M_{xy} \) represents the spatial distribution of the transverse magnetisation, the nuclei density reflecting the properties of an object can be measured.

**Relaxation and Image Contrast** When a 90° FR pulse is applied, the macroscopic magnetisation with magnitude equalling to the magnetisation of \( M_0 \) flips to the \( x-y \) plane. After the FR pulse is removed, the magnetisation returns to align with the static magnetic
field while the transverse magnetisation becomes to zero. The former process that the longitudinal magnetisation grows back to its original size is called longitudinal relaxation, while the later one that the transverse magnetization starts to disappear is called transverse relaxation. $T1$ and $T2$ are time constants for completing the two relaxation processes respectively. The longitudinal magnetisation $M_z$ grows exponentially by $T1$, while the transverse magnetisation $M_{xy}$ decays exponentially by $T2$. The values of them depend on the tissue composition, structure and surroundings.

During the relaxation process, a signal is generated and measured via a receiver coil. The measurable signal is approximately by:

$$Signal \propto M_0(1 - e^{-TR/T1})e^{-TE/T2},$$

where $TR$ refers to the time interval of exciting the nuclei by applying RF pulses, and $TE$ (called time echo) is the time to collect signal after excitation.

Figure 2.1: Example of Different Weighted MR Images. This figure is generated with the fMRI data provided by [BBH+13].

From Equation 2.3, we can see that short $TE$ reduces $T2$ effects, while short $TR$ reduces $T1$ effects. In consequence, by controlling the values of $TR$ and $TE$ we can get signals that are sensitive to $T1$ or $T2$ so as to generate different weighted images (as shown in Figure 2.1). The difference between $T1$ and $T2$ weighted images is due to the density
of fluids, the $T_1$ weighted image is ideal for examining solid organs, while $T_2$ weighted image is ideal for examining soft tissues.

**Spatial Information Encoding**  The signal which is achieved with the presence of static magnetic field and RF pulse is the ensemble of all the protons in the object, it has no spatial information. As the precession frequency of a nucleus is proportional to the magnetic field at its location, the spatial information can be achieved by using three gradient coils $G_x$, $G_y$, and $G_z$ in addition to the RF coils. The three gradient coils are designed to provide field gradient along three orthogonal directions, they are added to the static field $B_0$ with much smaller magnetisation in comparison to $B_0$. Considering a linear field gradient $G_x$ along $x$-axis is added to the static field, the magnetic field at position $x$ is $B(x) = B_0 + G_x x$, where its Lamor frequency turns to be $\nu(x) = \nu_0 + \gamma G_x x$.

With presence of the gradient field gradients, the spatial localisation is completed in three directions: slice selection, phase encoding, and frequency encoding. The slice selection works along $z$-axis, it selects a slice at a certain level with a certain thickness. As the RF can only excite the protons, whose Lamor frequency are equal to the RF frequency, the location of a selected slice is determined by the setting of RF frequency. In addition, to control the thickness of the slice, the RF pulse is set with a range of frequencies (i.e. a bandwidth of frequencies). Once a slice is selected, the spatial distribution is encoded by phase and frequency encodings. The frequency encoding separates signals at different positions along the $x$-axis, and each of the separated signals is the sum of all signals at the same $x$ position but at different $y$ positions. Phase encoding is involved to further separate the signals so that a two-dimensional image can be achieved.
**K-Space** The signal that is contributed from a small region between \( r \) and \( r + dr \) is \( M_r dr \), where \( r = (x, y) \). Then, the measured signal equals to:

\[
\text{Signal} = \int M_r dr.
\] (2.4)

At time \( t \), there is an additional phase \( 2\pi \gamma G_r t \) to the spin magnetisation, so the signal measured at time \( t \) is:

\[
s(t) = \int M_r e^{-i2\pi\gamma G_r t} dr,
\] (2.5)

where \( G_r = G r \) indicates the gradient field at position \( r \). Let \( k = \gamma G t \), the Equation 2.5 becomes to:

\[
s(k) = \int M_r e^{-i2\pi rk} dk.
\] (2.6)

Equation 2.6 states that the received signal \( s(k) \) is the Fourier transform of the object sampled at frequency \( k \). For this reason, \( s(k) \) is called \( k \)-space signal representing the 2D Fourier transform measure of the MR image. In the \( k \)-space, data points near the centre of the space present the overall information of the image (i.e. low frequency components), while the points in the edges of the space present the details of the image (i.e. high frequency components).

**Signal Sampling**

The generated \( k \)-space signal can be sampled following some specific paths. The path of samples in the \( k \)-space is referred to as a trajectory, which can have various shapes ranging from parallel grid lines to pseudo-random paths. Cartesian, echo planar imaging and spiral imaging are three popular trajectories that have been used in fMRI. The sampling trajectories are accomplished by manipulating the gradient field \( G \), and they are designed to meet hardware conditions. How the \( k \)-space is sampled is of great importance to the reconstruction step. In addition, as the change of the blood flow in response to the neural
activity remains only $6s \sim 8s$, the signal sampling methods must be fast enough to capture the significant change.

**Cartesian Sampling** Cartesian sampling trajectory which samples equal spaced lines of $k$-space has been most widely used in MRI technique. The Fast Low Angle Shot (FLASH) imaging arose in 1986 [HFM+86] is a low flip angel fast imaging technique that uses Cartesian sampling trajectory. A lower flip angel causes a smaller relaxation time, which reduces the time for generating MR signal so that the encoded signal used for reconstructing a image can be captured in a shorter time interval. FLASH was used before the emergence of fast gradient hardware technique. However, because of the hardware limitation, FLASH still needs $3s \sim 6s$ to measure a slice, which is impractical for fMRI application. In addition, because of its long acquisition time, it is very sensitive to hardware instability. With the improvement of hardware, other Cartesian based imaging methods emerged to accelerate the imaging speed, such as FSE (Fast spin echo) and HASTE (half-Fourier acquisition with a single-shot turbo spin echo) [Bux09]. The FSE technique shortens the imaging time by collecting multiple lines (i.e. shots) of $k$-space with one RF pulse. As with FSE, the HASTE is a multi-shot technique. However, it only measures half of the $k$-space, which works by taking the advantage of symmetry of the $k$-space.

**Echo Planar Imaging** Echo Planar Imaging (EPI) [CS12] has been the most widely used function MR imaging technique due to its remarkable acquisition speed. It can scan the whole brain with slice thickness equalling to 4mm in about $3s$. However, it introduces various sources of artefacts (e.g. geometric distortion and Nyquist ghosting) and therefore correction algorithms have to been involved to minimise the artefacts. As with Cartesian sampling trajectory, it samples the whole $k$-space line by line but in a back and forth scanning pattern. The EPI highly improved the spatial resolution in comparison to Cartesian
sampling, while its single-shot technique generated images with spatial resolution limited up to $128 \times 128$. In order to improve the spatial resolution, multi-shot EPI was proposed to cover more of the $k$-space. Partial EPI acquisition method which collects slightly more than half of the complete $k$-space data was introduced to further reduce the scan time.

**Spiral Imaging**  
Spiral imaging [Glo12] is another practical technique for fMRI application, the sampling trajectory of which spirals out from the centre of $k$-space. It can be implemented in a single-shot fashion, and higher spatial resolution images can be achieved via a multi-shot spiral imaging technique. Spiral imaging efficiently uses the available gradient field. It exhibits blurring artefacts rather than ghosts or geometric distortion. Spiral imaging has not been widely used by MR imagers due to its high computational complexity for image reconstruction process.

**Image Reconstruction**

The objective of image reconstruction in MRI is to reconstruct spatial brain images from the acquired encoded $k$-space signal. When the sampled signal covers the whole $k$-space in uniform grid (e.g. Cartesian and EPI), the image can be reconstructed by directly calculating the inverse Fourier transform of the sampled signal. The image reconstruction of spiral imaging technique which generates non-uniformly sampled $k$-space data is more complex. It requires a interpolations process which maps the samples onto rectangular grid before implementing the inverse Fourier transformation. This complex reconstruction process lead to long computation time.

When partial $k$-space sampling (e.g. HASTE, partial EPI) methods are applied, the missing data must be filled in first. The most direct way is to simply fill the uncollected data with zero. This method results in blurring images and the result is acceptable only if the collected $k$-space faction is close to 1. Homodyne [NNM91] is another standard partial
Fourier reconstruction method. It fills in the missing data based on the symmetry property of the Fourier transform that the real part of an image corresponds to the conjugate symmetric component of the transform. Also, it assumes the tissue varies slowly in phase direction. Both properties are constrained with a weighting function. The weighting function preweights the $k$-space data so that a uniform part of the image corresponding to a uniform weighting in the $k$-space. This method requires phase correction and works well only if the image in phase direction changes slowly. Projection onto convex sets (POCS) [CZLZ09], an iterative algorithm, is a more advanced and popular method for partial $k$-space reconstruction. POCS fills in the missing data by iteratively transforming between spatial and frequency domain. Compared to the Homodyne method, it is not limited by the changes of image phase and performs better when the sampled $k$-space fraction is small.

The above reconstruction methods require the sampled data covering at least half of the $k$-space so as to satisfy the Nyquist sample theory: the sampling frequency must be more than twice the bandwidth of the source signal to produce an errorless reconstruction. The recent researchers [LDP07, GBK08, WLD+06, LV09, LLAV11] managed to speed up the functional MRI imaging using Compressive Sensing (CS) technique which allows a exact reconstruction even when the samples are against the Nyquist sample theory. By applying CS methods, high quality functional MR images can be reconstructed with much fewer samples. This can highly reduce the number of samples, so the imaging speed can be further increased. The compressive sensing based reconstruction methods and corresponding sampling methods will be explained in Section 4.1 in Chapter 4.

### 2.2.2 Data Analysis

The major target of functional MRI analysis is to find the correlations between brain activity and target tasks (i.e. brain states). The BOLD activity signal is relatively weak
in comparison to the various sources of noise (e.g. system noise, physiological noise and etc.) involved in the sampling process. For this reason, preprocessing steps must be used to increase the signal-to-noise ratio so that results from later analysis operations are robust. Different preprocessing methods are provided by different softwares (e.g. SPM, FSL) and used by different laboratories. A standard preprocessing pipeline consists of steps, such as motion correction, slice timing correction, distortion correction, alignment and registration. These steps aim to remove various types of artefacts in the data and may prepare the data (e.g. normalisation) where necessary.

Recent fMRI studies fall into two complementary categories: brain mapping and brain reading. Both studies aim to understand the representation of information in brain. Brain mapping uses stimuli to predict the brain activity, while brain reading uses brain activity to predict the state of the stimuli. Univariate analysis and Mutivariate Pattern Analysis (MVPA) are two popular fMRI analysis techniques for accomplishing both types of studies. Most recently, brain connectivity study has been drawn increasing attentions. It manages to understand the brain functions in terms of interactions of activated brain regions, which is more approximate to the real mechanisms of the brain.

**Univariate Analysis**

Univariate analysis methods analyse each isolated voxels independently. They determine active brain regions by selecting the most statistically significant voxels in response to a target task, and are typically used for implementing brain mapping. Univariate analysis methods can also be used for implementing brain reading by using the detected significant voxels to predict the target task. General Linear Model (GLM) [FJT94] is the most popular univariate analysis method in the context of fMRI data. It sets up a model and fits
it to each voxel’s time-series data:

\[ x = E\beta + \epsilon \]  

(2.7)

where \( x \in \mathbb{R}^m \) is a column vector containing BOLD signal at a single voxel, and \( \epsilon \) refers to the zero-mean Gaussian noise with unknown variance \( \sigma^2 \). \( E \) is a \( m \times e \) design matrix describing the model that one row per time point and one column per explanatory variable. Each explanatory variable is initialised as a series of 1s and 0s, which describes a specific stimulus state. As the brain’s hemodynamic response is a delayed and blurred version of the input time-series, the explanatory variable is convolved with the hemodynamic response function (HRF) to get a better possible fit of the model to the data. The column vector \( \beta \in \mathbb{R}^e \), which consists of the parameters of interests, is estimated to minimise the error \( \epsilon \). If the voxel responds strongly to a stimulus \( i \), a large value will be found for \( \beta_i \).

To test if the estimated parameter is significant from zero (null hypothesis), \( T \) value is usually used. The \( T \) value is calculated by comparing the parameter to its uncertainty in its estimation, and other statistic values (e.g. \( p \)) can be converted from it via a standard statistical transformation. After achieving a statistic map (e.g. \( T \)), a thresholding method is applied to determine, at a given level of significance, the part of brain regions which were activated. Using a certain significance \( p \) to filter the voxels is the simplest method, but this introduces large number of false positives that is caused by the large number of tests (i.e. number of voxels). To reduce the number of false positives, correction methods are required. Bonferroni correction, a simple voxel-wise method, is typically used in fMRI data. It multiplies the significant value at a voxel by the number of tested voxels. To take the spatial smoothness of statistic map into consideration, a multiple comparison method, called Gaussian random field (GRF) theory, was proposed and has been widely used. The GRF method estimates the likelihood by which the voxels with particular statistic levels would appear by chance according to the local smoothness. This method is applicable
as, in univariate analysis, a spatial smoothing preprocessing step is normally applied to the data beforehand. The purpose of the spatial smoothing process implemented with a Gaussian kernel is to reduce the effect of high frequency variation in brain images so as to increase signal-to-noise ratio.

The above method is called first-level analysis, and it is only used for within session (separate experiments on the same subject) analysis. In order to accomplish multiple sessions or subjects analysis, a second-level analysis is introduced. A second-level analysis aligns brain images and combines independent statistic maps of the sections/subjects obtained through the first-level analysis.

**Multivariate Pattern Analysis**

In contrast to univariate analysis, Multivariate Pattern Analysis (MVPA) of fMRI attempts to informatively decode patterns of brain activities [HGF+01]. By measuring multiple voxels simultaneously, MVPA is more sensitive to patterns within brain activities and more robust to noises. As a result, MVPA can improve the predictive power in comparison to the univariate analysis methods, and therefore it is better for implementing the brain reading function. On the other hand, implementation of MVPA is more complex. Unlike univariate analysis which is implemented in SPM and FSL, implementations of MVPA are lacking. In addition, it directly analyses the fine-grained images (unsmoothed data), so the implementation of across subjects analysis is a big challenge.

**Curse of Dimensionality**  
Curse of dimensionality is a common problem in high dimensional data analysis. With a fixed number of samples, when more features are taken into consideration the predictive performance should be increased. However, after an optimal number of features is utilised, the increase of feature dimension leads to a significant degradation on the predictive performance. This is because the increased feature
2.2. Functional Magnetic Resonance Imaging

dimension results in a more complex model which describes noise more, so the model perfectly fits the training dataset but its predictive performance on unseen data is worse (called overfitting). This phenomenon is called the curse of dimensionality. In fMRI, for each individual subject, the number of features/voxels ($\approx 1$ million) is many times larger than the number of the samples ($100 \sim 200$). This large feature-to-sample ratio results in a significant curse of dimensionality problem, evidenced in a conventional predictive model (e.g. naive Bayes and linear support vector machine) overfitting the training dataset.

**Conventional Methods** Conventional machine learning methods complete the MVPA analysis via two separate steps: feature selection and predictive modelling. The feature selection process overcomes the curse of dimensionality problem by reducing the feature dimension before implementing the predictive modelling process. To further reduce the dimensionality, some conventional MVPA techniques performed the analysis on specific brain regions of interest (ROI). However, this requires prior information of the study, which is normally unknown in practice. In addition, a target task usually activates multiple brain regions. For these two reasons, analysis methods are preferably applied to analysing the entire brain region data.

**Feature Selection** Feature selection is normally realised using two techniques. One technique implements the dimensionality reduction via principle components analysis (PCA) or independent components analysis (ICA). The features/voxels are first projected on an alternative space. The resulting components are ranked according to their significance when projected onto this space, and a selection of the most significant components are used as features, thereby reducing the dimensionality of the data set. The other technique reduces the number of features in their original space. It assumes the useful features are sparse in data.

The latter technique has been more widely used in fMRI analysis. This is because its
assumption that the active brain regions in response to the target task are sparse compared
to the whole brain region is normally the case in fMRI studies. Features may be selected
from statistic maps generated by GLM or Independent Component Analysis (ICA). ICA
is a multivariate data analysis method which is developed to overcome the limitation of
GLM. As the HRF may vary between subjects or different brain regions, the GLM method
that employs a predefined HRF may not provide an ideal analysis as the selected HRF is
usually not precise in practice. ICA overcomes this limitation by learning the sources
of stimuli as well as noise (i.e. design matrix) from the data. This blind source separa-
tion method estimates the parameter $\beta$ as well as the design matrix $D$ in Equation 2.7
by analysing the voxels simultaneously, not independently. After selected a stimulus of
interests, the corresponding statistic map and significant voxels can be achieved via the
similar steps of GLM. The feature selection process implemented by GLM or ICA is
preferably to be performed on spatial unsmoothed data so that fine-grained information
remains. A more advanced method, called recursive feature elimination (RFE) [Rak03],
has been introduced to improve the feature selection by introducing cross-validation pro-
cess. It is a multivariate selection method that iteratively removes the least informative
features detected by the predictive model.

**Predictive Models** With the reduced number of features, most conventional MVPA
methods use linear predictive models, such as linear discriminant analysis (LDA), naive
Bayes, and linear support vector machine (SVM). These methods construct a predictive
model using individual voxels as predictors and time point volumes as samples, and they
compute a weighted sum of voxels to determine brain state. The value of a weight indi-
cates the relevance of its corresponding voxel to the specific brain state. Prediction-based
brain mapping can be then implemented using the selected relevant voxels. Other MVPA
methods employ nonlinear models, including nonlinear SVM and neural networks with
hidden layers. The predictors of the nonlinear models are voxels projected onto a dif-
ferent space. In consequence, brain maps cannot be constructed. Nonlinear predictive models can show good predictive power even when none of the input voxels are individually sensitive to the brain state; however these models overfit the training dataset more regularly compared to linear predictive models. Furthermore, the optimisation process of nonlinear models is more complex. This is because their complex functions introduce additional parameters, the optimal values of which have to be selected with additional layer of cross-validation process. In addition, many researchers prefer linear predictive models due to added value in the form of brain mapping [DVS⁺08, MEKT11, FDV08].

Many researches [FDV08, MKBK10, DVS⁺08] have been working on comparing the performances of the conventional predictive models. Among them, linear SVM has been demonstrated to have the best performance.

**Advanced Methods** Feature selection can help to overcome the overfitting problem, while the voxel reduction may result in loss of significant information which defeats the purpose of MVPA. More advanced methods [YSY⁺08, CCR⁺09, RSAM10, VGT12] employed linear sparse modelling to solve the dimensionality problem by exploring spatial sparsity of fMRI data, these methods provide an opportunity to analysis voxels of the entire brain directly. As with the conventional linear predictive modelling methods, they compute a weighted sum of voxels to determine the brain state but set most of the weights of the voxels to zero. This is based on the assumption that relevant voxels in response to a target task is sparse compared to the whole brain voxels. As linear sparse modelling only uses a small subset of input voxels for prediction, the overfitting problem can be remitted. In contrast to the conventional modelling methods whose feature selections are based on straightforward filtering methods (e.g. significant test), linear sparse modelling optimally chooses a subset of voxels without using any hard cutoff threshold (i.e. significant value), it selects the relevant voxels using the non-zero weights (i.e. coefficients of the estimated model parameter). The linear sparse modelling methods incorporate feature
selection and modelling into one single process so that relevant voxels are automatically selected. Brain maps constructed with the selected relevant voxels can be used to interpret neural activities.

In addition to the sparsity property, a “group effect” exists in the relevant voxels, which indicates that many relevant voxels are strongly correlated with each other. The main cause of this is the spatial correlation: the brain activity units are localised regions rather than voxels. The conventional linear sparse modelling methods (e.g. lasso [Tib96]) seek the minimal set of voxels that provides the best predictive performance, they only select a fraction of the closely correlated relevant voxels. However, the target of most MVPA studies is not only to predict well, but also desires that the obtained relevant voxels express more biologically information to provide an accurate and meaningful primary results for brain mapping. To overcome this problem, some advanced methods [CCR+09, RSAM10] added an additional penalty to increase the “group effect” and reduce the sparsity constraint. The fundamentals of these two methods are same, except that one [RSAM10] is proposed for solving the classification problem and the other one [CCR+09] is for solving the regression problem. The Bayesian method [YSY+08] controlled the “group effect” by adding spatial prior to the weights of voxels, and [VGT12] integrated lasso with clustering and boosting algorithms so as to obtain a more accurate and comprehensive relevant voxel selection. Among them [CCR+09, RSAM10] focused more on the predictive performance (i.e. brain reading), while [VGT12, YSY+08] focused on the relevant voxel selection (i.e. brain mapping). Because of the lack of ground truth in real fMRI datasets, these methods were tested on the simulation datasets, and desirable results were demonstrated.

**Brain Connectivity**

Most current activation studies focus on localising distinct brain regions in response to brain functions via either the univariate analysis or the MVPA methods. However, some
researchers started to attempt to understand brain functions by studying brain connectivity. The aim of this type of study is to understand how brain regions interact and work together to create brain functions, which better approximates the brain’s operational mechanism. It has the potential to provide more meaningful brain mapping as well as highly improve the brain reading performance, and therefore will play a major role in the study of understanding human brain. The current brain connectivity studies using fMRI technique centred on two different forms of connectivity: 1) functional connectivity and 2) effective connectivity; and their principles and existing methods have been reviewed by [Fri11, PMN11, LTK+12].

**Functional Connectivity**  Functional connectivity refers to the temporal correlations in activity between spatially remote brain units (i.e. regions or voxels) based on the fluctuations in their BOLD signal. When the brain regions are chosen as unit to be investigated, their signals are generated using either the mean of voxels within the region or the first eigenvariate of the region. There are no significant differences between these two methods. Early studies of functional connectivity used multivariate decomposition methods (e.g. PCA and ICA). These methods are simple and easily implemented, whereas the correlations are calculated for components of the brain units rather than the units themselves. As a result, the spatial correlation is difficult to be extracted.

In current times, full correlation analysis is the most commonly used method in practical clinical research. The time course of a chosen unit is correlated with the remaining units’ time courses in a unit-by-unit matter. The degree of the correlation of a unit pair is indicated by the value of their estimated correlation coefficients. A large value indicates a high correlation, and the most significant correlated pairs are selected use a hard cut-off threshold. However, a significant correlation between unit $A$ and $B$ obtained by this method can arise for a number of reasons [PMN11]. Firstly, it may be caused by a direct causation $A \rightarrow B$ or $B \rightarrow A$. Secondly, it may be caused by an indirect interaction that
the influence is mediated by a third unit, e.g. $A \rightarrow C \rightarrow B$. Thirdly, it may be caused by an influence resulted from the same source, e.g. $C \rightarrow A$ and $C \rightarrow B$. To correctly estimate the direct interaction, partial correlation analysis method is employed [Smi12]. As with the full correlation analysis, it estimates the correlation of two units in time course. However, before the estimation process, time course of a third unit is regressed out from each of the two. If correlation is truly influenced by a third unit, the correlation will no longer exist after the influence is removed.

**Effective Connectivity**  Effective connectivity refers to the causal influences that brain units exert to another. Effective connectivity is usually built upon functional connectivity. This is because if a correlation is present, it implies that a causal relation exists either between the two units or some unmeasured third unit. Effective connectivity requires the causal model to describe the causal interactions between two active brain units. It can simulate the direction of the information that flows around the brain network, which is more meaningful but harder to estimate. Granger causality and dynamic causal modelling (DCM) [Fri11] are two prevalent modelling methods that have been used for this study. The Granger causality, a multivariate autoregression modelling method that is originally developed for analysing the economic data, models the causal relation by checking if the time course of a unit is the time-shifted version of the time course of another unit. It does not need to specify an anatomical network, and can be applied on whole brain map analysis. On the other hand, the dynamic causal modelling is the first specific method tailored specifically to effective connectivity analysis. The model is comparatively more complex and requires a pre-specified hemodynamic response function. In addition, it has only been demonstrated to be practical on a very small number of units ($< 10$).
Chapter 3

Sparse Modelling

In signal processing, the relationship between a source signal \( x \in \mathbb{R}^n \) and its corresponding measurements \( y \in \mathbb{R}^m \) is often described by a linear model:

\[
y = \Phi x,
\]

where \( \Phi \in \mathbb{R}^{m \times n} \) is a measurement matrix that is constructed via a predefined project function (e.g. Fourier transform function) which is determined by physical constraints.

