Scalable Bayesian Methods for Large-scale Neuroimaging



Kan Keeratimahat

Kellogg College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2024

To my dearest grandmother, Naree Lohakul,
who nurtured every step of my journey and now rests serenely in heaven.

Abstract

The era of *Big Data* has revolutionised neuroimaging, a field increasingly driven by advancements in MRI technology, and the accumulation of large-scale datasets such as the UK Biobank study. With unprecedented access to high-resolution imaging and vast medical data, this thesis addresses three pivotal questions to bridge the gaps between data processing, analysis, and meaningful interpretation in neuroimaging.

First, as imaging hardware evolves, how can analytical paradigms adapt to increasingly granular data? Modern MRI scanners achieve remarkable temporal and spatial precision, yet widely used fMRI software struggles with rising autocorrelation in high-resolution data, leading to flawed inferences. This thesis highlights these limitations in tools like *FSL*, *SPM*, and *AFNI* and introduces a framework that integrates computational efficiency and robust theory to address this issue.

Second, how do we extract meaningful signals from increasingly complex and unparalleled data? High-resolution imaging poses the challenge of high dimensionality and limited sample size. To tackle this, we extend the Bayesian *Posterior Mean Screening* (He and Kang, 2021) framework with data-driven smoothing and variable screening, efficiently identifying critical features while preserving application relevance.

Finally, how can vast datasets improve prediction reliability? This thesis presents a unified framework combining Gaussian Processes for anatomical and functional priors, neural networks for scalability, and Bayesian statistics for uncertainty quantification. Extensions explore interactions across imaging modalities and uncover latent subgroups, offering deeper insights into brain-behaviour dynamics and their modulation by biological and socio-demographic factors.

Declarations

I, Kan Keeratimahat, hereby declare that, except where specific reference is made to the work of others, the content of this thesis is my own original work and has not been submitted, either in whole or in part, for consideration towards any other degree or qualification at any university or institution. Collaborative work is clearly acknowledged and appropriately cited within the text.

- The work presented in Chapter 3 is a collaborative effort involving multiple researchers and supervisors as listed at the beginning of the chapter. My contributions to this work included verifying the mathematical and statistical validity of the methods, interpreting experimental results, drawing scientific conclusions, and co-authoring the manuscript. This work, for which I am the first author, will soon be published along with the software developed by Dr. Afyouni, which is currently implemented in `MATLAB`.
- The work presented in Chapter 5 will shortly be submitted for publication. This work was presented as a poster for the Organisation for Human Brain Mapping (OHBM) 2024 conference.
- The work developed in Chapters 4,5,6 are part of a research project funded by the National Institute on Drug Abuse (NIDA), titled Scalable Bayesian Methods for Big Imaging Data Analysis. This project is co-led by Prof. Jian Kang and Prof. Timothy D. Johnson, both from the Department of Biostatistics at the University of Michigan.

Kan Keeratimahat

January 2025

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Professor Thomas E. Nichols, whose daring confidence and generosity provided me with the magnificent opportunity to pursue my academic aspirations in statistics and its applications. His consistent guidance and support throughout this PhD journey, whether during the isolation of COVID-19, in the department, or off-sites, have been invaluable. I am equally grateful to the Neuroimaging Statistics Oxford (NISOx) group, its members, and collaborators for fostering such a welcoming and stimulating environment during my time as a modest DPhil student.

I extend my sincere thanks to Professor Jian Kang and Professor Timothy D. Johnson for their significant support of this work through the NIH Grant [R01DA048993]. Their mentorship, alongside my supervisor's, has been instrumental throughout my PhD, and I am deeply appreciative of their wisdom and guidance.

I am profoundly thankful to the Oxford Thai Society, the Oxford Thai Foundation, and the communities of St Hugh's College and Kellogg College. Despite being thousands of kilometres away from home and in the midst of Britain's cold and grim climates, they welcomed me, nurtured me, and celebrated every milestone with me. Their warmth and camaraderie made Oxford feel like home.

Special thanks go to Mona, Mek, Boss, along with others from afar. Their unwavering encouragement and support through the ups and downs of this PhD journey have been nothing short of extraordinary. Despite the physical distance, they have stood by me as if completing this life chapter alongside me, reminding me never to give up, even in the darkest times.

Finally, and unconditionally, I owe my deepest gratitude to my father and mother for their endless love and trust in sending this boy on this journey to achieve his fullest potential. To my sister, who has been like a six-year-ahead trial version of me, always offering her guidance from her first-hand experiences, and to my Aunt Jang and Uncle Rut, who have cared for me as if I were their own son—I thank you from the bottom of my heart. This achievement is as much yours as it is mine.

Contents

1 Introduction **1**

2 Background **13**

2.1 Neuroimaging 13

2.1.1 Magnetic Resonance Imaging Background 13

2.1.1.1 The Study of Brain Function 13

2.1.1.2 Neuroimaging: Early Developments and the Role of MRI 14

2.1.1.3 Functional MRI and the BOLD Signal 15

2.1.1.4 Technological Advances and the Big Data Era of Neuroimaging 16

2.1.2 MRI Modelling and Modalities 17

2.1.2.1 BOLD fMRI and the Haemodynamic Response Function 17

2.1.2.2 T1-Weighted MRI (Structural MRI) 20

2.1.2.3 Resting-State fMRI (rs-fMRI) 20

2.1.2.4 Task-Based fMRI (t-fMRI) 20

2.1.2.5 Diffusion MRI (dMRI) 21

2.1.3 Pre-processing Pipeline 21

2.1.3.1 Brain Extraction 21

2.1.3.2 Distortion Correction 22

2.1.3.3 Motion Correction 23

2.1.3.4 Spatial Normalisation or Registration 23

2.1.3.5 Spatial Smoothing 24

2.1.3.6 Temporal Filtering 25

2.2 Statistical Modelling . . . 25
2.2.1 Regression . . . 26
2.2.1.1 Linear Model . . . 26
2.2.1.2 Generalised Linear Model (GLM) . . . 29
2.2.1.3 Interaction Terms in Regression . . . 31
2.2.2 Bayesian Statistics . . . 32
2.2.2.1 Bayesian Linear Model vs Frequentist Linear Model . 33
2.2.2.2 Approximate Posterior Distribution . . . 35
2.2.2.3 Gaussian Processes . . . 39
2.2.2.4 Gaussian Process Regression (GPR) . . . 42
2.2.3 Neural Networks . . . 44
2.2.3.1 Activation Function . . . 44
2.2.3.2 Backpropagation . . . 45
2.2.3.3 Stochastic Gradient Descent (SGD) . . . 45
2.2.4 Dimensionality Reduction . . . 46
2.2.4.1 Feature Extraction . . . 46
2.2.4.2 Variable Screening . . . 47
2.2.5 Time Series Modelling . . . 50
2.2.5.1 Basics of Time Series Data . . . 50
2.2.5.2 Modelling Approaches . . . 51
2.3 Current Study of Statistics in Neuroimaging . . . 53
2.3.1 Effect Discovery Based on Brain MRI . . . 53
2.3.2 Scalar-on-Image Regression . . . 54
2.3.3 Time Series fMRI Modelling . . . 55
2.3.3.1 Autocorrelation in fMRI Data . . . 55
2.3.3.2 Time Series fMRI Extraction . . . 57
2.4 Big Imaging Datasets . . . 60
2.4.1 UK Biobank Dataset . . . 61
2.4.2 NKI-Rockland Sample Dataset . . . 61
2.4.3 Human Connectome Project (HCP) Dataset . . . 62
2.4.4 Developing Human Connectome Project (dHCP) Dataset . . . 63
2.4.5 F-CON1000 Beijing Dataset . . . 63

2.5 Conclusion 64

3 Spectral Adjusted Least Square for Autocorrelated Noise 65

3.1 Introduction 65

3.2 Methods 67

3.2.1 Mass Univariate Time Series Linear Model 67

3.2.2 Estimation of Autocorrelation Matrix $\hat{\mathbf{\Omega}}_i$ 68

3.2.2.1 Restricted Maximum Likelihood 70

3.2.2.2 Methods of Moments 70

3.2.2.3 Nonparamaetric 71

3.2.2.4 Low Frequency Drift Models 71

3.2.2.5 Bias Correction for Design and Drift 73

3.2.2.6 Global-plus-Local Autocorelation Estimation 74

3.2.2.7 Computationally Efficient Whitening 75

3.2.3 Evaluations 76

3.2.3.1 Data 76

3.3 Results 78

3.3.1 Autocorrelation Across Choice of Smoothing, Tissue Types and Sampling Frequencies 78

3.3.2 Spectrum and False Positive Rate of Whitening Methods . . . 79

3.3.3 Subject Level Analysis and Sensitivity 84

3.3.4 Group Level Task Analysis 88

3.4 Conclusion 90

3.5 Code Availability 91

4 Bayesian High-Dimensional Variable Screening: Extensions to Posterior Mean Screening 92

4.1 Introduction 92

4.2 Background 97

4.3 Methodology 103

4.3.1 Constructing the Prior Covariance $\mathbf{\Lambda}$ 105

4.3.1.1 Data-guided method 105

4.3.1.2 Response-guided method 107

4.3.2 Shrinkage for Stabilising Prior Covariance $\mathbf{\Lambda}$ Estimation . . . 108
4.3.3 Gaussian Kernel Smoothing Estimation . . . 110
4.3.3.1 Determining the Smoothing Parameter . . . 111
4.3.4 Metrics . . . 112
4.3.4.1 Relative Importance Index . . . 112
4.3.4.2 Power, FPR, FDR in Hypothsis testing . . . 113
4.4 Analysis setup . . . 115
4.4.1 Data . . . 115
4.4.2 Model Configuration . . . 115
4.5 Simulation Study . . . 117
4.5.0.1 Simulation Setup . . . 117
4.5.0.2 Scenario 1: Single Contiguous Region in the ROI . . . 118
4.5.0.3 Scenario 2: Double Separated Regions in the ROI . . . 119
4.5.0.4 Scenario 3: Sparse Signals Across the ROI . . . 119
4.5.1 Smoothing Parameter . . . 119
4.5.1.1 Sensitivity . . . 121
4.5.1.2 Stability . . . 124
4.5.1.3 Screening Consistency . . . 125
4.6 Real-Data Analysis . . . 126
4.6.1 Predictive Accuracy . . . 126
4.6.2 Interpretability . . . 129
4.6.2.1 Region Importance . . . 129
4.6.2.2 Brain-Mapped Selected Voxels . . . 130
4.6.3 Computational Efficiency . . . 133
4.7 Discussion . . . 134
4.8 Code Availability . . . 138

5 Introduction to Bayesian Gaussian Process Neural Networks 139
5.1 Introduction . . . 139
5.2 Method . . . 141
5.2.1 Gaussian Process Neural Network . . . 141
5.2.2 Prior specifications and low-rank representation . . . 142
5.2.3 Posterior computation . . . 144

5.2.3.1 Architecture variants . . . 145
5.2.3.2 Minibatch training and Learning rate specification . 146
5.3 Bayesian Inference . . . 147
5.3.1 Stochastic Bayesian inference . . . 147
5.3.2 Adaptive Gaussian priors . . . 148
5.3.3 Measurements of interest . . . 150
5.3.3.1 Saliency map . . . 150
5.3.3.2 Posterior predictive interval . . . 150
5.3.4 GP construction and parameter tuning . . . 151
5.4 Dataset and Pre-Processing . . . 152
5.5 Results . . . 153
5.5.1 Simulation study . . . 155
5.5.1.1 Simulation set-up . . . 155
5.5.1.2 Performance . . . 156
5.5.2 Real-data study . . . 157
5.5.2.1 Performance . . . 157
5.5.2.2 Larger sample in real-data study . . . 159
5.5.2.3 Bayesian sampling . . . 159
5.5.2.4 Brain age gap diagnostics . . . 161
5.5.2.5 Saliency . . . 161
5.6 Discussion . . . 164
5.7 Code Availability . . . 166

6 Extensions to Gaussian Process Neural Network 167
6.1 Introduction . . . 167
6.2 Background . . . 171
6.2.1 Interaction Effects in Models . . . 171
6.2.2 Activation Function . . . 173
6.2.3 Gaussian Process Neural Network (GPNN) . . . 174
6.3 Methodology . . . 176
6.3.1 Combining Imaging and Non-Imaging Covariates at the Final Layer . . . 176
6.3.1.1 Example: Incorporating Sex as a Non-Imaging Covariate 176

6.3.2 Multi-class K-GPNN . . . 179
6.3.2.1 K-GPNN-Linear-GP . . . 180
6.3.2.2 K-GPNN-GP-Linear . . . 183
6.3.2.3 K-GPNN-GP-GP . . . 185
6.3.3 Imaging Bimodality GPNN . . . 186
6.3.3.1 Bimodality GPNN-Linear-GP . . . 187
6.3.3.2 Bimodality GPNN-GP-Linear . . . 189
6.3.4 Posterior Sampling . . . 190
6.4 Data . . . 191
6.4.1 Voxel-Based Morphometry (VBM) Data . . . 192
6.4.2 Default Mode Network (DMN) Data . . . 192
6.4.3 Non-Imaging Covariates . . . 193
6.4.4 Covariate Categorisation . . . 193
6.4.5 Response Variable Transformation . . . 194
6.5 Results . . . 195
6.5.1 Parameters set-ups . . . 195
6.5.2 Metrics . . . 195
6.5.3 Simulation study . . . 196
6.5.3.1 Simulation Set-up . . . 196
6.5.3.2 Simulation Results . . . 198
6.5.4 Real-data: Cognitive task . . . 202
6.5.4.1 Selecting number of latent classes "K" . . . 202
6.5.4.2 Predictive Accuracy . . . 205
6.5.4.3 Bayesian Sampling . . . 210
6.5.4.4 Model Representation . . . 215
6.5.4.5 Bimodality GPNN . . . 217
6.6 Discussion . . . 219
6.7 Code Availability . . . 222

7 Conclusion **223**

A Bayesian High-Dimensional Variable Screening: Extensions to Posterior Mean Screening **229**

A.1 Extended results tables 230

A.1.1 Simulations 230

A.1.1.1 Smoothing 230

A.1.1.2 Sensitivity and Stability 230

A.1.2 Real-data experiment 231

A.1.2.1 Predictive accuracy 231

A.2 Additional experiments 232

A.2.1 Alternative simulation setup 1 232

A.2.2 Alternative simulation setup 2 233

B Introduction to Gaussian Process Neural Network **236**

B.1 Bayesian Derivations 236

B.2 Covariance Kernel in Gaussian Processes 238

B.3 Additional figures 239

C Extensions to GPNN **241**

C.1 K-GPNN-GP-Linear 241

C.2 K-GPNN-GP-GP 242

C.3 GPNN-Linear-GP Bimodality 243

C.4 GPNN-GP-Linear Bimodality 245

C.5 Additional figures 247

Bibliography **250**

List of Figures

2.1 Brain data acquisition and imaging techniques 15
2.2 Different MRI acquisition techniques and their purposes 18
2.3 Neural acitivity and haemodynamtic response 19
2.4 Brain extraction process 22
2.5 Multiple brain functional networks extracted via ICA 59

3.1 Autocorrelation by brain tissues and datasets 79
3.2 Noise power spectrum and flatness 81
3.3 False positive rates by brain tissues and datasets 83
3.4 Subject-level sensitivity analysis 85
3.5 Subject-level z-scores 87
3.6 Group-level z-scores 89

4.1 Simulation study signals 120
4.2 Screening consistency results 127
4.3 Relative importance index of brain regions 131
4.4 Screened voxels in real data application 132

5.1 Visualisation of GPNN-GP-GP 143
5.2 K-L basis expansion for Gaussian Process 152
5.3 Target distribution in simulation study 154
5.4 Brain age gaps 162
5.5 Brain-mapped model saliency/coefficients 163

6.1 Non-imaging covariate distributions in simulation study 199
6.2 Real-data experiment results with 2 latent groups 204

6.3 Real-data experiment results with 3 latent groups 206
6.4 Brain-mapped model saliency/coefficients 218

B.1 Additional visualisation of model architectures 240

C.1 Additional visualisation of K-GPNN 248
C.2 Additional visualisation of GPNN Bimodality 249

Chapter 1

Introduction

The human brain, with its billions of neurons and trillions of synapses, is among the most intricate and enigmatic organs known to science. Its complexity underpins the vast array of cognitive, motor, and sensory functions that define human experience. For much of history, progress in understanding the brain was constrained by the inability to directly observe its structure and activity. Early methods, such as dissection and physiological observation, provided only limited insights, focusing on gross anatomy and basic functional responses. The advent of neuroimaging marked a pivotal moment in neuroscience, offering non-invasive tools to probe the living brain. Techniques like Computerised Axial Tomography (CAT) and Positron Emission Tomography (PET) paved the way for detailed anatomical and functional studies. However, these methods were hindered by invasive procedures, low resolution, and restricted accessibility. The introduction of Magnetic Resonance Imaging (MRI) in the 1970s revolutionised the field, providing a non-invasive means of visualising the brain's structure with unprecedented clarity.

Building on these foundations, modern MRI techniques now encompass a range of imaging modalities. Structural MRI (sMRI) captures detailed snapshots of grey and white matter anatomy, revealing patterns of cortical thickness and volume associated with health and disease. Functional MRI (fMRI) exploits changes in blood oxygenation levels to dynamically infer neural activity, enabling identification of brain regions involved in perception, memory, and decision-making. Resting-state fMRI (rs-fMRI) has illuminated intrinsic connectivity networks, such as the Default Mode Network

(DMN), which remains active during rest. Meanwhile, diffusion MRI (dMRI) maps the brain's white matter tracts, providing insights into structural connectivity and its disruptions in conditions such as stroke or neurodegeneration. Despite these advances, the increasing resolution and diversity of imaging techniques present significant challenges. Researchers must contend with the vast volume of data, inherent noise in imaging signals, and complex spatial and temporal correlations. Deciphering the brain remains an iterative process, requiring not only advancements in imaging technology but also sophisticated analytical frameworks to interpret its complexities.

In parallel with technological progress, large-scale neuroimaging datasets such as the UK Biobank, the Human Connectome Project (HCP), and the Adolescent Brain Cognitive Development (ABCD) study have enabled population-level studies of brain health. These datasets integrate tens of thousands of participants with behavioural, genetic, and socio-demographic data, offering unparalleled opportunities for studying the brain across diverse populations. However, their scale and complexity demand advanced analytical frameworks. Traditional methods often fall short in addressing high dimensionality, noise, and intricate dependencies, highlighting the need for innovative approaches.

One critical issue in neuroimaging analysis is handling temporal autocorrelation in task-based fMRI. Consecutive time points in fMRI signals are inherently correlated due to neural activity, haemodynamic lag, and scanner noise. If unaccounted for, this temporal dependence distorts statistical inferences, inflating false positive rates and reducing sensitivity. The issue is especially pronounced in fast-acquisition fMRI datasets, where repetition times (TR) below two seconds amplify temporal dependencies. Widely used AR(1) models, which assume a simple first-order autoregressive structure, often fail to capture the complexity of these dependencies, undermining the statistical validity of analyses (Bollmann et al., 2018). Various software packages, such as FSL, SPM, AFNI, and fMRIStat, implement distinct methods to address this, yet each has limitations. For example, FSL uses a spectral technique to model autocorrelation at voxel-level, but struggles with bias at the global level. SPM's AR(1)-based ReML approach assumes spatial homogeneity, which can oversimplify voxel-level variability (Corbin et al., 2018), while AFNI's voxel-wise ARMA model improves flexibility at the cost of high computational demands. These limitations

underscore the need for models that simultaneously address global and voxel-specific temporal dependencies.

In addition to temporal complexities, spatial correlations inherent in neuroimaging data present another significant challenge. Brain activity is organised into networks, such as the Default Mode Network (DMN), where neighbouring voxels and even distant regions exhibit correlated activity due to shared neural and haemodynamic responses (Smith et al., 2009). Ignoring these spatial dependencies can result in fragmented findings and inflated error rates, particularly during multiple comparison corrections (Friston et al., 1994). Bayesian frameworks that incorporate spatial priors, such as anatomical or functional relationships, improve robustness and interpretability (Smith and Fahrmeir, 2007; Goldsmith et al., 2014; He and Kang, 2021), but their computational demands can hinder application to large datasets. Addressing spatial and temporal complexities in a unified framework remains an ongoing challenge in neuroimaging analysis.

Dimensionality is a critical challenge in neuroimaging, where the number of predictors (p) often exceeds the number of subjects (n) by several orders of magnitude. High-resolution datasets, such as structural MRI scans with millions of voxels, exemplify this ultra-high-dimensional setting. Traditional statistical methods, like ordinary least squares (OLS) regression, are computationally infeasible and statistically invalid in these scenarios due to overfitting and identifiability issues. Even regularised approaches, such as LASSO and Elastic Net, struggle with instability, particularly when predictors are highly correlated or exhibit weak individual signals (Tibshirani, 1996; Zou and Hastie, 2005). The spatial correlations between neighbouring voxels further exacerbate these challenges, as traditional methods often fail to account for these dependencies. This complexity necessitates dimensionality reduction techniques to filter irrelevant predictors while retaining biologically meaningful features, enabling interpretable and efficient downstream analyses.

Variable screening methods have emerged as effective pre-processing tools to address dimensionality. Early approaches, such as mass univariate analysis (MUA), perform separate statistical tests for each voxel, assuming independence across predictors (Friston et al., 1994). However, MUA's failure to account for spatial dependencies often results in fragmented findings or inflated error rates. More advanced methods,

such as Sure Independence Screening (SIS) (Fan and Lv, 2008) and High-dimensional Ordinary Least Squares Projection (HOLP) (Wang and Leng, 2016), incorporate global correlations but overlook fine-grained spatial organisation. Recent developments, such as Conditional Sure Independence Screening (CSIS) (Barut et al., 2016), Partition-based Screening (PartS) (Kang et al., 2017), and Covariance-Insured Screening (CIS) (He et al., 2019), incorporate spatial priors to enhance robustness by leveraging inter-voxel dependencies or anatomical parcellations. Bayesian approaches, such as Posterior Mean Screening (PMS) (He and Kang, 2021), take this further by embedding spatial priors into a probabilistic framework, using Gaussian priors and alternative covariance structures to align with the brain's hierarchical organisation. These advancements have shown promise in capturing biologically relevant predictors, yet challenges related to computational scalability and flexibility remain, requiring further innovation in high-dimensional neuroimaging analysis.

A distinct challenge in neuroimaging lies in scalar-on-image regression tasks, where high-dimensional imaging data are linked to scalar outcomes such as cognitive scores, disease susceptibility, or brain age. These tasks demand models capable of handling the inherent non-linear relationships and spatial dependencies in neuroimaging data—challenges that traditional methods, such as ordinary least squares and regularised regression techniques, fail to adequately address. These models often struggle to account for the intricate spatial organisation of the brain or the complex interactions underlying cognitive and biological processes. In response, advanced machine learning methods, particularly neural networks and Gaussian processes, have emerged as transformative tools for capturing these associations while incorporating prior knowledge about brain structure and function.

Neural networks excel at modelling non-linear relationships in high-dimensional data, enabling the detection of subtle and intricate patterns. However, this flexibility often comes at the cost of interpretability, a crucial requirement in neuroscience. Gaussian processes, by contrast, naturally encode spatial dependencies and provide a probabilistic framework for uncertainty quantification, enhancing the robustness of predictions (Robert et al., 1999; Welling and Teh, 2011). The combination of these methods in hybrid frameworks has the potential to offer both the flexibility needed to model non-linear associations and the spatial coherence necessary for biologically

meaningful insights (Scheinost et al., 2019). These approaches are particularly valuable in applications like brain age prediction using voxel-based morphometry (VBM), where they can improve predictive accuracy while illuminating the spatial patterns underlying neurobiological ageing. By bridging computational power with biological interpretability, machine learning continues to pave the way for more reliable and nuanced analyses of neuroimaging data.

Beyond imaging data, integrating non-imaging covariates such as socio-economic status, education, and lifestyle factors is crucial for a comprehensive understanding of brain structure and function. These variables profoundly influence neurobiological processes through mechanisms like chronic stress, healthcare access, and environmental exposures (Klee et al., 2023; Cohen et al., 2024; Röhr et al., 2024). For example, socio-economic deprivation accelerates brain ageing and reduces cortical thickness, while education fosters neuroplasticity, mitigating cognitive decline (Marzola et al., 2023; Cai et al., 2014). Lifestyle factors such as alcohol consumption and smoking further shape brain structure, with alcohol linked to decreased grey matter volume and dementia risk, and smoking associated with cortical thinning and impaired cognition (Spindler et al., 2021; Daviet et al., 2022; Karama et al., 2015; Khlif et al., 2022). Despite their importance, traditional models often treat these covariates as static, neglecting their dynamic and interactive effects on neural processes. Unified frameworks that explicitly incorporate these interactions offer a nuanced understanding of individual differences in brain outcomes, particularly in ageing or disadvantaged populations.

Multi-modal imaging provides complementary insights into brain organisation and function. Structural MRI (sMRI) reveals anatomical integrity, such as cortical thickness and grey matter density, while resting-state fMRI (rs-fMRI) captures intrinsic functional connectivity, including networks like the Default Mode Network (DMN). Various studies demonstrate that integrating complementary modalities, such as anatomy and functional connectivity, can significantly improve prediction accuracy for outcomes like brain age, cognitive performance and disease progression by identifying modality-specific contributions and their interactions (Jirsaraie et al., 2023; Niu et al., 2020; Young et al., 2013; Liem et al., 2017b; Ajith et al., 2024). These integrated approaches provide a holistic perspective on brain mechanisms, offering deeper insights into their pathways and potential for personalised interventions.

The increasing complexity of neuroimaging datasets demands analytical methods that are accurate, robust, interpretable, and scalable. This thesis presents a comprehensive pipeline for neuroimaging statistical analysis, addressing challenges across the entire workflow. The proposed methods range from pre-processing time-series fMRI data, ensuring more accurate and reliable signals by correcting for temporal autocorrelation, to variable screening techniques that efficiently reduce dimensionality while retaining biologically meaningful predictors. Building on this foundation, advanced predictive models are introduced to capture non-linear relationships, integrate spatial dependencies, and quantify uncertainty, providing interpretable and reliable insights into complex brain data. By combining statistical precision with computational efficiency, this thesis offers a unified approach to neuroimaging analysis that enhances both methodological rigor and practical applicability.

Structure of the Thesis

This thesis is organised as follows, with each chapter addressing a specific aspect of neuroimaging analysis:

- **Chapter 2: Background**

  This chapter provides a comprehensive review of the field of neuroimaging, beginning with the historical development of imaging modalities and their contributions to understanding brain structure and function. The discussion progresses through major advancements, including the introduction of Magnetic Resonance Imaging (MRI) and its subsequent evolution into modalities such as structural MRI (sMRI), functional MRI (fMRI), and diffusion MRI (dMRI). Each modality is detailed in terms of its methodological underpinnings, key strengths, and limitations, alongside examples of their application in neuroscience research.

  The chapter also highlights the advent of large-scale neuroimaging datasets, such as the UK Biobank and Human Connectome Project (HCP), which have enabled population-level analyses of brain structure and function. These datasets

integrate imaging with behavioural, genetic, and socio-demographic data, presenting opportunities for comprehensive brain-behaviour studies. However, the chapter emphasises the analytical challenges posed by these datasets, including high dimensionality, spatial and temporal dependencies, and the integration of multi-modal data. This background sets the stage for the methodological contributions presented in subsequent chapters.

- **Chapter 3: Spectral Adjusted Least Square for Autocorrelated Noise**

  Temporal autocorrelation is a persistent issue in task-based fMRI analysis, where consecutive time points in fMRI data exhibit dependencies due to neural and haemodynamic processes as well as scanner noise. This chapter introduces the Spectral Adjusted Least Squares for Autocorrelated Noise (SALSA) method, developed to address these challenges and improve the validity of statistical inferences in fMRI studies.

  Methodology: SALSA employs a two-stage approach:

  1. **Global Autocorrelation Model:** A flexible, global model is used to capture long-range temporal dependencies across the entire brain, accounting for shared patterns in the data.
  2. **Voxel-Specific Adjustments:** Localised adjustments refine the autocorrelation estimates at the voxel level, addressing the residual dependencies not captured by the global model.

  SALSA also incorporates bias correction techniques to mitigate the effects of high-pass filtering and design-related artefacts. The method is benchmarked against existing approaches, such as those implemented in FSL, SPM, and AFNI, through simulations and real-data applications. Results demonstrate SALSA's ability to maintain false positive rates and improve statistical power, particularly in fast-acquisition fMRI studies.

- **Chapter 4: Bayesian High-Dimensional Variable Screening: Extensions to Posterior Mean Screening**

  The ultra-high dimensionality of neuroimaging data, where the number of predictors (p) far exceeds the number of subjects (n), necessitates effective variable

screening methods. This chapter extends the Posterior Mean Screening (PMS) framework, a Bayesian approach that incorporates spatial priors, to address key challenges in dimensionality reduction for neuroimaging datasets. While PMS foundationally employs Gaussian priors to encode spatial dependencies, our work focuses on advancing the framework to enhance its applicability and robustness.

Methodology: We extend PMS by addressing four critical aspects:

1. **Data-Guided Method:** Incorporating data-driven techniques to refine the prior covariance structure and better capture dependencies specific to the dataset.
2. **Response-Guided Method:** Integrating response-guided approaches, such as supervised principal component regression (SPCR), to improve variable selection by aligning with the outcome of interest.
3. **Robustness Improvement:** Enhancing the stability and performance of the prior covariance matrix by introducing a structural shrinkage.
4. **Smoothness Parameterisation:** Developing a principled approach for selecting and optimising the 3D spatial smoothness parameter.

Extensive simulations validate these extensions, demonstrating improved performance over existing methods like MUA, Ridge, HOLP and spatial PMS. Applications to real neuroimaging datasets further highlight the extended PMS framework's ability to identify relevant predictors in ultra-high-dimensional settings, facilitating interpretable and efficient downstream analyses.

- **Chapter 5: Introduction to Bayesian Gaussian Process Neural Network**

  Scalar-on-image regression tasks, such as brain age prediction, require models capable of capturing non-linear and spatial relationships while maintaining interpretability. This chapter introduces the Gaussian Process Neural Network (GPNN), a hybrid framework that combines the non-linear modelling capacity of neural networks with the spatial coherence of Gaussian processes. By integrating these strengths, GPNN provides a robust and biologically informed approach to scalar-on-image regression.

Methodology: The GPNN framework is structured as follows:

1. **Input Layer:** The architecture begins with an input layer corresponding to the entire 3D brain image, ensuring biological plausibility and alignment with neuroanatomical structures.
2. **Non-Linear Activation:** Between layers, non-linear activation functions capture complex voxel-level associations with scalar outcomes, addressing the inherent non-linearity of neuroimaging data.
3. **Gaussian Process Priors:** Priors are imposed on network weights to encode spatial dependencies, ensuring that spatially coherent predictions align with known brain organisation.
4. **Uncertainty Quantification:** Bayesian updates are incorporated to generate uncertainty estimates alongside predictions, enhancing reliability in both clinical and research contexts.

The GPNN framework is evaluated using T1-weighted voxel-based morphometry data from the UK Biobank, where it demonstrates robust performance in predicting brain age. Its predictions are accompanied by interpretable spatial patterns that offer insights into neurobiological ageing. To establish its superiority, GPNN is compared against traditional methods such as LASSO, Ridge regression, and Gaussian Process Regression (GPR). The results highlight GPNN's ability to outperform these methods in predictive accuracy, while also offering enhanced interpretability and uncertainty quantification, making it a versatile tool for scalar-on-image regression tasks.

- **Chapter 6: Extensions to Gaussian Process Neural Network**

  Building on the GPNN framework, this chapter incorporates socio-demographic covariates, latent subgroup structures, and multi-modal imaging data to provide a unified approach for modelling brain-behaviour interactions. These extensions aim to capture the dynamic interplay between biological and non-biological factors, enhancing predictive accuracy, interpretability, and uncertainty quantification.

  Methodology:

1. **Integration of Socio-Demographic Covariates:** Variables such as socio-economic status and education level are incorporated to model their direct effects and interactions with brain structure and function, capturing the influence of systemic inequalities and behavioural factors on neuroimaging outcomes.
2. **Latent Subgroup Identification:** A Bayesian framework identifies subgroups within the population, enabling region-specific and subgroup-specific analyses of brain-behaviour relationships. This approach also improves uncertainty quantification, yielding narrower credible intervals and lower overall uncertainty. Additionally, it allows for class-specific uncertainty quantification, offering deeper insights into how uncertainty varies across distinct latent subgroups.
3. **Multi-Modal Imaging:** A bimodal variant of GPNN integrates structural MRI (sMRI) and resting-state fMRI (rs-fMRI), leveraging the complementary strengths of anatomical and functional imaging. This integration enhances predictive performance by capturing distinct but related aspects of brain organisation, improving predictions of brain age and cognitive performance.
4. **Brain-Mapped Saliency Analysis:** Saliency maps not only reveal the contributions of individual imaging modalities but also highlight interaction effects, such as those between imaging modalities and specific non-imaging covariates (e.g. socio-demographic factors), latent classes, or other imaging modalities. These analyses provide deeper insights into how imaging features and external factors jointly contribute to predictions, enriching the interpretability of the results.