Signal reconstruction is a common goal in signal processing, the aim is to reconstruct the source signal from a series number of measurements by using Equation 3.1. When the number of measurements is greater than or equal to the dimension of the source signal (i.e. \( m \geq n \)), an accurate and unique solution of signal \( x \) can be efficiently calculated by using simple linear regression methods [SL12] (e.g. ordinary least squares). However, in many applications, the number of available measurements is much smaller than the dimension of the source signal (i.e. \( m << n \)). The number of measurements may be limited by the number of measurement devices (e.g. sensors), high measurement costs or low measurement speed. When we try to reconstruct the source signal from such a small number of measurements, the number of unknowns (i.e. coefficients \( \{x_i\}_{i=1}^n \) is much...
Chapter 3. Sparse Modelling

larger than the number of equations \( \{ y_j = \phi_{j1}x_1 + \phi_{j2}x_2 + \ldots + \phi_{jn}x_n \}_{j=1\ldots m} \), which gives an infinite number of solutions for \( x \). These underdetermined measurements, the number of which is much smaller than the dimension of the signal to be reconstructed, make the signal reconstruction challenging.

Sparse modelling is a powerful tool to handle this underdetermined signal reconstruction problem. It formulates the relation between \( x \) and \( y \) using the linear model (shown in Equation 3.1), and places a sparsity constraint on \( x \). The sparsity constraint is that \( x \) is either a sparse signal itself or can be sparsely represented in another domain. This constraint is easy to be satisfied as, in many applications, a high-dimensional source signal contains relatively little information compared to its ambient dimension. The high dimensional source signal is naturally either sparse itself or can be sparsely represented by a transformed signal \( w \in \mathbb{R}^n \):

\[
x = \Psi w,
\]

where \( \Psi \in \mathbb{R}^{n \times n} \) is referred to as the sparse matrix. \( w \) is a sparse signal if most of its coefficients \( w_i \) are zeros and the rests are non-zeros and it is \( S \)-sparse if the number of non-zero coefficients is \( S \), where the non-zero coefficients are called supports. When \( x \) is sparse itself, \( w \) equals to \( x \) and the sparse matrix is an identity matrix, whose diagonal elements \( \{ \psi_{ii} \}_{i=1}^n \) are ones and the rests are zeros. Sparse modelling reconstructs the source signal by first reconstructing the sparse signal \( w \), and using Equation 3.2 to calculate the source signal. According to Equations 3.1 and 3.2, the relation between the measurements and the sparse signal is formulated by a linear sparse model:

\[
y = \Theta w,
\]

where \( \Theta \in \mathbb{R}^{m \times n} \) is called a design matrix. In this signal processing process, the design matrix is constructed by the measurement matrix \( \Phi \) and the sparse matrix \( \Psi \) so that \( \Theta = \Phi \Psi \). When \( w \) is \( S \)-sparse with \( S << n \), all the information of signal \( w \) is contained
by the small number of non-zero coefficients. These non-zeros are the unknowns to be calculated and they are represented with a small number of linear equations:

\[ y_j = \theta_{s_1}w_{s_1} + \theta_{s_2}w_{s_2} + \ldots + \theta_{s_S}w_{s_S} \quad \text{with} \quad j = 1, \ldots, m, \quad (3.4) \]

where \( s \) refers to the indices of non-zeros coefficients \( w_i \) in \( w \). When \( S > m \) and the knowledge of which \( S \) coefficients are non-zeros is given, even when \( m << n \), an exact solution of \( w \) can be obtained by solving the linear equations (as shown in Equation 3.4) using simple linear regression methods. However, in the real world, we usually merely know a signal is sparse, while its exact sparsity \( S \) and which its coefficients are non-zero are not known. Therefore, it is difficult to obtain an accurate reconstruction using simple linear regression methods. To overcome this problem, advanced sparse signal reconstruction methods have been proposed, the development of which has grown very fast from the conventional convex optimisation methods to the compressive sensing methods [Don06].

In addition to the sparse reconstruction methods, the reconstruction accuracy is also determined by the information provided by the measurements. The measurements which contain more useful and less redundant information can lead to a more accurate reconstruction. Therefore, with a given budget (i.e., the number of available measurements), the optimal set of measurements are the ones that providing the most useful information about the source signal. Sparse modelling methods select the optimal set of measurements based on the prior knowledge of the signal \( x \), which often refers to its sparsity. As the measurements are determined by the measurement matrix (as shown in Equation 3.1), selecting optimal measurements is actually a process of designing the measurement matrix \( \Phi \). Various measurement design methods have been proposed for sparse modelling to boost the reconstruction accuracy [Bar07, Can06, CWF09].

Predictive modelling, a key component in machine learning, is the use of data to create a predictive model that can predict the probability of an outcome. Many recent studies
in biology and neuroscience formulated their prediction problems using the linear sparse model (as shown in Equation 3.3). In this predictive modelling process, the design matrix is constructed with \( m \) observations of \( n \) features corresponding to a response vector \( y \), and signal \( w \) is the model parameter that is estimated during the predictive modelling process. The coefficients of model parameter \( w \) refer to the weights of features: a non-zero coefficient \( w_i \) indicates that its corresponding feature contributes to the prediction and therefore is selected as a predictor. In those studies, the number of samples is often limited by the experiment conditions (e.g. physical constraints and experimental budgets), while the number of features (e.g. genes in genomic data analysis and voxels of brain image in fMRI data analysis) are always very large. This large feature-to-sample (\( n/m \)) ratio makes the estimation of the model parameter \( w \) an underdetermined problem. However, as the number of predictors (e.g. significant genes and voxels) is often small compared to the number of features (e.g. measured genes and voxels of the whole brain image), a sparse model parameter \( w \) is desired. Consequently, estimation of this high-dimensional model parameter turns to be a sparse signal reconstruction problem, and therefore the predictive modelling process shares efficient algorithms (i.e. sparse signal reconstruction methods) and theoretical results with the signal reconstruction process in signal processing.

Moreover, the target of predictive modelling in biological and neuroscience studies is not only the predictive performance, but also the interpretability of the predictors. For instance, in genomic data analysis, the selected predictors (i.e. significant genes) should have biological meaning, with which the biological difference in different groups of subjects can be accurately interpreted; in fMRI data analysis, the selected predictors (i.e. voxels of brain images) should be able to interpret the exact neural activity responding to a specific brain function. As the design matrix in predictive modelling is constructed with features and their observations, it may have special properties such as features with high correlations. This means it is difficult for the design matrix to satisfy conditions required by the compressive sensing reconstruction methods to provide exact or approxi-
mate reconstructions. When high correlations existing among predictors, the compressive sensing reconstruction methods can only reconstruct a fraction of their corresponding weights $w_i$, so only a subset of correlated predictors are detected. This inaccurate selection has little effect on the predictive performance but can highly affect the interpretation accuracy. In addition, as the ground truth for the predictors is unknown, it is impossible to assess the selected predictors using the selection accuracy. Conventional methods assess the selected predictors based on their predictive performance. On one hand, it is difficult to differentiate if the selections covering all predictors or only a subset of the correlated ones. On the other hand, [MB10] has shown that when the number of samples is small, the selection based on the predictive performance assessment is very sensitive to the variations in samples which introduces large number of false positive selections. To improve the accuracy of predictor selection, some stable signal reconstruction methods such as ensemble based methods [Bac08, MB10, DGH+06] and group selection methods [ZH05, FHT10] have been proposed.

In the rest of this chapter, I overview the existing sparse signal reconstruction and measurement design methods for sparse modelling. I first introduce some popular compressive sensing reconstruction methods in Section 3.1. Then, in Section 3.2, the state-of-the-art stable reconstruction methods are discussed. Finally, relevant measurement design methods are explained in Section 3.3.

### 3.1 Compressive Sensing Reconstruction Methods

Nyquist/Shannon Sampling theory is one of the central tenets of signal processing: the measurement frequency must be more than twice the bandwidth of the source signal so as to provide an errorless reconstruction. In traditional signal processing techniques, the source signal is uniformly sampled at Nyquiste rate (i.e. a frequency twice as the bandwidth of the source signal) resulting in a set of measurements. The source signal is re-
constructed from the measurements using a conventional reconstruction algorithm such as an inverse Fourier transform. When using the conventional reconstruction methods, the resolution of the reconstructed signal is proportional to the number of measurements, reconstructing a high resolution signal requires a large number of measurements. On one hand, this has high demand on measurement equipment, which may be hard to satisfy in some applications. On the other hand, when large number of measurements can be acquired, they are compressed to be stored or transmitted when the resources of storage and transmission capacity are limited. These measure-then-compress frameworks suffer from inherent inefficiencies: they must start with massive measurements even though most of them will be discarded after compression. This results in a waste of computation and storage. In 2006, an alternative theory, called Compressive Sensing (CS), was posed by Donoho [Don06]. The theory of CS shows that when the signal to be reconstructed is sparse, it is possible to reconstruct the signal from a considerably incomplete set of measurements, i.e. with a number of measurements much less than required by the Nyquist-Shannon theorem.

The CS reconstruction methods reconstruct the sparse signal $w$ by calculating the linear sparse model (as shown in Equation 3.3). When measurements $y$ and the design matrix $\Theta$ are given, it can find an exact or approximate solution for $w$ if the design matrix $\Theta$ satisfies the Restricted Isometry Property (RIP) for $(2S, \sqrt{2} - 1)$. A matrix $\Theta$ which satisfies the RIP with parameters $(S, \epsilon)$ for $\epsilon \in (0, 1)$ and all $S$-sparse vector $u \in \mathbb{R}^n$, it should satisfy:

$$ (1 - \epsilon) ||u||_2 \leq ||\Theta_{\text{Supp}} u||_2 \leq (1 + \epsilon) ||u||_2, $$  

(3.5)

where $\text{Supp} \subset \{1, \ldots, n\}$ indicates the indices of supports in $u$, and $\Theta_{\text{Supp}}$ is a $m \times |\text{Supp}|$ submatrix obtained by extracting the columns of $\Theta$ corresponding to the indices of supports in $u$, where $|\text{Supp}|$ denotes cardinality of $\text{Supp}$. This property essentially requires that every set of columns with cardinality less than $S$ are approximately orthogonal. An
important result is that if all the columns of a design matrix $\Theta$ are approximately orthogonal, then an exact reconstruction can be obtained [Can06].

The RIP condition is equivalent to requiring that the design matrix $\Theta$ has all its eigenvalues in $[\sqrt{1-\epsilon}, \sqrt{1+\epsilon}]$ [BDDW08, DM09]. That is

$$\sqrt{1-\epsilon} \leq \lambda_{\min}(\Theta) \leq \lambda_{\max}(\Theta) \leq \sqrt{1+\epsilon},$$

(3.6)

where $\lambda_{\min}$ and $\lambda_{\max}$ refer to the minimum and maximum singular values of $\Theta$ respectively. If a design matrix does not satisfy Equation 3.6, it definitely does not satisfy the RIP condition. Because checking whether a design matrix satisfies the RIP condition has high computational complexity, Equation 3.6 is normally used instead to check the required condition of the design matrix.

### 3.1.1 $\ell_p$ Optimisation Algorithms

Define the $\ell_p$ norm of the vector $w$ as $||w||_p = (\sum_{i=1}^{n} |w_i|^p)^{\frac{1}{p}}$, where $|w_i|$ indicates the absolute value of $w_i$. A classical approach directly calculates $w$ in Equation 3.3 by finding a least squares solution (i.e. $\ell_2$ minimisation):

$$\arg \min_{w \in \mathbb{R}^n} ||w||_2 \quad s.t. \quad y = \Theta w.$$  

(3.7)

Even though a convenient closed-form solution $w = (\Theta^{\prime}\Theta)^{-1}\Theta^{\prime}y$ can be obtained from Equation 3.7, it almost always gives an unsatisfactory result which is a non-sparse solution. Figure 3.1 illustrates the approximation process to minimise different $\ell_p$ norms by using a signal $w \in \mathbb{R}^2$ as an example. The approximation process can be considered as the growth of an $\ell_p$ sphere until it intersects with a solution plane, which is the set of all $w$ vectors that satisfy Equation 3.3. In this example, the solution plane is one-dimensional and the size of the support of $w$ is one. As the $\ell_2$ sphere is spherical (shown in Figure
3.1(a)), it normally picks points \( \hat{w} \) far from the coordinate axis, which yields a non-sparse member of the solution plane.

To enforce a sparsity constraint on the solution, \( \ell_0 \) norm is used:

\[
\text{arg min}_{w \in \mathbb{R}^n} \|w\|_0 \quad \text{s.t.} \quad y = \Theta w.
\]  

(3.8)

As shown in Figure 3.1(b), the \( \ell_0 \) sphere is a cross along the coordinate axis, therefore the solution found by \( \ell_0 \) norm is definitely the point on coordinate axis so that a desired sparse solution is obtained [Can06]. Unfortunately, solving the \( \ell_0 \) minimisation problem is an Non-deterministic Polynomial-time hard (NP-hard); the solution cannot be found using a standard computer in polynomial time. This makes the optimisation process extremely computationally expensive, so \( \ell_1 \) minimisation is infeasible to use in practice.

To maintain the sparsity constraint but reduce the computational complexity, \( \ell_1 \) optimisation was commonly used in the development of CS reconstruction algorithms. \( \ell_1 \) optimisation, which is typically referred to as ”Basis Pursuit (BP)”, is expressed as:

\[
\text{arg min}_{w \in \mathbb{R}^n} \|w\|_1 \quad \text{s.t.} \quad y = \Theta w.
\]  

(3.9)

Figure 3.1: Best Approximation of A Point in \( \mathbb{R}^2 \) by A One-dimensional Support Using \( \ell_p \) norm, where \( p = 0, 1, 2 \). Reproduce from [DDEK11].
3.1. Compressive Sensing Reconstruction Methods

The $\ell_1$ also produces a sparse solution (as shown in Figure 3.1(c)) and has strong reconstruction guarantees. Moreover, compared to the $\ell_0$ norm, it is a much easier optimisation target which can be solved in polynomial time, making it more useful in practice.

The above $\ell_p$ approximation methods deal with noise-free measurements, while in real world applications noise normally existing in the measurements $y$. To formulate the noisy problem, Equation 3.3 is rewritten as:

$$y = \Theta w + \eta.$$ \hspace{1cm} (3.10)

where $\eta \in \mathbb{R}^m$ is a column vector, a coefficient $\eta_j$ in which refers to the level of noise existing in the corresponding measurement $y_j$. The noise is different in various applications, which is normally caused by the measurement conditions such as measurement devices and environmental factors. In many practical application, $\eta$ is defined as zero-mean Gaussian variables and therefore each of its coefficients is a sample from a Gaussian distribution with zero mean and a covariance predetermined by the measurement conditions.

![Figure 3.2: Best Approximation of A Point in $\mathbb{R}^2$ by A One-dimensional Support Using Lasso.](image)

To deal with the reconstruction problem (Equation 3.10) with noisy measurements, Absolute Shrinkage and Selection Operator (Lasso) was proposed. It has been a prevalent optimisation target in CS, and is a modified optimisation of BP, therefore is also known as Basis Pursuit De-Noising (BPDN). In addition to the $\ell_1$ norm, it adds a least-square
penalty to handle the measurement noise by controlling the reconstruction error:

$$\arg \min_{w \in \mathbb{R}^n} \frac{1}{2} ||\Theta w - y||_2^2 + \lambda_1 ||w||_1,$$  \hfill (3.11)

where $\lambda_1$ is a positive number that controls the trade-off between sparsity and reconstruction error. The Lasso sphere is same as $\ell_1$ (as shown in Figure 3.2), while the solution plane is determined by the objective function $||\Theta w - y||_2^2$ which yields an ellipse plane. It produces a relax sparsity constraint on signal $w$. The Lasso sphere does not intersect with the solution plane on its coordinate axis, but at a point nearby. The intersect point is determined by the choose of $\lambda_1$, a larger value of $\lambda_1$ makes the interaction points move towards the coordinate axis.

Interior point methods were first developed for solving BP and Lasso by convex optimisation. Some of these interior point methods such as primal-dual and primal log-barrier approach are provided by the $\ell_1$ magic toolbox [CR05], where the primal-dual method is mainly used for solving BP and the primal log-barrier method is used for Lasso. These algorithms offer significant performances on providing accurate reconstruction when the number of measurements is underdetermined and the design matrix satisfies the RIP condition. However, they are not practical in most applications, since they are expensive to compute for high dimensional signals.

Iterative thresholding algorithms [BD09, DDD04] are fast alternative approaches to solve Lasso (Equation 3.11). They start from an initial signal estimate $w_0$, and iterate a gradient descent step followed by a hard/soft thresholding step until a convergence criterion is met. At each iteration, the gradient descent step updates the estimate of $w$ with the signal residual calculated by $\Theta^{-1}(y - \Theta w_{old})$, where $w_{old}$ is the estimate from last iteration; the thresholding step remains the significant supports with a present hard/soft threshold. These iterative thresholding algorithms are feasible to be applied to reconstruct high dimensional signals because the convergence criterion is fast to met and small storage space
is required. However, they suffer from the inherent problem of Lasso, that the probability of successful reconstruction is controlled by the choice of constant parameter $\lambda_1$ in Equation (3.11). It is difficult to choose an optimal value of $\lambda_1$ because we lack knowledge of the signal sparsity in real applications.

### 3.1.2 Greedy Algorithms

The greedy algorithm is another type of approach to solve the sparse approximation problem and requires the design matrix to satisfy the RIP condition. Similar to the iterative thresholding algorithms, greedy algorithms find the sparse solution step by step in an iterative fashion. They estimate the signal coefficients by iteratively identifying supports of the signal until a convergence criterion is met.

Orthogonal matching pursuit (OMP) [TG07] is the simplest and oldest greedy algorithm. It begins by finding a column of $\Theta$ which is most correlated with the measurements, where the selected one indicates that its corresponding signal coefficient $w_i$ is the support of $w$. At each iteration, it finds a new support by selecting the column most correlated with the current signal residuals, which yields a locally optimal solution. To find globally optimal solutions, some improved version of OMP have been proposed, such as regularised OMP (ROMP) [NV09], Stagewise OMP (StOMP)[DTDS12], Compressive Sampling Matched Pursuit (CoSaMP) [NT09] and Subspace Pursuit (SP) [DM09].

Different from the OMP algorithm which selects one column at each iteration, ROMP, OMP, CoSaMP and SP can select several columns at a time. Obviously, they can provide good approximation with smaller number of iterations. Also, they improve the probability of selecting correct supports at each iteration. Both StOMP and ROMP solve the noise-free problem, while ROMP performs better. ROMP does not require a presence threshold to select the supports, where the value of the threshold has significant impact on the reconstruction result. Also, it is the first greedy algorithm whose performance guarantee
is as strong as the convex optimization algorithms. CoSaMP is an improved version of ROMP, it is robust even when the measurements are contaminated with noise. The SP algorithm arose at the same time as CoSaMP. When matrix $\Theta$ satisfies the RIP condition, it allows a $S$-sparse signal to be exactly reconstructed from noiseless measurements, or approximately reconstructed from noisy measurements. The main difference between SP and CoSaMP is that, at each iteration, the columns selected by SP are half of those selected by CoSaMP. This makes the SP algorithm much faster. In addition, ROMP, CoSaMP and SP have similar flavour to OMP. They all need to determine the sparsity parameter $S$ beforehand, and the value of which has direct effect on the reconstruction performance. However, as the sparsity $S$ of a real world signal is unknown in practice, the reconstruction performance is unpredictable.

### 3.1.3 Sparse Bayesian Learning

Sparse Bayesian Learning (SBL) [Tip01], also known as Bayesian Compressive Sensing (BCS) [JXC08], is a more advanced compressive sensing method. It is distinct from the $\ell_p$ optimisation and greedy algorithms as: 1) there is no parameter to be specified (e.g. $\lambda_1$ in Lasso and $S$ in CoSaMP and SP); 2) it can empirically provide a useful sparse solution even when the design matrix does not satisfy the RIP condition [YWD+12]; 3) previous reconstruction algorithms result in point estimate of the sparse signal $w$, whereas SBL not only improves the accuracy over the point estimate but also formulates a posterior density function (pdf) for $w$. With the posterior density function, “error bars” are provided. The error bars can indicate the measure of confidence of the reconstructed signal as well as guiding the optimal design of additional measurements so as to reduce the uncertainty in $y$, with which the reconstructed signal can be modified.

SBL can reconstruct sparse signal with noisy measurements (shown in Equation 3.10). Under a common assumption of zero-mean Gaussian noise, where $\eta \sim N(0|\sigma^2)$, a Gaus-
sian likelihood model is achieved:

\[ p(y|w, \sigma^2) = (2\pi\sigma^2)^{-m/2} \exp\left(-\frac{1}{2\sigma^2}||y - \Theta w||^2\right). \]  

(3.12)

Given measurement \( y \) and assume \( \Theta \) is known, the quantities to be estimated are the supports of \( w \) and the noise variance \( \sigma^2 \). Here, it seeds a full prior distribution for \( w \) as well as \( \sigma^2 \), and the sparsity of \( w \) is constrained by placing a sparse distribution (e.g. Laplace and Student’s t distributions) on it. It is straightforward to see that when given \( y \), and assuming the likelihood function in Equation 3.12, the solution in (3.11) corresponds to a Maximum A Posterior (MAP) estimate for \( w \). MAP, which is known as Type-I method, finds the optimal solutions by maximising the posterior distribution of \( p(w|y) \propto p(y|w)p(w) \) with respect to \( w \).

SBL has been used to realise signal reconstruction in signal processing and predictive modelling. It was proposed with two algorithms: Sparse Bayesian Regression (SBR) and Sparse Bayesian Classification (SBC); where SBR can be used in both signal processing and predictive modelling while SBC is normally used in the later one. The framework of SBR is essentially to SBC, except that the values of their target variable \( y \) is different, where \( y \) takes continuous values in SBR but binary values in SBC.

Sparse Bayesian Regression

For the regression problem as well as the reconstruction problem in signal processing, \( t = y(\Theta; w) \) is usually sampled from a model with additive noise, whose conditional distribution takes the form:

\[ p(t|\Theta, w, \sigma^2) = N(t|y(\Theta; w), \sigma^2). \]  

(3.13)
In this probability framework, the sparsity of \( w \) is formulated by placing a sparse prior distribution on it. Laplace density is a common choice for this prior. However, as the Laplace prior is not conjugate to the Gaussian likelihood, the Bayesian inference cannot be performed in a closed-form. Therefore, the optimisation process is intractable. For this reason, a hierarchical prior model is used instead. It has similar properties to a Laplace prior, but allows closed-form solutions to the Bayesian inference. The hierarchical prior model is constructed with a zero-mean Gaussian prior distribution \( p(w|\alpha) = \prod_{i=1}^{n} N(w_i|0, \alpha_i^{-1}) \) placed over \( w \) and a Gamma prior \( p(\alpha|a, b) = \Gamma(\alpha_i|a, b) \) which is introduced on each hyperparameter \( \alpha_i \). By marginalising over the hyperparameters, the hierarchical prior on \( w \) is written as:

\[
p(w|a, b) = \prod_{i=1}^{n} \int_{0}^{\infty} N(w_i|0, \alpha_i^{-1}) \Gamma(\alpha_i|a, b) d\alpha_i.
\]

(3.14)

Now, the overall prior on \( w \) can be evaluated analytically, and it corresponds to the Student’s t distribution. This distribution will be strongly peaked about zero (i.e. it is a sparse prior) when appropriate values of \( a \) and \( b \) are used. Normally, \( a \) and \( b \) are set to zero which makes the priors non-informative and therefore results in uniform scale priors on \( w \). Similarly, an inverse Gamma prior \( p(\sigma^2|c, d) = \Gamma^{-1}(\sigma^2|c, d) \) is placed on the noise variance \( \sigma^2 \), and \( c \) and \( d \) are set to zeros.