The model is evaluated using data from the UK Biobank, demonstrating its capability to handle complex neuroimaging datasets. By combining advanced modelling techniques with interaction-focused saliency analysis and enhanced uncertainty quantification, the extended framework provides a robust foundation for personalised neuroimaging studies. It offers a deeper understanding of brain-behaviour dynamics while maintaining practical relevance through improved predictive reliability.

- **Chapter 7: Conclusion**

  This final chapter synthesises the contributions of this thesis, which spans methodological advancements across the neuroimaging pipeline, from time-series pre-processing to variable screening, predictive modelling, and uncertainty quantification. The work bridges the gap between traditional statistical methods and modern machine learning frameworks, addressing key challenges in high-dimensional neuroimaging data analysis.

Significance of the Thesis

This thesis addresses critical methodological challenges in neuroimaging analysis by developing a unified and comprehensive pipeline that integrates traditional statistical methods with modern machine learning frameworks. Through a combination of innovations in temporal modelling, advanced variable screening techniques, hybrid predictive modelling, and the integration of multi-modal imaging and non-imaging data, this work significantly enhances the ability to analyse and interpret high-dimensional neuroimaging datasets. The proposed methodologies ensure greater precision, scalability, and interpretability by capturing complex spatial and temporal dependencies and incorporating robust uncertainty quantification.

The key contributions of this thesis include:

1. **Time-Series fMRI Pre-Processing:** Developed a robust pipeline of whitening techniques to address temporal autocorrelation in task-based fMRI data, improving statistical validity to ensure more reliable inferences from task-based fMRI data.

2. **Dimensionality Reduction:** Extended the Posterior Mean Screening framework to incorporate data-driven, response-guided, and robust approaches, enhancing the efficiency and interpretability of variable screening in ultra-high-dimensional neuroimaging datasets.

3. **Predictive Modelling:** Introduced Gaussian Process Neural Networks, a hybrid framework capable of capturing non-linear relationships and spatial dependencies

while providing uncertainty quantification, advancing scalar-on-image regression tasks such as brain age prediction.

4. **Multi-Modal and Covariate Integration:** Enhanced GPNNs to integrate multi-modal imaging data, socio-demographic covariates, and latent subgroups, enabling exploration of dynamic brain-behaviour interactions and improving prediction accuracy and interpretability.

These contributions not only push the boundaries of neuroimaging analysis but also provide transformative tools for addressing longstanding methodological gaps. By combining statistical precision with computational innovation, this thesis lays a robust foundation for future studies, enabling deeper insights into the intricate relationships between brain structure, function, and behaviour, and paving the way for meaningful progress in neuroscience.

Chapter 2

Background

2.1 Neuroimaging

This section provides an overview of Magnetic Resonance Imaging (MRI) within the context of neuroscience. Section 2.1.1 traces the history of neuroscience, highlighting key advancements that ultimately led to the development of MRI as a pivotal tool in brain research. Section 2.1.2 discusses the primary MRI modalities commonly used to capture diverse aspects of brain structure and function. Finally, Section 2.1.3 details essential pre-processing steps required for MRI analyses, along with FMRIB Software Library (FSL) tools that facilitate these procedures.

2.1.1 Magnetic Resonance Imaging Background

2.1.1.1 The Study of Brain Function

The human brain is an exceptionally complex organ, consisting of approximately 86 billion neurons (Azevedo et al., 2009) interconnected by trillions of synapses, forming intricate neural networks. These networks underlie the myriad cognitive, motor, and sensory processes that define human behaviour and perception. The field of neuroscience is dedicated to understanding these brain functions, but progress in this domain has largely been driven by technological advancements that have enabled the direct study of brain activity and structure.

In modern neuroscience, the brain is studied at multiple levels, ranging from the molecular and cellular to the systems level. Historically, anatomical and physiological approaches were the primary means of exploring the brain. However, recent advances in neuroimaging have shifted the focus towards non-invasive methods such as Magnetic Resonance Imaging (MRI) that can provide both structural and functional insights. By combining multiple imaging modalities, such as structural MRI (sMRI) and functional MRI (fMRI), researchers can investigate how different brain regions are interconnected and how their activity patterns correlate with various cognitive functions. These tools have greatly expanded our ability to study brain-behaviour relationships, allowing for more detailed mapping of neural circuits and their roles in perception, memory, and decision-making.

2.1.1.2 Neuroimaging: Early Developments and the Role of MRI

The development of neuroimaging techniques has revolutionised the study of the brain (see Figure 2.1 for examples). Early methods, such as electroencephalography (EEG) in the 1920s (Berger, 1933), provided the first glimpse into electrical brain activity, offering a real-time window into neural oscillations. While EEG remains valuable for studying neural dynamics, its spatial resolution is limited, making it challenging to map brain structure in detail.

Subsequent advances in imaging led to the development of techniques like Computerised Axial Tomography (CAT) scans in the 1960s (Bhattacharyya, 2016), which allowed non-invasive imaging of the brain's internal anatomy through X-ray measurements. Shortly thereafter, Positron Emission Tomography (PET) was introduced, allowing scientists to observe brain function by monitoring radioactive tracers that reflect metabolic activity (Carlson and Birkett, 2021). While PET scans represented a significant advancement by generating functional brain images, they were invasive and had relatively limited spatial resolution compared to more contemporary imaging methods.

The introduction of Magnetic Resonance Imaging in the 1970s marked a pivotal moment in neuroimaging. MRI, based on the interaction of magnetic fields and radio waves with atomic nuclei, enabled high-resolution imaging of brain structures without the need for ionising radiation. This non-invasive technique rapidly gained popularity

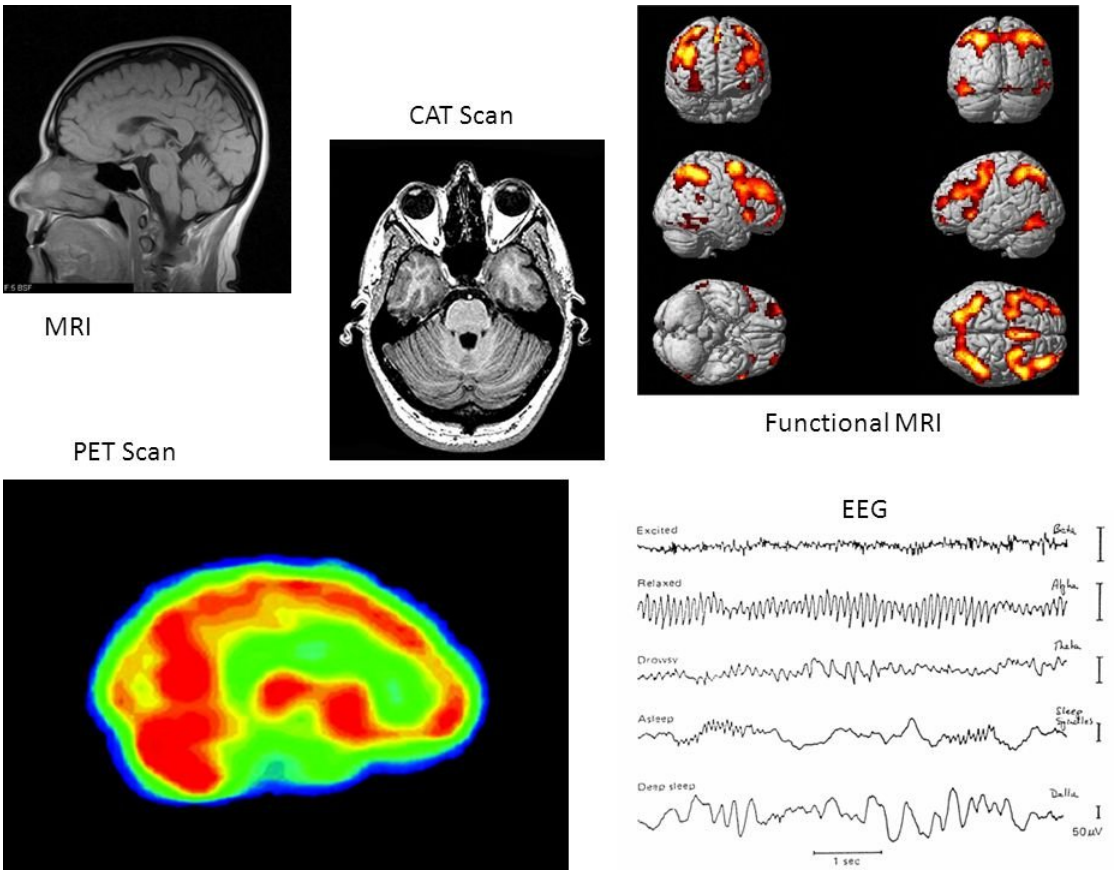


Figure 2.1: Examples of brain data acquired using various medical imaging techniques. Top (from left to right): MRI (detailed structural images based on tissue relaxation properties), CAT Scan (cross-sectional images using X-ray attenuation), and Functional MRI (dynamic brain activity via BOLD signal). Bottom (from left to right): PET Scan (metabolic activity using radioactive tracers) and EEG (electrical brain activity measured from the scalp). Images from @Psych_Review, `https://x.com/Psych_Review/status/1391824948285644800/photo/1`.

for anatomical studies, as it could provide detailed images of brain tissues, allowing for the detection of structural abnormalities in neurological conditions such as stroke, brain tumours, and degenerative diseases.

2.1.1.3 Functional MRI and the BOLD Signal

In 1991, the development of functional MRI by (Belliveau et al., 1991) introduced a new era of brain imaging, allowing researchers to investigate brain function dynamically. Functional MRI is built upon the principle of Blood Oxygen Level Dependent (BOLD) contrast, which measures changes in blood oxygenation levels as a proxy for neural

activity. When a brain region is more active, it consumes more oxygen, leading to an increase in local blood flow, which can be detected by the MRI scanner. This allows for real-time mapping of brain activity during various cognitive tasks, sensory stimuli, or even during rest.

The BOLD signal has since become the standard method for functional neuroimaging, enabling the identification of brain regions involved in processes such as vision, language, memory, and decision-making (Glover, 2011). By comparing the activity in different brain regions during task-based fMRI (t-fMRI) experiments, researchers can isolate the neural correlates of specific cognitive functions. For instance, the activation of motor regions during movement tasks or the engagement of memory-related structures during recall tasks has been widely studied using this technique.

In addition to task-based studies, resting-state fMRI (rs-fMRI) has emerged as a powerful tool for examining intrinsic brain networks. Resting-state networks, such as the default mode network (DMN), have been identified as key players in maintaining brain function during periods of rest (Smith et al., 2009). These networks represent coordinated patterns of activity across distant brain regions, suggesting that even in the absence of explicit tasks, the brain remains active in organising and maintaining its internal functional architecture.

2.1.1.4 Technological Advances and the Big Data Era of Neuroimaging

Recent advances in imaging technology and data acquisition have dramatically expanded the scope of neuroimaging research. Improvements in scanner resolution, data storage, and computational power have enabled the collection of increasingly large and detailed datasets, often referred to as "big data" in neuroscience (Bzdok et al., 2019). Large-scale projects like the UK Biobank (Allen et al., 2012), Human Connectome Project (HCP) (Van Essen et al., 2013a), and the Adolescent Brain Cognitive Development study (Casey et al., 2018) have amassed neuroimaging data from tens of thousands of participants, providing researchers with unprecedented opportunities to study brain structure and function across diverse populations.

These datasets have been instrumental in advancing our understanding of brain connectivity and in identifying biomarkers for neurological and psychiatric conditions. For example, the HCP has mapped the intricate networks of the brain using both

structural and functional imaging, providing insights into how different regions are connected and how these connections influence behaviour and cognition (Van Essen et al., 2013a). Similarly, the UK Biobank offers comprehensive imaging and phenotypic data that have been used to explore the genetic and environmental influences on brain health across the lifespan (Miller et al., 2016).

However, the rise of big data in neuroimaging has also introduced new challenges. Analysing such vast amounts of data requires advanced computational methods and statistical techniques to handle variability, noise, and complexity. Researchers are increasingly relying on machine learning and artificial intelligence to extract meaningful patterns from neuroimaging data, while ensuring that findings are reproducible and robust.

2.1.2 MRI Modelling and Modalities

MRI offers a versatile platform for investigating the brain, with different modalities tailored to uncover unique aspects of its structure and function. While structural MRI provides a detailed static snapshot of brain anatomy, fMRI captures dynamic changes linked to neural activity, primarily through the BOLD signal. These modalities vary in the dimensionality of the data they generate, ranging from static 3D images to dynamic 4D datasets or multi-directional information, as illustrated in Figure 2.2. Understanding the principles of BOLD imaging and its physiological underpinnings lays the groundwork for exploring these modalities.

2.1.2.1 BOLD fMRI and the Haemodynamic Response Function

Functional MRI primarily relies on the BOLD signal, which reflects the interplay between neuronal activity and blood flow, a phenomenon known as neurovascular coupling, as illustrated in Figure 2.3. When neurons become active, cerebral blood flow (CBF) increases locally to supply the necessary oxygen and glucose. The magnetic properties of haemoglobin, which differ between its oxygenated and deoxygenated states, form the basis of BOLD imaging. Deoxygenated haemoglobin is more magnetic than its oxygenated counterpart, creating measurable inhomogeneities in the magnetic

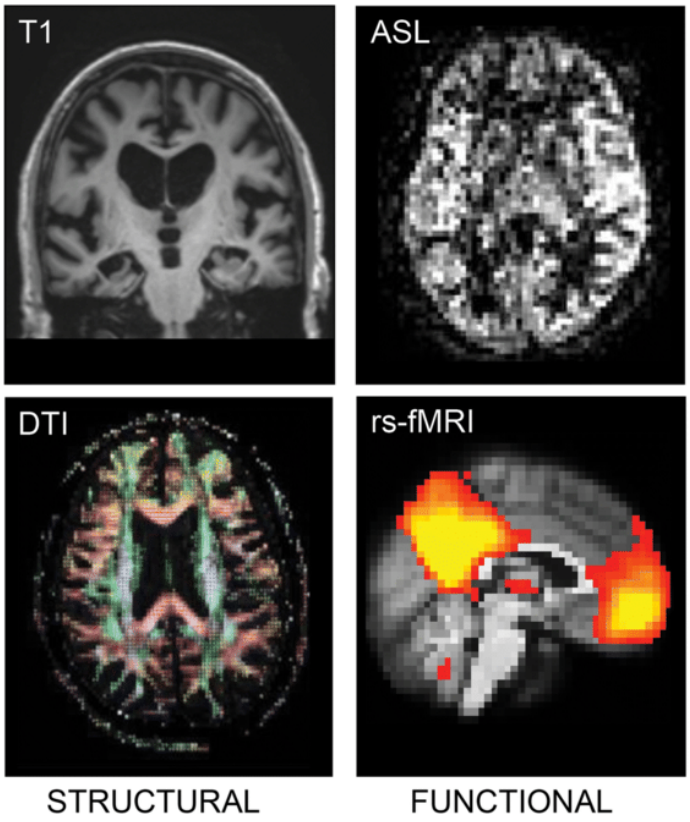


Figure 2.2: Example static slices of brain data acquired through different MRI techniques and processing methods from Ten Kate et al. (2018). Left column (Structural): T1-weighted (top, anatomical detail), DTI (bottom, white matter pathways). Right column (Functional): ASL (top, cerebral blood flow), rs-fMRI (bottom, resting-state brain activity).

field of the MRI scanner. These changes in magnetic properties, driven by shifts in oxygenated blood flow, constitute the BOLD signal.

The dynamics of the BOLD response following neuronal activation are characterised by the Haemodynamic Response Function (HRF). The HRF describes a typical sequence of events: an initial dip, a peak increase, and a gradual return to baseline. This response is modulated by physiological factors, including CBF, cerebral blood volume (CBV), and the cerebral metabolic rate of oxygen consumption (CMRO2). While the BOLD signal is an indirect proxy for neuronal activity, it remains a cornerstone of functional neuroimaging, enabling non-invasive measurement of brain activity.

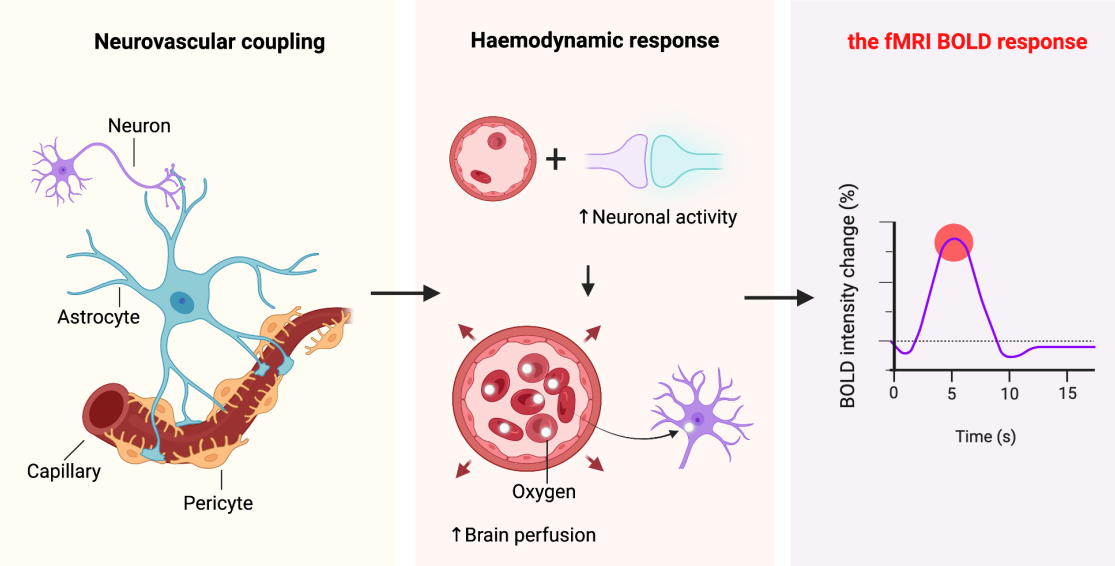


Figure 2.3: Illustration of the progression from neurovascular coupling to the fMRI BOLD response from Vasilkovska (2023). Left: Neurovascular coupling, where neuronal activation triggers astrocytes and pericytes to mediate changes in the capillary bed, increasing cerebral blood flow. Centre: The haemodynamic response, characterised by an increase in oxygenated blood supply and brain perfusion, following neuronal activity. Right: The resultant fMRI BOLD response, showing the percentage change in signal intensity over time, driven by variations in deoxygenated haemoglobin and depicted as a haemodynamic response function (HRF).

2.1.2.2 T1-Weighted MRI (Structural MRI)

T1-weighted MRI, commonly known as structural MRI, provides high-resolution, 3D volumetric images of the brain's anatomy by capturing variations in the longitudinal relaxation time (T1) of different tissues. This modality enables clear visualisation of grey matter, white matter, and cerebrospinal fluid, offering a detailed, static snapshot of brain structure at a single time point, without any temporal dimension. Structural MRI is crucial for studying brain morphology, including the measurement of cortical thickness and the identification of structural abnormalities such as atrophy related to neurodegenerative diseases.

A key application of T1-weighted images in this thesis is voxel-based morphometry (VBM), a whole-brain, voxel-wise technique used to assess anatomical differences across individuals, focusing on grey matter. VBM involves tissue segmentation and spatial normalisation, with each voxel reflecting either local grey matter concentration or volume adjusted by the deformation field. Values near 1 indicate minimal deformation, while deviations reflect local expansion or contraction. These VBM-derived maps form the basis for statistical analyses of group differences.

2.1.2.3 Resting-State fMRI (rs-fMRI)

Resting-state fMRI examines the spontaneous fluctuations in the BOLD signal when the subject is at rest, providing insights into intrinsic brain activity without any explicit task. These fluctuations are thought to reflect functional connectivity between spatially distinct regions of the brain that operate together as networks. Resting-state fMRI produces 4D data (3D spatial dimensions + 1 temporal dimension), capturing both spatial and temporal information, and is especially valuable for studying brain networks such as the default mode network and alterations in connectivity in psychiatric and neurological conditions.

2.1.2.4 Task-Based fMRI (t-fMRI)

In task-based fMRI, brain activity is measured while the subject performs a specific task, such as a cognitive or motor task. The goal of tfMRI is to link changes in the BOLD signal to task-induced neural processes, allowing for the identification of

brain regions responsible for particular functions. Like rs-fMRI, t-fMRI generates 4D data by collecting a series of 3D images over time. The analysis typically involves comparing the BOLD signal during task conditions to a baseline or control condition, using the HRF to model the expected neural response to task-related stimuli.

2.1.2.5 Diffusion MRI (dMRI)

Diffusion MRI (dMRI) is a modality that measures the diffusion of water molecules within brain tissue, providing insights into the microstructure of white matter tracts. This technique generates multi-directional 3D data by mapping how water diffuses in different directions at each voxel. Through the use of diffusion tensor imaging (DTI), dMRI enables the reconstruction of white matter pathways, allowing researchers to investigate structural connectivity and detect changes related to brain injuries or degenerative conditions. dMRI is particularly valuable for understanding the brain's wiring and how it may be affected by various neurological disorders.

2.1.3 Pre-processing Pipeline

Pre-processing is an essential step in the analysis of neuroimaging data, ensuring that the data are clean, aligned, and standardised before further analysis. In fMRI analysis, pre-processing typically includes several key steps such as brain extraction, distortion correction, slice timing correction, realignment, co-registration, spatial normalisation, spatial smoothing, and temporal filtering. These procedures are essential for minimising artefacts, ensuring anatomical consistency, and optimising the data for subsequent statistical analyses.

2.1.3.1 Brain Extraction

Brain extraction is a vital preprocessing step in both structural and functional MRI data analysis, designed to isolate brain tissue by removing non-brain elements such as the skull, scalp, and surrounding tissues, illustrated in Figure 2.4. The goal of this procedure is to ensure that subsequent steps, such as registration, segmentation, and spatial normalisation, focus exclusively on brain structures, minimising the influence of extraneous tissues. This step is especially important for enhancing the accuracy of

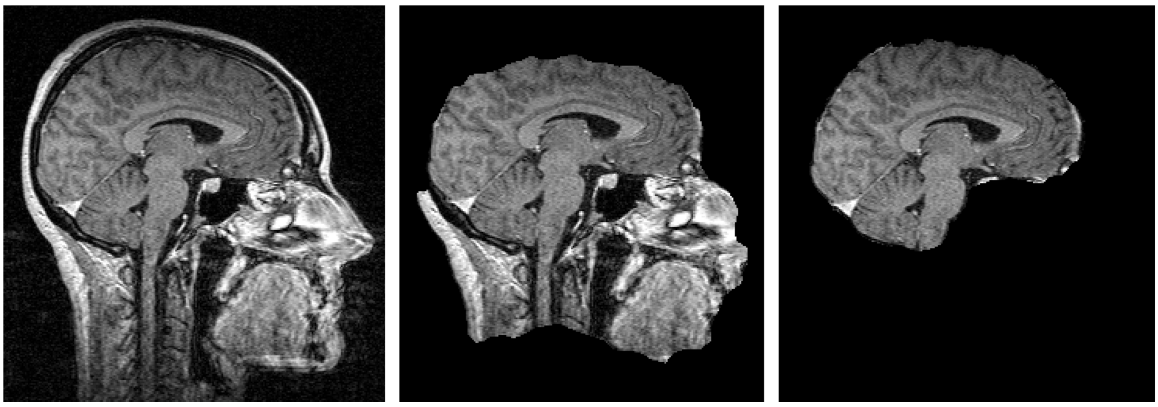


Figure 2.4: Illustration of the brain extraction process, isolating brain tissue by removing non-brain elements such as the skull, scalp, and surrounding tissues from *FSL Course Online Materials (2024)*. From left to right: the original structural MRI scan, intermediate stage during extraction, and the final brain-extracted image.

coregistration and spatial alignment processes subsequently. For instance, differences in non-brain structures can lead to misalignments during image registration, ultimately degrading the quality of downstream analyses.

A commonly used tool for this purpose is the Brain Extraction Tool (BET) from FSL, which applies a deformable surface model to fit the brain's boundaries and segment brain tissue from the surrounding areas. By generating a brain-extracted image, BET ensures that analyses focus only on brain regions, making it a key initial step in standard fMRI preprocessing pipelines.

2.1.3.2 Distortion Correction

Distortion correction is necessary to mitigate the spatial distortions caused by gradient non-linearities and magnetic field inhomogeneities during MRI acquisition. These distortions can significantly affect both structural and functional MRI data, particularly in regions near air-tissue interfaces, such as the frontal and temporal lobes, where magnetic susceptibility differences are more pronounced. If left uncorrected, these distortions can result in inaccurate anatomical representation, leading to errors in spatial alignment and functional localisation.

Distortion correction typically relies on a field map, which captures the spatial variation of the magnetic field during scanning. The field map enables software tools to estimate and correct the geometric distortions in the images by applying spatial

transformations to unwarp the data. Overall, this correction step is crucial for ensuring that the MRI data faithfully represents brain anatomy, which is essential for accurate image registration and subsequent analyses.

2.1.3.3 Motion Correction

Motion correction, often referred to as realignment, is a critical preprocessing step that compensates for the inevitable head movements that occur during an fMRI scan. Even subtle movements can cause significant artefacts, leading to the mixing of signals between different tissue types (such as grey matter and white matter) or complete signal loss in edge regions of the brain. Head motion can severely compromise the quality of the data, generating false activations in task-related fMRI analyses and masking true neural signals.

The motion correction process involves rigid-body transformations that adjust each volume in the fMRI time series to align with a reference volume, which is typically the first volume or a mean image of the entire time series. The rigid-body transformation consists of translations and rotations applied to ensure that each voxel corresponds to the same anatomical location across all time points. Tools like SPM and FSL offer distinct optimisation algorithms for this purpose—SPM uses a least-squares cost function, whereas FSL employs normalised correlation to optimise the alignment of brain volumes. This step significantly reduces motion-related noise and artefacts, enabling more accurate statistical analyses in subsequent stages of fMRI processing.

2.1.3.4 Spatial Normalisation or Registration

Spatial normalisation is the process of aligning individual brain images to a common anatomical space, such as the MNI152 template. The MNI152 template is a widely-used standard in neuroimaging, created by averaging brain scans from 152 healthy adults at the Montreal Neurological Institute (MNI), providing a consistent reference for comparison across individuals.

Within-subject alignment, such as aligning functional images to structural scans of the same subject, typically requires only rigid-body transformations, which account for translations and rotations without distorting the shape of the brain. FSL's FLIRT (FMRIB's Linear Image Registration Tool) is commonly used for linear registration.

However, between-subject alignment involves registering brain images from different individuals into a shared space, non-linear transformations are necessary to accurately match detailed anatomical structures. Tools like FSL's FNIRT (FMRIB's Non-linear Image Registration Tool) enable voxel-wise deformations, capturing local anatomical differences that cannot be resolved by affine methods alone. This step is crucial in group-level analyses, ensuring that corresponding brain regions across subjects are aligned to the same anatomical location.

2.1.3.5 Spatial Smoothing

Spatial smoothing is a key preprocessing step in neuroimaging, applied to reduce noise and increase the signal-to-noise ratio (SNR) in MRI data. This process is typically achieved by convolving the brain images with a 3D Gaussian kernel. Smoothing serves multiple purposes, such as enhancing the detection of true signals by reducing high-frequency noise, and improving statistical power in group analyses by compensating for anatomical variability between subjects.

The degree of smoothing is determined by the full-width at half-maximum (FWHM) of the Gaussian kernel, which defines the width of the kernel at half of its maximum height. FWHM is a standard measure in neuroimaging for controlling the amount of smoothing applied. The relationship between FWHM and the standard deviation (σ) of the Gaussian kernel is given by:

$$\text{FWHM} = 2\sqrt{2\ln 2} \cdot \sigma$$

where σ represents the spread of the Gaussian kernel. For practical applications, the smoothing parameter is often specified in millimeters to correspond to real-world spatial dimensions. For example, if each voxel is a 2 mm cube, a smoothing kernel with $\sigma = 2$ voxels translates to a 4.71 mm FWHM. The proper interpretation of these parameters is essential to ensure that spatial smoothing is applied consistently across analyses.

Spatial smoothing helps fulfil the assumptions required for statistical parametric mapping, where spatial correlation among neighbouring voxels is an expected feature of the data. By averaging voxel intensities within a specified spatial neighbourhood,

the method reduces the risk of overfitting and improves the detection of functionally relevant brain regions.

2.1.3.6 Temporal Filtering

Temporal filtering is another essential preprocessing step for functional MRI data, aimed at isolating the neural signals of interest from unwanted noise. Functional MRI time series often contain low-frequency drifts due to scanner instability, as well as high-frequency noise from physiological sources like respiration and cardiac cycles. These artefacts can obscure the true BOLD response, making it difficult to interpret neural activity patterns.

A band-pass filter is commonly applied to remove frequencies outside the expected range of the BOLD signal. Typically, a high-pass filter is used to eliminate slow drifts below 0.01 Hz, while a low-pass filter may be applied to suppress high-frequency noise above 0.2-0.3 Hz. Temporal filtering thus enhances the quality of the fMRI data by attenuating unwanted artefacts and preserving the frequency components most relevant to neural activity. In some cases, temporal derivatives of motion and other nuisance regressors are also included to better account for artefacts that cannot be addressed by filtering alone.

2.2 Statistical Modelling

This section presents the core statistical methodologies employed in this study, including an overview of their foundational assumptions. Section 2.2.1 introduces key concepts in linear modelling, covering the basic linear model for assessing linear relationships between response and predictor variables, as well as generalised linear regression for complex response variables and interaction effects within linear models. Section 2.2.2 provides essential background on Bayesian statistics, covering posterior predictive distributions, Markov Chain Monte Carlo methods, and Gaussian Processes—fundamental to Chapters 5 and 6 of this thesis. Section 2.2.3 then explores neural networks, another primary analytical tool for Chapters 5 and 6. Section 2.2.4 details feature extraction and variable screening methods, primarily used in Chapter

4. Finally, Section 2.2.5 covers the theory and methodology of time series modelling, laying the foundation for Chapter 3.

2.2.1 Regression

2.2.1.1 Linear Model

Linear regression is one of the fundamental techniques for statistical analysis, particularly useful in applied fields like neuroimaging. Suppose we have data consisting of n samples and p predictors. The goal is to model the relationship between the response variable $\mathbf{y} \in \mathbb{R}^n$ and the predictors represented by the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. The unknown parameters to be estimated are the coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, with an additional error term $\boldsymbol{\epsilon}$ that represents random noise in the data. This noise, $\boldsymbol{\epsilon}$, is assumed to be independent of the data, and each element ϵ_i, for $i = 1, \ldots, n$, is *independently and identically distributed* (iid) according to a normal distribution with mean zero and constant variance σ^2, i.e., $\epsilon_i \sim N(0, \sigma^2)$.

The linear model can be expressed as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$$

where $\mathbf{y}$ is the response vector, $\mathbf{X}$ is the design matrix of predictors, $\boldsymbol{\beta}$ represents the coefficients to be estimated, $\boldsymbol{\epsilon}$ is the error term and $\mathbf{I}_n$ is an identity matrix of size n.

To ensure that the relationship between the predictors and the response is properly captured, the following assumptions are made:

1. **Linearity:** The model assumes that the relationship between the predictors and the response is linear. The response y is modelled as a linear function of the predictors $x_1, x_2, \ldots, x_p$:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

   where ϵ is the random error term.

2. **Independence:** The observations are assumed to be independent of one another. This means that the error terms for different observations, ϵ_i and ϵ_j, are uncorrelated:

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \text{for } i \neq j$$

3. **Homoscedasticity (Constant Variance of Errors):** The variance of the error terms is assumed to be constant across all values of the predictors:

$$\text{Var}(\epsilon_i) = \sigma^2 \quad \forall\ i$$

   If this assumption is violated, heteroscedasticity occurs, leading to inefficient estimates and unreliable inferences.

4. **Normality of Errors:** The error terms are assumed to follow a normal distribution with mean zero and variance σ^2:

$$\epsilon_i \sim N(0, \sigma^2)$$

   This assumption is crucial for making valid statistical inferences, such as hypothesis testing and constructing confidence intervals.

5. **No Perfect Multicollinearity:** The independent variables should not exhibit perfect collinearity, meaning that no predictor is an exact linear combination of the others. If multicollinearity exists, the matrix $\mathbf{X}^\top\mathbf{X}$ becomes singular, preventing the estimation of the coefficients. Formally:

$$\det(\mathbf{X}^\top\mathbf{X}) \neq 0$$

In practice, two common methods for estimating the coefficients $\boldsymbol{\beta}$ are *Maximum Likelihood Estimation* (MLE) and *Ordinary Least Squares* (OLS), both of which can be derived analytically or solved numerically using techniques like gradient descent.

Maximum Likelihood Estimation (MLE) is a fundamental approach for esti-

mating the parameters of the linear model. Given the linear model:

$$\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

we seek to find the parameter estimates that maximise the likelihood of observing the data. The likelihood function is expressed as:

$$L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right).$$

Taking the logarithm of the likelihood function, we obtain the log-likelihood:

$$\ell(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2.$$

To find the maximum likelihood estimate for $\boldsymbol{\beta}$, we differentiate the log-likelihood with respect to $\boldsymbol{\beta}$:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0,$$

which gives the MLE estimator:

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}.$$

Ordinary Least Squares (OLS) estimates $\boldsymbol{\beta}$ by minimising the sum of squared residuals:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Taking the derivative with respect to $\boldsymbol{\beta}$ and setting it equal to zero:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0,$$

we derive the OLS estimator:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}.$$

These analytical calculations demonstrate that in a simple linear regression model, the coefficient estimates for $\boldsymbol{\beta}$ are identical whether derived through MLE or OLS. A key requirement in both derivations is that $n \geq p$, ensuring that $(\mathbf{X}^\top \mathbf{X})$ is of full rank and invertible. However, in neuroimaging analysis, it is common for $p \gg n$, a scenario in which regularisation techniques like ridge regression and LASSO are often employed.

Ridge regression addresses this issue by adding an L_2 penalty to the sum of squared residuals. The objective function is:

$$S_{\text{ridge}}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta},$$

where $\lambda \geq 0$ is the regularisation parameter. The ridge regression estimate for $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}.$$

This regularisation helps control overfitting and multicollinearity by shrinking the coefficients.

Least Absolute Shrinkage and Selection Operator (LASSO) applies an L_1 penalty to the residual sum of squares:

$$S_{\text{LASSO}}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1,$$

where $\|\boldsymbol{\beta}\|_1$ is the sum of the absolute values of the coefficients. Unlike Ridge, the LASSO estimate for $\boldsymbol{\beta}$ is typically solved numerically, and the L_1 penalty encourages sparsity in the coefficients, leading to some being exactly zero.

2.2.1.2 Generalised Linear Model (GLM)

While the linear model assumes a linear relationship between the predictors $\mathbf{X}$ and the response $\mathbf{y}$, it is limited in its ability to model data where the response variable does not follow a normal distribution or when the response is non-continuous. The Generalised Linear Model (GLM) extends the linear model by allowing for response

variables that follow any distribution from the exponential family (e.g. binomial, Poisson, etc.).

The motivation for using a GLM is to handle cases where the linear model's assumptions (e.g. normally distributed residuals or continuous response) are not applicable. In a GLM, we model the response variable $\mathbf{y}$ through a *link function* $g(\cdot)$, which relates the expected value of the response $\mathbb{E}[\mathbf{y}]$ to the linear predictors $\mathbf{X}\boldsymbol{\beta}$:

$$g(\mathbb{E}[\mathbf{y}]) = \mathbf{X}\boldsymbol{\beta}.$$

This generalisation allows for more flexible modeling of different types of response variables, such as binary or categorical outcomes.

One common application of a GLM is the **Logistic Regression** in binary classification problem, where the response $\mathbf{y}$ takes values in $\{0, 1\}$, as in logistic regression. Here, the probability of the response being 1 is modelled using the logistic (sigmoid) function as the link function:

$$\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})}$$

This can be written as:

$$\log\left(\frac{\mathbb{P}(y_i = 1 \mid \mathbf{x}_i)}{1 - \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

This is the *log-odds* (logit) of the probability that $y_i = 1$, and the objective is to estimate $\boldsymbol{\beta}$, typically through MLE, by maximising the likelihood of the observed data under the logistic model.

For categorical outcomes with more than two categories, logistic regression can be generalised to **multinomial logistic regression**. Suppose the response $\mathbf{y}$ takes values in $\{1, 2, \ldots, K\}$, where $K > 2$. The GLM framework generalises the binary logistic model to model the probability of each category as:

$$\mathbb{P}(y_i = k \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_k)}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j)}, \quad k = 1, \ldots, K-1$$

The final category's probability is determined as:

$$\mathbb{P}(y_i = K \mid \mathbf{x}_i) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j)}$$

Here, the goal is to estimate $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}$ using methods like MLE.

2.2.1.3 Interaction Terms in Regression

In many real-world scenarios, the effect of one predictor on the outcome may depend on the value of another predictor. In such cases, adding *interaction terms* to a regression model allows for more flexibility and better captures the underlying relationships between variables. Interaction terms are especially useful when the combined effect of two variables is not simply the sum of their individual effects but is influenced by their interplay. For example, in a study investigating the effect of medication and exercise on blood pressure, the effect of exercise on blood pressure may depend on whether a patient is taking medication. Without interaction terms, the regression model would assume the effect of exercise is the same regardless of medication use, which might not reflect the reality.

A simple demonstration is to consider a typical linear regression model with two predictors, the relationship between the predictors and the outcome is expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where y is the response variable, x_1 and x_2 are predictors, and β_1, β_2 are their respective coefficients.

To account for interactions between x_1 and x_2, we introduce an interaction term $x_1 x_2$, resulting in the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2) + \epsilon$$

Here, β_3 represents the interaction effect between x_1 and x_2. If β_3 is significantly different from zero, it indicates that the effect of one predictor (e.g. x_1) on y depends on the level of the other predictor (e.g. x_2). Interaction terms allow the model to account for this non-additive relationship between predictors.

Interaction effects can be diagnosed through visualisation techniques such as interaction plots, where the relationship between one predictor and the outcome is plotted for different levels of another predictor. These plots help identify whether the effect of one predictor varies across levels of the other.

For instance, an interaction plot for the model with two predictors x_1 and x_2 would display the predicted values of y for different levels of x_2 on the x-axis, with separate lines for different levels of x_1. If the lines are not parallel, this suggests the presence of an interaction effect.

Another useful tool is the residual plot, which can help detect interaction effects by showing patterns in the residuals. If there is an interaction that has not been accounted for in the model, the residual plot may exhibit systematic patterns, indicating a poor model fit. Incorporating the interaction term could resolve these patterns, leading to improved model performance.

2.2.2 Bayesian Statistics

Bayesian statistics provides a structured way to incorporate prior beliefs about parameters and update them based on observed data to obtain a posterior distribution. Unlike classical methods like OLS, which yield point estimates, Bayesian methods provide a full distribution over the parameters, reflecting uncertainty.

Consider the simple linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$$

In the Bayesian framework, we place a prior distribution on the parameters $\boldsymbol{\beta}$. Two common choices for the prior are:

- Uniform prior: $\boldsymbol{\beta} \sim U[a, b]$, this represents no prior information about $\boldsymbol{\beta}$.
- Gaussian prior: $\boldsymbol{\beta} \sim N(0, \tau^2 \mathbf{I}_p)$, where τ^2 is the prior variance.

The likelihood function, as in the classical linear model, is based on the assumption that the errors are normally distributed:

$$\mathbf{y} \mid \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

Using **Bayes' theorem**, the posterior distribution of $\boldsymbol{\beta}$ is proportional to the product of the likelihood and the prior:

$$p(\boldsymbol{\beta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}) \cdot p(\boldsymbol{\beta})$$

Under the linear regression setting, this posterior distribution will also be Gaussian, with the mean and variance influenced by both the likelihood (from the data) and the prior. The posterior mode (the maximum a posteriori, or MAP estimate) serves as the Bayesian point estimate of $\boldsymbol{\beta}$.

2.2.2.1 Bayesian Linear Model vs Frequentist Linear Model

If we assume a uniform (non-informative) prior on $\boldsymbol{\beta}$, the prior does not influence the posterior. In this case, the MAP estimate of $\boldsymbol{\beta}$ is equivalent to the MLE, which is the OLS estimate:

$$\hat{\boldsymbol{\beta}}_{\text{MAP}} = \hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

If we place a Gaussian prior on $\boldsymbol{\beta}$, the MAP estimate becomes:

$$\hat{\boldsymbol{\beta}}_{\text{MAP}} = (\mathbf{X}^\top \mathbf{X} + \frac{\sigma^2}{\tau^2} \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$$

This is equivalent to the Ridge regression estimator, where the penalty term λ in Ridge regression is related to the prior variance as $\lambda = \frac{\sigma^2}{\tau^2}$. Thus, the Bayesian interpretation of Ridge regression is that it corresponds to using a Gaussian prior on $\boldsymbol{\beta}$, with the strength of the regularisation controlled by the prior variance τ^2.

One of the key advantages of the Bayesian linear model is the ability to construct *credible intervals* for the parameters. In frequentist statistics, a 95% *confidence interval* for the parameter β_j is an interval $[a, b]$ constructed such that:

$$P(a \leq \hat{\beta}_j \leq b \mid H_0) = 0.95$$

Here, $\hat{\beta}_j$ is the estimator of β_j based on the data $\mathbf{y}$, and H_0 represents the assumption of repeated random sampling. This means that if we were to repeatedly sample data from the population, 95% of the intervals computed in this manner would

contain the true parameter value β_j. Importantly, the confidence interval does not provide a direct probability statement about the specific interval for the observed data.

In contrast, a 95% credible interval for the parameter β_j is an interval $[a, b]$ such that:

$$P(a \leq \beta_j \leq b \mid \mathbf{y}) = 0.95$$

This means that, given the data $\mathbf{y}$, there is a 95% probability that the true parameter β_j lies within the interval $[a, b]$. The credible interval is derived from the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$, which incorporates both the prior and the likelihood, providing a direct probabilistic interpretation of parameter uncertainty.

While credible intervals offer this direct probability interpretation, the Bayesian framework provides additional advantages when making predictions, particularly by incorporating prior knowledge and quantifying uncertainty in a more intuitive way.

Another significant advantage is the ability to compute the *posterior predictive distribution*, which allows for quantification of uncertainty in predictions for new subjects. The posterior predictive for a new data point $\mathbf{x}_{\text{new}}$ is given by:

$$p(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \mathbf{y}) = \int p(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \mathbf{y})\, d\boldsymbol{\beta}$$

This distribution takes into account both the uncertainty in the parameters $\boldsymbol{\beta}$ and the inherent variability in the data, providing a full probabilistic prediction for each new subject. This is particularly valuable in settings like neuroimaging or personalised medicine, where quantifying uncertainty in predictions is crucial for decision-making.

For example, we can assess the following three common properties of the posterior predictive distribution for a new data point $\mathbf{x}_{\text{new}}$, given the observed data $\mathbf{y}$:

- **Mean:** The mean of the posterior predictive distribution, representing the expected value of the new prediction y_{new}, is given by:

$$\mathbb{E}[y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \mathbf{y}] = \int \mathbf{x}_{\text{new}}^{\top} \boldsymbol{\beta} p(\boldsymbol{\beta} \mid \mathbf{y})\, d\boldsymbol{\beta}$$

This integrates over the posterior distribution of the model parameters $\boldsymbol{\beta}$ and provides the central prediction for y_{new}.

- **Variance:** The variance of the posterior predictive distribution reflects the uncertainty in both the data and the model parameters. It is given by:

$$\text{Var}(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \mathbf{y}) = \underbrace{\int \text{Var}(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \mathbf{y})\, d\boldsymbol{\beta}}_{\text{Data Variance}} + \underbrace{\text{Var}(\mathbf{x}_{\text{new}}^{\top} \boldsymbol{\beta} \mid \mathbf{y})}_{\text{Parameter Variance}}$$

  This combines the variance due to the inherent variability of the data and the uncertainty in the model parameters.

- **95% Posterior Predictive Interval:** The posterior predictive credible interval provides a measure of uncertainty for the new prediction y_{new}. A 95% credible interval is defined as:

$$P(a \leq y_{\text{new}} \leq b \mid \mathbf{y}, \mathbf{x}_{\text{new}}) = 0.95$$

  This means that, given the observed data $\mathbf{y}$ and the new input $\mathbf{x}_{\text{new}}$, there is a 95% probability that the predicted value y_{new} will lie within the interval $[a, b]$.

Together, these properties provide a comprehensive understanding of both the central prediction and the uncertainty around it, accounting for both the data variability and model uncertainty.

2.2.2.2 Approximate Posterior Distribution

In many practical cases, the posterior distribution may not have a closed-form solution or is computationally intractable due to the complexity of the likelihood or prior. When this occurs, we cannot directly compute the posterior distribution of the parameters. Instead, we rely on **Posterior Sampling** techniques to approximate the posterior by drawing samples from it.

The goal of posterior sampling is to generate a set of samples from the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$ that can be used to estimate the distribution itself, along with key descriptive statistics such as means, variances, and credible intervals.

In posterior sampling, we typically define two distributions:

- **Target distribution**: The posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$, which we aim to sample from.
- **Proposal distribution**: A simpler, known distribution $q(\boldsymbol{\beta})$, from which we can easily sample. This proposal distribution is used to explore the space of possible parameter values.

The idea is to use the samples from the proposal distribution to approximate the target posterior distribution.

One of the most widely used techniques for posterior sampling is **Markov Chain Monte Carlo** (MCMC). MCMC methods generate a sequence of samples that form a Markov chain, where each sample depends only on the previous one, and over time, these samples converge to the target posterior distribution. Even in high-dimensional models or models with complex priors, MCMC methods allow us to approximate the posterior distribution and compute summaries of interest, such as the posterior mean or credible intervals.

The **Metropolis-Hastings** algorithm is a foundational approach in this family, and more advanced methods, like Gibbs sampling and Hamiltonian Monte Carlo, build on these ideas to improve efficiency and convergence. The Metropolis-Hastings alogirthm works as follows:

1. **Initialise**: Start with an initial value for $\boldsymbol{\beta}_0$.
2. **Proposal**: At each step t, propose a new value $\boldsymbol{\beta}^*$ from the proposal distribution $q(\boldsymbol{\beta} \mid \boldsymbol{\beta}_t)$.
3. **Acceptance** Calculate the acceptance ratio:

$$\alpha = \min\left(1, \frac{p(\mathbf{y} \mid \boldsymbol{\beta}^*)p(\boldsymbol{\beta}^*)}{p(\mathbf{y} \mid \boldsymbol{\beta}_t)p(\boldsymbol{\beta}_t)} \cdot \frac{q(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}^*)}{q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}_t)}\right)$$

4. **Update** Accept the proposed sample $\boldsymbol{\beta}^*$ with probability α. If rejected, set $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t$.

   By iterating this process, the samples $\{\boldsymbol{\beta}_t\}$ will eventually approximate the posterior distribution.

Another foundational variant of MCMC method considered in this study is **Langevin Dynamics**. Langevin Dynamics is a method that originates from physics, where it is used to simulate the motion of particles under the influence of both deterministic forces and random noise. In the context of posterior sampling, Langevin dynamics can be used to propose new values for $\boldsymbol{\beta}$ by incorporating gradient information from the log-posterior, making it a more informed proposal strategy.

The update step in Langevin dynamics is given by:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}_t \mid \mathbf{y}) + \epsilon N(0, \mathbf{I}_p)$$

Here, $\nabla_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}_t \mid \mathbf{y})$ is the gradient of the log-posterior at the current parameter value, and ϵ is a step size. The term $\epsilon N(0, \mathbf{I}_p)$ adds random noise, ensuring that the algorithm explores the parameter space. This framework serves as the foundation for the following methods.

Metropolis-Adjusted Langevin Algorithm (MALA) combines Langevin dynamics with the Metropolis-Hastings algorithm to improve the sampling efficiency. The key idea is to use the Langevin dynamics update rule as a proposal distribution, but with a Metropolis-Hastings acceptance step to correct for any bias introduced by the approximation.

The proposal distribution in MALA is:

$$q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}_t) = N\left(\boldsymbol{\beta}_t + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}_t \mid \mathbf{y}), \epsilon^2 \mathbf{I}_p\right)$$

The acceptance probability is computed similarly to the Metropolis-Hastings algorithm:

$$\alpha = \min\left(1, \frac{p(\mathbf{y} \mid \boldsymbol{\beta}^*)p(\boldsymbol{\beta}^*)}{p(\mathbf{y} \mid \boldsymbol{\beta}_t)p(\boldsymbol{\beta}_t)} \cdot \frac{q(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}^*)}{q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}_t)}\right)$$

By using gradient information, MALA tends to converge more efficiently than traditional Metropolis-Hastings, particularly in high-dimensional spaces where the posterior landscape may be complex.

However, like other generic MCMC algorithms, when dealing with large datasets, MALA's reliance on full gradient computations for the log-posterior can be computationally prohibitive. **Stochastic Gradient Langevin Dynamics** (SGLD) addresses this issue by using stochastic approximations of the gradient, computed on randomly selected mini-batches of the data.

Let $\mathcal{D}_t$ represent a randomly sampled mini-batch of the data at iteration t. The SGLD update rule is then:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}_t \mid \mathbf{y}_{\mathcal{D}_t}) + \epsilon N(0, \mathbf{I}_p)$$

where $\nabla_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}_t \mid \mathbf{y}_{\mathcal{D}_t})$ is the gradient of the log-posterior evaluated using only the mini-batch $\mathbf{y}_{\mathcal{D}_t}$ rather than the full dataset. By randomly selecting a different mini-batch $\mathcal{D}_t$ at each iteration, SGLD introduces sufficient variability to explore the parameter space without requiring the full gradient computation at each step. Furthermore, the random noise term $\epsilon N(0, \mathbf{I}_p)$ ensures that the algorithm can escape local minima and explore the posterior distribution. Therefore, unlike MALA, SGLD lets us bypass the Metropolis-Hastings acceptance step from this sufficient variability.

MCMC provides a set of samples $\{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{\beta}^{(N)}\}$ from the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$, which can be used to approximate the descriptive statistics for the posterior predictive distribution. These approximations include the following:

- **Mean Estimate:** The mean of the posterior predictive distribution can be approximated as the sample mean of the MCMC samples:

$$\mathbb{E}[y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \mathbf{y}] \approx \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{\text{new}}^{\top}\boldsymbol{\beta}^{(i)}$$

  This provides an unbiased estimate of the true posterior predictive mean.

- **Variance Estimate:** The variance of the posterior predictive distribution can be estimated using the MCMC samples as:

$$\mathrm{Var}(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \mathbf{y}) \approx \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{x}_{\text{new}}^{\top} \boldsymbol{\beta}^{(i)} - \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_{\text{new}}^{\top} \boldsymbol{\beta}^{(j)} \right)^2$$

This captures both the variability in the data and the uncertainty in the model parameters.

- **95% Credible Interval Estimate using Empirical distribution:** The 95% credible interval can be approximated by first generating posterior predictive samples:

$$\{y_{\text{new}}^{(1)}, y_{\text{new}}^{(2)}, \ldots, y_{\text{new}}^{(N)}\} = \{\mathbf{x}_{\text{new}}^{\top} \boldsymbol{\beta}^{(1)}, \ldots, \mathbf{x}_{\text{new}}^{\top} \boldsymbol{\beta}^{(N)}\}$$

Then, order the samples and define the 95% credible interval as the 2.5th and 97.5th percentiles:

$$\mathrm{CI}_{95\%}(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \mathbf{y}) \approx \left[y_{\text{new}}^{(0.025N)}, y_{\text{new}}^{(0.975N)} \right]$$

- **95% Credible Interval using Gaussian Approximation:** Assuming the posterior predictive distribution is approximately Gaussian, the credible interval can be computed as:

$$\mathrm{CI}_{95\%}(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \mathbf{y}) \approx \bar{y}_{\text{new}} \pm \Phi_{0.975} \times \sqrt{\mathrm{Var}(\boldsymbol{y}_{\text{new}} \mid \mathbf{x}_{\text{new}}, \mathbf{y})}$$

where $\bar{y}_{\text{new}}$ is the sample mean of the posterior predictive samples, and $\Phi_{0.975}$ is the 97.5th percentile of the standard normal distribution.

2.2.2.3 Gaussian Processes

Gaussian Processes (GPs) provide a flexible and powerful framework for non-parametric regression and classification. They are particularly valued for their ability to model uncertainty in predictions by placing a prior distribution over functions, with the assumption that any finite set of function values follows a joint Gaussian distribution.

GPs have wide applications in fields such as machine learning, spatial statistics, and time series analysis.

A GP is a generalisation of the multivariate normal distribution to an infinite-dimensional function space. Formally, a GP is defined as:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$

where:

- $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ is the mean function of the process,
- $k(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}(f(\mathbf{x}), f(\mathbf{x}'))$ is the covariance function (kernel), which encodes the relationship between function values at points $\mathbf{x}$ and $\mathbf{x}'$.

For any finite set of inputs $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the corresponding function values $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n)]^\top$ follow a multivariate Gaussian distribution:

$$\mathbf{f} \sim N\left(\mathbf{m}, \mathbf{K}\right)$$

where $\mathbf{m} = [m(\mathbf{x}_1), m(\mathbf{x}_2), \ldots, m(\mathbf{x}_n)]^\top$ is the mean vector and $\mathbf{K}$ is the covariance matrix, with entries $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

GPs are used in various domains due to their ability to model complex, non-linear relationships while providing uncertainty estimates. Some practical applications include:

- **Regression**: GPs are commonly used in regression tasks where, given training data $\{\mathbf{X}, \mathbf{y}\}$, the posterior distribution of the function $f(\mathbf{x})$ for a new input $\mathbf{x}^*$ is Gaussian. This provides point estimates as well as confidence intervals for predictions.
- **Bayesian Optimisation**: In optimisation tasks where the function to be evaluated is expensive, GPs guide the selection of new evaluation points by balancing exploration and exploitation, making them ideal for Bayesian optimisation.
- **Spatial Modelling**: In geostatistics, GPs are used in kriging, a method for predicting spatial data. GPs account for spatial autocorrelation and provide uncertainty estimates for unobserved locations.

- **Time Series Forecasting**: GPs model time series data by using kernels that capture temporal dependencies. This makes GPs suitable for applications where uncertainty quantification in predictions is crucial.

The choice of kernel function $k(\mathbf{x}, \mathbf{x}')$ is crucial as it defines the properties of the Gaussian Process, such as smoothness and periodicity. Some commonly used kernels include:

- **Squared Exponential (RBF) Kernel:**

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2l^2}\right)$$

  where σ_f^2 controls the variance, and l is the length scale determining how quickly correlations decay.

- **Periodic Kernel:**

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{2\sin^2\left(\frac{\pi||\mathbf{x}-\mathbf{x}'||}{T}\right)}{l^2}\right)$$

  where T represents the period of the data, and l controls the smoothness of the periodic fluctuations.

- **Linear Kernel:**

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$$

  This kernel is suitable for modelling linear relationships between input variables.

- **Matérn Kernel:**

$$k(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||}{l}\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||}{l}\right)$$

  where ν controls the smoothness of the function, l is the length scale, and K_ν is the modified Bessel function of the second kind. The Matérn kernel is more flexible than the RBF kernel for modelling rougher functions.

The choice of kernel in a Gaussian Process is crucial for defining the properties of the function being modelled, such as its smoothness, periodicity, or linearity. While kernels like the squared exponential or Matérn kernel are commonly chosen based on prior knowledge of the data, more complex or computationally efficient kernel representations can be derived using theoretical approaches. One such approach is Mercer's theorem, which provides a way to approximate kernels through a finite set of basis functions.

Mercer's theorem offers a theoretical framework for approximating positive semi-definite kernels by decomposing them into an infinite series of eigenfunctions. For any kernel function $k(\mathbf{x}, \mathbf{x}')$ that satisfies Mercer's conditions, it can be expressed as:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')$$

where λ_i are the eigenvalues, and $\phi_i(\mathbf{x})$ are the corresponding orthonormal eigenfunctions of the kernel. These eigenfunctions form a complete basis, and the kernel is represented as a weighted sum of these basis functions.

This type of expansion is known as the **Karhunen–Loève expansion**. In practice, for computational reasons, this infinite sum can be truncated to a finite sum:

$$k(\mathbf{x}, \mathbf{x}') \approx \sum_{i=1}^{L} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')$$

where L is the number of eigenfunctions that sufficiently approximate the kernel. By truncating the expansion, we can reduce the dimensionality of the problem and make Gaussian Process regression computationally feasible for larger datasets.

2.2.2.4 Gaussian Process Regression (GPR)

Gaussian Process Regression (GPR) is a non-parametric, Bayesian approach to regression that treats the function $f(\mathbf{x})$ being learned as a random function, distributed according to a GP prior. The goal is to compute the posterior distribution over the function values $f(\mathbf{x}_*)$ for a new input $\mathbf{x}_*$, given observed data $\mathbf{X}$ and $\mathbf{y}$. Unlike traditional linear regression, GPR allows the data to dictate the complexity of the

model, making it especially useful for capturing non-linear relationships and providing uncertainty estimates for predictions.

A key advantage of GPR over parametric methods like linear regression is its ability to quantify uncertainty in predictions. By defining a prior over functions, GPR provides both a prediction for each new data point and a measure of uncertainty, which is crucial in applications like time series forecasting, geostatistics, and machine learning tasks such as hyperparameter optimisation. Additionally, GPR naturally adapts to complex and non-linear relationships without requiring manual feature engineering.

Given a training set of n input-output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top$ is the matrix of input data and $\mathbf{y} = [y_1, y_2, \ldots, y_n]^\top$ is the vector of observed outputs, the outputs are assumed to follow a joint Gaussian distribution:

$$\mathbf{y} \sim N(\mathbf{0}, \mathbf{K}_{\mathbf{X},\mathbf{X}} + \sigma^2 \mathbf{I}_n)$$

where $\mathbf{K}_{\mathbf{X},\mathbf{X}}$ is the covariance matrix computed using a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$, and σ^2 is the noise variance.

The joint distribution of the observed outputs $\mathbf{y}$ and the function value $f(\mathbf{x}_*)$ for a new input $\mathbf{x}_*$ is:

$$p(\mathbf{y}, f(\mathbf{x}_*) \mid \mathbf{X}, \mathbf{x}_*) = \int p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f}, f(\mathbf{x}_*) \mid \mathbf{X}, \mathbf{x}_*) \, d\mathbf{f}$$

which, under Gaussianity assumptions, corresponds to

$$\begin{bmatrix} \mathbf{y} \\ f(\mathbf{x}_*) \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{\mathbf{X},\mathbf{X}} + \sigma^2 \mathbf{I}_n & \mathbf{K}_{\mathbf{X},\mathbf{x}_*} \\ \mathbf{K}_{\mathbf{x}_*,\mathbf{X}} & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right)$$

where $\mathbf{K}_{\mathbf{X},\mathbf{x}_*}$ represents the covariance between the training points and the new input, and $k(\mathbf{x}_*, \mathbf{x}_*)$ is the variance at the new input.

The posterior predictive distribution for the new input $\mathbf{x}_*$ is obtained by conditioning on the observed data and marginalising over the latent function values $\mathbf{f}$. Since the prior and likelihood are both Gaussian, the posterior is also Gaussian, given by:

$$p(f(\mathbf{x}_*) \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(f(\mathbf{x}_*) \mid \mathbf{f}, \mathbf{X}, \mathbf{x}_*) p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) \, d\mathbf{f}$$

which corresponds to

$$f(\mathbf{x}_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_*) \sim N(\mu_*, \Sigma_*),$$

where the posterior mean is:

$$\mu_* = \mathbf{K}_{\mathbf{x}_*,\mathbf{X}}(\mathbf{K}_{\mathbf{X},\mathbf{X}} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}$$

and the posterior variance (uncertainty) is:

$$\Sigma_* = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{K}_{\mathbf{x}_*,\mathbf{X}}(\mathbf{K}_{\mathbf{X},\mathbf{X}} + \sigma_n^2\mathbf{I})^{-1}\mathbf{K}_{\mathbf{X},\mathbf{x}_*}$$

2.2.3 Neural Networks

Neural networks are models that approximate complex functions by stacking layers of linear transformations and non-linear activation functions. These models are widely used in tasks like classification, regression, and image recognition. Similar to linear regression, the weights (analogous to $\boldsymbol{\beta}$) are learned from data through optimisation techniques that minimise a loss function.

2.2.3.1 Activation Function

Activation functions introduce non-linearity, enabling neural networks to capture complex patterns. Without them, the network would behave like a linear model. Common activation functions include:

- **ReLU (Rectified Linear Unit)**: $\text{ReLU}(x) = \max(0, x)$, widely used due to its simplicity and ability to mitigate the vanishing gradient problem.
- **Sigmoid**: $\sigma(x) = \frac{1}{1+e^{-x}}$, often used in binary classification.
- **Softmax**: $\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$, used in the output layer for multi-class classification to convert logits into probabilities that sum to 1.

These functions allow the network to model non-linear relationships, which is crucial for complex tasks like image classification or language modeling.

2.2.3.2 Backpropagation

Suppose our problem has a loss function L in which we aim to minimise its value, a common approach in neural network is to **Gradient Descent** method. This gradient descent is done in an iterative manner, and when we have more than 1 layer as in the case of neural networks, it is used with a notion called **Backpropagation**. This applies applies the chain rule to propagate gradients backward through the network, layer by layer, enabling weight updates via gradient descent. In each layer ℓ, the network computes:

$$\mathbf{z}^{(\ell)} = \mathbf{W}^{(\ell)}\mathbf{a}^{(\ell-1)} + \mathbf{b}^{(\ell)}, \quad \mathbf{a}^{(\ell)} = h(\mathbf{z}^{(\ell)})$$

where z is the pre-activated neurons, $\mathbf{W}$ is the layer weight, a is the posted-activated neurons, b is the layer bias, and $h(\cdot)$ is an activation function.

The goal of backpropagation is to compute $\frac{\partial L}{\partial \mathbf{W}^{(\ell)}}$, where $L(\theta)$ is the loss function. For the output layer, the gradient of the loss with respect to the activations is computed directly. For hidden layers, the chain rule gives:

$$\frac{\partial L}{\partial \mathbf{W}^{(\ell)}} = \frac{\partial L}{\partial \mathbf{z}^{(\ell)}} \cdot \frac{\partial \mathbf{z}^{(\ell)}}{\partial \mathbf{W}^{(\ell)}}, \quad \frac{\partial L}{\partial \mathbf{z}^{(\ell)}} = \frac{\partial L}{\partial \mathbf{a}^{(\ell)}} \cdot h'(\mathbf{z}^{(\ell)})$$

Once the gradients are computed, the weights are updated using gradient descent:

$$\mathbf{W}_{t+1}^{(\ell)} = \mathbf{W}_{t}^{(\ell)} - \nu \frac{\partial L}{\partial \mathbf{W}^{(\ell)}}$$

where ν is the *learning rate*, a hyperparameter that controls the step size of each update. A larger ν speeds up learning but risks overshooting the minimum, while a smaller ν ensures stability but may slow convergence.

2.2.3.3 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) is a variant of gradient descent where the weights are updated using a mini-batch (or a single sample) of data rather than the full dataset. This allows for faster updates and better generalisation due to the stochastic nature of the updates. The weight update for mini-batch $\mathcal{D}_t \subset \mathbf{X}$ at step t is:

$$\mathbf{W}_{t+1}^{(\ell)} = \mathbf{W}_{t}^{(\ell)} - \nu \frac{1}{|\mathcal{D}_t|} \sum_{i \in \mathcal{D}_t} \nabla_{\mathbf{W}^{(\ell)}} L(\mathbf{W}^{(\ell)}, \mathbf{x}_i)$$

SGD introduces randomness into the gradient estimates, allowing the network to escape local minima and potentially converge faster. However, choosing the right learning rate ν is crucial for convergence.

Robbins and Monro (1951) describes the **Robbins-Monro** condition as a guideline for setting the learning rate ν in SGD to ensure convergence:

$$\sum_{t=1}^{\infty} \nu_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \nu_t^2 < \infty$$

This condition suggests that the learning rate should decrease over time, ensuring that updates become smaller as the algorithm converges. A typical approach is to set $\nu_t = \frac{\nu_0}{t}$, where ν_0 is the initial learning rate.

2.2.4 Dimensionality Reduction

When dealing with a large number of variables (high-dimensional data), dimensionality reduction helps filter out noise and focus on a smaller set of informative variables. There are two main approaches: *feature extraction* and *variable screening*.