Given measurements \( y \), and assuming the hyperparameters are known, the posterior distribution over \( w \) can be expressed as a multivariate Gaussian distribution:

\[
p(w|t, \alpha, \sigma^2) = \frac{p(t|w, \sigma^2)p(w|\alpha)}{p(t|\alpha, \sigma^2)} = N(\mu, \Sigma);
\]

(3.15)

(3.16)

with mean and covariance:

\[
\mu = \frac{1}{\sigma^2} \Sigma \Theta^t t,
\]

(3.17)

\[
\Sigma = \left( \frac{1}{\sigma^2} \Theta^t (\Theta + A)^{-1} \right),
\]

(3.18)
3.1. Compressive Sensing Reconstruction Methods

where $A = \text{diag}(\alpha_1, \alpha_2, \ldots \alpha_n)$. When using a type-I method to implement this optimisation process, it seeks to maximise $p(w, \alpha, \sigma^2|t)p(w, \alpha, \sigma^2)$ with respect to the unobserved variables (i.e. $w$, $\alpha$ and $\sigma$). The type-I method requires the unobserved variables to be marginalised over, whereas this is not feasible in this case. Therefore, [Tip01] used a Type-II method instead. The Type-II method estimates the hyperparameters by maximising their marginal likelihood, which is obtained by marginalising over variable $w$. With the uniformed priors, this gives:

$$p(t|\alpha, \sigma^2) = \int p(t|w, \sigma^2)p(w|\alpha)dw$$

$$= (2\pi)^{-m/2} \text{det}(C)^{-1/2} \exp \left( -\frac{1}{2} t'C^{-1}t \right) $$

where $C = \sigma^2 I + \Theta A^{-1}\Theta'$ and $\text{det}(C)$ refers to the determinant of it. By calculating differentiation of 3.19, equating to zero and rearranging, the hyperparameters are then expressed as:

$$\alpha_i = \frac{\gamma_i}{\mu_i^2};$$

$$\sigma^2 = \frac{||w - \Theta \mu||^2_{m - \Sigma_i \gamma_i}}{m - \Sigma_i \gamma_i};$$

where $\gamma_i \equiv 1 - \alpha_i \Sigma_{ii}$, and $\Sigma_{ii}$ is the $i$-th diagonal element of the covariance in Equation 3.18. Clearly, the hyperparameters and the mean and covariance of $w$ are functions of each other. Therefore, the optimisation problem can be implemented via an iteration processing between Equations 3.17-3.18 and 3.21-3.22 until a convergence criterion is satisfied.

**Sparse Bayesian Classification**

For the classification problem, the target variable becomes binary, that is $t \in \{0, 1\}$. To account for the change in $t$, a Bernoulli likelihood and a sigmoidal link function are used.
Chapter 3. Sparse Modelling

The likelihood distribution of target variable is represented as:

\[ p(t|w) = \prod_{i=1}^{n} \sigma\{y(\theta_i; w)\}^{t_i}[1 - \sigma\{y(\theta_i; w)\}]^{1-t_i}, \tag{3.23} \]

where \( \sigma\{y(\theta_i; w)\} = \frac{1}{1+e^{-y(\theta_i; w)}} \) is the logistic sigmoid function. Let \( y_i = \sigma\{y(\theta_i; w)\} \), the logistic sigmoid function bounds the value of \( y_i \) to be \([0, 1]\), and the value of the corresponding target variable \( t_i \) is determined with a discrimination threshold, where \( t_i = 1 \) if \( y_i \geq \text{threshold} \) and \( t_i = 0 \) if \( y_i < \text{threshold} \). Unlike regression, there is no noise element and the marginal likelihood can no longer be obtained analytically by integrating \( w \) from Equation 3.23. There is closed-form expression for neither the marginal likelihood nor the posterior distribution \( p(w|t, \alpha) \). However, as the posterior distribution \( p(w|t, \alpha) \) is proportional to \( p(t|w)p(w|\alpha) \), the MAP process can be implemented by maximising \( p(t|w)p(w|\alpha) \) or equivalently its logarithm:

\[ \log\{p(t|w)p(w|\alpha)\} = \sum_{i=1}^{n} (t_i \log y_i + (1 - t_i) \log (1 - y_i)) - \frac{1}{2}w^T A w. \tag{3.24} \]

By calculating the first and second derivatives of Equation 3.24 (based on the Laplace’s method [Mac92]), the posterior probability of \( w \) centred at \( \mu \) is defined as:

\[ \mu = \Sigma \Theta^T B t, \tag{3.25} \]

\[ \Sigma = (\Theta^T B \Theta + A)^{-1}. \tag{3.26} \]

Where \( B = \text{diag}(\beta_1, \beta_2, \ldots, \beta_n) \) with \( \beta_i = \sigma\{y(\theta_i)\}[1 - \sigma\{y(\theta_i)\}] \). Compared with the mean and covariance in the regression problem (in Equation 3.17-3.18), we can see that the inverse noise variance for \( \eta \) is actually indicated by \( \beta_i \). As each \( w_i \) associated with an independent value of \( \beta_i \) rather than sharing a single noise variance \( \sigma^2 \), this results in a data-dependent noise. In the same manner as for the regression process, the hyperparameter \( \alpha \) is calculated via Equation 3.21, and the optimisation problem is implemented via
3.2 Stable Reconstruction Methods

In the predictive modelling problem, the sparse signal refers to the model parameter. A non-zero coefficient indicates its corresponding feature is selected as a predictor. Therefore, the sparsity of signal $w$ is determined by the sparsity of corresponding predictors. The modelling process, which is used to estimate the model parameter, is a sparse signal reconstruction process indeed and can be implemented via the CS reconstruction methods. In contrast to the signal reconstruction problem in signal processing, the estimated model parameter cannot be assessed via its reconstruction accuracy as the ground truth for both the model parameter and the predictors are unknown. As the primary goal of predictive modelling is to construct a model that could most accurately predict the target value for a new input, predictive performance is a prevalent assessment to assess the estimated sparse model parameter and the selected predictors are the ones with the most significant predictive power.
Given a new dataset \( \{ t_{new_i}, \theta_{new_i} \}_{i=1}^{m_{new}} \), the predictive performance of a regression model is calculated by Mean Square Error (MSE) or Root Mean Square Error (RSEM), where

\[
MSE = \frac{1}{m_{new}} \sum_{i=1}^{m_{new}} (t_{new_i} - \hat{t}_{new_i})^2, \quad RMSE = \sqrt{MSE}
\]

and \( \hat{t}_{new_i} \) refers to the predicted target value. On the other hand, the predictive performance of a classifier is often calculated by

\[
Accuracy = \frac{N_c}{m_{new}}
\]

where \( N_c \) indicates the number of correct classifications. **Accuracy** states the proportion of correct classifications of all given classes. It is a prevalent measure of predictive performance in multiclass problem, where the number of classes is greater than or equal to two. While, there are some specific assessment methods for the binary classification problem. For instance, when a person has a disease, one may care about how often the classification result will be positive; and if a person does not has a disease, one may care about how often the classification result will be negative. These two questions are measured by **Sensitivity** and **Specificity** respectively. Let \( TP, TN, FP, \) and \( FN \) indicates the number of true positives, true negatives, false positives, and false negatives respectively.

The **Sensitivity** and **Specificity** are calculated as:

\[
Sensitivity = \frac{TP}{TP + FN}, \quad (3.28)
\]

and

\[
Specificity = \frac{TN}{FP + TN}. \quad (3.29)
\]

When the classifier has high sensitivity, one can be nearly certain that a patient doesn’t have the disease if the classification result is negative; and when the classifier has high specificity, one can be nearly certain that a patient has the disease if the classification result is positive. In other words, a robust classifier should have high sensitivity and specificity simultaneously.

Most binary classifiers output a continuous \( y \) value in \([0, 1]\) which refers to the probability
or score that represents the degree of an instance $\theta$ belonging to either of two classes. A final decision $t$ is made with a cutoff threshold which normally takes 0.5; that is the instance belongs to class 1 (i.e. positive) if the classifier output is above the cutoff threshold, and belongs to class 0 (i.e. negative) if below. The Sensitivity and Specificity depend on how well the groups are separated by the constructed classifier as well as the choose of the cutoff threshold. Figure 3.3 illustrates the effect of threshold on both Sensitivity

![Figure 3.3: The Effect of Threshold on Sensitivity and Specificity.](image)

and Specificity with different classifiers. When a classifier can perfectly separate two groups (i.e. positive vs. negative) (as shown in Figure 3.3(a)), a threshold in a large range (around 0.5) can make both Sensitivity and Specificity equal to 1, so it has little effect on the assessments. However, when the two groups are overlap, the Sensitivity and Specificity are vary with the value of threshold. That is with the increase of threshold the Sensitivity increases but the Specificity decreases. For this reason, Sensitivity and Specificity are not the perfect measures for predictive performance. To remove the effect of threshold, Receiver Operating Characteristic (ROC) curve is an alternative and more accurate measure. It is a graphical plot that illustrates the relation between true positive rate (TPR) (i.e. sensitivity) and false positive rate (FPR) (i.e. 1-specificity) at various threshold settings.

From Figure 3.4, we can see that when a classifier provides distributions for two groups
Chapter 3. Sparse Modelling

(a) Distributions of groups provided by different predictive models

(b) ROC Curve

Figure 3.4: Examples of ROC Curves for Different Predictive Models. Reproduced from [Sr09].

that overlap almost totally, its ROC curve is a diagonal line. In other words, this classifier provides classifications that are nearly random guesses. However, when the groups are better separated, with smaller overlapping areas, the ROC curve moves towards the upper left corner. Moreover, any classifier with ROC curve under the diagonal line provides classifications worse than a random guess. In such situations, using the negation of the classifier will give a curve above the diagonal. In addition, the area under the curve (AUC) is used as a measure of the predictive performance. A good model (like model \(a\) in Figure 3.4) has AUC near 1, and a bad one which provides results similar to randomly guess (like model \(c\) in Figure 3.4) is near 0.5.

To accurately assess the performance of a predictive model in generalisation to new data, it is critical that the available dataset is separated to train and test the predictive model. If the model is built and assessed using the same dataset, the assessment of the model will be biased. To assess a model in an unbiased matter, cross-validation is necessary, which often refers to \(K\)-fold cross-validation. The \(K\)-fold cross-validation separates a dataset into \(K\) blocks, and a predictive model is trained on \(K-1\) out of all the blocks and tested on the rest. This works repeated on all blocks, and the predictive performance is averaged over all the repetitions. When \(K\) is equal to the number of samples, the process is called leave-out-one cross-validation. This process can provide more unbiased assessment than \(K\)-fold, but is relatively computationally expensive. 10-fold cross-validation appears to
be a good compromise in many applications [PMN11].

Conventional methods only focus on the predictive performance of the predictive model (i.e. linear combination of weighted predictor), while more recent studies start to take the accuracy of predictor selection into consideration so that an accurate and meaningful interpretation can be made from the selected predictors (as explained in the beginning of this chapter). When using the linear sparse model (Equation 3.10) to formulate a predictive problem, the model parameter $w$ is estimated so as to provide the best predictive performance. This is calculated on an available dataset via a cross-validation process, and the predictors that are selected with non-zero coefficients of the model parameter are considered to have the best predictive power. However, this cannot guarantee the selection is accurate and approaches to the truth. When using predictive performance as the assessment to select predictors, the use of different subsets of dataset or modelling method can result in the selection of different subsets of predictors which have the same or similar predictive performances. This unstable selection can reduce our confidence in the interpretation. For this reason, the stability in predictor selection has been analysed in many applications. When using sparse modelling methods to select sparse predictors from a high dimensional dataset, the instability is mainly caused by two sources: 1) the correlation among predictors. When there are many highly correlated predictors, the CS sparse signal methods which manage to find the sparseness solution often select one out of them. This has small impact on the predictive performance, but results in many false negative selections (i.e. Type I error); 2) small number of samples of the high dimensional data. When the number of samples is limited, the selection is very sensitive to variations in the training dataset, so the reconstruction methods are hard to provide an accurate estimation of the model parameter. This makes the selected predictors involve large number of false positive selections (i.e. large Type II error). [EDZD06] suggests that when select significant genes for predicting outcome in cancer, it requires at least thousands samples to achieve a stable selection which is much larger than the feature dimension.
To reduce the instability with respect to sampling variations, the most intuitive way is to increase the number of samples. However, this is impractical in many real applications. For this reason, an alternative method which increases the number of samples by creating artificial samples has been proposed for this purposed. It first calculates the distribution of available samples, and then increases the sample size by generating artificial samples according to the distribution. The generated artificial samples are mainly used in two ways. The first one is to mix them with the original samples, and then executes the training and testing process [KDB+02]. The second way is to use the generated samples only in the test process [HC07]. In addition, there are more advanced methods that have been introduced to improve the selection accuracy via reducing the instability. These methods can be separated into two categories: ensemble based methods, and group selection methods.

### 3.2.1 Ensemble Based Methods

Ensemble methods have been widely used in predictive modelling which were proposed under the assumption that "two (or more) heads are better than one". They aggregate predictions of multiple predictive models with the goal of improving predictive performance. Bagging, boosting and random subspace are three typical ensemble methods [SD02]. In bagging, the training set is randomly sampled many times with replacement to construct sub datasets, and a predictive model is then constructed on each of the datasets. In boosting, predictive models are constructed in an adaptive way: the training dataset used for constructing a new predictive model is the weighted version of the original training dataset, where the weights of the samples are determined by the previous prediction result. Initially, equal weights are set on all the training samples, and the first predictive model is constructed with this dataset. Then, the predictive performance of the model changes the weights, where the incorrectly predicted samples get larger weights, and the next predictive model is boosted on the reweighted training datasets. This process sequentially gen-
erates a set of predictive models. In random subspace, predictive models are constructed on the datasets that are generated by randomly sampling the training dataset in feature space. All these three methods need an aggregation process to combine the predictions obtained from multiple predictive models, where a classification process often uses the simple majority voting method [SD02] that counts the largest number of predicted results that agree with each other, and a regression process often uses a linear combination of the results [MMSJS12]. [KK12] noticed that when working with noise-free data, boosting algorithm is considered stronger than the other two methods; while when deal with a noisy data, bagging and random subspace perform much better. To produce more robust predictive models, some researchers [TTLW06, KK12] combine all the three methods.

Recently, some ensemble predictor selection methods have been proposed to reduce the instability of predictor selection [Bac08, MB10, DGH’06]. As in the ensemble predictive modelling, these methods consist of two steps. The first step is to select sets of predictors with multiple sparse models that are constructed with subsets of the original dataset, and the second step is to aggregate the selected predictors to provide a more stable and accurate selection. [Bac08] proposed a stable selection method by aggregating boosting and lasso, and [MB10] introduced a randomised lasso method which was implemented by using both random subspace and bagging methods. These methods have demonstrated their desirable abilities of handling the instability selection caused by sampling variations, while the second method also reduced the instability with respect to the high predictor correlation. Because the random subspace method generate sub datasets by randomly sampling from features space, the feature correlations in each sub dataset are highly reduced. This means more correlated predictors are found with the multiple sparse models.

In addition, to reduce the instability, because of the use of different modelling algorithms, various ensemble methods have been proposed [TFRG09, YXS05, DG07]. The difference among these methods are mainly in the aggregation procedure, where [TFRG09] used the linear combination method, [YXS05] introduced a distance synthesis method and [DG07]
employed a Markov chain based rank aggregation method.

### 3.2.2 Group Selection Methods

Group selection methods aim to detect groups of correlated predictors by improving the estimate accuracy of model parameters, so they can reduce the selection instability cause by the first source (i.e. high correlations among features).

The elastic net proposed by Zou and Hastie [ZH05] is the first sparse modelling method that works for groups of predictors. It adds an $\ell_2$ norm penalty in addition to Lasso. The $\ell_2$ term encourages the selection of correlated predictors and removes the limitation on the number of selected predictors, optimising

$$
\arg \min_{w \in \mathbb{R}^n} \frac{1}{2} ||\Theta w - y||_2^2 + \lambda_2 ||w||_1 + \lambda_3 ||w||_2.
$$

Figure 3.5: Illustration of Elastic Net: (a) Spheres of Lasso, Ridge and Elastic Net, (b) Best Approximation of a Point in $\mathbb{R}^2$ by A One-dimensional Support Using Elastic Net. Reproduced from [ZF14].

It is easy to see that when $\lambda_2 > 0$ and $\lambda_3 = 0$, it is equivalent to Lasso; when $\lambda_3 > 0$ and $\lambda_2 = 0$, it becomes to ridge regression which is a modified optimisation of $\ell_2$ by adding an objective function $||\Theta w - y||_2^2$. Both Lasso and ridge are convex. Therefore, the elastic net which is the sum of them is also convex and its sphere is intermediate between the
Lasso and ridge (as shown in Figure 3.5(a)). Compared to Lasso, the elastic net introduces a relax sparsity constraint on signal \( w \), and its sphere moves towards the Lasso sphere only if the ratio between \( \lambda_2 \) and \( \lambda_3 \) becomes larger, where a larger \( \lambda_2 \) enforces a more sparse solution and the group parameter \( \lambda_3 \) controls the group effect by assigning similar weights (i.e. coefficients of model parameter) to correlated predictors. The optimal values of these two parameters are normally obtained via a cross-validation process. Figure 3.5(b) illustrates why the elastic net promotes grouping selection: the high correlation solution plane is more likely to hit the less sparse point, whereas the low correlation solution plane prefers the more sparse point.

The elastic net is a data driven method that the groups of predictors which have similar weights are learnt from a given dataset. However, when the prior knowledge of the predictor groups is given, group Lasso [YL06] and sparse group Lasso [FHT10] are preferred. They are expressed as:

\[
\text{arg min}_{w \in \mathbb{R}^L} \frac{1}{2} \sum_{l=1}^{G} \Theta_l w_l - y ||^2 + \lambda_4 \sum_{l=1}^{G} \Theta_l || w_l ||_2^2 + \lambda_5 \sum_{l=1}^{G} || w_l ||_2 + \lambda_6 || w ||_1
\]

respectively; where \( L \) indicates the number of groups and \( \Theta_l \) refers to the size of the corresponding group. Matrix \( \Theta_l \) is constructed by the features of group \( l \). The weight of a group feature is represented as \( w_l \in \mathbb{R}^{\Theta_l} \), and the weight of an individual feature within the group \( \Theta_l \) is indicated by \( w_{lm} \) with \( m = 1, \ldots, \Theta_l \).

The procedure of group Lasso acts like Lasso but at a group level, and when \( \Theta_l = 1 \) for \( l = 1, \ldots, \mathcal{L} \) it reduces to Lasso. Figure 3.6(a) shows how group Lasso promotes a sparse solution at the group level while non-sparse solution within groups. The features within a group are constrained by the \( \ell_2 \) norm which treats all features equally and does not encourage sparsity, while the sparsity is encouraged at the group level which is constraint
Figure 3.6: Illustrations of Group Lasso and Sparse Group Lasso for Two Groups with $w_1 \in \mathbb{R}^1$ and $w_2 \in \mathbb{R}^2$. Reproduced from [YL06, FHT10].

by the second term in Equation 3.31 which is equivalently to the $\ell_1$ norm of the group coefficients. The group Lasso only yields sparsity at the group level. When sparsity at both the group and within group levels are desired, sparse group Lasso is used. By introducing an additional $\ell_1$ penalty to group Lasso, it encourages a relax sparse solution which is similar to the elastic net but for the features within each separated group (as shown in Figure 3.6(b)). Consequently, it does not only allow the selection of group predictors but also sets various weights to individual features within each group.

In addition to the above convex minimisation methods, Bayesian approaches have been developed for group sparse modelling [BND12] which can yield ‘error bars’ (similar to SBL) for measuring the confidence of estimated model parameter. These methods are similar to group Lasso which consider the sparsity at group level. They constrain the sparsity of group features by placing a specific prior distribution on each group weight.
3.3 Measurement Design

[RFW+09, MSM09] proposed to use multivariate Laplace priors on separate groups, while [GO10, LCDH10] used Laplacian scale mixtures to construct the group sparse prior and the optimisation process was implemented using expectation-maximisation (EM). In addition, [BND12] proposed a general multivariate signal prior construction method suitable for group-sparse modelling, and allows the parameters of the model to be estimated via variational Bayesian (VB) method which is an optimisation method that can provide approximate estimates even when the posterior distribution of the model parameter is intractable, which cannot be handled by EM.

3.3 Measurement Design

In many real world applications, it would be great if one could design an efficient measurement protocol so that an exact or approximate reconstruction of an unknown signal could be obtained with measurements as few as possible. If this was the case, we could reduce the measurement duration as well as the measurement cost which are very important issues in many applications (e.g. neuroimaging, radar). When formulating the reconstruction problem using the linear sparse models in Equation 3.10, the design problem is to design an efficient measurement matrix Φ. The existing methods for sparse modelling mainly fall into two categories: random methods and informative methods. Both of them aim to design a measurement matrix so that the corresponding measurements can provide useful information for an accurate reconstruction.

3.3.1 Random Methods

Most existing CS reconstruction algorithms guarantee an exact or approximate reconstruction with a given limited number of measurements, only if the design matrix which is constructed with its corresponding measurement matrix and a sparse basis matrix sat-
Chapter 3. Sparse Modelling

satisfies the RIP condition (explained in Section 3.1). Therefore, the most intuitive way to implement the measurement design is to find the shortest measurement matrix (i.e. with the smallest number of rows) which makes the design matrix satisfy the RIP condition. However, it is difficult to check if the resulted design matrix satisfies RIP, especially for very large values of $S$ because of the high computational complexity. Hence, an alternative condition, called incoherence, is normally used. It has been demonstrated that a set of incoherent measurements can easily satisfy the RIP condition [Can06].

$$
\zeta(\Phi, \Psi) = \sqrt{n} \max_{i,j} |\phi_i \psi_j^\prime|, \text{ with } 1 \leq i \leq m \text{ and } 1 \leq j \leq n, \tag{3.33}
$$

where $\zeta(\Phi, \Psi)$ refers to the coherence between the measurement matrix $\Phi$ and a given sparse matrix $\Psi$, and it efficiently measures the largest correlation between any rows of both matrices $\Phi$ and $\Psi^\prime$. The value of the coherence has a range of $[1, \sqrt{n}]$. It is large if the matrices contain large number of correlated rows and otherwise is small. The role of the coherence is very simple: the smaller the coherence, the fewer samples are needed. This is because the measurements contains less redundant information if their corresponding measurement matrix is less coherent with the sparse matrix.

Random matrices are largely incoherent with most known sparse basis (e.g. wavelet projection and Fourier projection), and they are the prevalent measurement matrices used in the CS techniques. Gaussian matrix and Bernoulli matrix [Bar07] are the two most relevant examples of them. The entries of both measurement matrices are determined by sampling from a specified distribution, where the entries of a Gaussian matrix are chosen as i.i.d Gaussian random variables with expectation 0 and variance $1/m$, and the entries of a Bernoulli matrix are independent realisation of $+1/\sqrt{m}$ or $-1/\sqrt{m}$ with equal probability. The Gaussian and Bernoulli matrices provide the optimal conditions for sparse signal reconstruction with the smallest number of measurements. However, they are of limited use in practical applications because the form of the measurement matrix is often
3.3. Measurement Design

constrained by physical or other conditions of measurement, and therefore the assumed distributions (i.e. Gaussian and Bernoulli) may not be justifiable in practice. Random Fourier matrix [Can06] is another popular random measurement matrix in CS. It is constructed by uniformly and randomly selecting \( m \) rows from an the discrete Fourier matrix \( F \in \mathbb{C}^{n \times n} \):

\[
F_{i,j} = \frac{1}{\sqrt{n}} \exp \frac{2\pi i j}{n}.
\]  

(3.34)

Compared to the former random matrices, it is more applicable to real applications such as applications in signal processing which measure the frequency signal (Fourier transform coefficients) of the source signal \( x \).