2.2.4.1 Feature Extraction

Feature extraction reduces the dimensionality of the data by creating new features that capture the most important variation in the data. Common techniques include:

- **Singular Value Decomposition (SVD)**: SVD decomposes the data matrix $\mathbf{X}$ into three matrices, $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}^\top$, where $\mathbf{\Sigma}$ contains the singular values, representing the importance of each component. The first few components capture most of the variation, while later components can be discarded as noise:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

- **Principal Component Analysis (PCA)**: PCA transforms the data into a set of orthogonal components (principal components) that maximise the variance in the data. The idea is to project the data $\mathbf{X}$ onto a lower-dimensional subspace while retaining as much variance as possible.

  Mathematically, PCA finds a projection vector $\mathbf{w}$ that maximises the variance of the projected data $\mathbf{w}^\top \mathbf{x}$, subject to the constraint that $\mathbf{w}$ has unit norm ($\|\mathbf{w}\|_2 = 1$):

$$\text{maximise} \quad \text{Var}(\mathbf{w}^\top \mathbf{X}) = \mathbf{w}^\top \mathbf{S} \mathbf{w} \quad \text{subject to} \quad \|\mathbf{w}\|_2 = 1$$

  where $\mathbf{S}$ is the sample covariance matrix of the data. The solution is given by the eigenvectors of $\mathbf{S}$, with the eigenvalues representing the amount of variance explained by each principal component. By keeping only the components with the largest eigenvalues, PCA reduces the dimensionality while retaining the key structure of the data.

2.2.4.2 Variable Screening

Variable screening selects a subset of the original variables based on their relevance to the response variable, making it easier to interpret than feature extraction.

In classical settings where the number of samples n is greater than the number of variables p, several statistical criteria are used to evaluate and select variables:

- **Akaike Information Criterion (AIC)**: AIC balances model fit and complexity. It is defined as:

$$\text{AIC} = -2 \log L + 2p$$

  where L is the likelihood of the model and p is the number of parameters. A lower AIC indicates a better model.

- **Bayesian Information Criterion (BIC)**: BIC is similar to AIC but applies a stronger penalty for models with more parameters:

$$\text{BIC} = -2 \log L + p \log n$$

- **Likelihood Ratio Test**: Compares a more complex model (with more variables) to a simpler one. The test statistic is:

$$\Lambda = -2 \left(\log L_{\text{restricted}} - \log L_{\text{full}} \right)$$

where $L_{\text{restricted}}$ and L_{full} are the likelihoods of the simpler and more complex models, respectively. This statistic follows a chi-squared distribution under the null hypothesis that the simpler model is sufficient.

In high-dimensional settings, such as neuroimaging, where the number of variables far exceeds the number of samples, more specialised methods are required:

- **Mass Univariate Analysis (MUA)**: MUA is a benchmark method in neuroimaging for analysing high-dimensional data. It involves fitting a model for each variable (e.g. voxel or region of interest) individually to the response variable and performing hypothesis testing for each. For each variable x_j, we test the null hypothesis $H_0 : \beta_j = 0$ using a t-test. The test statistic is the *t-statistic*:

$$T = \frac{\hat{\beta}_j}{\hat{\text{SE}}(\hat{\beta}_j)}$$

where $\hat{\beta}_j$ is the estimated coefficient and $\hat{\text{SE}}(\hat{\beta}_j)$ is its standard error. The p-value for variable x_j is calculated as:

$$p_j = 2 \times P \left(T_{n-p} > |T| \right)$$

A common threshold for statistical significance is $p < 0.05$. Due to the large number of comparisons in neuroimaging, multiple comparisons correction, such as Bonferroni correction, is often applied to adjust p-values.

MUA is widely used due to its simplicity and direct interpretation, as each variable (e.g. voxel or feature) is evaluated individually. However, a key drawback of MUA is that it does not consider the overall structure of the data, as it treats each variable separately. This can result in the loss of important interactions or dependencies between variables. Consequently, single-model

approaches like regularisation-based methods (e.g. LASSO, Ridge) that fit all variables simultaneously are gaining interest, as they better capture the underlying structure of the data and often improve predictive performance.

- **Regularisation-Based Methods**: In cases where $p \gg n$, regularisation techniques such as LASSO and Ridge regression are employed for variable selection while fitting a model. These methods shrink the coefficients of less important variables toward zero, reducing model complexity and enhancing interpretability. Unlike simple model selection methods that discard variables entirely, regularisation-based methods retain all variables but apply penalties that control the magnitude of the coefficients.

  - **LASSO Regression**: LASSO regression performs variable selection and regularisation simultaneously by introducing an L1 penalty. The LASSO estimator $\hat{\boldsymbol{\beta}}_{\text{LASSO}}$ is obtained by solving the following optimisation problem:

    $$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

    The L1 penalty causes LASSO to shrink some of the coefficients to exactly zero, allowing for variable selection. Variables with non-zero coefficients, i.e.

    $$\boldsymbol{\beta}_{\text{screen}} = \left\{ j : |\hat{\beta}_{\text{LASSO},j}| > 0 \right\}$$

    are considered important.

  - **Ridge Regression**: Ridge regression applies an L2 penalty, leading to a less aggressive shrinkage of coefficients compared to LASSO. The Ridge estimator $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$ is given by:

    $$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^\top \mathbf{y}$$

    where λ controls the degree of shrinkage, and $\mathbf{I}$ is the identity matrix. Unlike LASSO, Ridge does not shrink coefficients to exactly zero, but it reduces their magnitude. To select important variables, a threshold γ can

be defined, and variables with Ridge coefficients exceeding this threshold are retained:

$$\boldsymbol{\beta}_{\text{screen}} = \left\{ j : |\hat{\beta}_{\text{Ridge},j}| > \gamma \right\}$$

This allows for the screening of less important variables while keeping those with larger coefficients.

2.2.5 Time Series Modelling

Time series modelling is essential for analysing data collected over time, such as neuroimaging data, financial data, or economic data. These models are designed to capture key temporal structures, including trends, cycles, and autocorrelations, to make informed predictions and infer relationships across time points.

2.2.5.1 Basics of Time Series Data

Time series data frequently reflect both *cycles* and *trends*. Trends represent long-term directional movements in the data, capturing non-stationary elements such as general upward or downward shifts over time. Cycles, on the other hand, are periodic fluctuations that may result from seasonal or other repeating patterns. Both cycles and trends are critical features to consider in time series analysis.

In addition, *lags* are an essential characteristic of time series data. A lag represents the influence of past observations on the current value of the series. This relationship leads to *autocorrelation*, which quantifies the degree of correlation between the time series and its lagged values. The autocorrelation at lag k is defined as:

$$\text{ACF}(k) = \frac{\text{Cov}(y_t, y_{t-k})}{\text{Var}(y_t)}$$

where $\text{Cov}(y_t, y_{t-k})$ is the covariance between the value of the time series at time t and its value at lag $t-k$. Modelling autocorrelation correctly is key to capturing the temporal dependencies in time series data and improving the quality of forecasts.

2.2.5.2 Modelling Approaches

Several classical models are commonly used for analysing time series data, including the AutoRegressive (AR) model, the Moving Average (MA) model, and the AutoRegressive Moving Average (ARMA) model, which combines both. These models are designed to capture the temporal dependencies and patterns within the data.

The *autoregressive (AR)* model is used to model a time series by expressing the current value as a linear function of its p preceding values. An AR(p) process can be described by the following equation:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \tag{2.1}$$

where $\phi_1, \ldots, \phi_p$ are the model parameters, and ϵ_t is white noise with mean zero and variance σ^2. This model captures how past observations influence the present value of the series.

In contrast, the *moving average (MA)* model expresses the current value of the time series as a linear combination of past residuals (errors). The MA(q) process is written as:

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \tag{2.2}$$

where μ is the mean of the series, $\theta_1, \ldots, \theta_q$ are the model parameters, and ϵ_t is white noise. This model is used to account for short-term shock effects in the data.

The *autoregressive moving average (ARMA)* model combines the autoregressive and moving average components into a single model. The ARMA(p, q) process is defined as:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{2.3}$$

This hybrid model captures both the long-term dependencies (via autoregression) and short-term shocks (via the moving average component) in the time series data.

We can estimate the time series coefficients for these model. To give an example, assume an AR(p) process given in Eq. (2.1), two common methods for estimation include **Maximum Likelihood Estimation (MLE)** and the **Yule-Walker equations**.

Let $\mathbf{y} = [y_p, y_{p+1}, \ldots, y_T]^\top$ be the vector of observations from time p to T, and define the design matrix $\mathbf{X}$ of lagged values as:

$$\mathbf{X} = \begin{bmatrix} y_{p-1} & y_{p-2} & \cdots & y_0 \\ y_p & y_{p-1} & \cdots & y_1 \\ \vdots & \vdots & \ddots & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-p} \end{bmatrix}$$

The corresponding AR(p) model can be written in matrix form as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\phi} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$ is the vector of errors. The log-likelihood function for the AR(p) model is:

$$\log L(\boldsymbol{\phi}, \sigma^2) = -\frac{T-p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\phi})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\phi})$$

Taking the derivative of the log-likelihood with respect to $\boldsymbol{\phi}$ and setting it to zero gives the MLE for $\boldsymbol{\phi}$:

$$\hat{\boldsymbol{\phi}}_{\text{MLE}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

This solution is equivalent to the OLS estimate for the autoregressive model parameters. After estimating $\hat{\boldsymbol{\phi}}_{\text{MLE}}$, the MLE of the residual variance σ^2 can be computed as:

$$\hat{\sigma}^2 = \frac{1}{T-p}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\phi}}_{\text{MLE}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\phi}}_{\text{MLE}}).$$

Alternatively, the Yule-Walker equations provide a method to estimate the AR(p) model parameters by relating the autocorrelations of the time series to the AR coefficients. Define the autocovariance matrix $\mathbf{R}$ as:

$$\mathbf{R} = \begin{bmatrix} \rho(0) & \rho(1) & \cdots & \rho(p-1) \\ \rho(1) & \rho(0) & \cdots & \rho(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(p-1) & \rho(p-2) & \cdots & \rho(0) \end{bmatrix}$$

where $\rho(k)$ represents the autocorrelation at lag k. The vector of autocorrelations $\boldsymbol{\rho}$ is:

$$\boldsymbol{\rho} = \begin{bmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(p) \end{bmatrix}$$

The Yule-Walker equations in matrix form are then given by:

$$\mathbf{R}\boldsymbol{\phi} = \boldsymbol{\rho}$$

Solving this system yields the AR coefficients:

$$\boldsymbol{\phi} = \mathbf{R}^{-1}\boldsymbol{\rho}$$

This matrix form makes the Yule-Walker equations efficient to compute, especially for higher-order AR models, and is a key method for parameter estimation in time series analysis.

2.3 Current Study of Statistics in Neuroimaging

The field of neuroimaging has seen significant advances in statistical methodologies aimed at uncovering brain-related effects and predicting clinical outcomes based on high-dimensional imaging data. Section 2.3.1 presents major approaches to effect discovery in brain MRI, which serve as the foundation for the analyses in Chapter 4. Section 2.3.2 covers scalar-on-image regression model and the notion of brain age gap, a critical method used in Chapters 5 and 6. Finally, Section 2.3.3 examines the handling of time series data in fMRI, which is central to the work in Chapter 3.

2.3.1 Effect Discovery Based on Brain MRI

Effect discovery in brain MRI aims to identify brain regions or features associated with outcomes such as cognitive function or mental disorders. Given the high-dimensional nature of MRI data, statistical methods must efficiently handle a large number of variables while controlling for false positives. Two widely used approaches are

mass univariate analysis and *regularisation-based methods*. These variable screening techniques are particularly useful when the objective is to discover interpretable biomarkers or regions associated with specific neurological conditions. By identifying a sparse set of relevant variables, these methods enhance interpretability while addressing the high-dimensional nature of MRI data and controlling for false positives.

Mass Univariate Analysis (MUA) fits separate models to each voxel or region of interest (ROI) in the brain, testing for associations with the outcome of interest. Each voxel is tested using t-tests or F-tests, and corrections for multiple comparisons e.g. Bonferroni, False Discovery Rate (FDR) and Family-Wise Error Rate (FWER) are applied to control for false positives. While MUA is computationally simple and highly interpretable, its primary limitation is that it treats each voxel independently, potentially missing complex relationships between brain regions.

Regularisation-based methods, such as LASSO, provide an alternative by shrinking many coefficients to zero, retaining only a subset of important features. These methods handle high dimensionality effectively and are capable of capturing interactions between variables, which MUA may overlook. By applying a penalty to the model, regularisation techniques reduce complexity, improve interpretability, and model interactions across brain regions.

The choice between MUA and regularisation-based approaches depends on the study's objectives. MUA is favoured for simple, localised effect discovery, while regularisation methods are better suited to uncovering distributed brain patterns and complex interactions.

2.3.2 Scalar-on-Image Regression

Scalar-on-image regression is a statistical method in neuroimaging used to predict scalar outcomes, such as age, cognitive scores, or disease severity, from high-dimensional brain images. This approach has been widely applied to understand how variations in brain structure or function are associated with cognitive and clinical outcomes.

The **scalar-on-image regression model** can be formalised as follows: let $\mathbf{X}(\mathbf{s})$ represent the imaging data at voxel locations $\mathbf{s} \in \mathcal{S}$, and let y be the scalar outcome. The model is expressed as:

$$y = \int_{\mathcal{S}} \mathbf{X}(\mathbf{s})\beta(\mathbf{s})\, d\mathbf{s} + \epsilon$$

where $\beta(\mathbf{s})$ represents the coefficient function that captures the relationship between imaging features and the scalar outcome, and ϵ denotes the error term.

Scalar-on-image regression has been widely used to investigate how changes in brain structure or function relate to cognitive and clinical outcomes. However, a major challenge in this field is the high dimensionality of the imaging data, where the number of voxels far exceeds the number of subjects. This imbalance makes it essential to apply regularisation and dimensionality reduction techniques, which help ensure the model remains interpretable by focusing on the most relevant brain regions while reducing noise and overfitting.

A key application of scalar-on-image regression is in estimating the *brain age gap* (BAG), where the difference between the predicted brain age and chronological age serves as a potential biomarker for neurodegenerative diseases or cognitive decline. In **brain age prediction**, the goal is to build a model that uses MRI data to estimate the brain's age by mapping voxel-wise or region-wise brain features to chronological age. This is typically achieved using regularised linear regression or machine learning techniques, such as random forests or support vector machines. BAG has been shown to correlate with various neurological conditions, including Alzheimer's disease, schizophrenia, and general cognitive decline.

2.3.3 Time Series fMRI Modelling

fMRI data is inherently a time series, capturing brain activity over time. However, fMRI signals often exhibit temporal autocorrelation and non-stationarity, complicating their analysis. Pre-whitening techniques are typically applied to reduce temporal autocorrelations and control false positive rates (FPR).

2.3.3.1 Autocorrelation in fMRI Data

A significant challenge in modelling fMRI data is the presence of temporal autocorrelation, where the fMRI signal at time t depends on its past values at time $t - k$. Ignoring this autocorrelation can inflate false positives in statistical tests, as it leads

to underestimated standard errors and overly confident test statistics. As a result, statistical inferences based on OLS regression become invalid in the context of fMRI.

Consider a $T \times I$ matrix of BOLD measurements, $\mathbf{Y}$, where each column represents the time series for voxel i, and a $T \times Q$ design matrix, $\mathbf{X}$, containing Q independent variables. The design matrix $\mathbf{X}$ typically includes task-related predictors of interest, alongside nuisance variables. These nuisance variables help mitigate spurious effects in the BOLD signal arising from non-neuronal sources of variability, which are common in fMRI data.

For example, nuisance variables may account for low-frequency drifts caused by scanner instability (handled via high-pass filters), physiological fluctuations (such as respiratory and cardiac cycles), and head motion (including temporal derivatives to capture rapid motion changes). Additionally, when analysing multi-site fMRI data, scanning site effects are often modelled to account for inter-machine differences. These nuisance regressors reduce the influence of artefacts, helping improve the accuracy of the model's inferences.

The mass univariate linear model is then expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}, \tag{2.4}$$

where $\mathbf{E}$ is a $T \times I$ matrix of random errors. If each column of $\mathbf{E}$, denoted $\boldsymbol{\epsilon}_i$, were uncorrelated with mean zero, i.e. $\mathbb{C}(\boldsymbol{\epsilon}_i) = \sigma_i^2\mathbf{I}$, the OLS estimate $\hat{\boldsymbol{\beta}}_i^{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}_i$ would be the Best Linear Unbiased Estimator (BLUE) with variance $\mathbb{V}(\hat{\boldsymbol{\beta}}_i^{\text{OLS}}) = \sigma_i^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

However, in fMRI, $\boldsymbol{\epsilon}_i$ is typically serially correlated, violating the OLS assumptions and leading to biased variance estimates $\mathbb{V}(\hat{\boldsymbol{\beta}}_i^{\text{OLS}})$. This serial correlation necessitates more robust estimation techniques to correct for these temporal dependencies and improve the precision of the estimated coefficients.

To handle serially correlated errors, Generalised Least Squares (GLS) is used. For voxel i, assume $\mathbb{V}(\boldsymbol{\epsilon}_i) = \sigma^2\boldsymbol{\Omega}_i$, where $\boldsymbol{\Omega}_i$ is a non-diagonal correlation matrix. Given a whitening matrix $\mathbf{W}_i$ that satisfies $\mathbf{W}_i\boldsymbol{\Omega}_i\mathbf{W}_i^\top = \mathbf{I}$, we can transform the model into a whitened form:

$$\mathbf{W}_i\mathbf{y}_i = \mathbf{W}_i\mathbf{X}\boldsymbol{\beta}_i + \mathbf{W}_i\boldsymbol{\epsilon}_i, \tag{2.5}$$

which simplifies to:

$$\mathbf{y}_i^* = \mathbf{X}_i^* \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i^*,$$

where $\mathbf{y}_i^*$, $\mathbf{X}_i^*$, and $\boldsymbol{\epsilon}_i^*$ are the whitened data, design matrix, and error terms, respectively.

The GLS estimate is:

$$\hat{\boldsymbol{\beta}}_i^{\text{GLS}} = (\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1} \mathbf{X}_i^{*\top} \mathbf{y}_i, \tag{2.6}$$

with sampling variance:

$$\mathbb{V}(\hat{\boldsymbol{\beta}}_i^{\text{GLS}}) = \sigma^2 (\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1}.$$

The test statistic t_i for voxel i based on a contrast vector $\mathbf{c}$ is:

$$t_i = \frac{\mathbf{c}\hat{\boldsymbol{\beta}}_i^{\text{GLS}}}{\sqrt{\hat{\sigma}^2 \mathbf{c} (\mathbf{X}^\top \boldsymbol{\Omega}_i^{-1} \mathbf{X})^{-1} \mathbf{c}^\top}}. \tag{2.7}$$

In practice, the true correlation structure $\boldsymbol{\Omega}_i$ is unknown, so an estimate $\hat{\boldsymbol{\Omega}}_i$ must be obtained, often through likelihood methods, moment estimation, or parametric modelling.

2.3.3.2 Time Series fMRI Extraction

In both task-based fMRI and resting-state fMRI, time series analysis is used to extract meaningful patterns of brain activity, often through functional connectivity networks that reflect brain regions with synchronous activity. One widely used method for extracting this information from fMRI data is *Independent Component Analysis* (ICA), which identifies spatially independent brain networks from fluctuations in the BOLD signal.

Independent Component Analysis (ICA) is a statistical technique used to decompose multivariate signals into additive, independent components. In the context of fMRI, ICA separates the observed BOLD signals into temporally independent components, each corresponding to different brain networks that exhibit distinct

temporal dynamics. The objective of ICA is to find components that are statistically independent from each other, representing different functional networks within the brain.

Mathematically, the observed fMRI signal at time t, denoted by $\mathbf{y}_t \in \mathbb{R}^n$, is assumed to be a linear mixture of independent sources $\mathbf{s}_t \in \mathbb{R}^m$, such that:

$$\mathbf{y}_t = \mathbf{A}\mathbf{s}_t$$

where:

- $\mathbf{y}_t$ is the observed BOLD signal across n voxels at time t,
- $\mathbf{A} \in \mathbb{R}^{n \times m}$ is the unknown mixing matrix,
- $\mathbf{s}_t$ are the independent source signals (i.e. the brain networks).

The goal of ICA is to estimate the matrix $\mathbf{W}$, which is the inverse of the mixing matrix $\mathbf{A}$, such that:

$$\mathbf{s}_t = \mathbf{W}\mathbf{y}_t$$

This transformation recovers the independent components $\mathbf{s}_t$, each corresponding to a distinct functional connectivity network.

ICA is applied to the 4D BOLD signal to uncover spatially distinct brain networks that show coherent fluctuations over time. This approach is particularly powerful in rs-fMRI, where the goal is to identify intrinsic functional networks without task-driven activity. In t-fMRI, ICA can also be used to distinguish task-related networks from noise or non-task-related fluctuations.

Figure 2.5 shows an example application of ICA in detecting several well-established brain networks that are extracted from fMRI data, including:

- **Default Mode Network (DMN)**: Active during rest, this network is associated with self-referential thought and includes regions such as the medial prefrontal cortex, posterior cingulate cortex, and inferior parietal lobule.

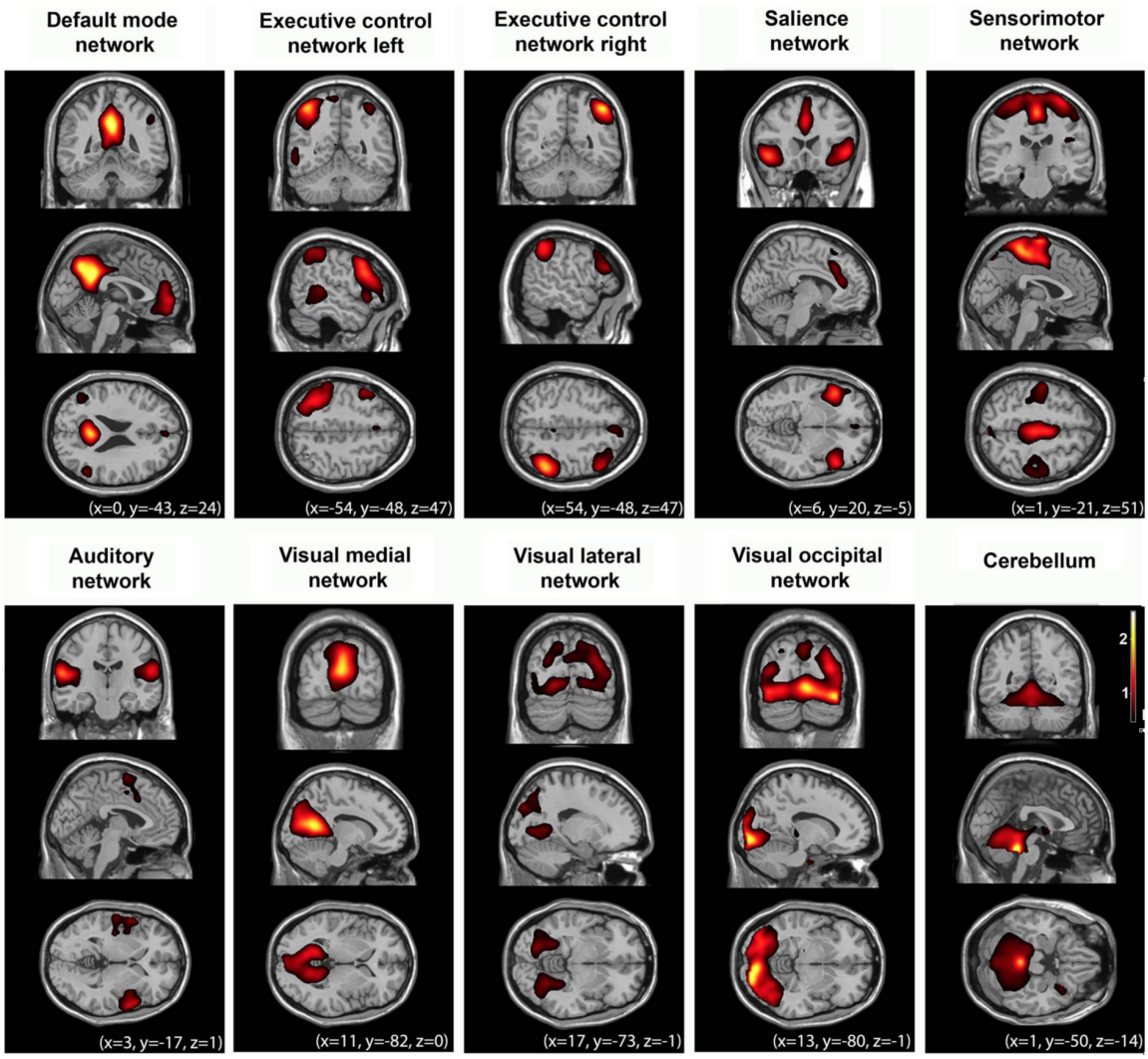


Figure 2.5: Multiple cerebral networks identified with fMRI during normal wakeful resting state using ICA from Heine et al. (2012).

- **Sensory-Motor Network (SMN)**: Involved in processing sensory inputs and motor outputs, this network includes the primary motor cortex, supplementary motor area, and sensory areas.
- **Visual Network (VN)**: Responsible for visual information processing, it includes regions such as the occipital lobe and primary visual cortex.
- **Auditory Network (AN)**: This network processes auditory stimuli and includes regions such as the superior temporal gyrus and Heschl's gyrus.
- **Frontoparietal Network (FPN)**: Associated with executive functions, such as working memory and decision-making, this network includes regions in the dorsolateral prefrontal cortex and the parietal cortex.
- **Salience Network (SN)**: Involved in detecting and responding to salient stimuli, this network includes the anterior insula and anterior cingulate cortex.
- **Cerebellar Network (CN)**: Involved in motor control and cognitive functions, this network includes the cerebellum and its connections to the cortex.

Each of these networks is identified by ICA as a spatial map of voxels that fluctuate together over time, reflecting the functional connectivity of the brain during rest or specific tasks.

2.4 Big Imaging Datasets

Advances in neuroimaging have allowed researchers to collect large-scale datasets that capture brain structure and function across vast and diverse populations. Traditional neuroimaging studies often rely on relatively small sample sizes, which limits the ability to generalise findings to broader populations and to explore subtle individual differences. In contrast, big imaging datasets, like those used in this study, provide comprehensive, high-resolution data on thousands of participants. This enables researchers to investigate complex brain-behaviour relationships, capture a wide range of demographic variability, and enhance statistical power for detecting patterns in brain function and structure.

These large datasets not only surpass the limitations of small sample studies in terms of statistical robustness and diversity but also allow for the integration of neuroimaging with genetic, behavioural, and clinical information. This integration supports a multi-dimensional approach to studying brain processes and neurodevelopmental, neurodegenerative, and psychiatric conditions. By leveraging such expansive datasets, researchers can build more reliable models, uncover population-level trends, and contribute to a deeper understanding of the human brain across different stages of life and health conditions.

2.4.1 UK Biobank Dataset

The UK Biobank is one of the largest and most comprehensive biomedical datasets in the world, designed to facilitate a wide array of health-related research. It includes extensive neuroimaging data from over 50,000 participants, with plans to extend this to 100,000 participants in the coming years. The dataset is comprised of a demographically diverse cohort of adults aged 40 to 69 years at the time of enrolment. Imaging modalities include structural MRI, resting-state functional MRI, and diffusion-weighted imaging, as well as task-based fMRI that covers a variety of cognitive domains.

In addition to the imaging data, the UK Biobank provides a wealth of behavioral, socio-demographic, clinical, and genetic information, making it an exceptional resource for studying the relationships between brain structure, function, and a wide range of health outcomes. The imaging protocols are standardised across all participants, ensuring consistency and enabling large-scale population studies. Researchers use this dataset to investigate various aspects of brain aging, neurodegenerative diseases, psychiatric disorders, and the genetic underpinnings of brain function. The UK Biobank's open access policy and its large sample size provide unparalleled opportunities for addressing fundamental questions in neuroscience and public health.

2.4.2 NKI-Rockland Sample Dataset

The Enhanced Nathan Kline Institute-Rockland Sample (NKI-RS) is an ongoing initiative aimed at creating a large-scale, community-based sample ($N > 1000$) spanning participants across the lifespan. Currently, the dataset includes comprehensive

neuroimaging data from 557 participants, aged 6 to 85 years, offering a diverse demographic profile. Both male and female subjects are well-represented, contributing to the generalisability of the findings. The imaging modalities include structural MRI and resting-state functional MRI with repetition times (TR) of 1400 ms and 645 ms. In addition, task-based fMRI data, such as the Checkerboard task, is available with a TR of 645 ms.

The NKI-RS also provides an extensive array of physiological, psychological, and genetic assessments alongside the neuroimaging data. Anonymised data are openly shared on a quarterly basis, aligned with the National Institute of Mental Health's strategic plan to advance psychiatric neuroscience. This dataset supports investigations into developmental trajectories related to risk and resilience throughout life. The rich phenotypic information, including cognitive assessments, psychological questionnaires, and behavioral measures, enables researchers to explore brain-behavior relationships in depth, facilitating hypothesis generation and testing.

2.4.3 Human Connectome Project (HCP) Dataset

The Human Connectome Project (HCP), a large-scale collaborative effort funded by the National Institutes of Health, aims to map the neural pathways of the human brain in unprecedented detail. The dataset includes high-resolution neuroimaging data from 1,200 healthy young adults aged between 22 and 35 years. Imaging modalities span structural MRI, resting-state fMRI with a TR of 720 ms, and task-based fMRI, including tasks such as the Motor (Left Hand) and Gambling Reward tasks, both with TRs of 753 ms. The dataset also includes Diffusion Tensor Imaging data for the analysis of white matter connectivity.

In addition to the imaging data, the HCP dataset includes an extensive set of behavioral, demographic, and genetic information, enabling the exploration of a wide range of topics related to brain structure, function, and connectivity. The cohort consists of a balanced mix of male and female participants, allowing researchers to perform population-level studies across various domains of brain science.

2.4.4 Developing Human Connectome Project (dHCP) Dataset

The Developing Human Connectome Project (dHCP) is an ambitious study led by King's College London, Imperial College London, and the University of Oxford, aimed at creating the first four-dimensional connectome of early brain development. The dataset includes high-resolution neuroimaging data from 55 infants, ranging from preterm to full-term neonates, with post-conceptional ages between 20 and 44 weeks. The imaging modalities include structural MRI and resting-state functional MRI with a TR of 392 ms.

The goal of the dHCP is to generate a dynamic map of human brain connectivity during the critical stages of early life, integrating imaging with clinical, behavioral, and genetic data. This resource offers a unique opportunity to investigate the neural processes underlying early brain development, particularly in infants with heightened genetic or environmental risks for conditions such as Autism Spectrum Disorder (ASD) or Cerebral Palsy. The open-source informatics structure of the dHCP facilitates broad access by researchers worldwide, enabling significant progress in understanding neurodevelopmental processes.

2.4.5 F-CON1000 Beijing Dataset

The F-CON1000 Beijing dataset is part of the 1000 Functional Connectomes Project (FCP), which seeks to establish a global repository of resting-state fMRI datasets from multiple sites. Contributed by researchers from Beijing Normal University under the supervision of Dr. Yufeng Zang, the dataset includes resting-state fMRI data from 82 healthy participants. Each participant underwent 225 scans during the fMRI acquisition with a TR of 2 seconds.

This dataset supports a wide variety of research applications, including the analysis of functional brain connectivity, comparative studies across different populations and sites, and the development of new analysis methods for resting-state fMRI. The F-CON1000 Beijing dataset is freely available to the scientific community, providing an invaluable resource for advancing our understanding of brain connectivity.

2.5 Conclusion

The content discussed hitherto forms the foundational knowledge essential for understanding the motivations and framework of this research. We began by exploring the motivations for advancing statistical methods in neuroimaging, tracing the history and development of MRI and fMRI, delving into their physical and biological mechanisms, and highlighting the valuable insights they offer into brain function and structure.

We then introduced the key statistical theories that underpin this thesis, detailing the broad pipeline of neuroimaging analysis. This included discussions on critical methodologies such as effect discovery in brain MRI, scalar-on-image regression, i.e. utilising MRI data to predict health outcomes, and time-series modelling of functional MRI. These methods form the core of the challenges and opportunities addressed in subsequent chapters.

Finally, we reviewed the datasets utilised in this research, underscoring the need for scalable, accurate, and adaptable analytical frameworks to effectively handle the complexity and scale of modern neuroimaging data. This comprehensive foundation sets the stage for the methodological advancements presented in the chapters to follow, driving forward the field of neuroimaging analysis.