3.3.2 Informative Methods

The random methods are easily implemented as the construction of the random matrices is simple and nonadaptive. However, because of the inherent property of random sampling that the measurements are selected by chance, obtaining the optimal measurement matrix is not always guaranteed. In addition, they only take the sparsity of the source signal into consideration, while the universality of random projections does not mean that they are universally optimal for every class of sparse signals. If a more precise prior knowledge of the source signal than only a simple sparsity assumption is available, it should be possible to design a more optimal measurement matrix for a specific class of signal. For instance, [SN08] noticed that a standard low-pass filtering (e.g. measurement based on PCA) often leads to a more accurate reconstruction of natural images than random matrix. Also, according to the properties of Fourier transform of real images that the low frequency signals contains more general information about images than the high frequency signals, [LDP07] mentioned that the measurement matrix that was designed to measure more low frequency signals than high frequency signal could result in a better reconstruction than random Fourier matrix.
Informative methods have long been used to design measurement protocol for linear systems [Lin89]. These methods are based on the Maximum Information (InfoMax) principle, by following which the measurements are designed to maximise mutual information between the measurement and the source signal. This method was originally introduced to solve the overdetermined problem, where \( m \geq n \). Later, [CWF09] adapted it to resolve the underdetermined issue (i.e. \( m << n \)). They showed that this design method was more like to produce optimal measurements than the random methods. The mutual information between a source signal \( x \) and its corresponding measurements \( y \) measures the information they share and it is formulated as:

\[
I(x; y) = \int_{x \in x} \int_{y \in y} p(x, \eta) \log \left( \frac{p(x, \eta)}{p(x)p(\eta)} \right),
\]

(3.35)

where \( p(x, \eta) \) is a joint probability of \( x \) and \( y \), when \( x = x \) and \( y = \eta \). An equivalent expression of the mutual information is:

\[
I(x; y) = h(y) - h(y|x),
\]

(3.36)

with

\[
h(y) = -\int_{\eta \in y} p(\eta) \log p(\eta)
\]

(3.37)

and

\[
h(x|y) = \int_{x, \eta \in y} p(x, \eta) \log \frac{p(\eta)}{p(x, \eta)}
\]

(3.38)

where \( h(y) \) is the differential entropy that measures the uncertainty in the measurements, and \( h(x|y) \) is the conditional entropy that quantifies the uncertainty in the realisation of \( x \) when the measurements are given. From Equation 3.10, we can see that when \( \Phi \) and \( x \) are given, the only factor that influences the value of \( y \) is noise \( \eta \). That is the conditional entropy is merely the entropy of the noise. Therefore, maximising the mutual information \( I(x; y) \) is equivalent to maximise the entropy \( h(y) \). Different from the RIP condition,
which provides a sufficient error bound for the reconstruction of the BP method, the InfoMax principle provides an optimisation criterion without assuming any reconstruction scheme. While, it strongly relies on the prior knowledge of the source signal, which is often encoded as probability density.

As discussed above, in many real applications, the possible measurements are often constrained by physical properties of the measurement process as well as practical implementations. In this situation, the measurement design process turns to construct the measurement matrix $\Phi$ by selecting sub rows from a candidate matrix $\Phi_{all}$ which contains all possible measurement locations. Since the probability distribution of a measurement is resulted from the multiplication of the measurement matrix $\Phi$ and the prior distribution of the source signal $x$, the optimal measurement matrix that maximises entropy $h(y)$ varies according to the prior distribution of the source signal. While, for many types of prior distribution, the optimisation process is often analytically intractable, it cannot be performed in a closed-form because of the complicated nature of the differential entropy.

The multivariate Gaussian distribution is a commonly used as a prior distribution, because its optimisation process can be found analytically. Assume a multivariate Gaussian distribution on the source signal $p(x) = N(\mu_x, \Sigma_x)$, then the distribution of the measurements is also a multivariate Gaussian distribution with mean $\mu_y$ and covariance $\Sigma_y$:

$$
p(y) = N(\mu_y, \Sigma_y) = N(\Phi \mu_x, \Phi \Sigma_x \Phi'),
$$

(3.39)

Thus, the entropy of $y$ is expressed as:

$$
h(y) = \frac{m}{2} (1 + \log(2\pi)) + \frac{1}{2} \log \det(\Phi \Sigma_x \Phi'),
$$

(3.40)

where $\det(\Phi \Sigma_x \Phi')$ is determinant of $\Phi \Sigma_x \Phi'$. As $\frac{m}{2} (1 + \log(2\pi))$ is constant, maximising
\( h(y) \) process is reduced to be:

\[
\arg \max_{\Phi \in \Phi_{all}} \log \det(\Phi \Sigma_x \Phi').
\]  \hspace{1cm} (3.41)

[JB09] indicates this optimisation process is equivalent to the maximum a posteriori probability (MAP) estimate of \( x \).

The design method based on this InfoMax principle can either be nonadaptive [Wei07] or adaptive [SNPS10, JXC08]. Given the prior distribution of the source signal, the non-adaptive method design the measurement matrix in an open-loop fashion (as shown in Figure 3.7), where the measurement design and the signal reconstruction processes are implemented independently. The measurement process is completed before applying any reconstruction method. On the contrary, the adaptive method designs the measurement matrix in a closed-loop fashion (as shown in Figure 3.8), where the measurements and signal reconstruction are processes in an iterative fashion. At each iteration, it reconstructs the signal with the current measurements and updates the prior distribution of the source signal; with this new prior distribution, the new measurement vector \( \phi_{new} \) is...
3.3. Measurement Design

selected to constitute the measurement $y_{new}$ for which the data is most uncertain. The adaptive process is stopped when either there is not any significant change in the uncertainty or the given budget $m$ is reached. In addition, as the adaptive method requires the distribution of the source signal to be updated during the reconstruction process, Bayesian reconstruction methods (e.g. Sparse Bayesian Learning [Tip01]) that can provide a pdf estimate of the reconstructed signal are required.
Chapter 4

Linear Dynamic Sparse Modelling for Functional MR Imaging

4.1 Introduction

The functional MR imaging technique images the brain according to changes in the blood oxygen level which are proxies for degrees of neural activity. An functional MR image is built up in units called voxels. Each voxel typically represents the average value of a million or more neurons, where the number of neurons represented is proportional to the size of the voxel ranging from 1mm to 5mm in the current fMRI technique. A higher spatial resolution image contains more voxels with smaller size. A smaller voxel which contains fewer neurons can make neural activity be recognised in more detail. As neuroscientists want to observe the brain deeply and clearly, high resolution functional MR images are often desired. However, since the scanning time directly rises with the number of voxels to be acquired, a higher resolution fMRI image takes longer to scan. This long scanning time leads to discomfort for the subjects in the scanner as well as reduces the temporal resolution of the brain image which causes significant information lost. For this reason,
an urgent problem for the fMRI technique is to optimise the image quality using a limited number of measurements; two fundamental problems need to be addressed: how to boost the reconstruction by improving the reconstruction algorithm, and how to gather more information via a well-designed measurement strategy.

**Reconstruction Algorithms**  With a limited number of measurements, the image quality of MRI has been greatly improved using an emerging technique known as Compressive Sensing (CS), which is a popular sparse signal reconstruction technique for sparse modelling (as discussed in Section 3.1 in Chapter 3). CS can reconstruct a signal accurately using underdetermined measurements as long as the signal can be sparsely represented in a specific domain [Don06]. The CS technique can be successfully applied to solve the reconstruction problem in the MRI applications due to two properties [LDSP08]: 1) medical images are often well modelled as being sparse in an appropriate transform domain (e.g. wavelet transform), and 2) MRI scanners acquire encoded measurements (e.g. in spatial-frequency encoding) instead of direct voxel measurements. The existing methods which utilise the CS techniques to solve the reconstruction problem in MRI mainly fall into three categories. Among them, the most direct method implements the imaging process by applying CS to each MR image separately [LDP07], and therefore the MR images are reconstructed independently. This method is simple but the quality of its reconstructed images is relevant low. Another method treats the entire sequence of MR images as a single spatio-temporal signal and performs CS to reconstruct it [GBK08, WLD+06]. This method can provide higher quality images, but real-time reconstruction is impossible. The most recent and advanced method [LV09, LLAV11] uses dynamic signal tracking techniques. This method greatly improves the reconstruction quality by utilising the correlations of sparse patterns between two time-adjacent MR images. Furthermore, it allows for real-time reconstruction.
Measurement Strategies  In addition to the reconstruction algorithm, the reconstruction quality is also determined by the measurement strategy. If a limited number of measurements carries more useful information about the signal, a higher quality image should be reconstructed. The most common measurement design scheme for the CS MR imaging technique is variable density random undersampling [LDP07] which chooses measurements according to a prior distribution. The prior distribution is calculated using distinct characteristics of signals in high and low frequency domains. Further, historical MR images were used as prior information to design measurement trajectories for new experiments [LLLZ12, RB11]. These design methods all depended on the prior knowledge of the measurements but worked in a nonadaptive way: once the prior knowledge was given, measurement trajectories were designed independently from the reconstruction process. In contrast to the above methods, Seeger et al. [SNPS10] proposed an adaptive measurement design method which was a Bayesian method that selected measurements iteratively. In each iteration step, the posterior distribution of a MR image was updated using previous measurements, which was implemented via a Bayesian based reconstruction method. The new measurement was selected to minimise the uncertainty of the posterior distribution.

Most of the above methods are investigated for improving the MR image quality. However, further improvement can be made in functional MRI, as it is a specialised application of MRI that has special properties, e.g. correlation exists between two time-adjacent functional MR images. In this chapter, I focus on the functional MR imaging study by considering the correlations of images in an fMRI sequence. My work extracts correlations from a key assumption that variations of functional MR images are sparse over time in the wavelet domain which is normally the case in fMRI studies. Based on the key assumption, I first introduce the concept of linear dynamic sparse model; it models an fMRI sequence as a linear dynamical system with an identity transition matrix, and the image variations which are presented by the system noise are assumed to be sparse. Then, a novel linear dynamic sparse modelling method is proposed to solve the fMRI sequence
reconstruction problem. This method consists of two processes: image reconstruction and measurement design. The corresponding algorithms for implementing both processes are investigated and proposed.

The image reconstruction process is implemented with a dynamic sparse signal tracking algorithm, called Hierarchical Bayesian Kalman Filters (HB-Kalman) [FKDY13]. The HB-Kalman algorithm employs the state-of-the-art compressive sensing reconstruction method, Sparse Bayesian Learning (SBL) [Tip01] (detailed in Section 3.1.3 in Chapter 3), to estimate the sparse variations between two adjacent images; while the classic Kalman filter update step is processed for image reconstruction. This algorithm can not only improve the point estimate of functional MR images, but also formulate a posterior density function (pdf) for each voxel. The “error bars” which is produced by the pdfs can indicate the measure of confidence of the reconstruction. My proposed measurement design algorithm, called Informative Measurement Design (IMD), designs the efficient measurement strategies by making use of the posterior distribution of the reconstructed previous adjacent image. It is the first measurement design method in functional MR imaging that utilises the correlations of fMRI images in a sequence. It calculates the prior distribution of the present image using both the posterior distribution of the previous adjacent image and the prior distribution of the image variations. After obtaining the prior distribution of an unknown image, the measurement design problem is to select $m$ feasible measurements, where $m$ is a given budget, and the measurements are determined to maximise the mutual information [CWF09] between the unknown image and measurements. As this problem is intractable, a novel approximation method is employed to solve it. Compared with the previous functional MR imaging methods, my approach makes better use of the prior information of the functional MR images so that the qualities of reconstructed images can be highly increased.

The remaining chapter is organised as follows. In Section 4.2, I first formulate the fMRI sequence reconstruction problem using a linear dynamic sparse model. I then illustrate
my linear dynamic sparse modelling method, and explain the algorithms of both reconstruction and measurement design in Section 4.3. Next, the experiment results of applying my method to a synthetic fMRI sequence are detailed in Section 4.4. Finally, Section 4.5 presents conclusions of my work.

4.2 Problem Formulation

4.2.1 Sparsity of Variations

MR images are usually piecewise smooth because of the structures of human organs and thus they are approximately sparse in the wavelet transform domain [LDP07, LV09, LLAV11] which has been widely used to compress natural images. Wavelet transform are based on small waves, called wavelets, the choice of which depends on the properties of a signal we wish to analyse. It decomposes a signal with sets of scaled and translated waves, and its discrete version (i.e. discrete wavelet transform (DWT)) is normally used in practice where the wavelets are discretely sampled. The scale factors reflect the frequency information of the signal: small scale factors measure the details and noise (i.e. high frequency components) of the signal; in contrary, large scale factors characterise the coarse structure (i.e. low frequency components) of the signal.

![Discrete Wavelet Transform (DWT) of a One Dimensional Signal.](image)

Figure 4.1: Discrete Wavelet Transform (DWT) of a One Dimensional Signal.

Figure 4.1 illustrates the DWT of a one dimensional signal. The DWT decomposes a
one dimensional signal \( S \) into coefficients \( A \) and \( D \) by using a low-pass(LP) and a high-pass(HP) filter respectively, where \( A \) refer to the coefficients of the low frequency components while \( D \) refer to the coefficients of the high frequency components. As an image

![DWT image](image1)

Figure 4.2: One-level Daubechies-4 DWT for an Functional MR Image.

is a two dimensional signal, it requires a 2D DWT which is implemented by applying the one dimensional DWT to the rows of the image and then applying the one dimensional DWT to the columns of the already horizontal transformed image. Figure 4.2(a) shows an example of one-level Daubechies-4 discrete wavelet transform (DWT) of an functional MR image. Daubechies-\( v \) is a common wavelet used in signal processing and MR imaging technique [LDP07, LLAV11]. It is easy to implement fast wavelet transform so that it is practical in real applications. A Daubechies wavelet can be decomposed into two functions: wavelet function and scaling function which represent the high pass and the low pass filter respectively. The smoothness of both functions are controlled by the parameter \( v \) which is called vanishing moment: a larger value of \( v \) results in more smooth wavelet and scaling functions. In this work, I empirically choose the Daubechies-4 wavelet which is normally used in the MR imaging technique [LDP07, LLAV11], and its functions are shown in Figure 4.2(b). The 2D DWT decomposes the original image into four sub images which characterise different components of the image. The sub image in the upper-
left contains the low frequency components that produces the final approximation image. It represents approximated version of the original at half the resolution. On the other hand, the details of the original image in vertical, horizontal and diagonal dimensions are represented by the sub images in upper-right, bottom-left and bottom-right respectively. Furthermore, the low frequency components can be further decomposed and therefore a two-level DWT is produced (as shown in Figure 4.3).

![Figure 4.3: Two-level Daubechies-4 DWT for an Functional MR Image.](image)

Rather than assuming the functional MR images are sparsely represented in the wavelet domain, in my work I assume that variations of functional MR images are sparse over time in the wavelet domain. I demonstrated this for a fMRI sequence [LLAV11] in Figure 4.4. The sparsity level is determined by $|n_d/n_t|$, where $n_t$ refers to the number of two-level Daubechies-4 2D DWT coefficients which support 96% energy of the functional MR image at time $t$, and $n_d = |n_t/n_{t-1}|$ refers to the number of DWT coefficients changes with respect to the previous frame. In most cases, the number of variations is less than 10% of the signal size, while in the worst case it is less than 13%
4.2. Problem Formulation

Figure 4.4: Example of Sparse Variations. The variations are calculated with 96% energy supports of DWT coefficients of images

4.2.2 Linear Dynamic Sparse Model

Linear dynamic model [KMK12] is a state-space model that describes the probabilistic dependence of a latent state and its corresponding observed measurements. It is characterised by a pair of equations: system equation and measurement equation (as shown in Figure 4.5).

![Linear Dynamic Model Diagram](image)

Figure 4.5: Linear Dynamic Model.

The system equation defines transition rules from the previous latent state to the current one, while the measurement equation maps the latent state space to the measurement space. My introduced linear dynamic sparse model for functional MR imaging is a special case of the linear dynamic model, where the system equation is modified to meet the sparsity constraint. Both the equations in my model are detailed below.
System Equation

Based on the assumption that variations of functional MR images are sparse, we can model an fMRI sequence as a linear equation with an identity transition matrix:

\[ w_t = w_{t-1} + \kappa_t, \quad (4.1) \]

where random variable \( w_t \in \mathbb{R}^n \) denotes the DWT coefficients of a functional MR image at time \( t \). For simplicity, I call \( w_t \) image in the rest of this chapter. Random variable \( \kappa_t \in \mathbb{R}^n \) denotes its sparse variations with respect to the previous image \( w_{t-1} \). To meet the sparsity constraint, a hierarchical sparseness prior is placed on \( \kappa_t \). Each element \( \kappa_{ti} \) of the variation \( \kappa_t \) is randomly sampled from a zero-mean Gaussian distribution \( N(\kappa_{ti}|0, \alpha_i^{-1}) \), the variance \( \alpha_i \) of which is randomly sampled from a Gamma \( \Gamma(\alpha_i|a, b) \). That is,

\[ p(\kappa_t|a, b) = \prod_{i=1}^{n} \int_0^{\infty} N(\kappa_{ti}|0, \alpha_i^{-1})\Gamma(\alpha_i|a, b)d\alpha_i. \quad (4.2) \]

After marginalising the hyperparameter \( \alpha_i \), the prior of \( \kappa_t \) corresponds to a product of independent Student’s t distribution. Tipping et al. [Tip01] demonstrates a strong sparse property of this hierarchical distribution.

Measurement Equation

The fMRI technique measures a subset of discrete Fourier Transform (DFT) coefficients of MR images. At each time \( t \), the measurement process can be modelled as:

\[ y_t = \Theta_t w_t + \eta_t, \quad (4.3) \]

where random variable \( y_t \in \mathbb{R}^m \), here called measurements, is a subset of DFT coefficients determined by the design matrix \( \Theta_t \in \mathbb{R}^{m \times n} \). Random variable \( \eta_t \in \mathbb{R}^m \) is
measurement noise. The design matrix $\Theta_t$ is formed by a subset of $m$ vectors selected from a given projection matrix $\Theta \in \mathbb{R}^{n \times n}$, which, in my case, is constructed by the DFT matrix with the inverse DWT matrix (detailed in Section 4.4.1). The budget $m$ is a given positive integer. It determines the number of frequencies to be measured.

### 4.3 Linear Dynamic Sparse Modelling

My proposed linear dynamic sparse modelling method aims to design the measurement strategy as well as reconstruct the image sequence by satisfying the linear dynamic sparse model. Figure 4.6 illustrates the framework of my method.

![Figure 4.6: Framework of fMRI Sequence Reconstruction.](image)

For each time instance, the measurement design method is first performed to select a subset of $m$ vectors from the projection matrix $\Theta$ by using the posterior distribution of the previous adjacent image $w_{t-1}$, where the selected vectors are used to form the design matrix $\Theta_t$. When the measurement $y_t$ is obtained by following the determined strategy, the
posterior distribution of the present image \( w_t \) can be calculated using the reconstruction algorithm, and it can be used in the next measurement design process. The framework is processed iteratively until the whole fMRI sequence is reconstructed.

### 4.3.1 Hierarchical Bayesian Kalman Filter

Kalman filter [WB95] is the most common technique for tracking a dynamic signal following the linear dynamic model. The conventional Kalman filter is based on the Gaussian assumption that both the measurement noise \( \eta_t \) and the system noise \( \kappa_t \) are chosen as zero-mean Gaussian variables with known covariance, i.e. \( \eta_t \sim \mathcal{N}(0, \sigma^2 I) \) and \( \kappa_t \sim \mathcal{N}(0, Q_t) \), where \( I \) is an identity matrix. Based on the Gaussian assumption, one has

\[
\begin{align*}
    p(w_t|w_{t-1}) &= \mathcal{N}(\mu_{t-1}, \Sigma_{t-1} + Q_t) \\
p(y_t|w_t) &= \mathcal{N}(\Theta_t \mu_t, \sigma^2 I)
\end{align*}
\]

where \( \mu_{t-1} \) and \( \Sigma_{t-1} \) refer to the mean and covariance of signal \( w \) at time \( t - 1 \) respectively. To track the dynamic signal \( w \), the Kalman filter continuously alternates between two steps: prediction and update. The prediction step predicts the prior state of the signal \( w_t \) by calculating the parameters of \( p(w_t|y_{t-1}) \) while the update step evaluates posterior state of the signal \( p(w_t|y_t) \) after observing the current measurement \( y_t \).

- **Prediction step:**

\[
p(w_t|y_{t-1}) = \int p(w_t|w_{t-1})p(w_{t-1}|y_{t-1})dw_{t-1} = \mathcal{N}(\mu_{t|t-1}, \Sigma_{t|t-1})
\]

- **Update step:**

\[
p(w_t|y_t) = \frac{p(y_t|w_t)p(w_t|y_{t-1})}{p(y_t)} = \mathcal{N}(\mu_{t|t}, \Sigma_{t|t})
\]
where the notation \( t | t - 1 \) means prediction at time instance \( t \) from the time instance \( t - 1 \). It can be shown that both distributions \( p(w_t | y_{t-1}) \) and \( p(w_t | y_t) \) are Gaussian. Given the current measurement \( y_t \), the globally optimal estimate of \( w_t \) can be determined by using the linear minimum mean square error (LMMSE) estimator [Sch91] in the update step. LMMSE is the conditional expectation of \( w_t \) given the measurement \( y_t \), where the conditional expectation is assumed to be a simple linear function of \( y_t \), \( E\{x|y\} = G_ty_t \). \( E\{\cdot\} \) refers to the expected value and \( G_t \in [0, 1]^{n \times m} \) which is called Kalman gain reflects relative certainty of the measurements and current state estimate. A low gain makes the filter follow the predictions which are obtained in the prediction step more closely. In other words, a gain of zero causes the measurements to be ignored while gain of one causes the predictions to be totally ignored. In addition, the LMMSE provides the estimate of the full distribution of the signal \( w_t \) so that Kalman filter can track the mean and the covariance of the signal \( w_t \) simultaneously.

The conventional Kalman filter can provide the estimate of a signal \( w_t \) without a sparsity solution. To meet the sparsity constraint, Hierarchical Bayesian Kalman filter [FKDY13] which is derived from the principles behind the Kalman filter and Sparse Bayesian Learning (SBL) is employed in my linear dynamic sparse modelling method for fMRI sequence reconstruction due to: 1) it uses the same model as described in Equations 4.1 and 4.3, 2) it tracks the mean and covariance of the image \( w_t \) which is necessary for implementing the measurement design process, 3) the employment of hyperparameters which is used to model the image variations promotes sparsity.

The main difference between the conventional Kalman filter and HB-Kalman is the distribution placed on the variable \( \kappa_t \) in Equation 4.1. Figure 4.7 illustrates this difference using their graphic models. The conventional Kalman filter places a Gaussian distribution on the variable \( \kappa_t \), while in HB-Kalman the distribution of \( \kappa_t \) is no longer a Gaussian but a hierarchical sparse distribution (Equation 4.2) which can promote the sparsity of the variable. The hierarchical sparse distribution which is constructed by a Gaussian distribution
\( \kappa_{ti} \sim \mathcal{N}(0, \alpha_{ti}^{-1}) \) with a gamma distribution \( \alpha_{ti}^{-1} \sim \Gamma(a, b) \) with \( i = 1 \ldots n \) is approximate to a Student’s t distribution (detailed in Equation 4.2) that promotes a strong sparsity property [Tip01]. It also places an inverse Gamma distribution prior \( p(\sigma^2|c, d) = \Gamma^{-1}(\sigma^2|c, d) \) on the noise variance \( \sigma^2 \) where \( c \) and \( d \) are set to small values. As opposed to the conventional Kalman filter where the covariance matrix \( Q_t \) of \( \kappa_t \) as well as the noise variance \( \sigma^2 \) are predefined with the knowledge of system and measurement noise, the covariance defined by \( A_t^{-1} \) where \( A_t = \text{diag}(\alpha_t) = \text{diag}([\alpha_1, \ldots, \alpha_n]) \) and the noise variance are learnt online from the measurement \( y_t \) and therefore an additional optimisation step is required.
4.3. Linear Dynamic Sparse Modelling

Similar to the conventional Kalman filter, the prediction and update steps need to be performed at each time instance. In the prediction step, one has to evaluate:

$$\begin{align*}
\mu_{t|t-1} &= \mu_{t-1}, \\
\Sigma_{t|t-1} &= \Sigma_{t-1} + A_t^{-1}, \\
y_{t|t-1} &= \Theta_t \mu_{t|t-1}, \\
y_{e,t} &= y_t - y_{t|t-1}.
\end{align*}$$

(4.8)

In the update step, one computes:

$$\begin{align*}
\mu_{t|t} &= \mu_{t|t-1} + G_t y_{e,t}, \\
\Sigma_{t|t} &= (I - G_t \Theta_t) \Sigma_{t|t-1}, \\
G_t &= \Sigma_{t|t-1} \Theta_t' (\sigma^2 I + \Theta_t \Sigma_{t|t-1} \Theta_t')^{-1};
\end{align*}$$

(4.9)

where $G_t$ is the Kalman gain at time $t$. Different from the standard Kalman filter, one has to perform the additional step of learning the hyperparameters $\alpha_t$ and $\sigma^2$. From Equation 4.8 we get $y_{e,t} = \Theta_t k_t + \eta_t$. Following the analysis in SBL [Tip01], maximising the likelihood $p(y_{e,t}|\alpha_t)$ is equivalent to minimising the following cost function:

$$L(\alpha_t) = \log \det(\Sigma_{\alpha_t}) + y_{e,t}' \Sigma_{\alpha_t}^{-1} y_{e,t},$$

(4.10)

where $\Sigma_{\alpha_t} = \sigma^2 I + \Theta_t A_t^{-1} \Theta_t'$ and $\det(\Sigma_{\alpha_t})$ refers to the determinant of it. The hyperparameter $\alpha_t$ and $\sigma^2$ can be directly estimated via the Expectation-Maximisation (EM) algorithm, while it is impractical to process the high dimension functional MR images due to its low computational speed. Therefore, a fast algorithm [TFO03] (described in Section 3.1.3 in Chapter 3) is used. Notice that as $\log \det(\Sigma_{\alpha_t})$ is a convex function while $y_{e,t}' \Sigma_{\alpha_t}^{-1} y_{e,t}$ is a concave function, the sum of them (i.e. the cost function $L(\alpha_t)$) is non-convex. Therefore the obtained estimate $w_t$ is generally suboptimal. However, this suboptimal solution is proved to be very useful in practice [KLD13]. The HB-Kalman algorithm is summarised in Algorithm 4.3.1.
Algorithm 4.3.1: HB-KALMAN($y_t$, $\Theta_t$, $\mu_{t-1}$, $\Sigma_{t-1}$)

\[ \begin{align*}
\mu_{t|t-1} & \leftarrow \mu_{t-1}; \\
y_{e,t} & \leftarrow y_t - \Theta_t \mu_{t|t-1}; \\
\alpha_t, \sigma^2 & \leftarrow \text{SBL} (y_{e,t}, \Theta_t); \\
\Sigma_{t|t-1} & \leftarrow \Sigma_{t-1} + \text{diag}(\alpha)^{-1}; \\
G_t & \leftarrow \Sigma_{t|t-1} \Theta_t' (\sigma^2 I + \Theta_t \Sigma_{t|t-1} \Theta_t')^{-1}; \\
\mu_{t|t} & \leftarrow \mu_{t|t-1} + G_t y_{e,t}; \\
\Sigma_{t|t} & \leftarrow (I - G_t \Theta_t) \Sigma_{t|t-1}; \\
\end{align*} \]

return ($\mu_{t|t}$, $\Sigma_{t|t}$)

**comment:** SBL is the optimisation algorithm described in [TFO03].

### 4.3.2 Informative Measurement Design

The reconstruction quality of a functional MR image is limited by the information obtained from measurements. According to [CWF09], information acquired from measurements can be quantified by the mutual information between the unknown image and measurements. The mutual information quantifies the extent to which uncertainty of the unknown signal is reduced when measurements are given. Furthermore, measurements are determined by a design matrix according to the measurement equation in Equation 4.3.