Chapter 3

Spectral Adjusted Least Square for Autocorrelated Noise

This chapter presents the development of a pre-whitening method to address the critical issue of inflated false positive rates in fMRI analysis, which can compromise the validity of subsequent statistical inferences. While this work does not employ Bayesian methods, as suggested by the thesis title and other chapters, it establishes a foundational framework for downstream fMRI analysis. Unfortunately, this methodology was not extended to the later chapters, which primarily focus on statistical analyses of post-processed data, and thus this chapter stands independently. This work was conducted collaboratively with Dr. Soroosh Afyouni (NDPH, University of Oxford), Prof. Mark Jenkinson (FMRIB, University of Oxford), Dr. Eugene Duff (FMRIB, University of Oxford), Prof. Mark Woolrich (Department of Psychiatry, University of Oxford), Prof. Stephen M. Smith (FMRIB, University of Oxford), and Prof. Thomas E. Nichols (NDPH, University of Oxford).

3.1 Introduction

Temporal autocorrelation is a long-standing challenge for task fMRI anlayses. Autocorrelation must be accounted for to ensure statistical efficiency and accurate standard errors (Worsley and Friston, 1995). Ignoring autocorrelation can produce inefficient

estimates (less than optimally-precise measures of task effects) and invalid inferences (inflated false positive rates). However, the accuracy of autocorrelation models is rarely verified (Monti, 2011). Olszowy et al. (2019) has shown that commonly used methods do not successfully whiten the noise, especially with fast (TR<2s) acquisitions, and Bollmann et al. (2018) showed typically used AR(1) models were insufficient for such acquisitions. As these fast acquisitions become increasingly common Chen et al. (2019), it is important that the statistical tools can fit this increased level and complexity of temporal autocorrelation.

The standard errors that are most affected by autocorrelation are first level standard errors. While some group level models ignore these standard errors (Mumford and Nichols, 2009), they are important for a number of reasons. Firstly, for any group fMRI analysis, optimal sensitivity requires the individual weighting using these first level variance estimates (Beckmann et al., 2003; Woolrich et al., 2004; Chen et al., 2012; Bianciardi et al., 2004). Secondly, there is a growing recognition of the importance of modelling the random effect of stimuli accounting for the trial-to-trial variation in response to face stimuli, when the faces are a random sample of all possible faces (Westfall et al., 2016). This model, which must disentangle subject and stimuli random effects variance, depends on precise estimates of error variance. And finally, single-subject task fMRI analyses remain important for clinical applications (Bullmore, 2012) including pre-surgical planning, and 'precision neuroimaging' (Michon et al., 2022), and these analyses depend on trustworthy statistic and p-values at the first level.

Widely used fMRI packages use different approaches to autocorrelation modelling. FSL uses a spectral technique using tapered and spatially smoothed autocorrelation coefficients of an arbitrary lag (Woolrich et al., 2001). fMRIStat, perhaps less well-known but still influential, uses a first-order autoregressive model using bias-corrected autocorrelation coefficients (Worsley, 2001). Although SPM initially used temporal colouring (Worsley and Friston, 1995), since SPM2, a Restricted Maximum Likelihood (ReML) method has been used for estimating a basis approximation to a AR(1) model (SPM-AR(1)) (Friston et al., 2002; Penny et al., 2011). Most recently, SPM has introduced a method called FAST for rapidly sampled task fMRI (Corbin et al., 2018) (SPM-FAST). SPM-AR(1) and SPM-FAST both only use a global (spatially homogeneous) model for autocorrelation, estimated from a subset of voxels showing

some evidence of significance. Finally, AFNI uses an Autoregressive Moving Average model (ARMA) of order (1,1) incorporated into the 3dREMLfit; this method uses ReML estimates of AR and MA for each voxel, using a grid search to improve scability.

In this work, we propose a new method, Spectral Adjusted Least Square for noise Autocorrelation (SALSA). SALSA draws on ideas from existing methods to produce a more accurate while still computationally efficient method. It uses a two-stage approach, first whitening with a global but highly flexible autocorrelation model to account for long-range dependence (like SPM-FAST), followed by a voxel-specific tapered autocorrelation estimate (like FSL). A bias-correction for the autocorrelation (like fMRIstat) is essential to account for the residual structure induced by the design and high-pass filter used for drift modelling, and we develop this bias correction to account for either basis drift models (AFNI, SPM) or kernel drift models (FSL). We demonstrate the accuracy of our approach and compare it existing methods on a variety of real datasets, including evaluation of false positive rates for task fMRI using resting fMRI as a source of null data, and single-subject and group analyses.

3.2 Methods

3.2.1 Mass Univariate Time Series Linear Model

Let $\mathbf{Y}$ be a $T \times I$ matrix of observations, where the i-th column is the time series BOLD measurement for voxel i and the t-th row is the observation at time point t. Denote by $\mathbf{X}$ the $T \times Q$ design matrix, comprising task predictors and nuisance regressors (e.g. drift basis, motion parameters, and temporal derivatives). The mass univariate linear model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E},$$

where $\mathbf{E}$ is a $T \times I$ matrix of additive random errors.

In the special case where each error column $\boldsymbol{\epsilon}_i$ has covariance $\sigma_i^2\mathbf{I}$, the ordinary least squares (OLS) estimator

$$\hat{\boldsymbol{\beta}}_i^{\mathrm{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}_i$$

is the Best Linear Unbiased Estimator (BLUE). However, BOLD time series typically exhibit serial correlation, violating the uncorrelated-errors assumption. As a result, the variance $\sigma_i^2(\mathbf{X}^\top\mathbf{X})^{-1}$ is biased and often underestimated, and the OLS estimator ceases to be optimal.

Generalised least squares (GLS) corrects for correlated errors by assuming $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \sigma^2\,\boldsymbol{\Omega}_i$, where $\boldsymbol{\Omega}_i$ is a non-diagonal correlation matrix. If there exists a whitening matrix $\mathbf{W}_i$ such that $\mathbf{W}_i\,\boldsymbol{\Omega}_i\,\mathbf{W}_i^\top = \mathbf{I}$, then we can have

$$\mathbf{y}_i^* = \mathbf{X}_i^*\beta_i + \epsilon_i^*, \tag{3.1}$$

where

$$\mathbf{y}_i^* = \mathbf{W}_i\,\mathbf{y}_i, \quad \mathbf{X}_i^* = \mathbf{W}_i\,\mathbf{X}, \quad \boldsymbol{\epsilon}_i^* = \mathbf{W}_i\,\boldsymbol{\epsilon}_i$$

which transforms the problem into one with uncorrelated errors. The GLS estimator is

$$\hat{\boldsymbol{\beta}}_i^{\mathrm{GLS}} = \left(\mathbf{X}_i^{*\top}\,\mathbf{X}_i^*\right)^{-1}\mathbf{X}_i^{*\top}\,\mathbf{y}_i^*, \tag{3.2}$$

having variance $\sigma^2\left(\mathbf{X}_i^{*\top}\,\mathbf{X}_i^*\right)^{-1}$. In practice, $\boldsymbol{\Omega}_i$ is unknown and must be estimated, which is the core focus of this work.

3.2.2 Estimation of Autocorrelation Matrix $\hat{\boldsymbol{\Omega}}_i$

Modelling the covariance matrix $\boldsymbol{\Omega}_i$ for fMRI presents particular challenges. First, the autocorrelation structure is complex and not adequately captured by simple models, and efforts are hindered by relatively short time series, often comprising only 100 to 200 observations. Second, it is not the raw data that are to be modelled but residuals of highly parameterised models; dependencies induced by the residual-forming process can substantially bias estimates of $\boldsymbol{\Omega}_i$. Thirdly, as the number of voxels for a subject can exceed hundreds of thousands, any whitening methods must be scalable and automated. Lastly, and uniquely to this time series application, precise autocorrelation estimates are required to support multiple-testing-corrected inferences over voxels. For example, while unbiased estimates of $\boldsymbol{\Omega}_i$ might be obtained, large sampling variability in $\hat{\boldsymbol{\Omega}}_i$ could corrupt the null distribution of test statistics t_i. Since

multiple testing corrections rely on accurate p-values far below the conventional α-level (e.g. 5%), even modest uncertainty in $\hat{\mathbf{\Omega}}_i$ can be problematic.

A feasible generalised least squares (fGLS) approach offers a practical solution by iteratively estimating the covariance structure. Initially, ordinary least squares (OLS) residuals, $\hat{\boldsymbol{\epsilon}}_i = \mathbf{y}_i - \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}_i$, are used to approximate $\hat{\mathbf{\Omega}}_i$. These estimates refine the parameter estimates via:

$$\hat{\boldsymbol{\beta}}_{\text{fGLS}} = (\mathbf{X}^\top\hat{\mathbf{\Omega}}_i^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\hat{\mathbf{\Omega}}_i^{-1}\mathbf{y}_i.$$

fGLS can be applied globally or locally. In global estimation, a single covariance structure $\mathbf{\Omega}$ is assumed across the brain, typically modelled as stationary (e.g. an autoregressive process), pooling information from all voxels. For local estimation, voxel-specific covariance matrices $\mathbf{\Omega}_i$ are derived independently, accommodating heterogeneity in autocorrelation across voxels. These approaches can be integrated, with global estimates providing initial parameter values (e.g. AR order) for local estimation.

In response to these challenges, several strategies are commonly used in fMRI time series modelling: likelihood-based methods, methods of moments, and non-parametric methods. Likelihood-based methods are effective at accounting for implicit dependencies in residuals, assuming the error term in Eq. (3.1) follows an underlying autoregressive (AR) or autoregressive moving average (ARMA) process. For example, the first-order autoregressive moving average process, ARMA(1,1), for voxel i is:

$$\boldsymbol{\epsilon}_{i,t} = \phi_i\boldsymbol{\epsilon}_{i,t-1} + \boldsymbol{\eta}_{i,t} + \zeta_i\boldsymbol{\eta}_{i,t-1}, \quad \boldsymbol{\eta}_i \sim \mathcal{N}(0,1), \tag{3.3}$$

where ϕ and ζ are the autoregressive and moving average coefficients, respectively. The correlation matrix $\mathbf{\Omega}_i$ of such a process is a Toeplitz matrix of size $T \times T$, where the corresponding diagonal of lag ℓ, $\Omega_{t,t\pm\ell}$, is given as (Box et al., 2015):

$$\begin{aligned}\omega_{i,\ell} &= \phi_{i,\ell-1}\omega_{i,1},\\ \omega_{i,\ell} &= \frac{(\phi_i + \zeta_i)(1 + \phi_i\zeta_i)\phi_{i,\ell-1}}{1 + 2\phi_i\zeta_i + \zeta_i^2}.\end{aligned}$$

This reduces to an AR(1) model if $\zeta = 0$. Knowing the underlying process in $\mathbf{\Omega}_i$,

and assuming normality, parameters of Eq. (3.3) can be jointly estimated with β by maximising the log-likelihood function over the parameter space $\theta = (\beta, \phi, \zeta)$:

$$\ell_{\mathrm{ML}} = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Omega}_{i,\theta}) - \frac{T}{2}\log(\sigma_i^2) - \frac{1}{2\sigma_i^2}(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i)^\top \boldsymbol{\Omega}_{i,\theta}^{-1}(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i).$$

However, for regression models involving time series, such likelihood estimates can exhibit substantial bias in finite samples (Cheang and Reinsel, 2000), as they do not account for the estimation of β.

3.2.2.1 Restricted Maximum Likelihood

Alternatively, Restricted Maximum Likelihood with non-identity covariance noise is suggested to reduce such biases where parameters of Eq. (2.4) is estimated by minimising the following restricted likelihood (Cheang and Reinsel, 2000; Harville, 1974)

$$\begin{aligned}\ell_{\mathrm{ReML}} = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Omega}_i) - \frac{T-Q}{2}\log(\sigma_i^2) \\ - \frac{1}{2\sigma_i^2}(\mathbf{y}_i - \mathbf{X}\hat{\boldsymbol{\beta}}_i^{\mathrm{GLS}})^\top \boldsymbol{\Omega}_{i,\theta}^{-1}(\mathbf{y}_i - \mathbf{X}\hat{\boldsymbol{\beta}}_i^{\mathrm{GLS}}) - \frac{1}{2}\log\det(\mathbf{X}^\top \boldsymbol{\Omega}_i^{-1}\mathbf{X}),\end{aligned}$$

where $Q = \mathrm{rank}(\mathbf{X})$ and $\hat{\boldsymbol{\beta}}_i^{\mathrm{GLS}}$ is estimate via Eq. (3.2). Despite the accuracy that ReML-based methods offer, they are not scalable in a sense that for higher order autoregressive process the parameter space expands astronomically whilst it is also computationally intensive to fit such models for more than a couple of hundreds of time series unless one constrains the parameter search to a low resolution ϕ and ζ admissible range, as done in AFNI.

3.2.2.2 Methods of Moments

Alternatively, methods of moments approaches directly estimate the covariance matrix $\hat{\boldsymbol{\Omega}}$ from residuals $\hat{\boldsymbol{\epsilon}}_i = \mathbf{R}\mathbf{y}_i$ as

$$\mathbf{R} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top,$$

assuming a parametric autoregressive model without using maximum likelihood optimisation. Yule-Walker is a widely-used methods-of-moments approach to estimate AR coefficients: For a given L^{th}-order autoregressive model, $r_{i,t} = \sum_{\ell=1}^{L} \phi_{i,\ell} r_{i,t-\ell} + \eta_{i,t}$ with $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I}\alpha^2)$, it can be defined as a Topelitz system of,

$$\boldsymbol{\omega}_i = \boldsymbol{\Psi}_i \boldsymbol{\phi}_i,$$

where $\boldsymbol{\Psi}_i$ is a $L \times L$ Toeplitz matrix of 0^{th} lag to $(L-1)^{th}$ lag autocorrelation coefficients and ω_i is a vector of 1^{st} lag to L^{th} lag autocorrelation coefficient.

Knowing that $\tilde{\boldsymbol{\Psi}}_i$ is full-rank and symmetric, its invertability is guaranteed hence, the AR coefficients up to order L can be obtained via $\boldsymbol{\Psi}_i^{-1} \boldsymbol{\omega}_i = \hat{\boldsymbol{\phi}}_i$. We note that parameters of ARMA models can also be estimated using methods of moments (Hannan and Rissanen, 1982; Hannan and McDougall, 1988), but since an AR model of sufficient order L can approximate any ARMA model (Box et al., 2015), we focus exclusively on AR models in our development.

3.2.2.3 Nonparamaetric

Lastly, in non-parametric methods, the whitening matrix is directly estimated using autocorrelation of the error term, $\boldsymbol{\epsilon}_i$, as a Toeplitz matrix of $T \times T$ where the ℓth diagonal is the ℓ-lag autocorrelation, ω_ℓ, of error at voxel i, $\boldsymbol{\epsilon}_i$, at lag ℓ;

$$v_{i,\ell} = \mathbb{E}[\boldsymbol{\epsilon}_{i,t} \boldsymbol{\epsilon}_{i,t-\ell}], \quad \omega_{i,\ell} = v_{i,\ell} / v_{i,0}, \tag{3.4}$$

where $\omega_{i,\ell} = 0$ for $\ell > L$. It is important to note that if not using likelihood methods, the true error term is not known and instead the autocorrelation matrix is estimated from the *residuals* $\hat{\boldsymbol{\epsilon}}_i$. The whitening process is often conducted in the Fourier domain using autocorrelation function (ACF) of the residuals. where the model in Eq. (2.4) is convolved with a prewhitening filter, $h(t)$.

3.2.2.4 Low Frequency Drift Models

A large share of the noise present in fMRI time series is slowly varying with $1/f$ spectral characteristic (Aguirre et al., 2002). Despite the unknown origin of such

noise, it greatly contributes to the autocorrelation present in the data. In theory, a single stochastic model could be used to fit all types of noise. However, in practice, the amplitude of drift noise is so large and often non-stationary that a composite approach is preferred. This involves using a deterministic model (such as a linear model or line smoother) to account for most of the low-frequency drift, while an autocorrelation model on the remaining error captures any remaining stationary temporal dependencies.

While the implementation of low frequency drift modelling varies greatly across different fMRI packages; all can be presented as a non-symmetric $T \times T$ filter matrix $\mathbf{K} = \mathbf{I} - \mathbf{G}$ where $\mathbf{G}$ is the drift fitting matrix.

Three software packages uses a linear drift basis $\mathbf{D}$ (see Table 3.1) so that $\mathbf{G} = \mathbf{D}(\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top$.

Table 3.1: Different strategies used to form matrix $\mathbf{D}$ including splines, polynomials or Discrete Cosine Transforms (DCT); one package uses a high pass filter (HPF) to remove drift from the data (see body text). The repetition time (TR) is denoted δ.

Method	**Software**	**Order/Cut-Off**	**Reference**
DCT	SPM	$2(\delta T)/128 + 1$	Penny et al. (2011)
Polynomials	AFNI	$1 + (\delta T/150)$	Saad et al. (2006)
Splines	fMRIStat	$(3\delta T)/360$	Worsley et al. (2002)
HPF	FSL	100s	Marchini and Ripley (2000)

FSL uses a Gaussian running-line smooth very similar to LOESS (Cleveland et al., 2017). We can create a matrix representation $\mathbf{G}$ that makes a prediction for each time point $1 \dots T$.

Consider a simple linear regression model for Y,

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}_K\boldsymbol{\beta}_K,$$

where $\mathbf{X}_k$ is a $T \times 2$ matrix $[\mathbf{1} \;\; (1 \dots T)^\top]$. For time point t, consider a weighting function that puts varying weights on time points t^* given their distance; larger values nearer t, lower values farther from t, with a controlling parameter c. Marchini and Ripley (2000) uses a Gaussian weights, which we write as a $T \times T$ diagonal matrix

$\mathbf{W}_{K,t}$ with diagonal entries,

$$\mathbf{W}_{K,t} = \frac{1}{\sqrt{2\pi c^2}} e^{-(t-t^*)^2/c^2}.$$

Then then weighted regression for time t has linear estimates $\hat{\boldsymbol{\beta}}_k = (\mathbf{W}_{K,t}\mathbf{X}_K)^+\mathbf{W}_{K,t}\mathbf{Y}$ and fitted value $\mathbf{X}_k(\mathbf{W}_{K,t}\mathbf{X}_K)^+\mathbf{W}_{K,t}\mathbf{Y}$. As a result we can form the running line smoother matrix $\mathbf{G}$ with its t^{th} row defined as

$$\mathbf{g}_t = \mathbf{X}_K(\mathbf{W}_{K,t}\mathbf{X})^+\mathbf{W}_{K,t}.$$

3.2.2.5 Bias Correction for Design and Drift

Methods of moments and nonparametric methods naively assume that residuals $\hat{\boldsymbol{\epsilon}}$ have the same dependency structure of the errors $\boldsymbol{\epsilon}$. Of course this is not the case as the residual covariance $\sigma_i^2\mathbf{R}\boldsymbol{\Omega}_i\mathbf{R}$ is not the same as the error's $\sigma_i^2\boldsymbol{\Omega}_i$. This may induce a bias in these methods as $\hat{\boldsymbol{\Omega}}_i$ will be capturing residual-induced structure.

Following Worsley (2000) and Worsley (2001), we propose a bias adjustment for autocovariances by equating the sample autocovariances to their approximate expectations while accounting for a temporal filter $\mathbf{K}$. In particular, we assume the residual forming matrix is $\tilde{\mathbf{R}} = \mathbf{R}\mathbf{K}$, and, crucially, we can no longer assume $\tilde{\mathbf{R}}$ is symmetric and idempotent. First, noting that the ℓ^{th}-lag autocovariance depends on the sum of all such lags, $\tilde{\boldsymbol{\epsilon}}_i^\top\mathbf{H}_\ell\tilde{\boldsymbol{\epsilon}}_i$, where $\mathbf{H}_\ell$ is a diagonal selector matrix with 1's on the ℓ^{th} upper diagonals and has the following expectation,

$$\begin{aligned}\mathbb{E}(\tilde{\boldsymbol{\epsilon}}_i^\top\mathbf{H}_\ell\tilde{\boldsymbol{\epsilon}}_i) &= \sigma^2\,\mathrm{tr}(\tilde{\mathbf{R}}^\top\mathbf{H}_\ell\tilde{\mathbf{R}}\boldsymbol{\Omega}_i) \\ &= \mathrm{tr}(\tilde{\mathbf{R}}^\top\mathbf{H}_\ell\tilde{\mathbf{R}})\omega_0 + \sum_{j=1}^{L}\mathrm{tr}(\tilde{\mathbf{R}}^\top\mathbf{H}_\ell\tilde{\mathbf{R}}(\mathbf{H}_j + \mathbf{H}_j^\top))\omega_\ell\end{aligned} \tag{3.5}$$

where the last equality arises from the fact that $\sigma^2\boldsymbol{\Omega}_i = \omega_0\mathbf{I} + \sum_{l=1}^{T-1}(\mathbf{H}_\ell + \mathbf{H}_\ell^\top)\omega_{i,\ell}$.

Equation 3.5, can be redefined in matrix form as $\hat{\boldsymbol{\omega}} = \mathbf{M}^{-1}\mathbf{a}$ where the estimated AR(L) sample autocovariance of the residuals $\mathbf{a}$ is given by $a_\ell = \sum_{k=\ell+1}^{\top}\tilde{r}_{i,k}\tilde{r}_{i,k-\ell}$ the

transformation matrix $\mathbf{M}$ is a symmetric invertible $L+1 \times L+1$ matrix where

$$m_{\ell j} = \begin{cases} \text{tr}(\tilde{\mathbf{R}}^\top \mathbf{H}_\ell \tilde{\mathbf{R}}), & j = 0. \\ \text{tr}(\tilde{\mathbf{R}}^\top \mathbf{H}_\ell \tilde{\mathbf{R}}(\mathbf{H}_j + \mathbf{H}_j^\top)), & 1 \leq j \leq L. \end{cases} \tag{3.6}$$

Following Eq. (3.4), the autocorrelation at lag ℓ is obtained by normalising the autocovariances with their value on lag 0; $\tilde{\omega}_\ell = \hat{\omega}_\ell / \hat{\omega}_0$.

3.2.2.6 Global-plus-Local Autocorelation Estimation

We propose a tissue specific fGLS procedure to partially whiten model in Eq. (2.4) where for each tissue type, a non-parametric $\mathbf{\Omega}_i$ is estimated, followed by voxel-specific autocorrelation estimation. The autocorrelation estimation of both the tissue and voxel-specific components is bias adjusted (Section 3.2.2.5) while incorporating the high pass filter (Section 3.2.2.4). For voxel i in tissue s, the model in Eq. (2.4) is

$$\tilde{\mathbf{W}}_{i,s} \tilde{\mathbf{W}}_{0,s} \mathbf{y}_{i,s} = \tilde{\mathbf{W}}_{i,s} \tilde{\mathbf{W}}_{0,s} \mathbf{X} \boldsymbol{\beta}_i + \tilde{\mathbf{W}}_{i,s} \tilde{\mathbf{W}}_{0,s} \boldsymbol{\epsilon}_i, \tag{3.7}$$

where $\tilde{\mathbf{W}}_{0,s}$ is a *global* whitening matrix for tissue s where the pooled autocorrelation across all voxels within tissue s is corrected via Eq. (3.5); specifically, OLS residuals are used to estimate autocorrelation at each voxel, which is then pooled for each lag ℓ over voxel. Similarly, $\tilde{\mathbf{W}}_{i,s}$ is the *local* whitening matrix where each time series of voxel i is whitened by adjusted ω_ℓ. Although it is common to reiterate the whitening until $\boldsymbol{\beta_i}$ converges (Cochrane and Orcutt, 1949), we only repeat the process twice (once globally and once locally) due to computational constraints.

Similar to Woolrich et al. (2001), Chatfield (2016), and Afyouni et al. (2019), we taper the autocorrelation to attenuate the effect of farther lags autocorrelations. We arbitrarily choose Tukey window to scale the adjusted autocorrelations by $\frac{1}{2}(1+\cos(\frac{\pi l}{M}))$ for $\ell <= M$ and zeroed for $k > M$. We validate the results across a range of M, including a case where M is chosen optimally as the first non-significant lag of voxels with highest autocorrelation length (Afyouni et al., 2019).

3.2.2.7 Computationally Efficient Whitening

Finding the whitening matrix for each voxel is computationally intensive, and can be substantially accelerated by use of Fourier methods. Let $\omega_{i\ell} = (\Omega_i)_{t,t\pm\ell}$ be the autocorrelation at voxel i for lag ℓ, and $\boldsymbol{\omega}_i = (1, \omega_{i2}, \ldots, \omega_{iT-1})^\top$ be the autocorrelation function (ACF) at voxel i.

Since the power spectral density (PSD) is related to the autocorrelation function by the Wiener-Khinchin theorem, the Fourier transform of $\boldsymbol{\omega}_i$ yields the PSD for the noise process at voxel i. Whitening can be carried out in the Fourier domain as follows. Let $\mathbf{x} = (x_0, \ldots, x_{T-1})$ be the length-T time series, and let $\mathbb{F}$ denote the discrete Fourier transform (DFT). The DFT of $\mathbf{x}$ is given by

$$\mathbb{F}[\mathbf{x}]_f = \sum_{\ell=0}^{T-1} x_\ell e^{-\frac{i2\pi f\ell}{T}}, \quad f = 0, 1, \ldots, T-1.$$

Whitening is carried out with weights $\mathbf{w}$ to scale each frequency f, such that the whitened signal is $\mathbb{F}^{-1}[\mathbf{w}\mathbb{F}[\mathbf{x}]]$.

The PSD is the squared modulus of the Fourier transform of the noise component $\boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i$ represents the i-th realisation of the noise process or residuals. The Wiener-Khinchin theorem states that the PSD and the ACF are Fourier pairs:

$$S_i(f) = \mathbb{E}\left(|\mathbb{F}(\boldsymbol{\epsilon}_i)|^2\right).$$

Hence, for a given ACF, we can compute $S_i(f)$, and the whitening weights are set by

$$w_{if} = S_i^{-1/2}(f).$$

We define a global whitening filter as $P(f)_{0,s} = \frac{1}{\sqrt{S(f)_{0,s}}}$, where $S(f)_{0,s}$ is the robust average of the PSDs of the residuals, obtained from the Fourier transform of their autocorrelation functions across all voxels. Similarly, we define the local whitening filter for voxel i as $P(f)_{i,s} = \frac{1}{\sqrt{S(f)_{i,s}}}$, where $S(f)_{i,s}$ is the PSD of the residuals at voxel i, computed as the Fourier transform of the globally whitened residuals.

3.2.3 Evaluations

We evaluate the performance of SALSA and other pre-whitening methods across multiple levels of analysis, focusing on autocorrelation patterns, spectral properties, and statistical robustness. To examine autocorrelation, we fit a mass univariate linear model using OLS and analyse residuals across brain tissue types, smoothing levels, and sampling rates. Key metrics include power spectra, autoregressive coefficients, and autocorrelation lengths, which enable a detailed assessment of the dependency structure in fMRI data.

To assess noise characteristics and statistical reliability, we analyse the spectral properties of residuals from different pre-whitening methods. This includes comparing power spectra and spectral flatness to evaluate the conformity of noise to white noise assumptions. False positive rates (FPR) are evaluated under the null hypothesis of no regressor effect to quantify the robustness of pre-whitening methods in suppressing autocorrelation and achieving reliable statistical inference.

At the subject level, we test sensitivity by superimposing synthetic task signals on resting-state fMRI data. This analysis measures the precision and recall of task signal detection across datasets with varying conditions, including temporal smoothing. We compare Z-statistics from SALSA and other pre-whitening methods using Tukey mean-difference plots to identify proportional biases and examine inter-method agreement.

For group-level analysis, we compare pre-whitening methods on task-fMRI datasets by evaluating Z-statistics, suprathreshold voxel percentages, and spatial distributions of extreme Z-scores. These analyses aim to quantify inter-method agreement and identify discrepancies in detecting task-related signals, particularly in extreme cases. By integrating insights across subject- and group-level evaluations, we aim to provide a comprehensive characterisation of pre-whitening methods.

3.2.3.1 Data

We conduct extensive empirical analyses using 1,412 resting-state scans across 8 datasets as described in Table 3.2, with sampling frequency varying from 0.5Hz to 2.5Hz to confirm that SALSA outperforms all existing methods in flattening the noise spectrum and obtaining a given false positive rate, especially in rapidly sampled fMRI scans. The datasets include resting-state and task-based fMRI data from various

studies, each contributing to a diverse collection of neuroimaging data, ensuring the robustness and generalisability of our findings.

Table 3.2: Summary of datasets used in this study, including the type of fMRI, number of subjects, number of scans per subject, repetition time (TR) and the source.

Dataset	**Type**	**#Subjects**	**#Scans per Subject**	**TR**	**Source**
F-CON1000 Beijing	resting-state	82	225	2s	F-CON1000: Beijing_Zang (Zang, Y.F.)
NKI Rockland	resting-state	557	404	1.4s	Nooner et al. (2012)
NKI Rockland	resting-state	557	900	0.645s	Nooner et al. (2012)
NKI Rockland (Checkerboard)	task	60	240	0.645s	Nooner et al. (2012)
HCP	resting-state	51	1200	0.753s	Van Essen et al. (2013b)
HCP (Motor Left Hand)	task	25	284	0.753s	Van Essen et al. (2013b)
HCP (Gambling Reward)	task	25	253	0.753s	Van Essen et al. (2013b)
dHCP	resting-state	55	2300	0.392s	Edwards et al. (2022)

The Nathan Kline Institute-Rockland Sample (NKI-RS) (Nooner et al., 2012) is an ongoing study aimed at creating a large-scale community sample of participants across the lifespan. It includes resting-state fMRI data with repetition times of 1400 ms and 645 ms from 557 participants, as well as task-based fMRI (e.g. Checkerboard task) with a TR of 645 ms. The dataset spans a broad age range (6 to 85 years) and offers extensive phenotypic information, facilitating research into brain-behaviour relationships.

The Human Connectome Project (HCP) dataset (Van Essen et al., 2013b) is a major initiative focused on mapping the neural pathways of the human brain. It includes high-resolution neuroimaging data from 1,200 healthy young adults, aged 22 to 35 years, with resting-state fMRI at a TR of 720 ms and task-based fMRI for tasks such as Motor (Left Hand) and Gambling Reward, both with TRs of 753 ms. This dataset also provides extensive behavioural, demographic, and genetic data, supporting studies of brain structure and function.

The Developing Human Connectome Project (dHCP) dataset (Edwards et al., 2022) aims to create the first 4-dimensional connectome of early life, including neuroimaging data from 55 infants, ranging from preterm to full-term neonates, with a TR of 392 ms. The data, collected from 20 to 44 weeks post-conceptional age, integrates imaging with clinical, behavioural, and genetic information, allowing for in-depth studies of early brain development, particularly in infants at risk for conditions such as Autistic Spectrum Disorder or Cerebral Palsy.

The F-CON1000 Beijing dataset, part of the 1000 Functional Connectomes Project (FCP), includes resting-state fMRI data from 82 subjects, with a TR of 2 seconds. This dataset facilitates comparative studies across different populations and sites, contributing valuable data to the investigation of functional brain connectivity.

3.3 Results

3.3.1 Autocorrelation Across Choice of Smoothing, Tissue Types and Sampling Frequencies

In order to evaluate the effect of autocorrelation across different brain tissue types we fit a mass univariate linear model using ordinary least square where the dependent variable is resting state and the independent variable is event related task randomised across subjects. Figure 3.1 shows the extend of autocorrelation on residuals of model fit via OLS via power spectrum, autoregressive model coefficients and autocorrelation lengths.

Knowing Wiener-Kinchin theorem which suggests that low frequency inflation on a spectrum indicates autocorrelation, Figure 3.1.A-D suggest that the presence of autocorrelation grows as the sampling frequency increases. However, the presence of the autocorrelation is not similar across tissue types; with CSF and grey matter suggesting a higher autocorrelation than white matter. Further, the smoothing is shown to consistently worsen the presence of the autocorrelation.

Further, the distribution of dependencies across tissue types and the effect of smoothing is confirmed on Figure 3.1.E-H where the first-order autoregressive coefficients of residuals are shown on an arbitrary brain slice averaged across 50 subjects for each of the four cohorts.

Lastly, in Figure 3.1.I-L we use autocorrelation length to show that the effect of smoothing and tissue types is not limited to early lags of the autocorrelation.

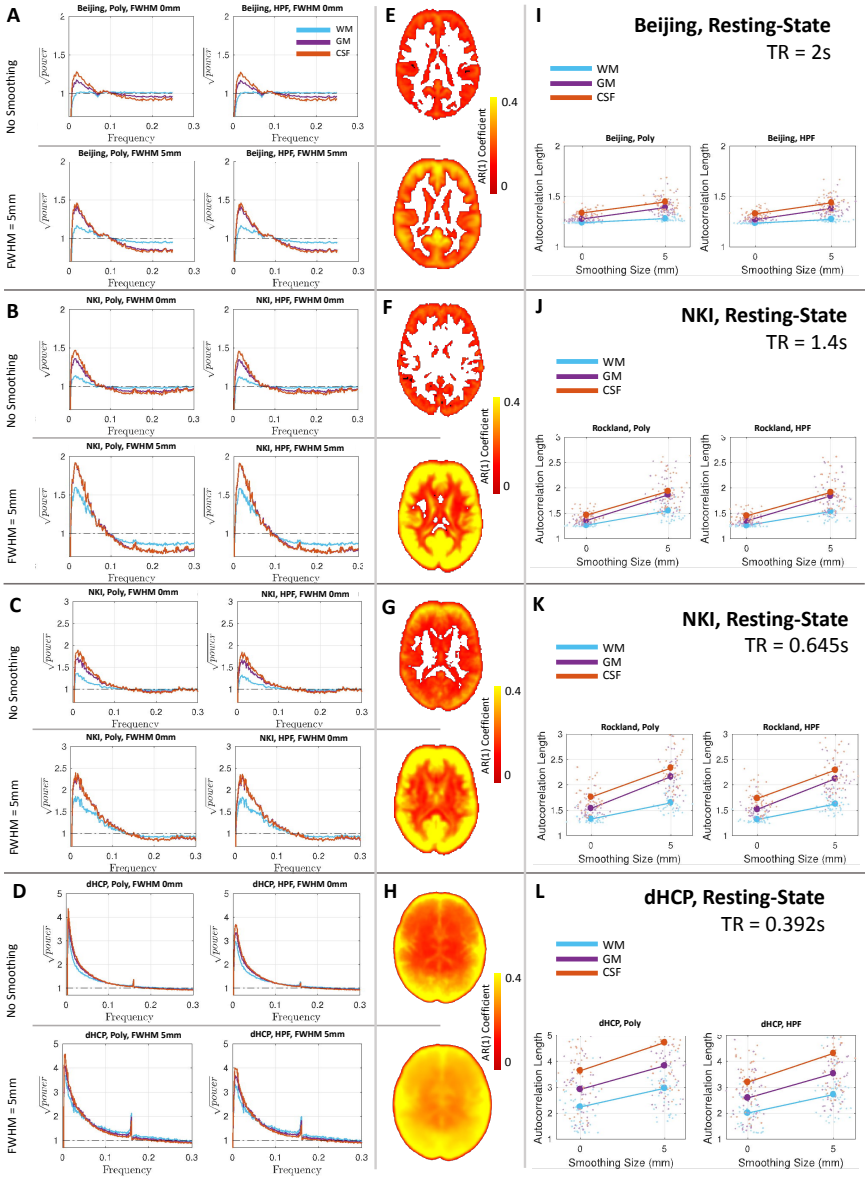


Figure 3.1: Difference in presence of autocorrelation across brain tissues and smoothing level across four different data sets; ordered from slowly sampled (Beijing) to rapidly sampled (dHCP) fMRI time series. Consistent across four data sets; CSF and grey matter demonstrates highest presence of autocorrelation. The autocorrelation presence increases when data spatially smoothed.

3.3.2 Spectrum and False Positive Rate of Whitening Methods

To better understand the characteristics of the noise process in fMRI data, we initially examine the residuals obtained from the OLS model fit for each dataset. Specifically, we apply five different pre-whitening methods, including SALSA, fit the GLM, and

then analyse the resulting residuals' power spectra.

The noise can be effectively assessed by examining its power spectrum and spectral flatness. Spectral flatness, defined in terms of the geometric mean and arithmetic mean of the power spectrum, characterises the deviation of the signals from true white noise behaviour. More specifically, it is given by

$$\text{Flatness} = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)} \tag{3.8}$$

where $x(n)$ represents the magnitude of bin number n. A white noise process has flatness of 1, opposed to a pure tone which has flatness of 0.

Figure 3.2 presents the power spectra and the average power spectra across all brain voxels and subjects based on each pre-whitening method and a Naive model for each dataset. The Naive model corresponds to fitting a OLS model directly onto the data without pre-whitening the noise. The power spectrum plots illustrate apparent increase in the strength of autocorrelation across all pre-whitening methods as smoothing is applied and as we increase the sampling rate. Notably, at lower frequencies, a spike is observed across many models and datasets, particularly evident in the Naive model. However, as the frequency increases, most models, except Naive, tend to exhibit flatness around 1, indicating the imposed assumption of white noise for pre-whitening methods, regardless of the sampling rate.

In the context of temporal smoothing, a smoothing method such as Full Width Half Maximum (FWHM) is commonly employed to obtain less noisy fMRI signals. We find that it may introduce stronger autocorrelation in the noise of the data, as observed from the higher peaks in the power spectra. This applies to all existing pre-whitening methods. However, our proposed method, SALSA, demonstrates tolerance to such smoothing, making it an advantageous choice as a pre-whitening technique in conjunction with smoothing procedures.

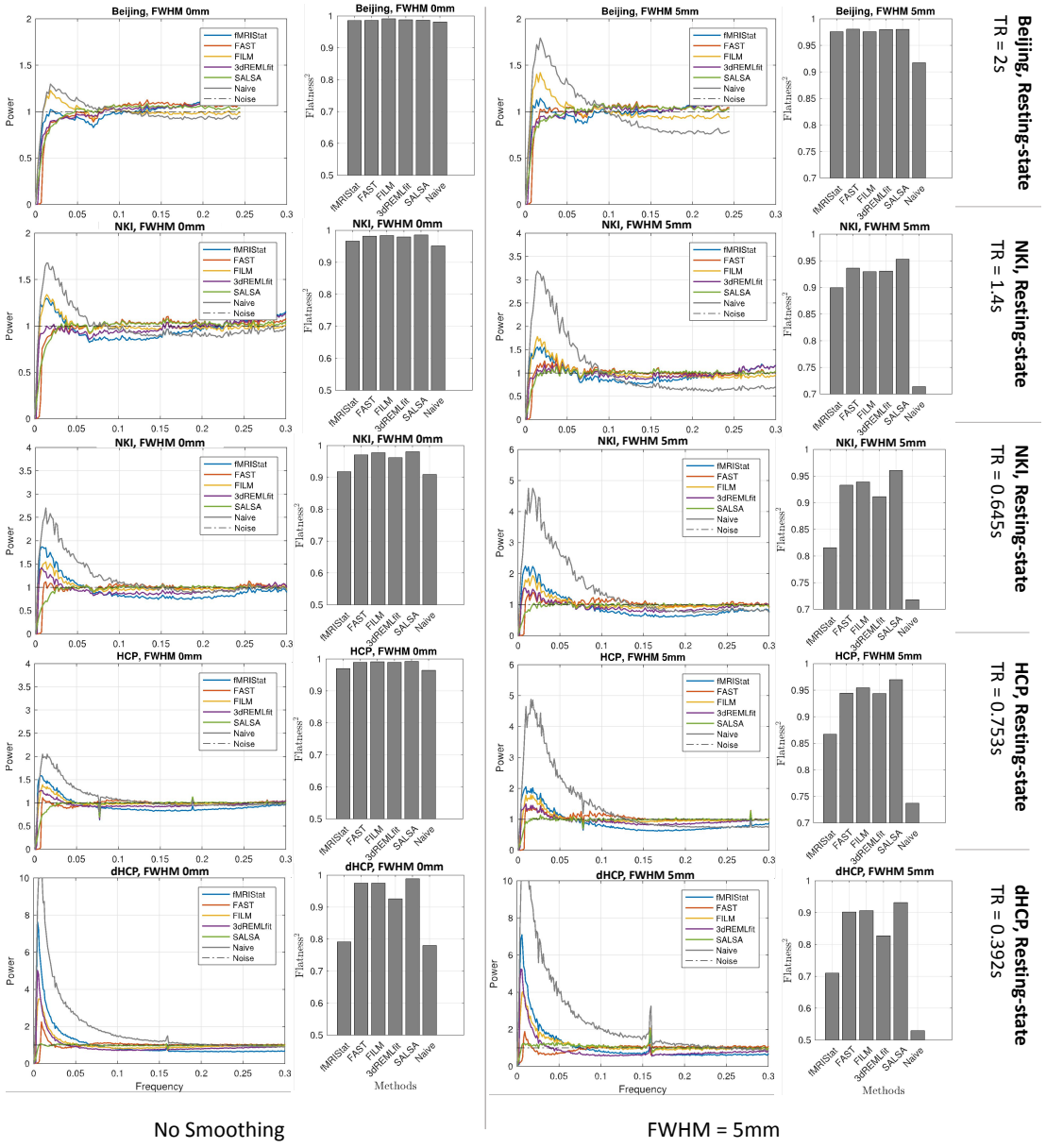


Figure 3.2: Power spectrum (odd columns) and flatness (even columns) of noise across four data sets; from slowly sample data set (Beijing, first row) to rapidly sampled data set (dHCP, last row). SALSA (green power spectrum) consistently outperforms other existing methods.

We can evaluate the False Positive Rate (FPR) where the null hypothesis assumes a zero regressor coefficient at a given voxel, which corresponds to no signal being explained by the regressor. Inflated FPR in fMRI analyses are widely acknowledged as a significant issue. Although there is limited literature on the topic of mitigating this problem with pre-whitening techniques, it's gaining popularity because higher

sampling rates are consistently associated with higher FPR, undermining the reliability of statistical analyses.

In Figure 3.3 A-E, we present the results of our FPR assessments. We observe that both smoothing and higher sampling rates tend to increase the FPR across most methods. However, SALSA stands out by consistently achieving the target FPR in most cases, demonstrating its robustness compared to other methods.

Furthermore, in Figure 3.3 F-G, we compare SALSA with existing methods in terms of suppressing the autocorrelation length. The outcomes reveal that SALSA outperforms these methods by effectively reducing the autocorrelation, resulting in a more homogeneously whitened noise across the voxels.

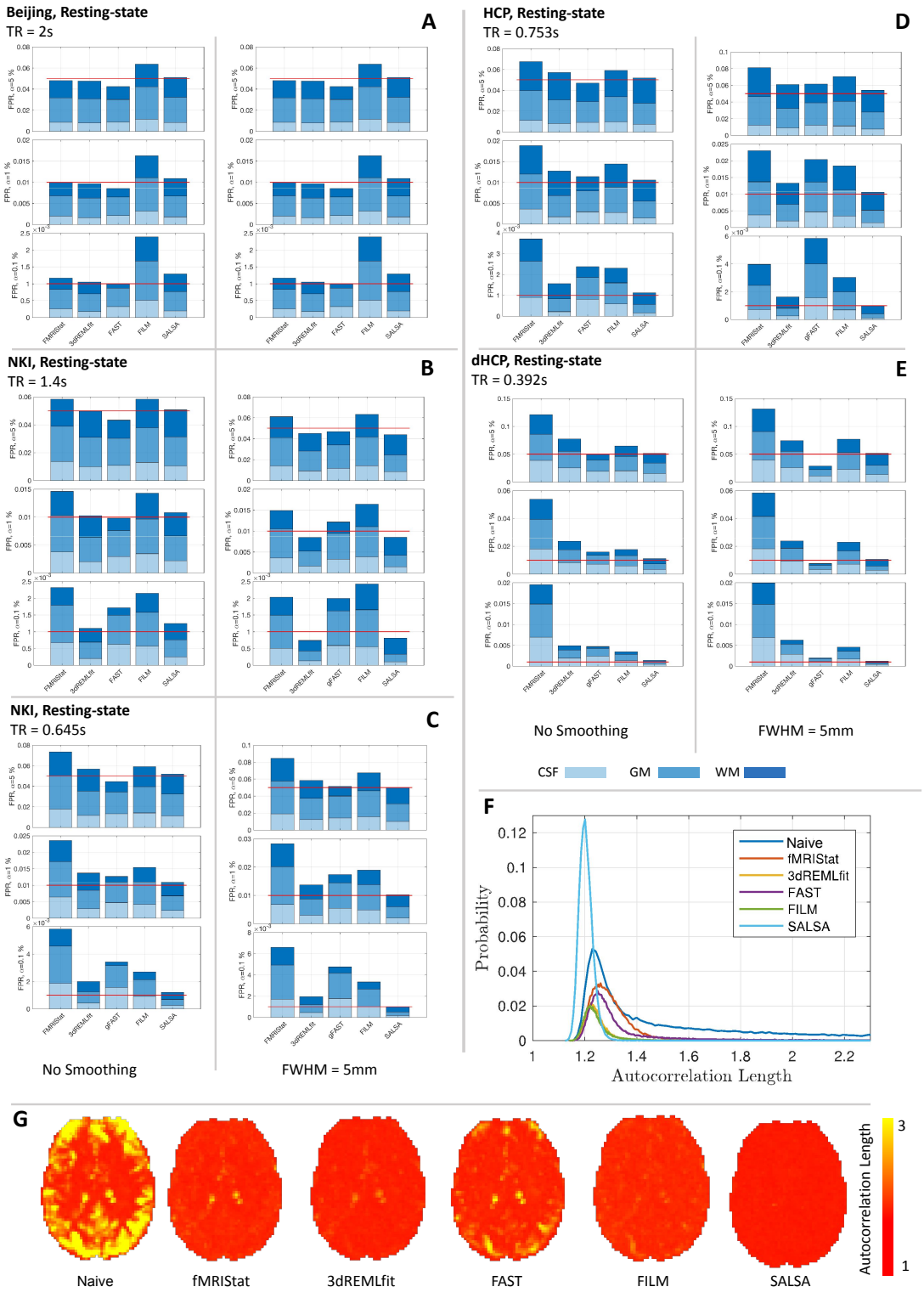


Figure 3.3: The false positive rate across three brain tissues (CSF,GM and WM) and four different data sets; slowly sampled data set (Beijing, A) to rapidly sampled data set (dHCP, C). SALSA almost consistently delivers desired α-level (horizontal red line). Histogram of ACL of noise time series (F) across 6 existing methods. SALSA outperforms existing methods in suppressing the autocorrelation length. The spatial distribution of ACL across existing methods (G) indicating SALSA results in a homogeneously whitened noise.

3.3.3 Subject Level Analysis and Sensitivity

An essential improvement showcased by SALSA is its ability to enhance the reliability of subject-level analysis, which is crucial, especially in pre-surgical applications of fMRI. In particular, we examine the efficacy of each pre-whitening method on the dHCP and HCP datsets. The analysis of infant brain fMRI data is known to be challenging, and thus, our aim is to show that SALSA can outperform the other existing methods even in a challenging condition.

To evaluate the sensitivity of signal detection after applying five different whitening methods, we treat the resting-state fMRI data from dHCP as the null and superimpose synthetic event signals to emulate task-fMRI signals. Next, we apply high-pass filtering to remove low-frequency components. In Figure 3.4 A, we present the pipeline used to construct these induced signals on the real resting-state fMRI data at a given voxel. This induced task signal design enables us to assess the pre-whitening performance based on sensitivity, as illustrated in Figure 3.4 B.

The results demonstrate that SALSA performs on par with fMRIStat and FILM and even outperforms FAST and 3dREMLfit in terms of sensitivity. However, upon closer examination, we discover more revealing insights. Figure 3.4 C displays the induced task signal characterized by spherical Gaussian signals with a cutoff point in a cubic shape, while Figure 3.4 D shows the detected voxels after applying four different pre-whitening methods.

Notably, both fMRIStat and FILM exhibit numerous false detections outside the designated box, indicating that their moderate sensitivity comes at the cost of a high false positive rate. In contrast, SALSA demonstrates considerably fewer false detections, highlighting its superiority in both precision and recall.

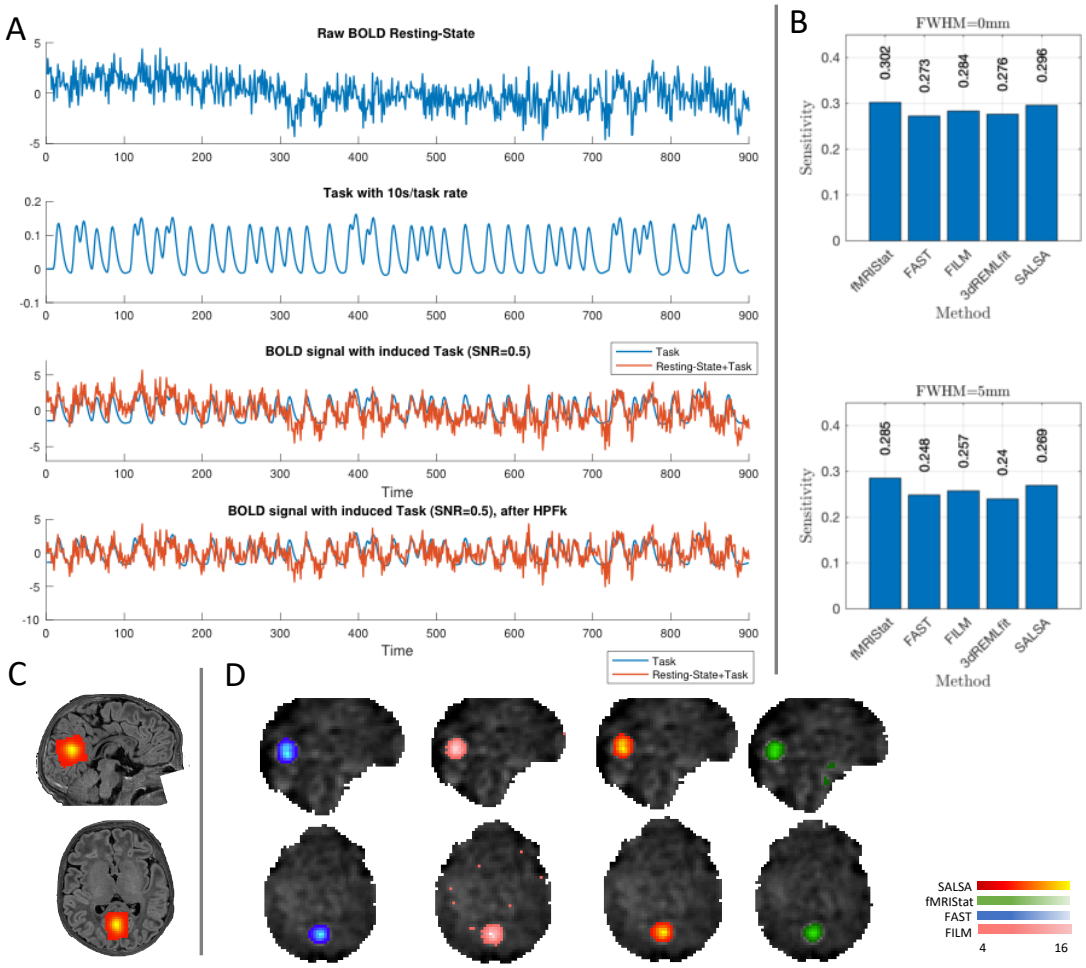


Figure 3.4: Illustration of the process to evaluate sensitivity of signal detection after applying various whitening methods on developing HCP dataset. (A) Depicts the construction of induced task signals on actual resting-state fMRI data, (B) Demonstrates how the sensitivity of detecting the induced task signals after applying distinct pre-whitening methods with/without data smoothing. (C) Presents the induced task signal characterised by spherical Gaussian signals with a cutoff point in a cubic shape, while (D) Displays detected voxels post application of distinct pre-whitening methods, revealing SALSA's superiority in precision and recall compared to fMRIStat and FILM.

We can investigate the agreement of derivation by SALSA and 4 other well-established pre-whitening methods. Figure 3.5 shows the Tukey mean-difference plots of Z-scores derived by SALSA against the other methods across three different t-fMRI dataset.

Comparing SALSA to FAST, we observe that the mean difference line between the two methods is approximately 0, indicating a consistent agreement between SALSA

and FAST. However, we can see that as the magnitude of the measurements become large, the variability of the difference between two methods also become large. This indicates a potential proportional bias between the methods.

In contrast, when comparing SALSA to fMRIStat and FILM, we see that SALSA consistently yields lower the magnitude of Z-statistics compared to the two methods. This concludes that SALSA tends to result in more conservative estimates compared to the two methods. This is well illustrated in the NKI Checkerboard example where we can clearly see the z-scores of FILM outshadow the z-scores of SALSA across several brain regions.

Lastly, comparing SALSA to 3dREMLfit shows a consistent agreement between the two methods, as observed from the flat mean difference line around 0. However, the case of HCP Gambling Reward dataset reveals that, despite having a flat mean difference line approximating 0, the substantial variability of the differences between the methods highlights the need for careful analysis.

In addition, we assess the effect of applying FWHM smoothing prior to the pre-whitening methods, which appears to encourage more extreme z-scores generally, i.e. higher autocorrelation, which can be observed from the brain-plotted z-scores across all three datasets. In particular, there are apparent improvements in methodological agreement for SALSA with FAST and 3dREMLfit according to the flatter mean difference line.

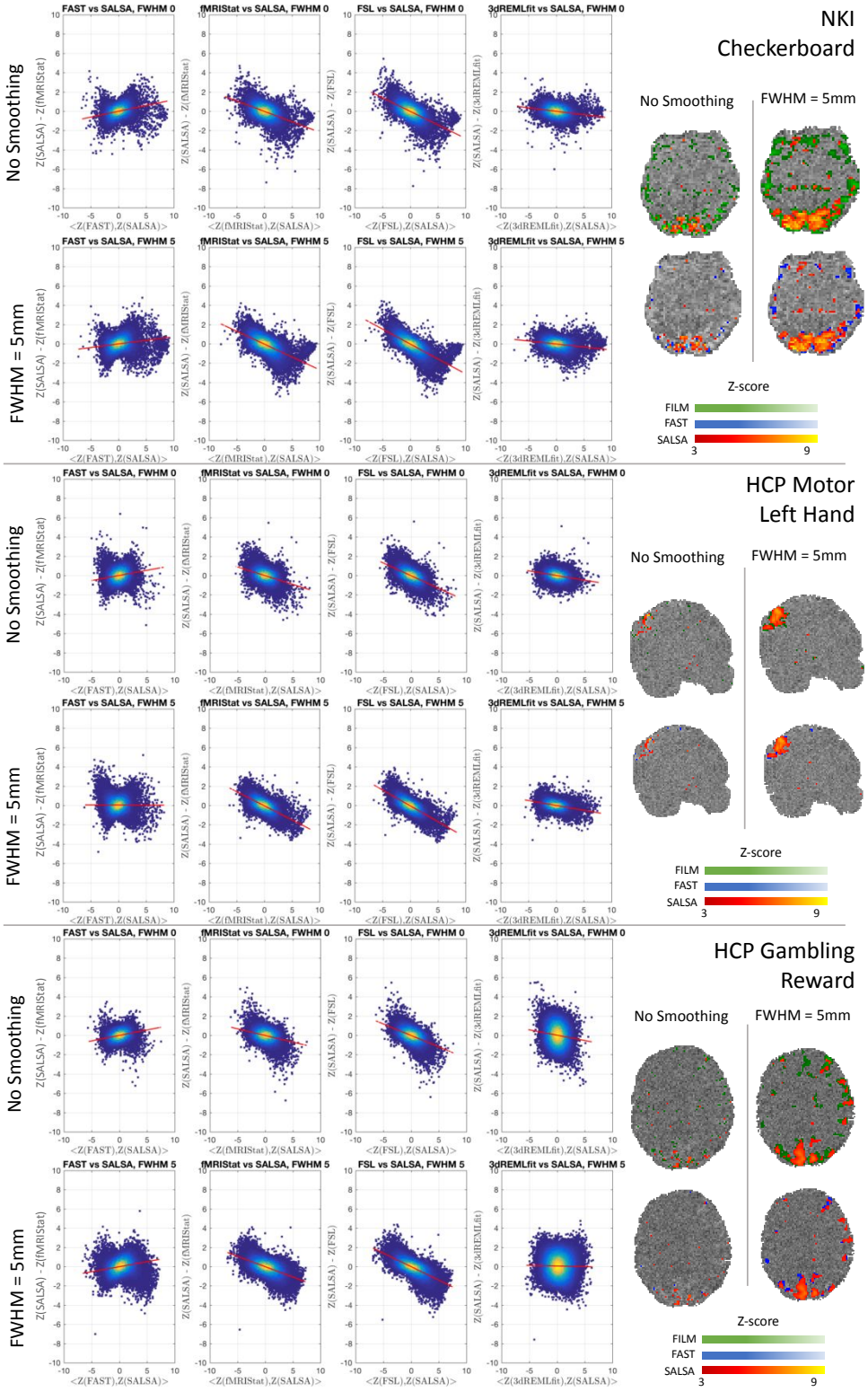


Figure 3.5: Comparative analysis of Z-score derivation between SALSA and other pre-whitening methods across diverse t-fMRI datasets. The Tukey mean-difference plots showcase the concordance and discrepancies in Z-scores among the methods. The comparisons encompass SALSA against FAST, fMRIStat and FILM, and 3dREMLfit, revealing distinct patterns of agreement and variability between these methods.

3.3.4 Group Level Task Analysis

Similar to the previous section, we can investigate the agreement between SALSA and the other four well-established methods on a group level. Figure 3.6 shows the Tukey mean-difference plots of Z-scores derived by SALSA against the other methods across three different t-fMRI dataset, along with the percentage of suprathresholded voxels and brain-plotted extreme Z-scores corresponding to each pre-whitening method.

In the context of NKI Checkboard dataset, we observe a clear agreement between SALSA and the other methods based on the mean difference line around 0. However, there is a slight note in the lower discrepancy variability among the high z-score, showing that they are similarly detect extreme values. The brain visualisations of thresholded z-scores show that all methods have similar region of detection, showing a consistent agreement across the brain in the group level.

On the contrary, the HCP Motor Left Hand dataset shows a spurious result for SALSA vs 3dREMLfit. Specifically, the right tail of the Tukey mean difference plot indicates that when considering the high extreme end of the z-scores derived by SALSA and 3dREMLfit, 3dREMLfit results in a lower z-score compared to SALSA for more extreme Z-scores, which shows a positive disagreement between the two methods.

Lastly, in the context of HCP Gambling Reward, FAST, fMRIStat and FILM all have consistent, low-varied agreement with SALSA, showing that SALSA is comparable to the well-established methods on the group level analysis. However, there is a clear disagreement between SALSA and 3dREMLfit which require further investigation.

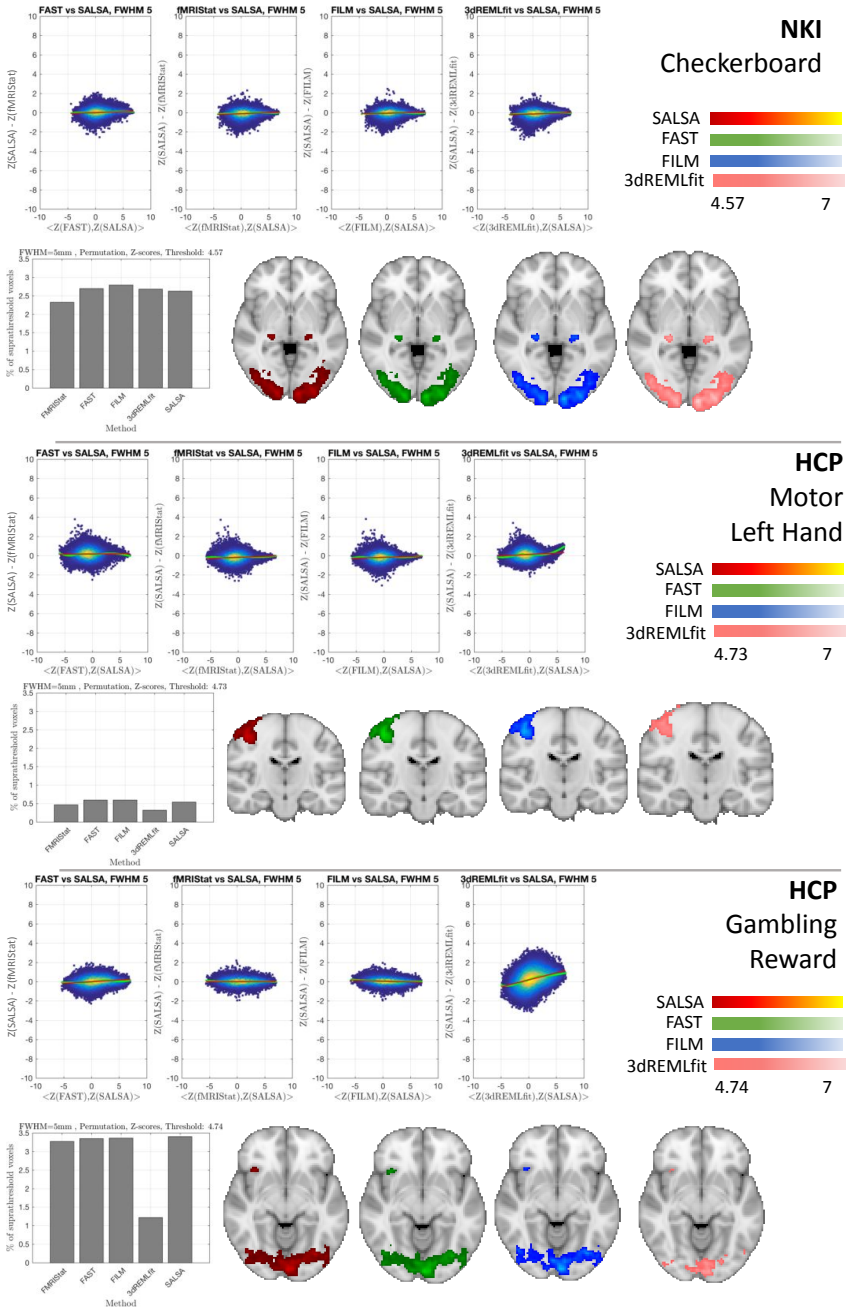


Figure 3.6: Tukey mean-difference plots of group-level Z-scores derived by SALSA and established pre-whitening methods are presented across three t-fMRI datasets. The figure also includes the percentage of suprathresholded voxels and brain-plotted extreme Z-scores for each method. The NKI Checkboard dataset showcases strong agreement between SALSA and other methods, with minimal variability in detecting extreme values. In the HCP Motor Left Hand dataset, a notable positive discrepancy emerges between SALSA and 3dREMLfit at extreme Z-scores. For the HCP Gambling Reward context, SALSA exhibits consistent, low-variance agreement with FAST, fMRIStat, and FILM, while a distinct discordance is observed with 3dREMLfit, warranting further investigation.

3.4 Conclusion

Our study investigates noise properties of fMRI data and introduces SALSA, a novel pre-whitening technique, to mitigate autocorrelation biases. Spanning tissue types, pre-whitening methods, and group analyses, our findings shed light on crucial aspects of fMRI analysis, enhancing its accuracy and reliability.

Tissue-specific influences on autocorrelation are explored, with CSF and grey matter demonstrating higher autocorrelation than white matter. The supporting evidence from power spectra, autoregressive coefficients, and autocorrelation lengths (Figure 3.1) underscores these observations.

Importantly, SALSA's exceptional resilience to temporal smoothing and high sampling rate sets it apart from FAST, fMRIStat, and FILM, which exhibit increased susceptibility to such influences. This distinctiveness positions SALSA as a robust pre-whitening choice.

Crucially, SALSA's consistency in achieving desired false positive rates, surpassing other methods. This underscores SALSA's efficacy in preserving statistical integrity in fMRI analysis. The superior sensitivity of SALSA over FAST and 3dREMLfit in detecting signals, as demonstrated through synthetic event signal analysis, augments its utility in subject-level analysis, offering promise in critical applications such as pre-surgical fMRI.

Group-level analysis further validates SALSA's potential, exhibiting alignment with established methods across diverse datasets and tissue types. However, noteworthy disparities surface, particularly the positive discrepancy between SALSA and 3dREMLfit, warranting focused exploration.

In conclusion, our study reveals complex fMRI noise dynamics, with SALSA emerging as a potent tool for mitigating autocorrelation biases. Its versatility and reliability hold potential to elevate neuroimaging research, ensuring greater precision in functional brain exploration even as imaging technology advances. As our understanding evolves, these findings stand to significantly advance insights into human brain dynamics.