Given the budget $m$ (i.e. the number of DFT coefficients to be measured), the measurement design problem is to select a subset of $m$ vectors from the projection matrix $\Theta$ so as to maximise the mutual information $I(w_t; y_t)$ between the unknown image and measurements, which is defined as follows:

\[ I(w_t; y_t) = h(y_t) - h(y_t|w_t), \quad (4.11) \]
where \( h(y_t) \) refers the marginal entropy that measures the uncertainty about the signal \( y_t \), while \( h(y_t|w_t) \) is called conditional entropy that measures the amount of uncertainty remaining about \( w_t \) after \( y_t \) is known. After \( y_t \) is given which means the design matrix is determined, the only variable that influences the value of \( w_t \) is noise \( \eta_t \) and thus the conditional entropy \( h(y_t|w_t) \) is merely the entropy of noise \( \eta_t \), which is an invariance to the design matrix \( \Theta_t \). We therefore can find the optimal subset of vectors by maximising the entropy \( h(y_t) \) of the measurements \( y_t \) instead, where \( h(y_t) = \sum_{j=1}^{m} -p(y_{tj}) \log p(y_{tj}) \).

Using the system equation (Equation 4.1) and the measurement equation (Equation 4.3), we obtain:

\[
y_t = \Theta_t(w_{t-1} + \kappa_t) + \eta_t. \tag{4.12}
\]

Now, the measurement design problem addresses the solution of the following optimisation problem:

\[
\Theta_t = \arg \max_{\Theta_t} h(\Theta_t(w_{t-1} + \kappa_t) + \eta_t) \quad s.t. \Theta_t \text{ is formed by } m \text{ row vectors of } \Theta,
\]

with

\[
h(\Theta_t(w_{t-1} + \kappa_t) + \eta_t) = \sum_{j=1}^{m} -p(\Theta_{tj}(w_{t-1} + \kappa_t) + \eta_t) \log p(\Theta_{tj}(w_{t-1} + \kappa_t) + \eta_t).
\]

The posterior distribution of \( w_{t-1} \), provided by the HB-Kalman reconstruction algorithms, is a multivariate Gaussian distribution with mean \( \mu_{t|t-1} \) and covariance \( \Sigma_{t|t-1} \). As explained in the system Equation 4.2.2, I place a Student’s t sparse prior on each element of \( \kappa_t \). To make the prior non-informative, I set the hyper-parameters \( a \) and \( b \) close to zero.

Given the posterior distribution of \( w_{t-1} \) and the prior distribution of \( \kappa_t \), the distribution of \( \Theta_t(w_{t-1} + \kappa_t) + \eta_t \) can be determined. However, as the calculation of closed-form of the sum of a Gaussian random variable and a Student’s t random variable is analytically intractable, we cannot provide the expression of the distribution to calculate the marginal
entropy $h(\Theta_t(w_{t-1} + \kappa_t) + \eta_t)$.

Figure 4.8: The Relationship between Student’s t Distribution and Gaussian Distribution.

Nevertheless, Seeger et al. [SW10] suggests that a Student’s t distribution $\text{Student}(v)$ is a super-Gaussian distribution that can be approximated in terms of a Gaussian distribution, where parameter $v$ controls the shape of the distribution. Figure 4.8 shows an example of the relationship between the Student’s t and Gaussian distributions. We can see that the Student’s t distribution becomes closer to the Gaussian distribution as $v$ increases. When $v$ is large enough it can be well approximated by a Gaussian distribution. Note that a larger value of $v$ generates a taller curve with longer tail which makes most of the samples concentrate around zero, therefore a sparser solution of $w$ can be obtained. As I desire a sparse solution of $w$ which requires a large value of $v$ indeed, we can use a zero-mean multivariate Gaussian distribution to approximate the sparse prior of $\kappa_t$, where $\kappa_t \sim \prod_{t}^{n} \mathcal{N}(0, \iota)$. The constant value $\iota$ is determined by the level of variations $\kappa_t$. The higher the level, the larger the value of $\iota$ should be.

As $\Theta_t(w_{t-1} + \kappa_t) + \eta_t$ is an affine transformation of $(w_{t-1} + \kappa_t) \sim \mathcal{N}(\mu_{t-1}, \Sigma_{t-1} + \text{diag}(\iota))$, it has a multivariate Gaussian distribution with mean $\Theta_t\mu_{t-1}$ and covariance...
\( \Theta_t(\Sigma_{t-1} + \text{diag}(\iota))\Theta_t' + \sigma^2 I \). Now, the marginal entropy of \( \Theta_t(w_{t-1} + \kappa_t) \) can be written as:

\[
h(\Theta_t(w_{t-1} + \kappa_t)) = \frac{m}{2} (1 + \log(2\pi)) + \frac{1}{2} \log \det(\Theta_t(\Sigma_{t-1} + \text{diag}(\iota))\Theta_t' + \sigma^2 I),
\]

where \( \det \) refers to the determinant of the matrix. As the value of the term \( \frac{m}{2} (1 + \log(2\pi)) \) is a constant which is an invariance to the design matrix, the optimisation process can be reduced to:

\[
\Theta_t = \arg \max_{\Theta_t} \log \det(\Theta_t(\Sigma_{t-1} + \text{diag}(\iota))\Theta_t' + \sigma^2 I)
\]

\[
s.t. \Theta_t \text{ is formed by } m \text{ row vectors of } \Theta.
\]

(4.16)

Let \( \hat{\Sigma} = \Sigma_{t-1} + \text{diag}(\iota) \), according to the properties of the determinant, we can write the objective function \( \log \det(\Theta_t\hat{\Sigma}\Theta_t' + \sigma^2 I) \) as:

\[
\log \det(\Theta_t\hat{\Sigma}\Theta_t' + \sigma^2 I) = \log \det(\sigma^{-2}\hat{\Sigma}(\sigma^{-2}\Theta_t\Theta_t' + \hat{\Sigma}^{-1})),
\]

\[
= \log \det(\sigma^{-2}\hat{\Sigma}) \log \det(\sigma^{-2}\Theta_t\Theta_t' + \hat{\Sigma}^{-1}),
\]

\[
= \log \det(\sigma^{-2}\hat{\Sigma}) \log \det(\sigma^{-2}\sum_{j=1}^{m}\theta_j'\theta_j + \hat{\Sigma}^{-1}).
\]

Because term \( \log \det(\sigma^{-2}\hat{\Sigma}) \) is an invariance to the design matrix, the optimisation process can be further modified as:

\[
\Theta_t = \arg \max_{\Theta_t} \log \det(\sigma^{-2}\sum_{j=1}^{m}\theta_j'\theta_j + \hat{\Sigma}^{-1})
\]

\[
\text{s.t. } \Theta_t \text{ is formed by } m \text{ row vectors of } \Theta.
\]

(4.20)

Solving this optimisation problem usually has high computational complexity [JB09] for high dimensional signal which is the case in this study. Hence, an approximation approach [SBV10] which highly reduces the computational complexity is employed. From Equation 4.20, we can see that the optimal subsets of vectors \( \{\theta_j\}_{j=1}^{m} \) can be obtained in a
greedy fashion which finds one optimal vector at a time. Let \( m_s \) and \( m_u \) denote the current selected and unselected projection vectors respectively and \( U = \sigma^{-2} \sum_{i \in m_u} \theta_i' \theta_i + \hat{\Sigma}^{-1} \), the next optimal vector \( \theta_{m^*} \) is the solution of the following optimisation problem:

\[
m^* \leftarrow \arg \max_{j \in m_u} \log \det(\theta_j' \theta_j + U).
\] (4.21)

As \( U \) is a invertible matrix and \( \theta_j \) is a row vector, then

\[
\log \det(\theta_j' \theta_j + U) = \log \det(U)(1 + \theta_j U^{-1} \theta_j').
\] (4.22)

By removing the invariance \( \log \det(U) \) and constant 1, the optimisation problem in Equation 4.21 can be simplified as:

\[
m^* \leftarrow \arg \max_{j \in m_u} \theta_j U^{-1} \theta_j',
\] (4.23)

The approximation approach [SBV10] that I employed is a greedy algorithm. It finds the optimal solution by solving the problem in Equation 4.23, and therefore it can accomplish the selection with \( m \) iterations.

**Algorithm 4.3.2:** IMD(\( \Sigma_{t|t-1} \), \( \Theta \), \( m_I \), \( m \))

\[
l \leftarrow 1; U_l \leftarrow \Sigma_{t|t-1}^{-1}; m_u^l \leftarrow m_I; m_s^l \leftarrow \emptyset;
\]

**for** \( l \leftarrow 1 \) **to** \( m \)

\[
\begin{align*}
  m^* & \leftarrow \arg \max_{j \in m_u^l} \theta_j U_l^{-1} \theta_j'; \\
  m_s^l & \leftarrow m_s^{l-1} \cup m^*; \\
  U_l^{-1} & \leftarrow U_{j-1}^{-1} - \frac{U_{j-1}^{-1} \theta_{m^*} \theta_{m^*}' U_{j-1}^{-1} \theta_{m^*}'}{1 + \theta_{m^*}' U_{j-1}^{-1} \theta_{m^*}'};
\end{align*}
\]

**return** \( (\Theta_{m_l}) \)

The approximation approach is detailed in Algorithm 4.3.2, where \( m_I \) refers to the initial
candidate set, and \( m_s^l \) and \( m_u^l \) indicate the selected and unselected projection vectors before iteration \( l \) respectively. In addition, \( U_l \) indicates the value of \( U \) after iteration \( l \), the value of which is calculated by:

\[
U_l^{-1} = (U_{l-1} + \theta_m^* \theta_m^*)^{-1},
\]

\[
= U_{l-1} - \frac{U_{l-1}^{-1} \theta_m^* \theta_m^* U_{l-1}^{-1}}{1 + \theta_m^* U_{l-1}^{-1} \theta_m^*}.
\]

My proposed method, Informative Measurement Design (IMD), not only uses the posterior distribution of the previous signal to model the uncertainty of the current unknown signal but also involves a sparse prior of the variation signal to further modify the uncertainties. The design matrix is constructed by \( m \) numbers of projection vectors selected from the projection domain, and the determined measurements can improve the reconstruction accuracy.

**Algorithm 4.3.3:** LDSM(\( \Theta, \mu_{t-1}, \Sigma_{t-1}, m, \nu \))

\[
\begin{align*}
n & \leftarrow \text{number of rows of } \Theta; \\
m_I & \leftarrow 1 \text{ to } n; \\
\Sigma_{t|t-1} & \leftarrow \Sigma_{t-1} + \text{diag}(\nu); \\
\Theta_t & \leftarrow \text{IMD} (\Sigma_{t|t-1}, \Theta, m_I, m); \\
y_t & \leftarrow \text{measurement obtained following } \Theta_t; \\
\mu_t, \Sigma_t & \leftarrow \text{HB-Kalman} (y_t, \Theta_t, \mu_{t-1}, \Sigma_{t-1}); \\
\text{return} & (\mu_t, \Sigma_t)
\end{align*}
\]

With the HB-Kalman and IMD algorithms detailed in Algorithms 4.3.1 and 4.3.2 respectively, I summarise my propose Linear Dynamic Sparse Modelling (LDSM) method in Algorithm 4.3.3.
4.4 Experimental Results

In this work, I performed experiments on a synthetic fMRI sequence rather than a real fMRI sequence. This is because the target of functional MR imaging is to reconstruct high quality image without noise, while a real fMRI sequence involves many different sources of noise which do not provide a correct reconstruction target. In order to simulate the real functional MR imaging process, I added a BOLD signal to a rest brain sequence to simulate the brain activity and added noise to the measurement to simulate the noisy measurement. The rest brain sequence (TR/TE=2500/24.3ms, 90 degree flip angle, 3mm slice thickness, 22cm FOV, 64 × 64 matrix, 90 volumes) used by Lu et al. [LLAV11] was acquired by a 3T whole-body scanner and a gradient-echo echo-planar imaging (EPI) acquisition sequence. The BOLD contrast signal having 144 voxels convolved with a bi-Gamma hemodynamic response (HDR) was created to represent a 30s on/off stimulus, and the measurements were generated with additive noise (SNR = 0.7).

Two experiments were conducted to reconstruct the first 25 volumes of the image sequence with $m = 0.3n$ measurements for $t > 1$. The reconstruction accuracy is evaluated according to the normalised error, defined as $e(t) = ||w_t - \hat{w}_t||_2/||w_t||_2$. In the first experiment, with $m$ measurements which were selected with three common methods respectively (i.e. random sampling, variational density sampling and low pass filter sampling), I compared the reconstruction accuracies obtained using the HB-Kalman algorithm and the SBL algorithm [Tip01], and demonstrated that HB-Kalman performed better. Then, in the second experiment, I used the HB-Kalman to reconstruct the fMRI sequence. I applied my proposed measurement design method to select $m$ measurements, and compared it against other sampling techniques.
4.4. Experimental Results

4.4.1 Implementation Details of the fMRI Sequence Reconstruction

For the sequence of functional MR images, let $X_t$ denote the $64 \times 64$ dimension MR image at time $t$, where $t = 1 \ldots 25$ and $n = 64 \times 64$. Let $Y_{f,t}$ and $W_t$ denote its $2D$ Discrete Fourier Transform (DFT) and Daubechies-4 $2D$ discrete wavelet transform (DWT) coefficients respectively, where $Y_{f,t} = F X_t F$ and $W_t = E X_tE'$. $F$ and $E$ refer to the DFT and DWT matrices. To make this possible with the reconstruction algorithm, I firstly transform these matrices to 1D with $y_{f,t} = vec(Y_{f,t})$ and $w_t = vec(W_t)$, where $vec(\cdot)$ denotes the vectorisation of the matrix in the bracket. Then the MRI measurements on the DFT transform of image $X_t$ can be expressed as

$$y_t = B_t y_{f,t} = B_t \Theta w_t = \Theta_t w_t, \quad (4.26)$$

where $\Theta = F_{1D}E'_{1D}$, with $F_{1D} = F \bigotimes F$ and $E'_{1D} = E' \bigotimes E' \bigotimes$ (\bigotimes refers to the Kronecker product). At time $t$, the measurement $y_t \in \mathbb{R}^m (m \ll n)$ is achieved by capturing $m$ number of Fourier coefficients of the image, and the locations to measure are determined by the $m \times n$ dimension matrix $B_t$, which contains a single 1 at a different location in each row and in which all other entries are 0.

Given the observation $y_t$ and projection matrix $\Theta_t$, an exact or approximate solution for $w_t$ can be found via reconstruction techniques [Tip01, FKDY13] by satisfying Equation 4.26. Then, the functional MR image can be reconstructed by $x_t = E'_{1D}w_t$. Figure 4.9 illustrates the transformations of functional MR images between DWT and DFT coefficients and the implementation of the reconstruction in the up and bottom of the figure respectively.
Chapter 4. Linear Dynamic Sparse Modelling for Functional MR Imaging

Figure 4.9: Implementation of the Functional MR Image Reconstruction. Note that the red point in the bottom-left image indicates the measurements.

Figure 4.10: Selected Points in $k$-space with Three Different Measurement Methods. Note that the white points indicate the selected ones.

4.4.2 Comparing Different Reconstruction Algorithms

In this experiment, I compare the performances of HB-Kalman and SBL when using the measurements selected with three different sampling methods: random measurement, low pass filtering and variational density sampling. As discussed in Section 3.3.1 in Chapter 3,
random sampling method which samples a completely random subset of $k$-space (i.e. DFT image) is the most popularly used measurement method for CS reconstruction method. On the other hand, Low Pass (LP) filter and Variational Density (VD) [LDP07] sampling methods have been widely used in the MRI technique, which make the use of the $k$-space property as described in Section 2.2 in Chapter 2. The property is that most of the energy of images is concentrated close to the $k$-space origin (i.e. low frequency components of the images). LP sampling method only selects the low frequency components, while VD method is the combination of LP and random sampling which randomly samples the points in $k$-space but sampling more near the $k$-space origin and less in the periphery of $k$-space. Figure 4.10 illustrates the measurements selected with three different methods respectively.

![Figure 4.10: Measurements selected with three different methods.](image)

(a) Random Sampling  
(b) LP Sampling  
(c) VD Sampling  
(d) Boxplot

Figure 4.11: Reconstruction Errors (HB-Kalman vs. SBL).
SBL reconstructs the image sequence by performing a simple SBL process on each MR image. SBL is comparable to HB-Kalman as it is a compressive sensing method satisfying the requirement of my method: it can estimate a posterior distribution of an unknown image. Both methods carry out the reconstruction process with a limited number of measurements. Figure 4.10(a)-(c) compares the reconstruction errors for SBL and HB-Kalman when using random sampling, LP sampling and VD sampling respectively. We can clearly see that HB-Kalman has much better reconstruction performance than SBL when using the measurements selected with random sampling and VD sampling methods, while its performance is very close to SBL when using LP sampling methods.

![Figure 4.12: Sparsity of Image $w_t$. $|\text{supp}(w_t)|$ refers to the 96% energy support of DWT coefficients of image at time $t$.](image)

When using the random sampling method, SBL algorithm generates nearly random guesses, while its reconstruction performance is highly improved when using the other two measurement methods which focus more on the selection of the low frequency components. This is because the DWT coefficients are not very sparse (as shown in Figure 4.12, $|\text{supp}(w_t)| \approx 32\%n$), the under-determined measurements ($m = 30\%n$) cannot provide
enough information of the unknown signal to produce a high quality reconstruction result. The low frequency components characterise the coarse structure of the image, while the high frequency components characterise the details of the image. When selecting more low frequency components and less high frequency components, a simple reconstruction method (i.e. simple SBL) which reconstructs each image independently can provide more general information of the image that results in smaller reconstruction error.

By contrast, HB-Kalman has the worst performance when only selecting the points close to the $k$-space (i.e. LP sampling method as shown in Figure 4.11(b)). It uses the knowledge of the preceding image as a prior to predict the present function, and the measurements are used to modify the prediction. It is obvious that the images in a fMRI sequence has similar coarse structures. When the coarse structure of the images are not accurate enough, the low frequency components help to improve it. However, when the coarse structure of a image is relevantly accurate, sampling only the low frequency components will provide redundant information for the reconstruction and cannot further improve the reconstruction accuracy. On the other hand, if we samples some high frequency components which provides details of the image, the uncertainty of the image can be reduced so that the reconstruction error is reduced. For this reason, HB-Kalman performs better when using the measurements which contain high frequency components such as random sampling and VD sampling methods. It is worthwhile to point out that HB-Kalman has a decreasing trend of reconstruction errors in the number of frames, as it reconstructs the fMRI sequence sequentially and causally.

From Figure 4.11(d) we can see that HB-Kalman provides more accuracy reconstruction than SBL, and especially when using random sampling and VD sampling methods it has remarkable performance. Among these two methods, random sampling method results in lower error median but larger quartiles. Figure 4.13 compares the reconstruction error of these two methods. Compared with the VD sampling method, the random sampling method produces worse reconstruction for the first 8 frames but better for the rests. This
is because of the same reason I mentioned above, the VD sampling method samples more low frequency components so that it can provide more general information of the image than the random sampling method at beginning. However, as most of its measurements are redundant, the reconstruction error cannot be further improved, so it remains relevantly stable. On the other hand, the random sampling method samples more high frequency components and provides less redundant measurements so that it generates high quality image after the coarse structure is relevantly accurate.

4.4.3 Comparing Different Measurement Algorithms

The above results demonstrate that the HB-Kalman reconstruction algorithm performs better on the fMRI application in comparison to the simple SBL algorithm. I therefore use HB-Kalman to implement the reconstruction process, and focus on comparing the reconstruction performances by utilising random sampling, VD sampling and the IMD method. The constant value in Equation 4.16 was empirically set to $\lambda = 1e^2$. 

![Figure 4.13: Reconstruction Errors (Random Sampling vs. VD Sampling).](image)
4.5 Discussion and Conclusion

The results, shown in Figure 4.14, demonstrate a significant improvement in the reconstruction accuracy from random sampling and VD sampling to the IMD method. The reconstruction error of the IMD method is in average 48% less than when using VD sampling (37.3% vs. 72.3%), and 48.5% than when using random sampling (37.3% vs. 71.7%).

Furthermore, Figure 4.15 shows the visually reconstructed results generated by the three methods. Both the random sampling and VD sampling result in more blurry and noisy functional MR images. Meanwhile, the IMD method is able to provide more detailed functional MR images, which is very important in fMRI techniques (e.g. activity pattern detection).

4.5 Discussion and Conclusion

In this chapter, I propose a Linear Dynamic Sparse Modelling (LDSM) method for solving functional MRI sequence reconstruction problem. Based on a key assumption that vari-
Reconstructions of functional MR images are sparse over time in the wavelet domain, the LDSM method models a fMRI sequence as a linear dynamic sparse model. By using the linear sparse model, the prior information of the unknown fMRI image can be extracted from the previous fMRI image and the sparse variations. The prior information, which expresses certainty and uncertainty of an unknown image, can be employed to boost the reconstructed image quality. Firstly, the uncertainty of the image can be used to guide the measurement so that more useful information can be obtained. Secondly, even when the number of measurements is underdetermined, a high quality image can still be generated by involving its prior information in the reconstruction process. For this reason,
the reconstruction and measurement design algorithms that use the linear dynamic sparse model are preferred.

Compressive Sensing (CS) techniques, as explained in Section 4.1, have been extensively studied to provide good solutions for static MR image reconstruction problem. However, performing CS on each image would not be an optimal approach for functional MR imaging, as it underused the properties of fMRI. In this work, the HB-Kalman filter reconstruction algorithm is employed. The algorithm uses a Bayesian Compressive Sensing method, Sparse Bayesian Learning, to extend the Kalman filter which has been popular used in traditional tracking. It is based on the linear dynamic sparse model and meets the sparsity constraint. Benefiting from the hierarchical Bayesian model, the posterior distributions of reconstructed images are provided, this satisfies the requirement of the measurement design process. By using the HB-Kalman filter, a functional MRI sequence is reconstructed causally and recursively. The experiment results demonstrated that HD-Kalman requires fewer measurements than those required for CS to provide an approximate or exact reconstruction; HB-Kalman resulted in smaller reconstruction error than simple CS using 30% of measurements.