3.5 Code Availability

The code supporting this research, developed by Dr. Soroosh Afyouni during his tenure as a postdoctoral research fellow with the NISOx group, including that used for exploratory data analysis, model developments, simulation study design and results, as well as the real-data study implementation and results, is openly accessible and can be found at:

https://github.com/NISOx-BDI/SALSA

Chapter 4

Bayesian High-Dimensional Variable Screening: Extensions to Posterior Mean Screening

This chapter proposes multiple advancements to a Bayesian high-dimensional variable screening method, *Posterior Mean Screening*, enhancing its applicability to neuroimaging-specific prior knowledge and medical outcome variables. This work was conducted under the supervision of Prof. Jian Kang, a co-author of the original Posterior Mean Screening methodology, in collaboration with Prof. Timothy D. Johnson and Prof. Thomas E. Nichols. Additional details regarding this chapter are provided in Appendix A.

4.1 Introduction

Neuroimaging studies often face the challenge of analysing data with a large number of predictors, p, relative to a comparatively small number of subjects, n. With the rapid advancement in imaging technology, modalities such as structural MRI, functional MRI (fMRI), and diffusion tensor imaging (DTI) now offer high-resolution data capturing diverse brain features: structural information, neural activity, and white matter connectivity, respectively. As a result, the dimensionality of the data, p, can easily

exceed n, the sample size, by several orders of magnitude. For instance, voxel-based data from high-resolution MRI scans can contain millions of voxels, representing detailed brain anatomy, while the number of subjects often ranges from hundreds to a few thousand. This "ultra-high-dimensional" setting poses a significant challenge for traditional statistical and machine learning methods, which typically require $n > p$ for stable model estimation and reliable inference.

The primary objective of this study is to develop and evaluate variable screening methods for identifying relevant voxels, effectively filtering out those that do not significantly contribute to the response. This pre-selection step is crucial for neuroimaging studies, where dimensionality reduction can greatly enhance model interpretability and computational efficiency. By retaining only the most relevant variables, we enable the use of linear models for further analysis, which remain highly valuable due to their interpretability, computational simplicity, and robustness. Linear models facilitate straightforward inference, making them a powerful analytical tool despite the complexity and high dimensionality inherent in neuroimaging data.

The large-scale nature of neuroimaging data, particularly in cases where $p \gg n$, limits the applicability of many traditional methods. Classical approaches like Ordinary Least Squares (OLS) estimation assume $n > p$ for identifiability, while variable selection criteria, such as likelihood ratio tests, Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC), are computationally impractical in this high-dimensional context. Even regularised regression techniques, such as LASSO and Elastic Net, often struggle with computational feasibility or require extensive tuning for stability in ultra-high dimensions (Tibshirani, 1996; Zou and Hastie, 2005). Although less-biased methods like adaptive LASSO (Zou, 2006) have been developed to achieve consistent selection under general design conditions, their applicability is limited in scenarios where the predictor dimension exceeds the sample size, as is common in neuroimaging. These limitations emphasise the need for advanced techniques capable of efficiently handling ultra-high-dimensional neuroimaging data.

Historically, the mass univariate analysis (MUA) approach has been widely used in neuroimaging studies for effect discovery and predictor selection. MUA conducts separate statistical tests for each voxel, assuming independence across voxels. However, this approach fails to account for spatial dependencies and correlations, which are

biologically relevant in the brain. The brain's functional and structural organisation suggests that contiguous regions, or specific networks, share relationships with cognitive and behavioural responses. Consequently, MUA's assumption of voxel independence is unrealistic and limits its utility in neuroimaging (Friston et al., 1994). Ignoring spatial correlation may lead to fragmented findings and inflated error rates, especially when adjusting for multiple comparisons. This has led to the development of more robust and spatially aware methods in neuroimaging analysis.

In recent years, ultra-high-dimensional variable screening techniques have emerged as promising pre-processing steps for dimension reduction before fitting a model like OLS. Notable methods include Sure Independence Screening (SIS) (Fan and Lv, 2008), which ranks predictors based on marginal correlations with the outcome, and Forward Regression (Wang, 2009), which incrementally adds predictors based on a pre-defined criterion. While these methods address dimensionality reduction, they have limitations in neuroimaging and genomics applications where marginal correlation-based screening may fail due to the heterogeneity of the data and weak individual signals. Specifically, SIS and similar methods struggle to detect predictors with weaker marginal associations when the predictors are highly inter-correlated, a common issue in neuroimaging where spatially proximal voxels may exhibit substantial correlation. This shortcoming has motivated the development of variable screening methods that are robust to predictor dependencies.

One such advancement is the High-dimensional Ordinary Least Squares Projection (HOLP) method (Wang and Leng, 2016), which uses the generalised inverse of the design matrix rather than relying solely on marginal correlations. HOLP has demonstrated improved performance in high-dimensional settings, particularly when predictor independence is not guaranteed, as it leverages the global correlation structure among predictors. However, while HOLP provides enhanced stability, it is not specifically tailored for neuroimaging data, where the complex spatial organisation and known regions of functional interest present unique challenges and opportunities.

In neuroimaging studies, brain data often comprises large, spatially contiguous regions containing over 200,000 voxels that can be grouped based on anatomical or functional brain regions. Existing studies frequently identify specific regions or voxels associated with cognitive behaviours or disease states, which could inform and

improve variable screening processes. This context raises important questions about how to incorporate such prior knowledge, such as pre-identified regions of interest (ROIs) or spatial dependencies, into variable screening frameworks. Several recent methods attempt to address this need. For example, Conditional Sure Independence Screening (CSIS) (Barut et al., 2016; Hong et al., 2018) includes predetermined important features directly in the screening process, and Partition-based Screening (PartS) (Kang et al., 2017) leverages spatial partitions in generalised linear models to improve screening by grouping voxels based on anatomical proximity. Additionally, Covariance-Insured Screening (CIS) (He et al., 2019) incorporates prior information on inter-feature dependencies, addressing correlations among voxels. However, these methods are rooted in frequentist approaches and typically focus on incorporating only one type of prior information, limiting their flexibility in complex neuroimaging contexts.

Bayesian variable selection methods have also been proposed to incorporate prior knowledge in neuroimaging studies, offering a probabilistic framework to account for uncertainties. Bayesian approaches, such as those by Smith and Fahrmeir (2007), Goldsmith et al. (2014), and Kang et al. (2018), enable fully Bayesian inference on model parameters and provide posterior inclusion probabilities for each predictor. While Bayesian models offer a rigorous approach to incorporating prior information, they are computationally demanding, especially for ultra-high-dimensional neuroimaging data, making them infeasible for large-scale applications. These methods often require intensive sampling techniques like Markov Chain Monte Carlo (MCMC), which can be prohibitively slow for the scale of modern neuroimaging studies.

This focus on ultra-high-dimensional neuroimaging challenges has led us to explore recent advances in Bayesian variable screening, specifically the Posterior Mean Screening (PMS) method (He and Kang, 2021). PMS introduces an innovative approach to variable screening by applying a Bayesian framework, where the coefficients are assigned a Gaussian prior. This Gaussian prior not only aligns intuitively with the structure of neuroimaging data but also acts as a conjugate prior, greatly simplifying Bayesian computations. The original PMS model specification allows flexibility beyond basic spatial priors; it enables the integration of region-specific mean and covariance structures, which incorporate more complex prior knowledge into posterior calculations.

Empirical studies have demonstrated that PMS outperforms traditional screening methods in neuroimaging, particularly in identifying relevant predictors amid complex spatial and functional relationships. For instance, PMS has demonstrated superior performance in predicting intelligence quotient (IQ) and sensibly selecting voxels associated with IQ using resting-state fMRI data from the Autism Brain Imaging Data Exchange (ABIDE) Study (Di Martino et al., 2014). This methodological simplicity and adaptability provide a robust foundation for further exploration, which is the primary focus of this chapter.

Our investigation focuses on optimising the specification of the prior covariance matrix within the PMS framework, starting from a baseline informed by local spatial knowledge, specifically correlations among neighbouring voxels. We extend this by introducing data-dependent priors, motivated by the principle that in the era of large-scale neuroimaging, static or uninformative priors can be significantly improved by leveraging structure within the data itself. In particular, we consider low-rank representations (e.g. PCA-based projection matrices) and empirical sample covariances estimated from small, representative subsets of the data. These approaches provide efficient yet informative approximations of the variability structure in the full dataset, under the assumption that large neuroimaging samples yield sufficiently generalisable local structure. Furthermore, we propose a response-guided variant of PMS based on Supervised Principal Component Regression (SPCR), aligning the prior structure with patterns relevant to the outcome variable. This results in a more tailored and potentially more powerful screening procedure, enhancing both interpretability and predictive performance across varied neuroimaging tasks.

We further explore shrinkage-based methods to reduce noise and mitigate bias that may arise from sample limitations, thereby increasing the robustness of the PMS framework. Through simulation studies, we demonstrate that these approaches maintain stability and accuracy across varying conditions, underscoring the adaptability and reliability of PMS in high-dimensional neuroimaging applications.

Additionally, we propose a novel approach for determining the data-driven smoothing parameter for the Gaussian smoothing kernel, a widely used method in neuroimaging to enhance the signal-to-noise ratio. This smoothing process is particularly critical for MRI data, which often exhibit high levels of noise, and serves as a foundational

enhancement for ultra-high-dimensional variable screening. Our findings demonstrate that PMS, when equipped with optimised neighbouring correlation and smoothing parameter configurations, achieves superior performance compared to methods without smoothing, while maintaining robust and consistent results across varying conditions.

In this chapter, we present the motivation for our methodological extensions developed from the foundational principles of linear regression, combined with a Gaussian prior on the regression coefficients. Our primary focus is on modelling the prior hypervariance to enhance flexibility and predictive performance. We propose two key configurations: a data-guided approach and a response-guided approach, both aimed at optimising the construction of the prior covariance matrix. Furthermore, we introduce two methodological refinements to improve the robustness and general applicability of these approaches: shrinkage estimation for prior covariance matrices and an adaptive 3D smoothing technique specifically tailored to ultra-high-dimensional neuroimaging data. Through a comprehensive set of simulation studies, we evaluate the proposed methods against established benchmarks, demonstrating their superior sensitivity, stability, and screening consistency. Finally, we validate the practical utility of our approaches using real-world data, highlighting their enhanced predictive accuracy and interpretability in variable selection tasks.

4.2 Background

In this background section, we establish notation for our data structure. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ represent the data matrix, where n denotes the number of samples (i.e. subjects) and p represents the number of variables (e.g. voxels). Let $\mathbf{y} \in \mathbb{R}^n$ denote the vector of response variables.

Mass Univariate Analysis (MUA) is a foundational method in neuroimaging for analysing high-dimensional data due to its simplicity and computational efficiency. In MUA, each variable (e.g. voxel or region of interest) is analysed independently by fitting a separate model to the response variable and performing hypothesis testing for each variable. For a given predictor x_j, the null hypothesis $H_0 : \beta_j = 0$ (i.e. the predictor does not provide significant information about the response) is tested using a t-test. The test statistic for this hypothesis test is the *t-statistic*:

$$T = \frac{\hat{\beta}_j}{\hat{\mathrm{SE}}(\hat{\beta}_j)} \tag{4.1}$$

where $\hat{\beta}_j$ is the estimated coefficient for predictor x_j, and $\hat{\mathrm{SE}}(\hat{\beta}_j)$ denotes its standard error. This approach provides an efficient means of testing associations between individual predictors and the response variable across a large number of features.

To identify significant associations, the resulting t-statistics are typically converted into p-values, and a significance threshold (e.g. $\alpha = 0.05$) is applied to control for false positives. However, in high-dimensional settings, it is often necessary to adjust for multiple comparisons to control the family-wise error rate or the false discovery rate.

For the purposes of variable screening, rather than focusing solely on significance testing, we rank variables by the absolute values of their t-statistics. The top-ranked variables, based on the largest absolute t-statistic magnitudes, are selected for further analysis, as these are likely to show stronger associations with the response variable. This ranking-based approach provides a practical means of prioritising relevant predictors in high-dimensional neuroimaging studies, where a vast number of variables need to be evaluated efficiently.

All other variable screening methods and developments discussed in this chapter build upon the linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.2}$$

where ϵ_i for $i = 1, \ldots, n$ are independently and identically distributed (i.i.d.) according to a normal distribution with mean 0 and unknown variance σ^2.

In cases where $n > p$, the Ordinary Least Squares (OLS) regression method is commonly used to estimate the coefficients $\boldsymbol{\beta}$, as it provides an unbiased and optimal estimator in this setting. The OLS estimator, denoted $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, is computed as:

$$\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \tag{4.3}$$

assuming that the matrix $\mathbf{X}^\top \mathbf{X}$ is invertible. However, when dealing with high-dimensional neuroimaging data, where $p \gg n$, the matrix $\mathbf{X}^\top \mathbf{X}$ becomes singular (non-invertible), rendering the OLS estimator undefined. This challenge necessitates

modifications to traditional OLS to allow for stable and reliable estimation in high-dimensional contexts.

A widely used solution to address this singularity issue is Ridge regression, a regularised version of OLS that incorporates a penalty term to stabilise the coefficient estimates. Ridge regression achieves this by adding a positive constant r to the diagonal of $\mathbf{X}^\top\mathbf{X}$, effectively shrinking the coefficient estimates and mitigating issues of multicollinearity. The Ridge regression estimator, $\widehat{\boldsymbol{\beta}}^{\text{Ridge}}(r)$, is given by:

$$\widehat{\boldsymbol{\beta}}^{\text{Ridge}}(r) = (\mathbf{X}^\top\mathbf{X} + r\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} \tag{4.4}$$

$$= \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + r\mathbf{I}_n)^{-1}\mathbf{y} \tag{4.5}$$

where $r > 0$ is the ridge penalty parameter, and $\mathbf{I}_p$ and $\mathbf{I}_n$ are identity matrices of dimensions p and n, respectively. The second form of the Ridge estimator in Eq. (4.5) is derived using the Sherman-Morrison-Woodbury formula, which allows the inversion of a modified matrix by expressing it in terms of the inverses of smaller matrices. Specifically, the Sherman-Morrison-Woodbury formula states that for invertible matrices A and D,

$$(A + UDV)^{-1} = A^{-1} - A^{-1}U(D^{-1} + VA^{-1}U)^{-1}VA^{-1}. \tag{4.6}$$

Applying this formula, setting $A = r\mathbf{I}_p, U = \mathbf{X}^\top, D = \mathbf{I}_n, V = \mathbf{X}$ in Eq. (4.4), enables us to rewrite the Ridge estimator in a form that remains computationally feasible even when $p > n$, as it avoids direct inversion of the high-dimensional matrix $\mathbf{X}^\top\mathbf{X}$ and instead operates on the lower-dimensional matrix $\mathbf{X}\mathbf{X}^\top$.

The high-dimensional ordinary least squares projection (HOLP) method is a variable screening technique designed to address the challenges posed by high-dimensional data. HOLP is motivated by both OLS and Ridge regression, providing an efficient and stable method for feature selection in cases where the number of variables p vastly exceeds the number of observations n. In high-dimensional neuroimaging studies, where the data dimensionality can be orders of magnitude larger than the sample size, traditional approaches such as OLS often fail due to the ill-conditioned or singular

nature of $\mathbf{X}^\top\mathbf{X}$. HOLP offers a solution by leveraging the generalised inverse of the design matrix to approximate variable screening without requiring a direct inverse of $\mathbf{X}^\top\mathbf{X}$.

The variable screening statistic in HOLP, denoted by $\widehat{\boldsymbol{\beta}}^{HOLP}$, is computed as:

$$\widehat{\boldsymbol{\beta}}^{HOLP} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}. \tag{4.7}$$

This formulation avoids the computational limitations of OLS by projecting the response $\mathbf{y}$ onto the column space of $\mathbf{X}$ through the generalised inverse. HOLP ranks variables based on the magnitudes of the entries in $\widehat{\boldsymbol{\beta}}^{HOLP}$, with larger values suggesting a stronger association with the response variable. By selecting variables with the largest magnitudes, HOLP efficiently identifies relevant predictors, making it particularly advantageous in ultra-high-dimensional contexts, such as neuroimaging, where screening out irrelevant features is essential for both interpretability and computational feasibility.

To explicitly connect HOLP with OLS and Ridge regression, note that as $r \to 0$ in the Ridge regression formulation, we can approximate:

$$(\mathbf{X}^\top\mathbf{X})^{+}\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y},$$

$$\text{OLS} \approx \text{HOLP},$$

where $(\mathbf{X}^\top\mathbf{X})^{+}$ denotes the Moore–Penrose generalised inverse of $\mathbf{X}^\top\mathbf{X}$. This result shows that HOLP approximates OLS by circumventing the direct inversion of $\mathbf{X}^\top\mathbf{X}$, thus enabling stable and efficient variable screening even when $p \gg n$.

All of the previously discussed approaches, including OLS, Ridge regression, and HOLP, are frequentist methods. However, in the neuroimaging field, there is often a wealth of prior information available, such as anatomical and functional brain data, that can enhance the robustness and interpretability of variable screening algorithms. Incorporating such domain-specific knowledge can provide valuable guidance for the model, potentially improving clinical relevance. In light of this, He and Kang (2021) introduces the Posterior Mean Screening (PMS) framework, a Bayesian variable

screening approach that incorporates prior information via a multivariate Gaussian prior on $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}, \tau^2 \boldsymbol{\Lambda}), \tag{4.8}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$ is the prior mean, and $\boldsymbol{\Lambda} = (\lambda_{i,j}) \in \mathrm{Sym}_+^p$ is the prior covariance matrix, with $\mathrm{Var}(\beta_j) = \tau^2 \lambda_{j,j}$ and $\mathrm{Cov}(\beta_j, \beta_k) = \tau^2 \lambda_{j,k}$. Here, $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ represent prior information that can capture regional or anatomical knowledge, while τ^2 is a parameter that controls the overall variability of the prior distribution for $\boldsymbol{\beta}$.

The prior distribution in (4.8) is conjugate, meaning that the posterior distribution of $\boldsymbol{\beta}$ also follows a multivariate normal distribution. Specifically, given the prior and data, the posterior distribution of $\boldsymbol{\beta}$ is:

$$(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, \bullet) \sim N_p(\boldsymbol{\nu}, \mathbf{K}),$$

where $\bullet = \{\sigma^2, \tau^2, \mathbf{X}, \mathbf{y}\}$, $\boldsymbol{\nu} = (\theta \boldsymbol{\Lambda}^{-1} + \mathbf{X}^\top \mathbf{X})^{-1}(\theta \boldsymbol{\Lambda}^{-1} \boldsymbol{\mu} + \mathbf{X}^\top \mathbf{y})$, and $\mathbf{K} = (\theta \boldsymbol{\Lambda}^{-1} + \mathbf{X}^\top \mathbf{X})^{-1} \sigma^2$. Here, $\theta = \sigma^2/\tau^2 > 0$ represents the precision of the prior information on the structure of the predictors.

In the PMS framework, He and Kang (2021) proposes using the posterior mean, $\boldsymbol{\nu}$, as the variable screening statistic, as it naturally incorporates the prior knowledge encoded by $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. Specifically, the PMS estimator is given by:

$$\widehat{\boldsymbol{\beta}}^{PMS} = (\theta \boldsymbol{\Lambda}^{-1} + \mathbf{X}^\top \mathbf{X})^{-1}(\theta \boldsymbol{\Lambda}^{-1} \boldsymbol{\mu} + \mathbf{X}^\top \mathbf{y}). \tag{4.9}$$

Calculating $\widehat{\boldsymbol{\beta}}^{PMS}$ directly can be computationally demanding due to the inversion of $\mathbf{X}^\top \mathbf{X}$. However, the authors show that this computation can be simplified by re-expressing $\widehat{\boldsymbol{\beta}}^{PMS}$ as:

$$\widehat{\boldsymbol{\beta}}^{PMS} = \boldsymbol{\mu} + \boldsymbol{\Lambda} \mathbf{X}^\top (\mathbf{X} \boldsymbol{\Lambda} \mathbf{X}^\top + \theta \mathbf{I}_n)^{-1}(\mathbf{y} - \mathbf{X} \boldsymbol{\mu}). \tag{4.10}$$

This reformulation reveals that PMS reduces to HOLP under non-informative priors, specifically when $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Lambda} = \mathbf{I}_p$, and $\theta = 0$. In this case, PMS relies solely on

the data without incorporating additional prior information, thus aligning with the HOLP framework in its simplest form.

The PMS framework opens a broad avenue for incorporating prior information at multiple levels, allowing for increased robustness and adaptability in neuroimaging analyses. In its simplest form, PMS can incorporate a non-informative prior, with $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Lambda} = \mathbf{I_p}$, and $\theta = 0$. This approach serves as a useful benchmark to evaluate the impact of adding prior knowledge, particularly when little prior information is available or when assessing whether specific prior information is beneficial or misleading. The non-informative prior also enhances computational efficiency, as the simplicity of the prior correlation matrix reduces complexity.

Another straightforward method for incorporating prior information, introduced in the original PMS paper, involves defining the prior correlation matrix $\boldsymbol{\Lambda}$ based on spatial correlations, such as the smoothing relationships between each voxel and its neighboring voxels. This approach leverages anatomical or spatial continuity, which is particularly valuable in neuroimaging where neighboring voxels often have correlated activity patterns.

In many neuroimaging studies, specific voxels or features may already be known to contribute meaningfully to the outcome of interest. For such cases, we denote $\mathcal{S}$ as the set of indices corresponding to these prior-selected features. We can specify $\boldsymbol{\mu}$ such that $\mu_j = 0$ for $j \notin \mathcal{S}$, while assigning values to $\boldsymbol{\mu}_{\mathcal{S}} = (\mu_j : j \in \mathcal{S})$. In this setting, Section 3.1 of He and Kang (2021) shows that the variable screening statistics for the prior-selected features retain the benefits of the PMS framework. Although some signal may still appear for non-selected features, it will be relatively weak and asymptotically bounded by a polynomial order of n, thereby prioritising the selected features.

A more complex level of prior incorporation involves specifying group-level importance, which is advantageous when features can be naturally grouped, and distinguishing individual importance within each group is challenging. This approach is particularly applicable to brain imaging data, where brain atlases, such as the Automated Anatomical Labeling (AAL) system (Tzourio-Mazoyer et al., 2002), partition the brain into distinct anatomical or functional regions. For instance, the AAL atlas divides the brain into over 90 regions. In high-resolution brain templates with over

200,000 voxels, modelling each individual voxel may be infeasible, making region-based grouping more practical. Formally, suppose the features are partitioned into m groups. Let $\mathbf{B} = (b_{g,j})$ be an $m \times p$ group indicator matrix, where $b_{g,j} = 1$ indicates that feature j belongs to group g. Each feature is assigned to one group, so the columns of $\mathbf{B}$ sum to one, i.e., $\mathbf{1}_m^\top \mathbf{B} = \mathbf{1}$. Let $\bar{\mu}_g$ represent the importance level for group g, and denote $\bar{\boldsymbol{\mu}} = (\bar{\mu}_1, \ldots, \bar{\mu}_m)^\top$. Then, the full prior mean vector, $\boldsymbol{\mu}$, is defined to assign equal importance within each group, i.e. $\boldsymbol{\mu} = \mathbf{B}^\top \bar{\boldsymbol{\mu}}$. Section 3.2 of He and Kang (2021) demonstrates that this group-level prior knowledge maintains the benefits of the PMS framework.

Additionally, more sophisticated approaches, such as Combined PMS (CPMS), allow the integration of multiple types of prior information. CPMS applies each type of prior knowledge separately to generate PMS statistics, then selects the maximum statistic for each variable across all prior types, ensuring that the most informative priors contribute to variable selection.

4.3 Methodology

The primary focus of this work is not on specifying a prior for the mean vector $\boldsymbol{\mu}$; thus, without loss of generality, we assume $\boldsymbol{\mu} = \mathbf{0}$. Instead, we focus on constructing various forms of the prior covariance (or correlation) matrix $\boldsymbol{\Lambda}$ for the Gaussian prior on $\boldsymbol{\beta}$. While the potential of utilising Principal Component Analysis (PCA) to inform the construction of $\boldsymbol{\Lambda}$ was briefly noted in (He and Kang, 2021), it remained largely unexplored due to concerns over PCA's susceptibility to capturing noise and its tendency to overfit. In this work, we rigorously examine these challenges and propose robust frameworks to address them, focusing specifically on four key aspects: (1) data-guided method, (2) response-guided method, (3) robustness improvement, and (4) smoothness parameterisation.

In place of the spatial priors used in the original PMS paper, we propose efficient data-driven methods to serve as a priori information for constructing $\boldsymbol{\Lambda}$. Here, we employ a representative subset of the data, significantly smaller than the full dataset, to build an approximate prior covariance matrix that captures functional relationships

within the data, offering a practical and computationally feasible alternative to traditional spatial priors as initially proposed in He and Kang (2021).

In its original configuration, the PMS approach establishes a neighbour correlation-based prior covariance structure, relying on both the neighbouring correlation coefficient and a pre-specified neighbour distance. Consider two voxels in a standard coordinate system, $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^{3\times 1}$, where $\mathbf{v} \sim \mathbf{v}'$ if $\mathbf{v}$ and $\mathbf{v}'$ are neighbours. The PMS framework sets the prior covariance kernel by:

$$\mathbf{\Lambda}(\mathbf{v}, \mathbf{v}') = \begin{cases} \exp\big\{ -\sqrt{\rho(\mathbf{v})}\sqrt{\rho(\mathbf{v}')}\|\mathbf{v}-\mathbf{v}'\|^2\big\} I\big\{\|\mathbf{v}-\mathbf{v}'\| \le d\big\}, & \text{if } \mathbf{v} \sim \mathbf{v}' \\ 0, & \text{otherwise,} \end{cases} \tag{4.11}$$

where $\rho(\mathbf{v})$ serves as the smoothness parameter for voxel $\mathbf{v}$, and d is the maximum distance within which two voxels are considered covarying. This construction yields a block-diagonal matrix structure, optimising computational efficiency for matrix operations.

The parameter $\rho(\mathbf{v})$ is determined based on the minimum distance among all neighbouring voxels $\mathbf{v}'$ and a neighbouring correlation parameter $\eta_{\mathbf{v}}$, defined as:

$$\rho(\mathbf{v}) = \frac{-1}{\min_{\mathbf{v}\sim\mathbf{v}'} \|\mathbf{v}-\mathbf{v}'\|^2} \log(\eta_{\mathbf{v}}). \tag{4.12}$$

For simplicity in both interpretation and computation, the authors in He and Kang (2021) recommend parameterising $\rho(\mathbf{v})$ using a global prior correlation for neighbouring voxels, denoted by η. Thus, for all voxels $\mathbf{v}$, we set:

$$\rho(\mathbf{v}) = \frac{-1}{\min_{\mathbf{v}\sim\mathbf{v}'} \|\mathbf{v}-\mathbf{v}'\|^2} \log(\eta), \tag{4.13}$$

This global parameterisation simplifies the prior covariance kernel's expression, making it linearly dependent on the global neighbouring correlation η.

4.3.1 Constructing the Prior Covariance Λ

4.3.1.1 Data-guided method

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the data matrix, where n represents observations and p represents variables. Principal Component Analysis (PCA) identifies a linear transformation that maximises the variance of the projected data. This is done by solving the following optimisation problem:

$$\max_{\mathbf{w}} \quad \text{Var}(\mathbf{X}\mathbf{w}) \quad \text{subject to } \|\mathbf{w}\| = 1,$$

where $\mathbf{w}$ is a weight vector. Solving this problem leads to the eigenvalue decomposition:

$$\mathbf{X}^\top \mathbf{X}\mathbf{w} = \xi \mathbf{w},$$

where ξ is the eigenvalue associated with eigenvector $\mathbf{w}$. The principal components are the eigenvectors corresponding to the largest eigenvalues, with each component capturing maximum variance under orthogonality constraints.

PCA achieves dimensionality reduction by projecting the data onto the first k principal components, where $k < p$. This projection is represented as:

$$\mathbf{X}_{\text{reduced}} = \mathbf{X}\mathbf{W}_k,$$

where $\mathbf{W}_k \in \mathbb{R}^{p \times k}$ contains the first k eigenvectors. Selecting k components allows us to retain a target variance level, $\alpha\%$, such that:

$$\frac{\sum_{i=1}^{k} \xi_i}{\sum_{i=1}^{p} \xi_i} \geq \frac{\alpha}{100},$$

where ξ_i are the eigenvalues. The projection matrix, given by

$$\mathbf{P} = \mathbf{W}_k \mathbf{W}_k^\top, \tag{4.14}$$

maps $\mathbf{X}$ onto a lower-dimensional space, preserving at least $\alpha\%$ of the total variance.

The efficient computation of PCA is achieved through Singular Value Decomposition (SVD), where $\mathbf{X}$ is decomposed as:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

with orthonormal matrices $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{p \times p}$, and $\mathbf{\Sigma}$ as a diagonal matrix of singular values. The largest singular values capture the most variance. Using SVD eliminates the need to explicitly compute $\mathbf{X}^\top\mathbf{X}$, a significant advantage when $\mathbf{X}$ is high-dimensional, as in neuroimaging. Moreover, computational optimisations allow us to compute only the top k singular values and vectors, which is both efficient and practical for large datasets.

The denoised projection provided by PCA is ideal for constructing a robust covariance matrix, enhancing stability in further analysis. One approach is to use the low-dimensional representation $\mathbf{X}_{\text{reduced}}$ to compute a robust covariance matrix:

$$\mathbf{\Sigma} = \frac{1}{n}\mathbf{X}_{\text{reduced}}^\top \mathbf{X}_{\text{reduced}}.$$

Alternatively, an equivalent formulation can be derived directly using the principal components and their corresponding eigenvalues. Let $\mathbf{\Xi}_k$ be a diagonal matrix of the top k eigenvalues, and $\mathbf{W}_k$ be the matrix of the corresponding eigenvectors. The robust covariance matrix $\mathbf{\Sigma}$ can be expressed as:

$$\mathbf{\Sigma} = \mathbf{W}_k \mathbf{\Xi}_k \mathbf{W}_k^\top. \tag{4.15}$$

This matrix formulation reconstructs the covariance structure using only the most significant components, with $\mathbf{W}_k$ providing the principal directions and $\mathbf{\Xi}$ scaling them by their variances. This approach is particularly advantageous for noise reduction and enhances the stability and interpretability of downstream analysis.

In our proposed PMS framework, we set $\mathbf{\Lambda}$ to either $\mathbf{P}$ or $\mathbf{\Sigma}$, both of which are positive semi-definite and are symmetric, making them suitable as prior covariance matrices.

The projection matrix $\mathbf{P}$ captures the dominant signal directions of high variance in the data while filtering out noise from lower-variance components. but does not retain the exact magnitude of variance in each direction. By using $\mathbf{P}$ as a prior covariance for $\boldsymbol{\beta}$, we incorporate the assumption that the true coefficients align with the main structure of $\mathbf{X}$. This approach is computationally efficient and less complex,

as it disregards individual variance contributions. However, $\mathbf{P}$ may lack flexibility when applied to data with heterogeneous variance across components, since it does not account for the specific variance magnitudes in each direction. This approach directs the estimation towards the principal subspace, promoting sparsity outside of this primary structure and enhancing interpretability by focusing on directions most strongly associated with the data. Additionally, the noise-filtering properties of $\mathbf{P}$ improve robustness, as the prior emphasises components with higher signal-to-noise ratios, which is especially valuable in high-dimensional settings where noise can be substantial.

An alternative choice is the estimated covariance matrix $\boldsymbol{\Sigma}$. The matrix $\boldsymbol{\Sigma}$ captures both the main directions of variability and the magnitude of variance in each component. Using $\boldsymbol{\Sigma}$ as a prior covariance for $\boldsymbol{\beta}$ provides a flexible structure that weights each direction by its variance contribution, placing greater emphasis on components with larger eigenvalues. This allows for more precise estimation of $\boldsymbol{\beta}$, especially when certain directions in $\mathbf{X}$ are known to vary significantly. By scaling each direction according to its variance, $\boldsymbol{\Sigma}$ enables nuanced regularisation that adapts to the specific variance structure of $\mathbf{X}$, offering an accurate representation of the data's structure, albeit with a modest increase in computational complexity.

4.3.1.2 Response-guided method

Supervised Principal Component Regression (SPCR) is an extension of traditional PCA that incorporates response variable information to enhance the selection of principal components relevant to prediction. Unlike standard PCA, which maximises variance in the predictor space without reference to the outcome, SPCR integrates supervised learning to ensure that the identified components are directly informative for predicting the response variable $\mathbf{y}$.