In functional MR imaging, measurements are achieved by following predefined acquisition trajectories. As described in Chapter 2, the early trajectory used in MR imaging is Cartesian. It generates images with few artifacts, but its long acquisition time against the speed requirement of fMRI. Then a rapid acquisition trajectory, Echo Planar Imaging (EPI), was started to be concerned in 1977 [CS12]. Compared with Cartesian, the acquisition speed is highly increased. However, it results in longer readout duration so that artefacts are introduced. More recently, spiral trajectory has regained interest with applications of functional MR imaging [Glo12]. It makes fast and efficient use of gradient hardware, and introduces less artefacts by reducing the readout duration. The conventional reconstruction methods for spiral trajectory [KPK+00, KBN00, YSM+02] require interpolation of the raw data and consume long computational time, e.g. several hours
and sometimes even days. The most resent reconstruction methods, based on Compressive sensing, are promising to overcome the computational limitation. They work with underdetermined measurements and require the measurements to be incoherent; random sampling is usually used as it can provide low incoherent measurements. In addition, VD sampling which is a random sampling method but utilises the $k$-space property is more commonly used in the MRI technique. To satisfy the requirement of reconstruction methods, the development of advanced trajectories has been continued driven by neurosciences, and more powerful and higher field strength systems have become available [LLDP05, WPO+02, ZPES14]. My proposed IMD method aims to find a small number of measurements that are maximally informative about the signal. Compared with random sampling and VD sampling methods, my method can generate more informative measurements, with which higher quality images are achieved. My method has potential to be developed by modifying the spiral trajectory. The spiral trajectory enables sparse acquisition methods, and the candidate measurements provided by it are individual voxels rather than parallel lines of $k$-space that provided by Cartesian trajectories. In addition, multi-interleave perturbed spiral trajectory [LLDP05] can cover approximately the full $k$-space which is desired by my method.

The IMD method is an extension of the Bayesian method of Seeger et al. [SNPS10] that utilises correlations of adjacent images in an fMRI sequence. This is the first study to explore the benefits of this for designing measurements. Two approximation techniques are used in this study to resolve the intractability of the measurement design problem. One uses a zero-mean multivariate Gaussian distribution to approximate the Student’s t distribution, which makes the calculation of the prior distribution of a MR image tractable. The other uses a greedy algorithm to reduce the computational complexity of the optimisation problem. The experiment results demonstrate that my proposed method can highly improve the quality of reconstructed functional MR images. However, the theoretical bounds of the approximation techniques are still unknown. Also, a learning algorithm
which can enable dynamic modification of the hyperparameters of variations using the information from reconstructed images needs to be explored in the future.

My reconstruction method is derived from the standard Kalman filter method assuming that each voxel is a stationary signal in time series, while fMRI times series are always at least weakly non-stationary [GMBT+97, HJBC10]. In addition, I model the measurement noise using an inverse Gamma distribution; whereas, in real application, the measurement contains varies sources of noise (e.g. system noise, physiological noise and etc.) and they are normally non-stationary signals. For this reason, my method is not robust enough to handle the real functional MR imaging problem. In order to make my method applicable in real application, I will modify my model with a non-stationary method which assigns time-vary distributions on both $w$ and $\eta$. 
Chapter 5

Robust Linear Sparse Modelling for fMRI Data Analysis

5.1 Introduction

Conventional fMRI analysis focuses on investigating the interpretation of neural activity with univariate analysis, such as General Linear Model (GLM) [FJT94]. The univariate analysis methods work on isolated voxels and they determine active brain regions with the most statistically significant voxels in response to a target task. In contrast to univariate analysis, Multivariate Pattern Analysis (MVPA) of fMRI attempts to informatively decode patterns of brain activities [HGF+01]. By measuring multiple voxels simultaneously, MVPA is more sensitive and informative to the brain activity and robust to noise. Several classical machine learning methods, such as linear discriminant analysis (LDA), naive Bayes, and support vector machine (SVM), have been widely used in MVPA [MKBK10] to construct predictive model using individual voxels as predictors and time point volumes as samples. When all voxels from the whole brain are used to find functional structure, the number of voxels is several orders of magnitude of the time points; and this is especially
true for individual subject analysis. Learning predictive models from such fMRI dataset is an underdetermined problem and classical machine learning methods may easily become overfitting. Previous MVPA methods attempt to reduce the dimension of input voxels by introducing feature selection methods beforehand or building predictive model on a small brain region of interests (ROI), while the input voxel reduction may result in loss of significant information that defeats the purpose of MVPA.

More advanced methods [YSY+08, CCR+09, RSAM10, VGT12] employed sparse modelling to solve the dimensionality problem by exploring spatial sparsity of fMRI data. They are based on the assumption that relevant voxels in response to a target task are often sparse compared to the number of whole brain voxels which is normally the case in fMRI studies. Compared with the conventional MVPA methods, they have shown distinct advantages: 1) they can directly work on the whole brain without introducing feature selection methods beforehand, 2) they only use a small subset of input voxels for prediction so that overfitting problem can be alleviated, 3) when a linear sparse model is used, neural activity can be interpreted by studying the selected voxels. Compressive Sensing (CS) [Don06] has been widely used to implement the sparse modelling with underdetermined measurements. However, most CS methods can successfully estimate the sparse model parameter only if the number of samples is large enough for estimating the supports (i.e. non-zero coefficients) of the model parameter, where the lower bound of the number of samples is proportional to the numbers of input voxels and supports; the design matrix must be well conditioned that satisfies the Restricted Isometry Property (RIP) condition. However, in fMRI studies, these are all violated: the number of samples (usually several hundreds) is normally less than the lower bound value and spatial correlations of neighbouring voxels cause the constructed design matrix to against the Restricted Isometry Property (RIP) condition. Some methods [YSY+08, CCR+09, RSAM10, VGT12] have been proposed to handle the modelling problem with poor conditioned design matrix. Some of these methods [CCR+09, RSAM10] focused on the predictive performance,
while the some others [VGT12, YSY+08] considered more on the relevant voxel selection. Note that a good selection does not always lead to a good prediction.

Along with the sparse modelling random subspace method to overcome the overfitting problem resulted from high dimensionality, random subspace has also been proven to alleviate the overfitting problem [TTLW06] by being integrated with some classical machine learning methods, e.g. linear discriminant analysis (LDA), naive Bayes, and support vector machine (SVM). The random subspace method generates a set of subspaces by randomly sampling a small subset of features, and a predictive model is built for each subspace; the final prediction is operated by aggregating the predictions produced by subspace models. In this chapter, I focus on improving the predictive power of the linear sparse model. I propose a robust sparse modelling method for fMRI analysis; it is implemented by incorporating the linear Sparse Bayesian Learning (SBL) with the random subspace method. I will show that with the implementation of random subspace method, the performance of linear SBL will be improved based on the following facts: 1) the correlation among features in a subspace can be tremendously reduced by random selection, and therefore better conditioned matrix can be produced; 2) discrepancy between the sample size and the voxel size as well as the support size in a subspace is highly reduced. As a consequence, linear SBL can provide robust predictive models for subspaces, and a final strong predictive model can be constructed via an aggregating process. Moreover, benefiting from the linear sparse model, predictive maps are provided so that interpretation of neural activity can be investigated.

The rest of this chapter is organised as follows. In Section 5.2, I introduce linear sparse modelling methods for fMRI MVPA study. In Section 5.3, I describe my proposed method; and experiment results of applying my model to a real fMRI data for binary classification tasks are detailed in Section 5.4. In the final section, I make a conclusion of this work.
5.2 Linear Sparse Modelling in Multivariate Pattern Analysis

Sparse modelling avoids overfitting problem by constructing predictive models using a small subset of high dimensional features. A sparse model is expressed as:

\[ y = \varphi(\Theta)w, \quad (5.1) \]

where \( \Theta \in \mathbb{R}^{m \times n} \) composes of \( m \) samples and \( n \) features corresponding to a response vector \( y \in \mathbb{R}^m \), \( \varphi(\Theta) \) is a fixed feature-space transformation, and \( w \) contains the coefficients of the model to be estimated.

The results of fMRI analysis show that active brain regions responding to a target task are just a small part of the entire investigated area (e.g. whole brain). Motivated by this observation, the decoding problem can be formulated by a linear sparse model, with which direct relevance of each voxel to the response can be obtained and no transformation of \( \Theta \) is needed. This gives:

\[ y = \Theta w, \quad (5.2) \]

where the non-zero coefficients of \( w \in \mathbb{R}^n \) indicate their corresponding voxels are relevant to the state \( y \). The total number of the relevant voxels, \( S \), should be far less than the total number of the voxels, that is \( S \ll n \).

Given \( y \) and \( \Theta \), the decoding problem is to learn the model parameter \( w \). Constrained by the fMRI imaging technique, the number of the samples \( m \) is always limited \( (m \ll n) \), making the estimation of \( w \) difficult. The Compressive Sensing (CS) (explained in Section 3.1 in Chapter 3) technique offers an opportunity for solving this problem with sparse constraint on \( w \) as long as the number of samples is large enough, e.g. \( m \gg \zeta \log(n/S) \), where \( \zeta \) is a small constant [Bar07]. In addition, most compressive sensing algorithms
[CRT06, TG07, NV10] have been proven to provide an accurate sparse solution if the
design matrix satisfies the Restricted Isometry Property (RIP) condition which requires
the every set of columns of the design matrix is approximately orthogonal.

However, in the fMRI study, the spatial correlation of voxels leads to poor conditioned
design matrix. When some columns of the design matrix $\Theta$ are highly correlated, some
sets of columns of the design matrix are not approximately orthogonal which against the
condition of RIP. Therefore, it is difficult for most CS methods to provide even an approx-
imate estimate of the sparse model parameter. I here used a real fMRI data (explained in
Section 5.4.1) as an example to evaluate the design matrix for fMRI study. To meet the re-
quirement of most CS methods that the columns of design matrix have zero mean and unit
standard deviation, the design matrix constructed by the real fMRI dataset was properly
normalised that:

$$
\frac{1}{m} \sum_{i=1}^{m} \Theta_{ij} = 0 \quad \text{and} \quad \left(\frac{1}{m} \sum_{i=1}^{m} \Theta_{ij}^2\right)^{\frac{1}{2}} = 1, \quad 1 \leq j \leq n. \quad (5.3)
$$

Because it has a combinatorial complexity to check whether the design matrix satisfies the
RIP condition using Equation 3.5 in Chapter 3, I used an alternative condition (detailed
in Chapter 3) instead:

$$
\sqrt{1 - \epsilon} \leq \lambda_{\text{min}}(\Theta) \leq \lambda_{\text{max}}(\Theta) \leq \sqrt{1 + \epsilon} \quad (5.4)
$$

where $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$ refer to the minimum and maximum singular values of $\Theta$ respec-
tively. From the Equation 5.4, we can achieve the lower bound of $\epsilon$:

$$
\epsilon \geq \lambda_{\text{max}}(\Theta)^2 - 1. \quad (5.5)
$$

The maximum singular values of the constructed design matrix with the real fMRI data
are extremely large ( > 10 ). I got $\epsilon > 30$ which against the RIP constraint ($\epsilon \in (0, 1)$), so I can say the design matrix does not satisfy the RIP condition. This is a counter-example to show that the design matrix constructed by the fMRI images does not satisfy the RIP condition.

### 5.3 Random Subspace Sparse Bayesian Learning

In this work, I adapt the Sparse Bayesian Learning (SBL), a state-of-the-art CS method, to implement the linear sparse modelling. The SBL algorithm has been demonstrated that it can empirically provide a useful sparse solution even when the design matrix is in poor condition and no penalty parameters need to be defined via cross-validation. However, in fMRI analysis, the large relevant-to-sample ratio ($S/m$) within subject still gives a challenge. For a high resolution fMRI image ($n \approx 200,000$), the number of relevant voxels is as large as several thousands. The number of samples for each individual subject is usually several hundreds, which is much smaller than the lower bound required for an accurate estimation. To overcome this problem, my approach integrates SBL with the random subspace method to construct a robust predictive model which is composed of multiple predictive models.

Random subspace (RS) method aims to generate multiple predictive models, from which a strong aggregated predictive model can be produced. It starts with generating $L$ subspaces, each of which contains $M$ voxels randomly sampled (with replacement) from the whole input space. Then, a linear sparse model is constructed in each subspace. In each subspace, both the relevant-to-sample ratio and the spatial correlation can be reduced due to the decreased size of the relevant voxels and the random sampling process respectively. Hence, the SBL has high probability to provide an accurate estimate of the model parameter with which the relevant voxels can be more correctly selected. However, even when the relevant voxels can be correctly detected, the majority of the predictive models
are weak as each of them only involves partial information. For this reason, an aggregating process is employed to produce a final strong predictive model by combining the predictive models of subspaces which make all the detected relevant voxels contribute to the final prediction. Here, I use the linear combination method for implementing the aggregation process:

\[
PredictiveModel^*(\Theta) = \frac{1}{L} \left( \sum_{l=1}^{L} PredictiveModel_l(\Theta) \right).
\]  

(5.6)

where \(\{PredictiveModel_l(\Theta)\}_{l=1}^{L}\) indicates the predictions from a series of weak predictive models \(\{PredictiveModel_l\}_{l=1}^{L}\), and the aggregated prediction result is represented as \(PredictiveModel^*(\Theta)\) which is a continuous value in \([0, 1]\). When this method is used to solve the regression problem, \(PredictiveModel^*(\Theta)\) is the final predictive result and Equation 5.6 is equivalent to the weighted combination method [ZZ11] with equal weight for each subspace. On the other hand, when my method is used to solve the classification problem, it requires a cutoff threshold to make the final decision (i.e output integer values to indicate the classes). When the threshold equals to 0.5 which is normally used in the classification problem, the aggregation process actually follows the majority voting rule [TTLW06] which counts the largest number of the predicted results that agree with each other:

\[
PredictiveModel^*(\Theta) = sgn\left\{ \sum_{l=1}^{L} PredictiveModel_l(\Theta) - \frac{L-1}{2} \right\},
\]  

(5.7)

where \(\text{sgn}\{F\}\) refers to the sign function, when \(F \geq 0\) it outputs 1, and 0 otherwise.

The final result benefits from the random sampling process, as it can highly reduce the probability of involving biased voters.

Figure 5.1 shows the framework of my method, random subspace SBL with linear model (RS-SBL), by illustrating how my method works in both training and prediction phases. In the training process, the RS-SBL method generates a set of predictive models along
with their selected voxels from a training dataset. In the prediction process, the predictive models that are learned from training process are applied to a prediction dataset, where the final predictions are obtained by the aggregating method. Note that both training and prediction datasets are constructed with preprocessed whole brain fMRI images.

Algorithm 5.3.1: RS-SBL for Training\((L, M, \{\Theta_{\text{train}}, y_{\text{train}}\}, F_I)\)

\[
\text{for } l \leftarrow 1 \text{ to } L \\
\quad \left\{ \\
\quad \begin{align*}
F_l & \leftarrow \text{randomly sample } M \text{ voxels from } F_I; \\
\Theta_{\text{train}, F_l} & \leftarrow \text{select the } F_l \text{ columns of } \Theta_{\text{train}}; \\
\text{PredictiveModel}_l & \leftarrow \text{SBL} (y_{\text{train}}, \Theta_{\text{train}, F_l}); \\
\end{align*}
\quad \right\}
\]

\text{return } (\text{PredictiveModel}, F)

\text{comment: SBL} – the optimisation algorithm described in [TFO03].
Algorithm 5.3.2: RS-SBL for Prediction($L, \{\Theta_{test}, y_{test}\}, F$)

for $l \leftarrow 1$ to $L$

\[
\Theta_{test_{F_l}} \leftarrow \text{select the } F_l \text{ columns of } \Theta_{test};
\]

\[
p_l \leftarrow \text{PredictiveModel}_l(\Theta_{test_{F_l}})
\]

\[
\text{PredictiveModel}^*(\Theta_{test}) \leftarrow \text{aggregating}\{p_l, 1 \leq l \leq L\};
\]

return ($\text{PredictiveModel}^*$)

The details of RS-SBL of training and prediction are summarised in Algorithms 5.3.1 and 5.3.2 respectively, where $F_I$ refers to the input voxels (i.e. voxels of whole brain image), and $\{\Theta_{train}, y_{train}\}$ and $\{\Theta_{test}, y_{test}\}$ refer to the datasets used for training and prediction respectively.

In fMRI studies the relevant voxels appears as sparse in the brain, and thus most of the information for prediction is contained by a small number of voxels and the rest will contribute only noise to the prediction. If subspace only samples noise voxels, the constructed predictive model would produce predictions that are completely incorrect. Having such predictive models will highly reduce the predictive accuracy of the final aggregated predictive model. In addition, the relevant voxels should be covered by the subspaces so that important information is not lost. Therefore, we want each subspace contains at least one relevant voxel and all relevant voxels are sampled at once by the subspaces. These two factors are controlled by the selection of $L$ and $M$.

Lumila et al. [KRP+10] introduced two criteria to meet our requirements: usability and coverage. A subspace is supposed to be usable as long as it contains at least one relevant voxel. The usability criterion measures the probability of getting all subspaces usable which is expressed as:

\[
P_u = (1 - (1 - \frac{S}{n})^M)^L.
\] (5.8)
On the other hand, the coverage criterion measures the probability of having subspaces that cover all relevant voxels which is represented as:

\[ P_c = (1 - (1 - \frac{M}{n})^L)^S. \]  \hspace{1cm} (5.9)

The larger value of these two probabilities should be able to result in less biased predictive models and less relevant information loss so that an accurate final predictive model can be produced. Therefore, I select the optimal setting of \( L \) and \( M \) from the candidates which provide both probabilities equal to 1.

### 5.4 Materials and Experiments

#### 5.4.1 Dataset

The fMRI data used in my experiments were provided by Human Connectome Project (HCP) [BBH+13] (see reference for more details). HCP has provided task-evoked fMRI datasets for seven different tasks which were designed to activate a variety of cortical and subcortical brain areas. In my work, I adapted my algorithm to analyse the relationship processing task, which was conducted to study the active brain regions for processing internally and externally generated information. The subjects engaged in the task had to perform two different types of processing, called relation processing and matching processing. In the relation processing, the subjects were first asked to identify the dimension differed across a pair of objects, and then decided if the other pair of the objects on the screen differed along the same dimension. In the match processing, the subjects only needed to answer whether an object matched either of the other two objects on a suggested dimension. The relation processing were supposed to involve processes of generation of both internal and external information. The match processing, whereas, only involved the
later one.

The HCP project collected data from 77 participants, who aged between 22 and 35 with no previous reported disease that might influence brain functions. The whole-brain EPI (TR/TE = 720/33.1 ms, 52 degree flip angle, 2 slick thickness, 208x180(RO x PE) FOV, 91x109x91 voxels) was acquired by a 3T WU-Minn HCP scanner with a modified 3T Siemens Skyra system. For each subject, two individual sessions for each subject were acquired, and each lasted about 167 s. Each session consisted of 3 blocks for each of the two processing tasks, as well as 3 blocks for a resting baseline during which subjects stayed relaxed. The raw fMRI volumes were preprocessed via the HCP fMRIVolume Preprocessing Pipeline [GSW⁺13] including standard preprocessing steps, such as gradient unwarping, motion correction, grand-mean intensity normalisation, and etc. The final “minimally preprocessed” 4D high spatial resolution brain images were generated by registering the original images to the standard MNI125 space so that volume-based analysis can be conducted across sessions or even subjects. In addition, to get benefits from the high resolution acquisition no smoothing process was applied.

5.4.2 Experimental Protocol

Under the relationship processing experiment, I focus on three classification tasks within individual subjects to check whether the subject: 1) did relation processing or stayed relaxed, 2) did match processing or stayed relaxed, 3) did relation processing or match processing. I performed the experiments on the preprocessed whole brain fMRI data of a group of 10 unrelated subjects which were selected from 77 candidates. For each subject, I performed training process on one session and prediction on the other. Each session consisted of 232 samples and each of the three respond vectors (relation, match, and rest) had approximately one third of the samples. As a result, there were about 155 training samples for each classification task. As the total number of voxels (≈ 228,000) was far
more than the number of samples, training classification models was a underdetermined problem.

As in these experiments, I focus on constructing binary classifiers to cognitive task of the input fMRI data, I used the classification algorithm of SBL, Sparse Bayesian Classification (SBC) (detailed in Section 3.1.3 in Chapter 3), with linear model as my key method; where \( y(\Theta; w) \) is \( \Theta w \) in Equation 5.2. I compared the performance of my proposed algorithm, RS-SBC, to three other MVPA methods:

- **Sparse Bayesian Classification with linear model (SBC).** This is the fundamental classification algorithm of my method, where the design matrix is constructed by the features in voxel space other than any projected spaces (e.g. projected using linear kernel function).

- **Support Vector Machine with linear kernel (SVM-Lin).** SVM has been the most popular classifier for fMRI data analysis, and the most useful version is the one with linear kernel [MKBK10]. The parameter of linear kernel is optimised by 10-fold cross-validation in my experiment.

- **Sparse Bayesian Classification with linear kernel (SBC-Lin).** SBC has been demonstrated having similar (or even better) classification performance to SVM on some applications (e.g. hyperspectral image classification [DE07]). However, no comparison has been made on the fMRI analysis. I constructed this competitor using SBC with linear kernel to make it comparable to the above SVM classification method. Here, the transformation matrix in (5.1) was defined as \( \rho(\Theta) = \Theta \Theta' \). The estimated sparse vector \( w \) was no longer the weights of voxels but the weights of features in a projected space. The weights of voxels were needed to be calculated by \( w^* = \Theta' w \), and relevant voxels had to be selected with a determined threshold.

All the competitors and my proposed algorithm were applied on whole brain fMRI im-
ages rather than a subset of voxels selected with a feature selection algorithm. The classification performance was evaluated by classification accuracy as well as the Receiver Operating Characteristic (ROC).

In my experiments, I also investigated some popular compressive sensing based MVPA algorithms: iterative algorithm [NT09], approximate message passing [DMM10], and elastic net [CCR+09]. As the classification accuracies were as bad as random guess, I do not report the results. Furthermore, to investigate the interpretation of the predictive maps generated by my algorithm, the predictive maps were compared to reference active maps. The active maps were computed by General Linear Model (GLM) implemented in SPM8 analysis tool [ABC+08] without spatial smoothing process.

5.4.3 Results

I selected the value of $L$ and $M$ according to Equations 5.8 and 5.9 by defining $S$ and $n$ in advance. As the ground truth for the number of relevant voxels $S$ is unknown, I determined it by the number of significant voxels detected by GLM. For the 10 subjects, the GLM method resulted in around 1000 significant voxels for both Relation vs. Rest and Match vs. Rest tasks, and less than 100 significant voxels for Relation vs. Match task. I set $S = 1000$ and $S = 50$ (determined by the average of significant voxels across subjects) for the first two and the last tasks respectively. Then, I calculated the distributions of the usability and the coverage (as shown in Figure 5.2). From Figure 5.2 we can see that the large values of both $L$ and $M$ can yield large values of the criteria. As there is no upper bound of the value of $L$ and it has been demonstrated that large values of the criteria give a robust classifier [KP10], I chose 16 equally spaced candidate pairs of $(L, M)$ which make $P_u = 1$ and $P_c = 1$. The local optimal values of $L$ and $M$ were chosen using 10-fold cross-validation for the training dataset. Due to the randomness of sampling, the results were averaged over 50 experiments.
5.4. Materials and Experiments

(a) Usability

(b) Coverage

(c) Usability

(d) Coverage

Figure 5.2: Surfaces for Usability and Coverage with $L = 1 \sim 100$ and $M = 1 \sim p$, (a)-(b) $S=1000, n=228000$ and (c)-(d) $S=50, n=228000$.

Classification Performance

Figure 5.3 shows the average classification accuracies across 10 subjects using four different classifiers. Each of the three classification tasks contained two experiments: using session 1 to predict session 2 and vice versa. The results demonstrate that my proposed method has the highest classification accuracy, while SVM has the worst performance as it is hard to handle overfitting with a huge number of voxels.

SBC-Lin performs better than SBC in the first two classification tasks. However, its classification accuracy is closer to and even worse than SBC for some subjects in the last one. For the first two tasks, although the number of relevant voxels is small compared to the total number of the input voxels, compressive sensing theory cannot detect all relevant ones with such a small sample size. Also, because of the existing correlation
among voxels, only a fraction of the relevant voxels can be detected by SBC. Therefore, inaccurate estimate was obtained, and the number of relevant voxels found by SBC was only several dozen which was extremely sparse compared to the significant voxels found by GLM. On the other hand, as the relevant voxels have high proportion in the input voxels, SBC-Lin which involves more relevant voxels for prediction can provide more

Figure 5.3: Classification Accuracies Across 10 Subjects with Four Different Classifiers: SVM-Lin, SBC-Lin, SBC and RS-SBC.
accurate prediction. However, the relevant voxels are much sparser in the last task, as the difference between the relation and match tasks only lies in the internal information generation region. Because of the sparse relevant voxels, SBC-Lin overfits the training dataset, and therefore produces worse classification performance.

Compared with SBC, RS-SBC can further increase classification accuracy by detecting more relevant voxels. This is because the number of correlated voxels can be highly reduced by random subspace, so the correlated relevant voxels have high probability to be detected by SBC in each subspace. On average, the classification accuracy is improved 15% over the SBC method, and 3% better than the SBC-Lin method. The number of relevant voxels determined by RS-SBC is larger than SBC but much smaller than SBC-Lin. In average, approximately 1500 and 70 voxels are found for the first two and last tasks respectively. Along with the highest average accuracy, my method can construct the most stable classifier with the smallest variance across all the cases.

As mentioned in Section 5.3, the aggregation process normally produces a continuous value for the classification result of a brain image and a threshold is required to determine which class the image belongs to. I calculate the classification accuracy by using majority voting rule which determines the class using threshold equaling to 0.5. To remove the effect of threshold on the evaluation, I also compare the classification performance of the four different classifiers using ROC curves (shown in Figure 5.4) with Area Under Curve (AUC) values (listed in Table 5.1). For each task, the ROC values is calculated with the classification results averaged over 10 subjects with both experiments. The ROC curves show similar evaluation results to the classification accuracy: 1) my method (i.e. RS-SBC) has much better classification performance than others, 2) SVM performs worst, 3) the classification performance of SBC-Lin is better than SBC in the first two tasks but is reduced towards SBC in the last task.
Chapter 5. Robust Linear Sparse Modelling for fMRI Data Analysis

Figure 5.4: ROC Curves for Four Different Classifiers.

(a) Relation vs. Rest

(b) Match vs. Rest

(c) Relation vs. Match

Table 5.1: AUC Values for Four Different Classifiers.
Predictive Map

I illustrate some examples of ‘individual’ predictive maps obtained by my method and active maps generated by the GLM analysis. The GLM analysis, implemented by SPM first-level analysis, returns the active maps containing voxels with the Familywise Error (FWE) correction value smaller than 0.05 (empirical value). The FWE correction was used instead of \( p \) value as it can reduce the probability of false positives. My method produced predictive maps by selecting voxels that were supposed to be relevant on average across 50 repeated experiments. The maps of both methods were created by registering relevant voxels to the subject’s T1w structural image. The active maps of GLM are shown in the top of the figures, and the predictive maps are shown in the bottom.

Figure 5.5 and 5.6 show the predictive maps of Relation vs. Rest and Match vs. Rest of one subject respectively. We can see that relevant voxels detected by different methods are not identical, but similar brain regions are clearly outlined. Both methods find voxels located in prefrontal and visual cortex. The prefrontal is typically for controlling complex cognitive processes, such as reasoning and working memory; and it is involved here for processing internally and externally generated information. The visual cortex is activated for viewing the objects shown on the screen. Unlike the GLM method, my method does not find the most significant voxels but the voxels jointly contributing to the relation processing task. In addition, the determined relevant voxels by my method are overly sparse.

As discussed in Section 5.4.3, regions involved in the Relation vs. Match task are very small. This is because the task is potentially only related to the brain regions for processing internally generated information. For some subject (e.g. Figure 5.7), both my algorithm and GLM method locate the similar activity regions. However, for some others (e.g. Figure 5.8), no significant voxels are detected by GLM. The SPM uses random-field theory to calculate the FWE correction. The random-field employs the spatial correlation
of the voxels in a given statistic image, it estimates how unlikely a voxel with particular statistic level would appear by chance according to the local smoothness. Using unsmoothed datasets in my experiments, the significant voxels may be too isolated to be detected using FWE. Therefore, I applied GLM to three smoothed datasets which were generated by smoothing the original brain images using 3D Gaussian kernels of FWHM = 4 mm, 6 mm, and 8 mm respectively. I then used different values of FWE ranging from 0.001 to 0.1 to filter the voxels. However, GLM still cannot detect any significant voxel. As a consequence, I can say that there is no significant voxel detected by GLM because there is no voxel individually carrying information about the cognitive state. On the contrary, my method can find relevant voxels that jointly response to the state, with which a
5.5 Conclusions

In this chapter, I propose a robust sparse modelling method (RS-SBL) to generate brain behaviour predictive models as well as predictive maps from whole brain fMRI images. With limited sample size, the state-of-the-art compressive sensing technique, Sparse
Figure 5.7: An Example of ‘Individual’ Predictive Map vs. Active Map for Relation vs. Rest Task.

Bayesian Learning, is used as the key technique of RS-SBL to determine sparse relevant voxels during predictive model construction. That is feature selection and modelling are integrated into one single step. The dependency of the choice of significant value in traditional feature selection methods can be therefore removed. My work is the first attempt to integrate random subspace method with SBL. By randomly sampling small subsets of features in voxel space, the spatial correlation and feature-to-sample ratio in each subspace are largely reduced so that multiple robust classifiers can be constructed. Therefore, aggregating the multiple subspace classifiers returns a strong classifier.

I have shown that the performance of my method outweights three other methods when
analysing the fMRI dataset provided by HCP. Among those methods, SVM-Lin and SBC-Lin involve feature space transformation, so all voxels are used for classification. Hence, relevant voxels were difficult to be selected. In addition, as all voxels were used for prediction, the predictive models tended to overfit the training data. SBC, on the other hand, directly detected relevant voxels, whereas the resultant predictive map was extremely sparse and its predictive performance was poor. My method, benefiting from the implementation of random subspace, was able to provide meaningful predictive maps with the strongest predictive power. The setting of parameters of my model was controlled by the sparsity of the relevant voxels, which is unknown in real application. I here used the GLM analysis result to estimate the sparsity, this might not result in a selection of the global
optimal parameter settings. However, my method still showed good performance, and better results are expected with global optimal parameters.

My constructed predictive maps highlight the consistent activated brain regions detected by the GLM method, while my results are overall sparse. My method provides the most meaningful predictive maps maintaining the strongest predictive power. However, these sparse solutions may contain noise voxels and discard some relevant ones. In order to enable an accurate interpretation of brain activity from my constructed predictive map, post-processing steps are required. For instance, use a smooth operator (e.g. Gaussian filter) to smooth the detected sparse regions. This can remove the isolated voxels which have high probability to be noise and distribute the activated regions across wide areas so that across subjects analysis can be applied.

In this work, I only concern about classification problem. My method can also be adapted to solve the regression problem by replacing SBC with SBR. Moreover, with the application of the extended SBC method [PDG10], multi-task classification can be implemented. My method has the potential to be further improved by adding some ensemble methods, such as bagging method. The bagging method can be applied to each subspace, where multiple predictors from a subspace are generated with bootstrapped samples.
Chapter 6

Balancing Stability and Predictive Performance of fMRI Models

6.1 Introduction

The early MVPA studies focus on how well brain states can be predicted. Recently, more and more studies consider its function of voxel selection so that neural activity responding to the brain states can be interpreted, which is an important factor for neuroscientific discovery. Linear sparse modelling has been a popular technique in MVPA studies, as it can be used for implementing predictive model as well as selecting relevant voxels from input voxels using the sparse model parameter. A model is considered to be robust for interpreting neural activity if the selected voxels are all relevant to the specific brain state. The conventional linear sparse modelling methods [YSY+08, CCR+09, RSAM10, VGT12] select the voxels by considering their predictive powers; the selected ones are those that provide the most accurate prediction. However, because the number of samples is always considered to be underdetermined compared to the number of input voxels, if only take the predictive performance into consideration, the selected voxels are specific
to particular dataset and irrelevant voxels may be wrongly selected. In addition, most existing linear sparse modelling methods manage to find the sparsest number of voxels used for prediction. Because of the correlations existing among relevant voxels, these methods can only detect a subset of them. This can introduce a large number of false negative selections while the predictive is almost unaffected. In consequence, with such wrongly selected voxels, even if the predictive performance can be guaranteed, the neural activity is misunderstood. For this reason, an urgent problem for MVPA is to select voxels that can interpret real neural activity as well as provide accurate predictions.

Biomarker discovery which aims to select biomarkers to differentiate diseases from normal states uses similar analysis methods as fMRI MVPA analysis. Both biomarker discovery and fMRI MVPA face the same problem as: 1) the training datasets usually have high feature-to-sample ratio; 2) the selected predictors are expected to be predictive and meaningful, where the predictors are always sparse compared to the high dimensional features; 3) correlations existing among predictors. In order to control the robustness of predictors, researchers (e.g. [AHdP+10, ZRS08]) introduced the concept of stability to biomarker discovery techniques. They denoted that a feature is considered to be stable if it is consistently selected when using a selection method with different sample sets. They demonstrate that if the stability is higher, the selected predictors are more robust to noise so that the probability of selecting noisy predictors is highly reduced. However, if an arbitrary fixed set of predictors is chosen, the stability is perfect but the predictive performance is very poor. This is because some real predictors which are unstable is discarded if only taking the stability into consideration. To deal with this, Kirk et al. [KWB+13] investigated strategies to balance the robustness and predictive performance of biomarkers by optimising both stability and predictive performance simultaneously.

To improve the accuracy of relevant voxel selection of MVPA methods, in this chapter, I introduce the concept of stability to the fMRI MVPA analysis which has not been considered in this field before. I explore the advantages of bringing stability into voxel selection
and propose a novel multivariate voxel selection method which selects the relevant voxels by considering their stability and predictive power simultaneously. The method is implemented by wrapping a proposed selection strategy around my novel sparse modelling method, Random Subspace Sparse Bayesian Learning (RS-SBL) (detailed in Chapter 5). My method aims to select voxels that can accurately discriminate different brain states as well as enable precision interpretation of brain activities. By using my selection strategy, which combines stability and prediction accuracy assessments, the probabilities of RS-SBL of both selecting irrelevant voxels and unselecting relevant voxels are highly reduced, whereas only small reduction of predictive accuracy is made.

This chapter is organised as follows. In Section 6.2, I first explain the voxel select function of RS-SBL and then describe my proposed method; experimental results of testing my method on both simulation and real datasets are detailed in Section 6.3. In the final section, I make a conclusion of this work.

## 6.2 Methods

In this section, I first detail the voxel selection function of the Random Subspace Sparse Bayesian Learning (RS-SBL) method. I then introduce the assessment of stability and how it is combined with the assessment of predictive performance. Finally, I describe my proposed multivariate voxel selection framework that adopts the combined assessment method.

### 6.2.1 RS-SBL for Multivariate Voxel Selection

Instead of focusing on boosting the predictive power of the RS-SBL method, I make special effort on the multivariate voxel selection process which aims to select relevant voxels in response to a target task by analysing multiple voxels simultaneously.
Figure 6.1 shows the framework for the relevant multivariate voxel selection function of the RS-SBL method. Given an input fMRI dataset $D = \{\Theta, y\}$ with $m$ samples, where $\Theta \in \mathbb{R}^{m \times n}$ composes of $m$ samples and $n$ voxels corresponding to a vector $y \in \mathbb{R}^m$ which indicates the brain states, the RS-SBL method selects a subset of voxels $T = \{\tau_i\}_{i=1}^{q^*}$ as the relevant ones together with a weight vector $\lambda_T$, where $q^*$ refers to the number of voxels selected by a sparse modelling method (e.g. RS-SBL). The selected voxels which are the non-zero elements of the linear models are achieved by aggregating the voxels selected from all subspaces. The weight of each selected voxel is proportional to the frequency of this voxel being selected by sparse modelling during the random sampling process (i.e. generating subspaces). It is defined as:

$$\lambda_T = \frac{Na_{\tau_i}}{Nb_{\tau_i}}, \quad \text{with} \quad i = 1...q^*;$$

(6.1)

where $Na_{\tau_i}$ and $Nb_{\tau_i}$ indicate the times of voxel $\tau_i$ selected by subspaces and linear sparse models of the subspaces respectively. In order to reduce the influence of the sampling randomness, RS-SBL is repeated $R$ times and the weights of the selected voxels are averaged over repetitions. The final selected voxels $T_{final}$ are the top $q$ voxels with the largest weights. The implementation details of RS-SBL Multivariate Voxel Selection (MVS) is shown in Algorithm 6.2.1, where $F_I$ refers to the input voxels and $w_I$ refers to
the model parameter for subspace $l$.

After the voxel selection process, the Sparse Bayesian Classification with Linear kernel (SBC-Lin) method (explained in Section 5.4.2 in Chapter 5) uses the selected voxels to make prediction. SBC-Lin is employed here as it has similar or even better predictive performance than SVM method which is the most popular predictive model for fMRI data analysis. In addition, no parameter needs to be optimised via cross-validation.

Algorithm 6.2.1: MVS($L, M, D = \{\Theta, y\}, q, F_I, R$)

$n \leftarrow$ the number of elements in $F_I$;

$average \leftarrow \{0\}^n$;

for $r \leftarrow 1$ to $R$

\[
\begin{align*}
T_r, \lambda_T & \leftarrow \text{RS-SBL} (L, M, \{\Theta, y\}, F_I); \quad \text{(Algorithm 6.2.2)} \\
I_T & \leftarrow \text{index of elements in } T_r \text{ in } F_I;
\end{align*}
\]

do

\[
\begin{align*}
weights & \leftarrow \{0\}^n; \\
weights(I_T) & \leftarrow \lambda_T; \\
average & \leftarrow average + weights;
\end{align*}
\]

average $\leftarrow$ average/R;

$I_{final} \leftarrow$ index of the largest $q$ elements in average;

$T_{final} \leftarrow F_I(I_{final})$;

return ($T_{final}$)
Algorithm 6.2.2: RS-SBL($L, M, D = \{\Theta, y\}, F_i$)

\[
T \leftarrow \emptyset;
\]

for \( l \leftarrow 1 \text{ to } L \) do

\[
\begin{align*}
F_i & \leftarrow \text{randomly sample } M \text{ voxels from } F_i; \\
\Theta_{F_i} & \leftarrow \text{select the } F_i \text{ columns of } \Theta; \\
\end{align*}
\]

\[
\begin{align*}
w_l & \leftarrow \text{SBL } (y, \Theta_{F_i}); \\
I_l & \leftarrow \text{index of the nonzero coefficients of } w_l; \\
\tau S_l & \leftarrow F_i(I_l); \\
T & \leftarrow T \cup \tau S_l;
\end{align*}
\]

\[
T \leftarrow \text{remove the duplicated elements in } T;
\]

for each \( \tau_i \in T \) do

\[
\begin{align*}
\mathcal{N}_{a_{\tau_i}} & \leftarrow \text{the times } \tau_i \text{ appears in } \tau S; \\
\mathcal{N}_{b_{\tau_i}} & \leftarrow \text{the times } \tau_i \text{ appears in } F; \\
\lambda_{\tau_i} & \leftarrow \frac{\mathcal{N}_{a_{\tau_i}}}{\mathcal{N}_{b_{\tau_i}}}; \\
\end{align*}
\]

return \((T, \lambda_T)\)

comment: SBL – the optimisation algorithm described in [TFO03].

### 6.2.2 Combining Stability and Predictive Performances

The conventional method for evaluating the selected relevant voxels assesses the performance of their predictive models when model parameters (e.g. \( L, M \) and \( q \)) are set to be optimal. Given a dataset \( D = \{\theta_i, y_i\}_{i=1}^m \) with \( m \) samples and a predictive model \( F_\chi \) with parameter \( \chi \), the predictive power is estimated by the predictive accuracy \( C \), which is proportional to the times that the predicted and observed states are the same:

\[
C(D; f) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(\theta_i) == y_i),
\]

(6.2)
where $D$ in this equation indicates a test dataset, and $f$ is the predictive function obtained by fitting predictive model $F$ on a training dataset; $\mathbb{I}(Z)$ is an indicator function, that $\mathbb{I}(Z) = 1$ if $Z$ is true and $0$ vice versa.

The common approach for assessing the predictive performance uses cross-validation so as to produce an unbiased assessment. It calculates the probability of correct prediction by averaging the predictive accuracies achieved via cross-validation:

$$P(\{\text{model accuracy}\} | \mathcal{F}_X) = \frac{1}{K} \sum_{k=1}^{K} C_k,$$

where $K$ refers to the number of folds for cross-validation. Because of the small number of samples, the selection is very sensitive to variations in the training samples, so the irrelevant voxels are incorrectly selected. Using cross-validation, different voxels are selected for different partitioned samples, and therefore unstable selection is obtained. To overcome this problem, I introduce the concept of stability to MVPA. Stable relevant voxels are the ones that are consistently selected across cross-validations, and the irrelevant voxels are excluded via stability assessment due to their instability. For any subset of the voxels $V \subseteq \{v_1, \ldots, v_n\}$, its stability is quantified by the probability that the voxels in $V$ are among the voxels selected by a MVPA MVS (e.g. RS-SBL MVS) method [KWB+13]:

$$P(\{\text{select } V\} | \mathcal{F}_X) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}(V \subseteq T_k).$$

On the other hand, if only the stability is taken into consideration, the predictive performance is poor as some unstable relevant voxels are unselected. For this reason, a more reliable way to evaluate the selected voxels is to use the performances of both the prediction and stability. As the assessments of both the prediction (Equation 6.3) and the stability (Equation 6.4) require cross-validation, it is natural to combine them via their joint probability. If I assume the number of the relevant voxels is $q$, Equations 6.3 and 6.4...
are redefined as:
\[ P(\{\text{model accuracy}\}|\mathcal{F}_\chi, q) = \frac{1}{K} \sum_{k=1}^{K} C_{qk}, \]  
(6.5)

and
\[ P(\{\text{select } V_q\}|\mathcal{F}_\chi, q) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}(V_q \subseteq T_{qk}). \]  
(6.6)

Then the joint probability of Equations 6.5 and 6.6 is obtained via a simple multiplication calculation:
\[ P(\{\text{select } V_q \& \text{ model accuracy}\}|\mathcal{F}_\chi, q) = \frac{1}{K} \sum_{k=1}^{K} C_{qk} \mathbb{I}(V_q \subseteq T_{qk}). \]  
(6.7)

### 6.2.3 Proposed Method

Adopting the combined assessment method in Equation 6.7, I propose a novel MVPA relevant voxel selection method that the selected voxels are stable in cross-validation process while returning a robust predictive model.

Figure 6.2: Framework of Proposed Method

Figure 6.2 illustrates the framework of my proposed method. Given parameters \( \chi = \{L, M\} \) and \( q \), multivariate voxel selection (RS-SBL) works on a subset samples \( D_k \), which is randomly sampled from the whole input fMRI dataset, \( D \). After determining the \( q \) most relevant voxels \( T_{qk} \), the predictive model (i.e. SBC-Lin) is applied on the
remaining dataset \( D_k = D \setminus D_k \). The predictive accuracy is defined as \( C_{qk} = C(D_k, f_{qk}) \) in Equation 6.2. In this work, the predictive model \( \mathcal{F}_\chi \) in Equation 6.7 is composed of the multivariate voxel selection model \( g_\chi \) and the predictive model \( h \). The model \( g_\chi \) is trained with the whole input voxels \( \{v_1, \ldots, v_n\} \), while the predictive model \( h \) is fitted only with the voxels \( T_q \) selected by \( g_\chi \).

The optimisation process in the framework finds an optimal set of relevant voxels \( V_q \) from \( K \)-fold cross-validation with the optimisation of both stability and predictive accuracy. This is achieved by maximising the joint probability (Equation 6.7). My method returns a set of selected relevant voxels \( V_{Optimal} \) with the associated \( P_{SC_{Optimal}} \) score, where \( P_{SC}(V_q) = P(\{select\ V_q &\ model\ accuracy\}|\mathcal{F}_\chi, q) \). By searching all possible values of parameter \( \chi \) and \( q \), the optimal set of relevant voxels with the highest \( P_{SCI} \) score is selected. I summarise my proposed method in Algorithm 6.2.3

---

**Algorithm 6.2.3: PROPOSED METHOD**

\( D = \{\Theta, y\}, L, M, q \)

\[ K \leftarrow 10; R \leftarrow 50; \]

\[ n \leftarrow \text{the number of columns of } \Theta; \]

\[ F_I \leftarrow 1 \text{ to } n; \]

\[ \text{for } k \leftarrow 1 \text{ to } K \]

\[ \begin{align*}
D_k &\leftarrow \text{randomly sample rows of } D; \\
D \setminus k &\leftarrow D \setminus D_k; \\
T_{qk} &\leftarrow \text{MVS} (L, M, D_k, q, F_I, R); \quad (\text{Algorithm 6.2.1}) \\
C_{qk} &\leftarrow \text{SBC-Lin} (D_k, D \setminus k, T_{qk}, F_I); \quad (\text{Algorithm 6.2.4})
\end{align*} \]

\[ V_{Optimal}, P_{SC_{Optimal}} \leftarrow \text{Optimisation} (T_q, C_q, q); \quad (\text{Algorithm 6.2.5}) \]

\[ \text{return } (V_{Optimal}, P_{SC_{Optimal}}) \]
Algorithm 6.2.4: SBC-LIN($D_k = \{\Theta_k, y_k\}, D_{\setminus k} = \{\Theta_{\setminus k}, y_{\setminus k}\}, T_{qk}, F_I$)

$I_T \leftarrow \text{index of elements in } T_{qk} \text{ in } F_I$;
$\Theta_{k_{\text{train}}} \leftarrow \text{select the } I_T \text{ columns of } \Theta_k$;
$\text{basis}_k \leftarrow \Theta_{k_{\text{train}}} \Theta_{k_{\text{train}}}'$;
$h \leftarrow \text{SBL} (y_k, \text{basis}_k)$;
$\Theta_{\setminus k_{\text{test}}} \leftarrow \text{select the } I_T \text{ columns of } \Theta_{\setminus k}$;
$f_{qk} \leftarrow h (\Theta_{\setminus k_{\text{test}}})$;
$C_{qk} \leftarrow C (D_{\setminus k}, f_{qk}); \ (Equation \ 6.2)$

return $(C_{qk})$

*comment:* SBL – the optimisation algorithm described in [TFO03].

Algorithm 6.2.5: OPTIMISATION($T_q, C, q$)

$i \leftarrow 0$;
$T_{\text{unique}} \leftarrow \text{remove the duplicated elements in } T_q$;

for any $V_q \subseteq T_{\text{unique}}$

\[ i \leftarrow i + 1; \]

\[ V\{i\} \leftarrow V_q; \]

\[ P_{SC_i} \leftarrow P(\{\text{select } V_q \text{ & model accuracy}\}|F_X, q); \ (Equation \ 6.7) \]

$I_{\text{Optimal}} \leftarrow \text{index of the element with maximum value in } P_{SC}$;

$V_{\text{Optimal}} \leftarrow V\{I_{\text{Optimal}}\}$;

$P_{SC_{\text{Optimal}}} \leftarrow P_{SC_{I_{\text{Optimal}}}}$;

return $(V_{\text{Optimal}}, P_{SC_{\text{Optimal}}})$

Note that, in the Optimisation process, selecting the $q$ voxels which provide the maximum joint probability from the voxel set $\{v_1, \ldots, v_n\}$ is a heuristic optimisation process. It is intractable for high dimensional feature space which is exactly the case in fMRI analy-
sis. For this reason, an approximation method is applied to optimise the following target function:

\[
P_{SCI}(V_q) = \frac{1}{q} \sum_{v_i \in V_q} P_{SC}(v_i),
\]

(6.8)

where \( P_{SC}(v_i) = P(\{ \text{select } V_i \& \text{ model accuracy} \}|F_\chi, q) \), and \( v_i \) is the \( i^{th} \) element of the voxel set \( V_q \). This approach margins the selection and accuracy probability associated with individual voxels rather than voxel set (e.g. Equation 6.8).

### 6.3 Experiment Results

I tested my method on both simulation and real fMRI datasets. As the simulation datasets provide the ground truth for the relevant voxels, the accuracies of both the prediction and relevant voxel selection can be obtained. By contrast, the real dataset can only calculate the former one. In my experiments, I compared my method with other two optimisation approaches, which were implemented via assessing the prediction accuracy and stability respectively:

- **Stability.** It selects the relevant voxels with the highest stability score which is defined as \( P_{SI}(V_q) = \frac{1}{q} \sum_{v_i \in V_q} \sum_{k=1}^{K} \mathbb{I}(v_i \subseteq T_{qk}) \). Same as \( P_{SCI} \), it is an approximation of stability probability (Equation 6.4) of voxel sets by margining the probabilities associated with individual voxels.

- **Accuracy.** It is the conventional optimisation strategy to select the relevant voxels using predictive models. Under different parameter settings (e.g. \( \chi, q \)), it calculates the accuracy score by \( P_e = P(\{ \text{model accuracy} \}|F_\chi, q) \) and selects the \( q \) most frequently selected voxels across \( K \) repetitions.

The optimal values of the parameters and voxels are selected by maximising the individual score of each approach.
6.3.1 Simulation Datasets

I tested my methods on simulation datasets whose ground truth (i.e. $w$) is known. The simulation datasets were generated in the same manner as it was in [VGT12] with $n = 8000$ and $m = 100$. In order to generate a design matrix $\Theta$ reflecting the key characteristic of fMRI image which is spatial correlation of voxels, I constructed $\Theta$ using a smoothed i.i.d Gaussian random matrix with a $2D$ Gaussian filter of standard deviation $\sigma = 0.8$. The $S$ non-zero elements of $w$ are grouped into different spatial clusters, and their values are randomly chosen from $\{-0.5, 0.5\}$. The state vector $y$ is then generated by the linear model (i.e. Equation 5.2 in Chapter 5) with additive noise ($SNR = 0.9$). Three datasets were generated with different number of non-zero elements, $S = 16, 54, 128$ respectively.

In my experiment, the number of repetitions $R$ was set to 50 and the parameter $K$ in cross-validation was fixed to 10. The candidate values of parameter $L$ and $M$ used in random subspace process were selected from $L \in \{10, 40, 70, 100\}$ and $M \in \{1/5n, 1/3n, 7/15n, 3/5n\}$ respectively. By defining real non-zero elements to be true positives, the predictive performances of these approaches was presented by accuracy and Receiver Operating Characteristic (ROC) and the selection accuracies were represented by false positive and false negative rates from 10-fold cross-validation.

Performance Comparison of Different Parameter Settings

I use the simulation dataset 1, the sparsity of which equals to 16 (i.e. $S = 16$), to compare the performance of different methods under different parameter settings.

Example results with selected settings of $L$ and $M$ are presented in Figure 6.3. The number of selected voxels, $q$, in MVS ranges from 1 to 30. It is clear that, with a given value of $L$ and $M$, the scores of my method and stability show similar trends: they both peak at $q = 16$, which is the true sparsity value of the dataset. Different from these
two scores, the accuracy score becomes relatively stable when $q > 5$. This difference is caused by the strong correlation between relevant voxels. The sparse modelling method seeks the minimal set of voxels that provide the best prediction performance, where only one out of the closely correlated relevant voxels is usually chosen. The random subspace method provides high probability of selecting the whole correlated voxel, and the selected relevant voxels are ranked by their predictive power. However, different training dataset $D_k$ results in different voxel ranks because of the sampling variations. As a result, when $q < S$, various subsets of correlated voxels are chosen across repetitions; and the level of stability increases with $q$ because the selected relevant voxels becomes covering the whole correlated relevant voxel set. Even though the selected voxels are unstable, the prediction accuracy can still be maintained with partial correlated relevant voxel set. When $q > S$, the stability decreases since some irrelevant voxels are wrongly selected.
Chapter 6. Balancing Stability and Predictive Performance of fMRI Models

<table>
<thead>
<tr>
<th>Dataset 1 Stability</th>
<th>(L, M)</th>
<th>q</th>
<th>P_c</th>
<th>FN</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(40,1600)</td>
<td>16</td>
<td>0.94375</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(70,2667)</td>
<td>16</td>
<td>0.94375</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(100,2667)</td>
<td>16</td>
<td>0.94375</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(100,3733)</td>
<td>16</td>
<td>0.94375</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>(10,1600)</td>
<td>14</td>
<td>0.971875</td>
<td>0.125</td>
<td>0</td>
</tr>
<tr>
<td>My Strategy</td>
<td>(100,3733)</td>
<td>16</td>
<td>0.94375</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.1: Optimal Results Achieved with Highest Scores for Simulation Dataset 1.

Moreover, at the highest score point (optimal results shown in Table 6.1), both the stability and my method detect 16 relevant voxels correctly with the predictive accuracy of 0.944. On the other hand, the accuracy method only detects 14 relevant voxels but with higher predictive accuracy, 0.972. Both the accuracy and my method find an optimal pair of (L, M), while the stability method finds four pairs having the highest score. In addition to predictive accuracy, I draw ROC curves (shown in Figure 6.4) to compare the predictive performance of the three different methods. Even though the stability methods finds more than one optimal settings, they select the same set of relevant voxels, so same predictive
models are constructed resulting in the same ROC values. Similarly, as my method finds the same set of relevant voxels as the stability method, the same ROC curve is obtained which is slightly worse than the one of the accuracy method.

### Comparison of Different Sparsity

The number of non-zero elements in the dataset used above is very small ($S = 16$) compared to the dimension of input voxels. The results show that the relevant voxels can be correctly detected with false positive rate equal to 0. In order to check the performance of my method on datasets with different number of relevant voxels, I use another two simulation datasets with $S = 54 q = 1 \ldots 70$ and $S = 128 q = 1 \ldots 150$ respectively.

<table>
<thead>
<tr>
<th>Dataset 2</th>
<th>Stability</th>
<th>$q$</th>
<th>FN</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>All pairs</td>
<td>2</td>
<td>0.963</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(10,4800)</td>
<td>48</td>
<td>0.333</td>
<td>0.222</td>
<td></td>
</tr>
<tr>
<td>(100,2667)</td>
<td>49</td>
<td>0.148</td>
<td>0.056</td>
<td></td>
</tr>
<tr>
<td>(70,3733)</td>
<td>7</td>
<td>0.167</td>
<td>0.222</td>
<td></td>
</tr>
<tr>
<td>(70,3733)</td>
<td>59</td>
<td>0.167</td>
<td>0.259</td>
<td></td>
</tr>
<tr>
<td>My Strategy</td>
<td>(100,1600)</td>
<td>47</td>
<td>0.130</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset 3</th>
<th>Stability</th>
<th>$q$</th>
<th>FN</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(100,1600)</td>
<td>115</td>
<td>0.211</td>
<td>0.110</td>
<td></td>
</tr>
<tr>
<td>(40,4800)</td>
<td>137</td>
<td>0.273</td>
<td>0.344</td>
<td></td>
</tr>
<tr>
<td>(40,4800)</td>
<td>148</td>
<td>0.250</td>
<td>0.406</td>
<td></td>
</tr>
<tr>
<td>(40,4800)</td>
<td>149</td>
<td>0.250</td>
<td>0.414</td>
<td></td>
</tr>
<tr>
<td>My Strategy</td>
<td>(100,1600)</td>
<td>129</td>
<td>0.172</td>
<td>0.180</td>
</tr>
</tbody>
</table>

Table 6.2: Comparison of Selection Accuracy. Optimal Results Achieved with Highest Scores of Three Different Methods (Simulation Datasets 2 and 3).

Tables 6.2 and 6.3 show the optimal results from the three methods with the two simulation datasets for comparing the predictive performance and selection accuracy respectively. From Table 6.2 we can see that the stability method can highly reduce the false positive rate but introduce high false negative rate (e.g. dataset 2). This because some significant relevant voxels are unstable across repetitions. The accuracy method does not only select more relevant voxels, but also irrelevant voxels. Among all three methods, my
Chapter 6. Balancing Stability and Predictive Performance of fMRI Models

<table>
<thead>
<tr>
<th>(L, M)</th>
<th>q</th>
<th>$P_c$</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all pairs except (10,1600)</td>
<td>2</td>
<td>0.898</td>
<td>0.9671</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(10,4800)</td>
<td>48</td>
<td>0.984</td>
<td>0.9989</td>
</tr>
<tr>
<td>(100,2667)</td>
<td>49</td>
<td>0.984</td>
<td>0.9991</td>
</tr>
<tr>
<td>(70,3733)</td>
<td>7</td>
<td>0.984</td>
<td>0.9999</td>
</tr>
<tr>
<td>(70,3733)</td>
<td>59</td>
<td>0.984</td>
<td>0.9996</td>
</tr>
<tr>
<td>My Strategy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(100,1600)</td>
<td>47</td>
<td>0.970</td>
<td>0.9984</td>
</tr>
<tr>
<td><strong>Dataset 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(100,1600)</td>
<td>115</td>
<td>0.947</td>
<td>0.9944</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(40,4800)</td>
<td>137</td>
<td>0.964</td>
<td>0.9973</td>
</tr>
<tr>
<td>(40,4800)</td>
<td>148</td>
<td>0.964</td>
<td>0.9978</td>
</tr>
<tr>
<td>(40,4800)</td>
<td>149</td>
<td>0.964</td>
<td>0.9976</td>
</tr>
<tr>
<td>My Strategy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(100,1600)</td>
<td>129</td>
<td>0.951</td>
<td>0.9951</td>
</tr>
</tbody>
</table>

Table 6.3: Comparison of Predictive Performance. Optimal Results Achieved with Highest Scores of Three Different Methods (Simulation Datasets 2 and 3).

Figure 6.5: ROC Curves of Three Different Methods with Optimal Parameter Settings for Simulation Datasets 2 and 3.

Method provides the most accurate selection which selects more relevant voxels and less noises compared to the stability and accuracy methods respectively. On the other hand, my method only makes a smaller reduction in predictive performance than the stability compared to the accuracy method with 1.39% vs. 5.25% and 1.70% vs. 1.77% for the predictive accuracy $P_c$ and AUC value respectively. In other words, my method makes the
trade-off between relevant voxel selection accuracy and predictive performance. Compared to the accuracy method, my method can highly reduce the false positive and false negative rates with a small reduction in predictive performance. Moreover, the optimal values of parameters selected by my method are more stable that only one combination is selected. The other two methods, by contrast, may select more than one optimal parameter settings. Under these settings, the stability method selects voxel set consistently, while the accuracy method returns different sets of voxels making it difficult to decide which one to choose. In addition, different sets of voxels can result in different constructed predictive models which have various sensitivities to the cutoff threshold. Therefore, with the optimal parameter settings, the accuracy method produces different ROC curves (shown in Figure 6.5), while all of them are better than the ROC curves of both stability and my methods.

Figure 6.6 presents some examples of scores of the two datasets. Comparing with the dataset 1, these two datasets introduce more correlated relevant voxels that the whole set of relevant voxels are more difficult to be detected via linear sparse modelling method because of the small number of samples. Even though the selection of whole relevant voxels is enabled by using the random subspace method, a subset of the correlated relevant voxels are more frequently selected than the rest across repetitions. This is because when correlated relevant voxels are sampled in a subspace, only one out of them is usually selected by the linear sparse modelling and the linear sparse modelling usually select the fixed ones. In consequence, when strong correlations existing among relevant voxels, the stability score easily peaks at a smaller sparsity level (e.g. $q = 2$ in Figure 6.6(a)) than the real one. On the other hand, although the score of my method has similar trend as the one of stability; my method which benefits from the accuracy assessment can detect more relevant voxels by finding the highest score with the optimal setting of parameters (e.g. $q = 47$ in the left panel of Figure 6.6(a)).
Figure 6.6: Examples of Scores of Three Different Methods under Different Parameter Settings (Simulation Datasets 2 and 3).

### 6.3.2 Real Datasets

The relationship processing datasets provided by HCP (detailed in Chapter 5) are used to conduct the real dataset experiments. Different from the experiments in Chapter 5, I here only focus on analysing the third classification task (i.e. Relation vs. Match) within individual subjects. The third one is the core of the relationship processing task, and it potentially activates small brain regions which are involved for processing internally generated information. I performed experiments on the whole brain fMRI data of two unrelated subjects, where the brain data have been preprocessed via the HCP fMRIVolume Preprocessing Pipeline [BBH+13]. For each subject, the dataset is constructed by
6.3. Experiment Results

combining the data obtained from two sessions. As a result, there are 310 samples and around 228,000 voxels for each dataset.

In this experiment, for both datasets, the SB-SBL process is repeated 50 times and 10-fold cross-validation is used. I select the values of parameter $L$ and $M$ from $L \in \{10, 30, 50\}$ and $M \in \{1/10n, 13/60n, 1/3n\}$ respectively. The number of the selected voxels $q$ is in range from 1 to 100, which is considered to be large enough to detected the relevant voxels (Chapter 5).

<table>
<thead>
<tr>
<th>$(L, M)$</th>
<th>$q$</th>
<th>$P_c$</th>
<th>$AUC$</th>
<th>$P_{SI}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1 Stability</td>
<td>(50,49654)</td>
<td>3</td>
<td>0.8278</td>
<td>0.9333</td>
</tr>
<tr>
<td></td>
<td>(30,76390)</td>
<td>3</td>
<td>0.8349</td>
<td>0.9412</td>
</tr>
<tr>
<td></td>
<td>(10,76390)</td>
<td>49</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>(10,76390)</td>
<td>50</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>(10,76390)</td>
<td>66</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>(10,76390)</td>
<td>67</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>(10,49654)</td>
<td>88</td>
<td>0.9976</td>
<td>0.9980</td>
</tr>
<tr>
<td>My Method</td>
<td>(50,22917)</td>
<td>8</td>
<td>0.9009</td>
<td>0.9577</td>
</tr>
<tr>
<td>Subject 2 Stability</td>
<td>(50,22932)</td>
<td>1</td>
<td>0.6604</td>
<td>0.7023</td>
</tr>
<tr>
<td></td>
<td>(30,76440)</td>
<td>57</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>(10,76440)</td>
<td>70</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>(10,22932)</td>
<td>84</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>(50,49686)</td>
<td>85</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>(30,76440)</td>
<td>89</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>(50,76440)</td>
<td>94</td>
<td>0.9976</td>
<td>0.9997</td>
</tr>
<tr>
<td></td>
<td>(50,49686)</td>
<td>97</td>
<td>0.9976</td>
<td>0.9998</td>
</tr>
<tr>
<td>My Method</td>
<td>(50, 22932)</td>
<td>29</td>
<td>0.9623</td>
<td>0.9946</td>
</tr>
</tbody>
</table>

Table 6.4: Optimal Results Achieved with Highest Scores of Three Different Methods (Subject 1 and 2).

The optimal results are shown in Tables 6.4. Similar to the simulation results, among all three methods, the stability method selects the fewest voxels with the highest stability score but the lowest predictive power. Conversely, the accuracy method results in the best predictive performance with highest predictive accuracy and AUC value, but it selects the most voxels. However, the accuracy method’s small stability scores suggest that its selected voxels contains lots of noise. Compared to the accuracy method, my method results
Chapter 6. Balancing Stability and Predictive Performance of fMRI Models

Even though the stability method finds two optimal parameter setting with the same number of relevant voxels \( q = 3 \) for Subject 1, the selected voxels are different. As these voxels have different predictive power, different \( P_c \) and ROC values are obtained.

In addition, in both simulation and real datasets experiments, my method selects the smallest \( M \) with the largest \( L \) from their candidates as the optimal parameter settings. Figure 6.8 and 6.9 present the examples of scores with different \( L \) and \( M \) values. It is clear that when using the largest value of \( M \), the scores of the stability and my methods peak at very small values of \( q \). This is because if a larger value of \( M \) is selected, more correlated relevant voxels are sampled in each subspace. As discussed in Section 6.3.1, the strong correlation existing among voxels can make the RS-SBL method select a small subset of relevant voxels more frequently than the rest across repetitions. Moreover, a large value of \( L \) can increase the probability of selecting the whole relevant voxels. Consequently, large value of \( L \) and small value of \( M \) are preferred for accurate voxel selection. The
6.3. Experiment Results

Figure 6.8: Examples of Scores of Three Different Methods under Different Parameter Settings (Subject 1).

The selection of $L$ and $M$ has significant effects on the results of the stability and my methods; but, the results of the accuracy method are not influenced in the same way. [KP10] proposed criteria for selecting $L$ and $M$. It demonstrated when the sparsity is unknown, large values of $L$ and $M$ are preferred for accurate prediction.

Figure 6.10 and 6.11 illustrate the predictive maps constructed with the selected voxels by different approaches for subject 1 and 2 respectively. From the figures we can see that, among the three methods, accuracy method finds more voxels but introduces unstable voxels which have high probability to be noise (e.g. the circled isolated voxels). The stability method reduces the probability of incorrect selection, whereas it also regards some relevant voxels as irrelevant ones which produce the worse predictive performance. Conversely, my approach is robust to noise and finds more relevant voxels than the stability
Figure 6.9: Examples of Scores of Three Different Methods under Different Parameter Settings (Subject 2).

assessment approach. The brain regions located by my method are the prefrontal cortex and primary visual cortex, which have been demonstrated to be involved in the process of internally generated information [CRGG03].

6.4 Conclusion

In fMRI analysis, MVPA has been the most popular method for the brain state prediction. In the early days, the MVPA studies mainly focus on the predictive power. Nowadays, it becomes popular to explore the interpretation of brain activity provided by the model predictors. Linear sparse modelling is an ideal approach, as the generated model parameter is directly related to the voxels. A non-zero coefficient of the model parameter
indicates its corresponding voxel is the relevant one in response to a target state. The conventional MVPA methods detect the relevant voxels only based on their predictive
power. The selected voxels tend to be specific to particular datasets especially when the feature-to-sample ratio of dataset is very high, which is always the case in fMRI analy-
sis. In consequence, the interpretation of brain activities is not perfect by exploring those
selected voxels.

In this chapter, I adopt the RS-SBL method on multivariate voxel selection and introduce
the concept of stability to the selection process. The experimental results demonstrated
that when using the predictive accuracy assessment, the sparse model selected voxels with
the highest predictive power, but also contained the irrelevant ones. By employing the as-
essment of stability, the number of false positive selections was highly reduced, but it
also reduced the number of true positive selections as well as the predictive performance.
My method combining both two assessments highly reduced the false positive and false
negative rates as well as maintained the predictive performance. The successful applica-
tion of my method on both simulation and real datasets indicates the potential of using
my MVPA method on real fMRI data to understand brain functions.
Chapter 7

Conclusion

7.1 Contributions

In this thesis, I propose novel sparse modelling methods for both the imaging and data analysis in fMRI techniques. The concrete contributions of this thesis are as follows:

Imaging

- I propose a linear dynamic sparse modelling method which is composed of both measurement design and reconstruction processes systematically boosting the functional MR image quality reconstructed with a small number of measurements. This method models an fMRI sequence as a linear dynamic sparse model based on the assumption that variations of functional MR images are sparse over time in the wavelet domain. To implement the reconstruction process, I employed the Hierarchical Bayesian Kalman filter (HB-Kalman) algorithm. This algorithm which was designed to track a dynamic sparse signal follows the linear dynamic sparse model. It can highly improve the image quality compared to the conventional reconstruction methods (e.g. compressive sensing). To implement the measurement
design process, I propose an Informative Measurement Design (IMD) method. The IMD method addresses the measurement design problem of selecting \( m \) feasible measurements such that the mutual information between the unknown image and measurements is maximised, where \( m \) is a given budget and the mutual information is extracted from the linear dynamic sparse model. The experimental results demonstrated that my method succeeded in boosting the quality of functional MR images.

Data Analysis

- I propose a novel linear sparse modelling method which integrates predictive modelling and relevant voxel selection into one process to implement the Multivariate Pattern Analysis (MVPA) of fMRI datasets. My method was developed based on the key assumption that active brain regions (i.e. relevant voxels) in response to a target task are sparse compared to the whole brain area, which is normally the case in fMRI analysis. By introducing sparsity in the modelling process, the overfitting problem which is caused by the large feature-to-sample ratios of training datasets is remitted. To estimate the sparse model parameters, I employed a state-of-the-art sparse signal reconstruction method, called linear Sparse Bayesian Learning (SBL). However, the design matrix is poor conditioned because of the high correlation existing among brain voxels. When with the poor conditioned design matrix, the linear SBL method is not able to provide an accurate estimation for the model parameter. To overcome this problem, I integrate linear SBL with a random subspace method. The random subspace method separates a training dataset into subspaces. In each subspace, the correlation and feature-to-sample ratio of the separated training dataset are largely reduced; therefore the linear SBL method has high probability to accurately estimate the model parameter. By aggregating the predictive models constructed within subspaces, a powerful predictive model can
be obtained. The experimental results from a real fMRI dataset demonstrated that when analysing the whole brain data my method had significant predictive performance compared to three other popular MVPA methods. In addition, the relevant voxels which were selected using the estimated model parameter located in informative brain areas.

- I propose a novel relevant voxel selection method so as to improve the selection accuracy of a linear sparse modelling method. By using a simulation dataset which provided ground truth for the relevant voxels, I demonstrated that the classical method in fMRI which selects the relevant voxels based on their predictive power, introduced a large number of false positive selections. To boost the selection accuracy, I introduced the concept of stability into the selection process. I found that if only emphasised on the selection stability in the selection process, even though the probability of selecting irrelevant voxels was reduced, the predictive accuracy was sacrificed and many false positive selections were introduced. I therefore proposed a method to combine the stability assessment with the predictive performance assessment, and designed a method to select voxels by considering their predictive accuracy and the level of stability simultaneously. My designed selection method was implemented by integrating the combined assessment with my proposed sparse modelling method, RS-SBL. The experimental results of both simulated and real fMRI datasets demonstrated that my method simultaneously reduced the number of false positive and false negative selections and maintained the prediction performance.

7.2 Future Work

My proposed methods have been demonstrated to provide novel solutions for imaging and data analysis in fMRI studies. However, they can be further improved to be more
intelligent, and modified to address other fMRI problems. The related further work is shown as follows:

**Imaging**

- *Algorithm Improvement*: My linear dynamic sparse modelling method showed outstanding reconstruction performance from the conventional compressive sensing measurement and reconstruction methods. Meanwhile, the measurement design algorithm can be further improved by: 1) investigating the theoretical bounds of the approximation techniques; 2) developing an intelligent learning algorithm to dynamically select the hyperparameters of the variations, the values of which have effect on the selection of measurements which determine the image reconstruction accuracy. These two issues will be considered in the development of my algorithm.

- *Other Applications*: Functional MRI data compression is another key component in fMRI technique. I will investigate the performance of my method when adapting it to implement fMRI data compression. With a powerful fMRI device, the brain can be imaged into nearly a million voxels. During an fMRI experiment, many different experimental tasks are performed, and several hundreds of scans are taken for each subject and tens to hundreds of subjects are usually involved. The obtained massive fMRI datasets leads to a significant challenge for data storage and transmission, and the dataset are required to be efficiently compressed to overcome the challenge. My method has the potential to do so. The small number of measurements which are generated during the imaging process can be stored and transmitted instead of the reconstructed high resolution fMRI images; therefore less storage space and transmission capacity are required. When a high quality of fMRI sequence is required for analysis, it can be reconstructed using the HB-Kalman algorithm.
Data Analysis

- **Algorithm Improvement**: In my current design, I only focus on the binary classification problem by using Sparse Bayesian Classification (SBC) which is the classification algorithm of SBL. However, my method has the potential to address other prediction problems by replacing SBC with different modelling methods. In future, I will adapt my method to solve the regression problem by using Sparse Bayesian Regression (SBR) algorithm instead of SBC and I will extend my method to achieve multi-task classification by employing an extended version of SBC. Moreover, my method has the potential to be improved by involving other ensemble methods, such as bagging and boosting, and I will investigate their performances in future work.

- **Other Applications**: In this thesis, my linear sparse modelling method was designed to implement the MVPA analysis of fMRI. In my future work, I will adapt it to realise functional connectivity study, which has gained increasing interest in fMRI studies. Functional connectivity analysis aims to find brain region interactions which can be used to differentiate between brain functions. Similar to the univariate analysis in fMRI data analysis, the state-of-the-art analysis method in this study first calculates the correlations of any pair of voxels, and then selects the most significant interactions by using a hard cutoff threshold. My method uses voxel pairs as features and their calculated correlations as samples. It can build a powerful predictive model in terms of brain connectivity as well as automatically select the relevant connections which can be used for interpretation of neural activity.
Bibliography


[CZLZ09] Jiaming Chen, Lu Zhang, Jianhua Luo, and Yuemin Zhu. MRI reconstruction from 2D partial k-space using POCS algorithm. In *Bioinformatics and*


[LLLZ12] Duan-duan Liu, Dong Liang, Xin Liu, and Yuan-ting Zhang. Under-sampling trajectory design for compressed sensing MRI. In *Engineering*