The SPCR procedure begins by fitting a regression model, such as OLS or Ridge regression, using the predictor matrix $\mathbf{X}$ to predict $\mathbf{y}$. This initial regression yields a set of fitted coefficients, $\hat{\boldsymbol{\beta}}$, representing the predictive importance of each feature in $\mathbf{X}$ with respect to $\mathbf{y}$:

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ denotes the residuals. Based on the estimated coefficients $\hat{\boldsymbol{\beta}}$, the predictors in $\mathbf{X}$ are re-weighted to emphasise features most relevant to $\mathbf{y}$. Specifically, the re-weighted predictor matrix $\mathbf{X}'$ is constructed as:

$$\mathbf{X}'_j = \sqrt{|\hat{\beta}_j|}\,\mathbf{X}_j,$$

where $\mathbf{X}'_j$ and $\mathbf{X}_j$ denote the j-th columns of the re-weighted and original data matrices, respectively. This transformation scales each predictor by the square root of its absolute coefficient value, thereby amplifying the variance of features that contribute more to the response and diminishing the variance of less influential features. This approach effectively adjusts the variance structure in the predictor space so that features with higher predictive relevance exhibit greater variance in $\mathbf{X}'$, guiding PCA towards components that capture predictive information rather than merely maximising variance.

PCA is then performed on the re-weighted predictor matrix $\mathbf{X}'$, allowing for the extraction of principal components that are more aligned with the response variable. The projection matrix

$$\mathbf{P}' = \mathbf{W}'_k \mathbf{W}'^{\top}_k, \tag{4.16}$$

and approximated covariance matrix

$$\boldsymbol{\Sigma}' = \mathbf{W}'_k \boldsymbol{\Xi}'_k \mathbf{W}'^{\top}_k \tag{4.17}$$

can subsequently be derived from $\mathbf{X}'$ as described in Section 4.3.1.1. These matrices incorporate the predictive relevance of each feature, enhancing the robustness of PCA for supervised tasks by focusing on the components most associated with the response variable.

4.3.2 Shrinkage for Stabilising Prior Covariance $\boldsymbol{\Lambda}$ Estimation

The derived matrices discussed in Sections 4.3.1.1 and 4.3.1.2 are valid covariance matrices, as they satisfy the conditions of symmetry and positive semi-definiteness. However, in high-dimensional applications, simply ensuring positive semi-definiteness

may be insufficient. When these matrices are near-singular, i.e. they have very small eigenvalues, numerical instability can arise during inversion, a common step in many statistical methods. This issue, coupled with the potential for estimation errors due to limited observed data, can lead to overfitting and poor performance in downstream tasks, such as variable selection or prediction.

Shrinkage offers an effective solution by combining the sample matrix $\mathbf{S}$ with a structured estimator $\mathbf{F}$, which introduces some bias but reduces variance and stabilises the estimation. The shrunk estimator $\mathbf{S}'$ is given by:

$$\mathbf{S}' = \delta\mathbf{S} + (1 - \delta)\mathbf{F},$$

where $\delta \in [0, 1]$ is the shrinkage constant that controls the balance between $\mathbf{S}$ and the more regularised $\mathbf{F}$. This approach trades off a slight increase in bias for reduced estimation error, making it especially useful in high-dimensional settings. An optimal shrinkage coefficient can be derived for large covariance matrices, as detailed in Ledoit and Wolf (2004), though this is beyond the scope of our current study.

By combining $\mathbf{S}$ with $\mathbf{F}$, shrinkage effectively shifts the eigenvalues of the matrix, reducing the likelihood of near-zero eigenvalues and mitigating singularity issues. This behaviour is similar to the regularisation effect of Ridge regression, where adding a penalty term shifts eigenvalues positively, thereby improving matrix conditioning and ensuring invertibility.

In settings where linear-model-motivated variable screening assumes independence among predictors, a common choice for $\mathbf{F}$ is a diagonal matrix. In our case, setting $\mathbf{F} = \mathbf{I_p}$, where $\mathbf{I_p}$ is the identity matrix of dimension p, corresponds to the covariance structure assumed in the HOLP method, which treats predictors as uncorrelated. Using $F = \mathbf{I_p}$ introduces enough structure to stabilise the estimate without imposing overly restrictive assumptions on the data.

The shrunk estimator is therefore constructed as:

$$\mathbf{S}' = \delta\mathbf{S} + (1 - \delta)\mathbf{I_p}, \tag{4.18}$$

producing a hybrid estimator that balances the sample covariance's ability to capture the data's structure with the stability imparted by the identity matrix. This formu-

lation reduces overfitting and enhances robustness, resulting in a reliable estimator that combines the strengths of both the sample covariance and the HOLP-inspired structured approach.

4.3.3 Gaussian Kernel Smoothing Estimation

A complementary approach to constructing a robust prior covariance matrix $\mathbf{\Lambda}$ is spatial smoothing, as discussed in Section 2.1.3.5. Spatial smoothing can be applied prior to the frameworks outlined in Sections 4.3.1.1, 4.3.1.2, and 4.3.2. By smoothing, we enforce dependency among voxel effects based on spatial contiguity, which aligns with the biological reality of MRI data, where neighbouring voxels often exhibit correlated signals. This method introduces a bias based on spatial proximity, effectively reducing the variance of differences across contiguous voxels and increasing the robustness of the data.

We specifically examine Gaussian kernel smoothing, a standard practice in neuroimaging due to its smooth, decaying characteristics. Gaussian kernel smoothing computes a weighted average of neighbouring voxels, assigning higher weights to voxels closer to the centre. In neuroimaging, Gaussian smoothing is widely used for several reasons:

- Increasing the Signal-to-Noise Ratio (SNR)
- Reducing overfitting by smoothing overly sharp images
- Enhancing spatial correlation to meet assumptions in statistical analyses

Smoothing helps reduce noise in voxel intensities, improves signal detection, and is often a preprocessing step for statistical analysis to satisfy spatial correlation assumptions.

The 3D Gaussian distribution centred at (x_0, y_0, z_0) has the probability density function:

$$G(x, y, z) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_z} \exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2} - \frac{(z-z_0)^2}{2\sigma_z^2}\right) \tag{4.19}$$

where σ_x, σ_y, and σ_z are the standard deviations along the x, y, and z axes.

In this study, we simplify the analysis by assuming isotropic smoothing, setting $\sigma = \sigma_x = \sigma_y = \sigma_z$. This homogeneity assumption acts as a form of regularisation, smoothing equally across all directions without favouring any specific orientation. This assumption is practical, as brain tissue generally does not exhibit directional bias in signal propagation.

Under this isotropic Gaussian kernel, we can express the spatial correlation induced between two voxels v and v' as:

$$\text{Corr}(v, v') = \exp\left(-\frac{\|v - v'\|^2}{2\sigma^2}\right). \tag{4.20}$$

This correlation structure reflects the decreasing influence of voxel v on voxel v' as their Euclidean distance $\|v - v'\|$ increases.

4.3.3.1 Determining the Smoothing Parameter

Selecting an appropriate value for the smoothing parameter σ is crucial to balance the signal and noise in spatial data. In practice, many prefer to interpret σ in terms of the full-width-half-maximum (FWHM), which represents the width of the Gaussian curve at half its maximum value. FWHM provides an intuitive metric for estimating the effective spread of smoothing in terms of distance, whether in voxels or millimetres, which is particularly useful for known effects in neuroimaging. However, for our purposes, we aim to derive σ in a data-driven manner, allowing for a more generalisable algorithm that adapts to various modelling contexts and effect studies.

Our proposed approach estimates σ through the averaged lag-1 spatial correlation of a 3D image, Q, which represents independent signals between the predictor data and the response variable. For instance, Q can be derived by fitting the HOLP model on the predictors $\mathbf{X}$ against the response $\mathbf{y}$, motivated by a linear and independent effect assumption. Alternatively, a simpler and computationally efficient method is to calculate voxel-wise correlations between individual voxels and the response variable, corresponding to a MUA approach.

Once we obtain Q, we shift it by one voxel along each spatial dimension (x, y, and z), generating three shifted images: Q^1_x, Q^1_y, and Q^1_z. We then calculate the correlation

of the original image Q with each of these shifted images, yielding correlation values c_x^1, c_y^1, and c_z^1 for each direction:

$$c_x^1 = \text{Corr}(Q, Q_x^1), \quad c_y^1 = \text{Corr}(Q, Q_y^1), \quad c_z^1 = \text{Corr}(Q, Q_z^1). \tag{4.21}$$

Finally, we compute the average correlation $\bar{c}$ across the three directions:

$$\bar{c} = \frac{1}{3}(c_x^1 + c_y^1 + c_z^1). \tag{4.22}$$

This averaged correlation, $\bar{c}$, serves as a data-driven estimate of spatial dependency, specifically capturing the correlation between each voxel and its immediate neighbours (with a voxel distance of 1).

Using this definition, we substitute $\bar{c}$ into Eq. (4.20) to express $\bar{c}$ in terms of σ:

$$\bar{c} = \exp\left(-\frac{1}{2\sigma^2}\right), \tag{4.23}$$

which can then be rearranged to solve for σ:

$$\sigma = \frac{1}{\sqrt{-2\log(\bar{c})}}. \tag{4.24}$$

This formulation provides a direct and adaptable means of estimating σ, allowing it to be calibrated according to the spatial correlation observed in the data. By grounding σ in the empirical neighbour-wise correlation $\bar{c}$, this approach ensures that the smoothing parameter is tailored to the intrinsic spatial structure of the image, enhancing robustness and generalisability across different studies.

4.3.4 Metrics

4.3.4.1 Relative Importance Index

To evaluate the similarity among feature-screening models, we assess the spatial distribution of the top p highest-ranked variables across brain regions. Directly comparing rankings between models may be impractical due to small, potentially insignificant differences in variable ranking (e.g., between the jth and $(j+2)$th variables). Thus, we define a cut-off, p, and treat the top p variables as equally

important. For each region, we compute a *Relative Importance Index* to capture how prominently the region contributes to the top p variables selected by the models.

The Relative Importance Index (RII) for Region m is defined as:

$$\text{RII for Region } m = \frac{\%\text{ of Region } m \text{ within the top } p \text{ selected voxels}}{\%\text{ of Region } m \text{ within the whole brain mask}}. \tag{4.25}$$

An index around 1 indicates that the feature-screening model selects voxels proportionally to the region's size within the brain mask. An index much greater than 1 suggests that voxels in Region m are overrepresented among the top p selected variables, indicating a stronger association with the response variable, and vice versa.

To further assess interpretability, the selected voxels can be visualised on a brain template, allowing us to observe the spatial patterns of selection across the brain. This qualitative examination offers insights into the coherence and potential anatomical relevance of the regions highlighted by each model.

4.3.4.2 Power, FPR, FDR in Hypothsis testing

The metrics used in this study are based on standard definitions from hypothesis testing:

- **True Positives (TP):** The number of relevant variables (true signals) correctly identified by the screening method, i.e. variables for which the null hypothesis H_0 is rejected correctly when H_1 is true.
- **True Negatives (TN):** The number of irrelevant variables correctly identified as irrelevant, i.e. variables for which the null hypothesis H_0 is correctly not rejected when H_0 is true.
- **False Positives (FP):** The number of irrelevant variables incorrectly identified as relevant, i.e. variables for which the null hypothesis H_0 is rejected incorrectly when H_0 is true.
- **False Negatives (FN):** The number of relevant variables incorrectly identified as irrelevant, i.e. variables for which the null hypothesis H_0 is incorrectly not rejected when H_1 is true.

These quantities allow for the calculation of several performance metrics that assess the screening method's effectiveness.

The *power* of a hypothesis test refers to the probability of correctly rejecting the null hypothesis H_0 when the alternative hypothesis H_1 is true, which is the complement of making a Type II error (false negative). Mathematically, the power is expressed as:

$$\text{Power} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \mathbb{P}(\text{Reject } H_0 \mid H_1 \text{ is true})$$

This metric, also referred to as recall or sensitivity, measures the proportion of true signals (relevant variables) successfully recovered by the screening method when H_1 is true. A higher power indicates the method is effective at detecting relevant signals, and is particularly important in applications where missing true signals could lead to significant consequences.

The false positive rate (FPR) is the probability of incorrectly rejecting H_0 when H_0 is true (making a Type I error). It is given by:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ is true})$$

FPR represents the proportion of irrelevant variables incorrectly identified as relevant. It provides a measure of the screening method's tendency to make Type I errors. However, it is worth noting that in the context of variable ranking, rather than voxel significance, FPR may be less relevant, as the primary concern is often the ordering of variables rather than their binary classification.

The false discovery rate (FDR) measures the proportion of variables identified as relevant that are actually false positives. It is defined as:

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}}$$

The FDR provides an indication of the precision of the screening method, with lower values indicating a higher confidence that identified signals are indeed true.

4.4 Analysis setup

4.4.1 Data

This chapter uses data from the UK Biobank, comprising a random sample of 4,000 subjects, each with volumetric-based morphometry (VBM) images and corresponding age information (the same sample subjects used in Chapter 6). Each VBM scan is a 3-dimensional image of dimensions $91 \times 109 \times 91$ with voxel resolution of 2 mm^3.

For the purposes of this study, we use a 12-region bilateral grey-matter brain atlas derived in Tziortzi et al. (2011) to structure our analysis of regional variable selection patterns. This atlas serves as a motivating example, offering a structured framework for investigating region-specific effects and providing interpretive context in later sections. While the choice of atlas does not influence the setup of variable screening methods, it adds anatomical context to the interpretation of regional effects and selection patterns.

The anatomical regions in this atlas are defined as follows:

Region Number	**Anatomical Area**	**Region Number**	**Anatomical Area**
1	Occipital Lobe	7	Parietal Lobe
2	Left Insular Cortex	8	Basal Ganglia
3	Right Insular Cortex	9	Thalamus
4	Temporal Lobe	10	Ventral Cerebellum
5	Frontal Lobe	11	Right Dorsal Cerebellum
6	Cingulate Cortex	12	Left Dorsal Cerebellum

Table 4.1: Anatomical regions in the 12-region grey-matter brain atlas.

4.4.2 Model Configuration

Table 4.2 provides an overview of the acronyms used for each variable screening method in the Results section of this chapter.

The parameter θ, referenced throughout this chapter, represents the penalty term. In the case of PMS, θ reflects the precision of the prior information on predictor structure; for Ridge and HOLP, it provides numerical stability during matrix inversion. We set $\theta = 0.0001$ to minimally impact the computation of each method, though ideally, this parameter should be optimised for each specific application.

Acronym	Description
Our proposed variable screening methods	
PMS:PCA	PCA (90% variance), forming $\mathbf{\Lambda}$ as a projection matrix given by Eq. (4.14)
PMS:SPCR	SPCR (90% variance), forming $\mathbf{\Lambda}$ as the proposed projection matrix given by Eq. (4.16)
PMS:SPCR-S	SPCR (90% variance) with the proposed data smoothing, forming $\mathbf{\Lambda}$ as a projection matrix given by Eq. (4.16)
PMS:SPCR-S-Cov	SPCR (90% variance) with the proposed data smoothing, forming $\mathbf{\Lambda}$ as a covariance matrix with shrinkage given by Eq. (4.17)
PMS:SPCR-S-Proj	SPCR (90% variance) with the proposed data smoothing, forming $\mathbf{\Lambda}$ as a projection matrix with shrinkage given by Eq. (4.16)
Variable screening literature benchmarks	
HOLP	HOLP variable selection method
PMS:Spatial	PMS variable selection method using spatial characterisation as in Eq. (4.11) to form $\mathbf{\Lambda}$
Ridge	Ridge variable selection
Ridge-S	Ridge variable selection with the proposed data smoothing
MUA	Mass univariate analysis
Non-screening benchmark	
Overall Ridge	Ridge regression applied to all data without prior variable screening

Table 4.2: Acronyms for variable screening and selection methods used in the analysis.

The smoothing parameter, σ, defined in Section 4.3.3.1 is derived using voxel-wise correlations between individual voxels and the response variable. It is consistently applied across all methods with smoothing (indicated by "-S") in the Results sections, except in Section 4.5.1. 00 The structural bias matrix, F, defined in Section 4.3.2, is assumed to be the identity matrix, $\mathbf{I_p}$, in all experiments. The shrinkage parameter, δ, is fixed at 0.1 for the simulation analyses and 0.3 for the real-data analysis.

For PMS configuration, we employ the default settings proposed by the `PMS`[1] package, specifying a neighbour degree of 1 in millimetre coordinates and using a global correlation η estimated via our method in Section 4.3.3.1 with Eq. (4.22). This corresponds to setting $d = 2\sqrt{2}$ in Eq. (4.11) and $\min_{\mathbf{v}\sim\mathbf{v}'} \|\mathbf{v} - \mathbf{v}'\|^2 = 4$ in Eq. (4.13).

[1] `https://github.com/kangjian2016/PMS/`

4.5 Simulation Study

For our simulation study, we selected a sample of 1,000 random subjects from the UK Biobank, focusing on their 3D VBM images. To ensure computational efficiency while preserving the real-world anatomical structure and noise present in neuroimaging data, we restricted our analysis to the central region of the brain, encompassing approximately 15,847 voxels and 18 regions. Figure 4.1 illustrates the field of view considered in this study. Synthetic signals were introduced into specified voxels or ROIs with varying structural configurations and intensities. The response variable was constructed as the sum of values from designated voxels, allowing us to evaluate the effectiveness of different variable screening methods in signal discovery task.

4.5.0.1 Simulation Setup

The simulation study was conducted as follows:

1. Extract VBM images from 500 random subjects, focusing on the central region of the brain.

2. Define the ROI as the central brain region, encompassing 1,867 voxels (approximately 11.8% of total brain volume) as shown in Figure 4.1a.

3. Analyse the voxel intensity statistics:

   - Inside the ROI: Mean = 0.4, Variance = 0.081
   - Outside the ROI: Mean = 0.6, Variance = 0.114

4. Apply HOLP to data within ROI using the age of the subjects as a response variable to obtain the coefficient vector $\hat{\boldsymbol{\beta}}^{\text{HOLP}}$ of length 1,867.

5. Replace voxel values in the ROI for each subject with independent samples drawn from a uniform distribution $U[0, 1]$, expressed as:

$$X_i(v) = w_i(v) \quad \text{where} \quad w_i(v) \sim U[0, 1],$$

for subject i and voxel v within the ROI. This is done to mimic the intensity characteristics within the ROI and to disassociate any structural correlation of the ROI from other parts of the brain.

6. For each simulation scenario, select a subset $s \leq n$ of *true effect* voxels within the ROI, define the set of *active* predictors S, and assign the *true* coefficient values $\boldsymbol{\beta}^S$. Set $\beta^S(\mathbf{v}) = 0$ for voxels not in S, and $\beta^S(\mathbf{v}) = \hat{\beta}^{\text{HOLP}}(\mathbf{v})$ for $v \in S$.

7. Construct a noise-free synthetic response variable $\boldsymbol{y}^{(0)}$ by computing $\mathbf{X}\boldsymbol{\beta}$, with variance $\frac{1}{12}\sum_{v \in S} \beta(\mathbf{v})^2$. This setup yields $R^2 = 1$.

8. Add Gaussian noise $N(0, \zeta^2)$ to each $y_i^{(0)}$. Adjusting ζ^2 controls R^2 as:

$$R^2 = \frac{\text{var}(\boldsymbol{y}^{(0)})}{\text{var}(\boldsymbol{y}^{(0)}) + \zeta^2}.$$

   To achieve a desired R^2, set:

$$\zeta^2 = \frac{\text{var}(\boldsymbol{y}^{(0)})}{R^2} - \text{var}(\boldsymbol{y}^{(0)}).$$

   To compute the corresponding Signal-to-Noise Ratio (SNR), simply compute $SNR = \frac{R^2}{1-R^2}$.

9. Assess the Power of detection for varying signal strengths.

For this simulation, with number of evaluation (training) subjects $n = 1,000$, we designate 300 voxels as the true effect, i.e. $|S| = 300$.

4.5.0.2 Scenario 1: Single Contiguous Region in the ROI

In the simplest spatially contiguous scenario, a contiguous block of 300 voxels is selected, indexed from 1 to 300 within the ROI, was designated to generate the signal (Figure 4.1b). With $||\boldsymbol{\beta}^S||_2 = 4.01$ and variance of 0.042, this scenario serves as a baseline, testing the methods' ability to recover signals within a well-defined anatomical structure.

4.5.0.3 Scenario 2: Double Separated Regions in the ROI

In this scenario, two spatially distinct regions of contiguous voxels are selected, indexed from 1 to 150 and 1001 to 1150 within the ROI (Figure 4.1c). This scenario introduces spatial complexity, reflecting real-world scenarios where disconnected brain regions are associated with a common outcome. The $||\boldsymbol{\beta}^S||_2 = 3.72$ and variance is 0.040.

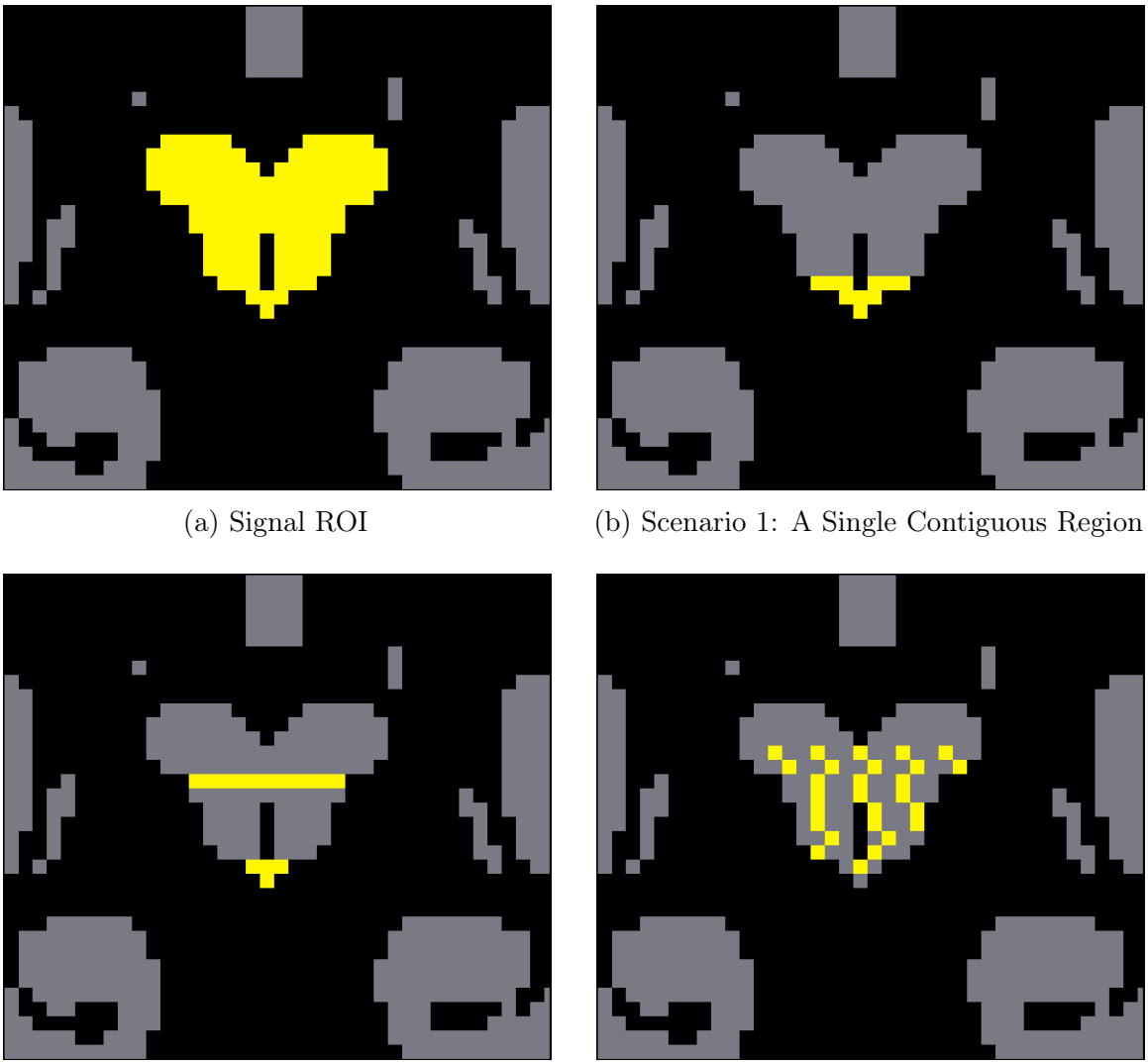
4.5.0.4 Scenario 3: Sparse Signals Across the ROI

In this scenario, we introduce sparse signals by selecting every 5th voxel from the 1st to the 1,500th voxel within the ROI (Figure 4.1d). This produces an unconventional, non-contiguous, scattered signal pattern. The $||\boldsymbol{\beta}^S||_2 = 4.64$ and variance is 0.039.

4.5.1 Smoothing Parameter

This section presents the evaluation of the Gaussian kernel smoothing parameter σ estimated using the proposed method from Section 4.3.3.1. We specifically examine two variants of the methodology: (1) averaged absolute correlations $\sigma^\dagger$, and (2) averaged correlations σ^*. It is important to note that the averaged absolute correlation approach is only applicable to Scenario 2, as Scenario 1 consists exclusively of positive lag-1 correlations in all directions, and Scenario 3 lacks lag-1 correlations due to its sparse signal configuration.

Table 4.3 shows the median power at $R^2 = 0.9$ across different smoothing parameters (σ) for Scenario 1 (Single Contiguous Region). Notably, PMS:SPCR-S-Cov and SPCR-S-Proj benefit disproportionately from higher levels of smoothing, as evidenced by the substantial increase in median power at $\sigma = 2$. This effect likely arises from the enhanced spatial correlations introduced by smoothing, which amplifies contiguous signals in the data. However, this level of over-smoothing adversely affects the PMS:Spatial method by reducing its ability to precisely localise the true signal within the region of interest. By contrast, the data-driven smoothing parameter, $\sigma^* = 0.8$, identified by our proposed estimation approach, provides a balanced improvement across methods, avoiding both the loss of localisation accuracy due to excessive smoothing and the under-smoothing that fails to adequately enhance signal structure. Additionally, Ridge-S performs poorly under low smoothing conditions,



(a) Signal ROI

(b) Scenario 1: A Single Contiguous Region

(c) Scenario 2: Double Separated Regions

(d) Scenario 3: Sparse Signals

Figure 4.1: Selected voxels (yellow), and unselected voxels (grey) included in simulation analysis. (a) represents the all voxels used in the ROI to compute all $\hat{\boldsymbol{\beta}}^{\text{HOLP}}$ of size 1,867 voxels. (b) - (d) represents the 300 selected voxels used in each respective scenario.

which aligns with findings in a later section showing that Ridge struggles when no smoothing is applied. This highlights Ridge-S's dependency on smoothing to achieve sufficient power in detecting signals within a highly correlated spatial structure.

Table 4.4 presents a similar analysis for Scenario 2 (Double Separated Regions). In this scenario, the data-driven smoothing parameters, $\sigma^* = 0.55$ and $\sigma^\dagger = 0.72$, demonstrate distinct advantages based on their estimation approach. The parameter $\sigma^\dagger$, derived from averaged absolute correlations, is somewhat biased upwards due to

non-negativity, but appears effective in maximising power for SPCR-S and Ridge-S, where spatially separated signals benefit from a moderate degree of smoothing. However, it fails to perform well on the rest of the methods.

Interestingly, the PMS:Spatial maintains robust performance at the data-driven smoothing parameters, but experiences a marked drop in median power at $\sigma = 2$, indicating that excessive smoothing disrupts its sensitivity to spatially distinct signals. Overall, σ^* yields offers a balanced smoothing level that supports consistent performance across a range of methods.

In Scenario 3, where signals are sparse and scattered throughout the brain, lag-1 correlation is not applicable due to the absence of spatial continuity. This lack of connectivity suggests that minimal smoothing is essential to preserve the integrity of isolated signals. Numerical experiments, omitted here for brevity, revealed that median power consistently decreases as the degree of smoothing increases across all variable screening methods. These findings confirm that strong smoothing, while useful for contiguous signals, reduces detection power in sparse configurations, highlighting the need for minimal smoothing to optimise power in scattered signal scenarios.

Method	Median Power			
	$\sigma = 0.25$	$\sigma^* = 0.8$	$\sigma = 1$	$\sigma = 2$
PMS:SPCR-S	11.17	54.83	63.33	58.00
PMS:SPCR-S-Cov	43.17	45.67	43.33	72.00
PMS:SPCR-S-Proj	43.33	49.17	45.33	73.67
PMS:Spatial	43.50	54.00	53.00	7.67
Ridge-S	0.00	33.67	36.67	26.00

Table 4.3: Median Power across different smoothing parameters (σ) for Scenario 1: A Single Contiguous Region. σ^* is derived from averaged correlations as proposed in Section 4.3.3.1.

4.5.1.1 Sensitivity

Sensitivity in variable screening methods reflects each method's ability to accurately identify relevant voxels contributing to the response variable. We simulate the setup from Section 4.5.0.1 across varying R^2 values, selecting the top $|S| = 300$ predictors and reporting median power and FPR, across 10 random runs. This analysis assesses each method's sensitivity and accuracy with varying R^2 levels: (1) $R^2 = 1$ as a

Method	Median Power				
	$\sigma = 0.25$	$\sigma^* = 0.55$	$\sigma^\dagger = 0.72$	$\sigma = 1$	$\sigma = 2$
PMS:SPCR-S	11.33	26.67	33.67	37.83	27.67
PMS:SPCR-S-Cov	40.83	40.17	35.83	25.50	41.67
PMS:SPCR-S-Proj	40.50	40.50	35.33	26.67	42.83
PMS:Spatial	40.83	42.00	40.00	36.33	4.17
Ridge-S	0.00	8.17	24.50	28.33	17.00

Table 4.4: Median Power across different smoothing parameters (σ) for Scenario 2: Double Separated Regions. σ^* is derived from averaged correlations as proposed in Section 4.3.3.1, and $\sigma^\dagger$ is derived from averaged absolute correlations.

high-signal benchmark, (2) $R^2 = 0.9$ representing strong signal, and (3) $R^2 = 0.5$ indicating balanced signal and noise.

It is important to note that, due to the fixed selection of $p = 300$ predictors across all methods, the False Positive Rate (FPR), False Discovery Rate (FDR), and Power metrics provide overlapping performance insights. The FPR reaches a maximum of 1.93%, as observed in the Ridge method where the Power is 0%. This upper bound reflects the scenario in which all selected variables are false positives, given a total of 15,547 non-signal voxels. Similarly, the FDR is directly related to Power, calculated as $100\% - \text{Power}$. Consequently, FPR and FDR metrics are omitted in this simulation analysis to focus on the more informative Power metric.

Table 4.5 presents the median power for Scenario 1 (Single Contiguous Region) across different R^2 levels. Here, PMS:Spatial is the top performer, particularly in high-signal scenarios, reaching a power of 57 at $R^2 = 1$, providing a strong benchmark when signal quality is ideal. However, in high-noise settings ($R^2 = 0.5$), our proposed methods PMS:SPCR-S-Cov and PMS:SPCR-S-Proj outperform PMS:Spatial, demonstrating resilience and robustness where signal detection is challenging.

Ridge consistently shows poor results with zero power across all R^2 values, underscoring its limitations in this setting, while MUA also performs poorly, confirming its limited utility for contiguous signals. Our proposed smoothing-augmented methods, PMS:SPRC-S-Cov and PMS:SPCR-S-Proj, suggesting them as a valuable option for variable screening in noise-prone environments.

Table 4.6 shows the median power results for Scenario 2 (Double Separated Regions). PMS:Spatial demonstrates the best performance overall, especially at higher R^2 levels, reaching a power of 42.5 at $R^2 = 1$. In the high-noise scenario ($R^2 = 0.5$), HOLP

slightly outperforms PMS, underscoring its suitability for complex signal detection tasks in noisy data.

Our proposed PMS:SPCR-S-Cov and PMS:SPCR-S-Proj methods show performance comparable to PMS:Spatial and HOLP across all R^2 levels, with the smoothing and Cov/Proj augmentations consistently enhancing detection power for PMS:PCA (see Appendix A.1.1.2 table A.4) and PMS:SPCR. These improvements suggest that smoothing supports more effective capture of dispersed signals, as seen in the strong performance of PMS:SPCR-S-Cov and PMS:SPCR-S-Proj. Ridge and MUA again exhibit poor detection power in this setting, reaffirming their limitations with spatially separated signal regions.

Table 4.7 presents the median power results for Scenario 3 (Sparse Signals) across various R^2 levels. HOLP consistently achieves the highest power, confirming its robustness with sparse signals lacking spatial continuity. MUA also performs well, often surpassing most of our proposed methods due to its lack of spatial assumptions, which aligns with the scattered signal distribution in this scenario.

Among our methods, PMS:PCA shows notably higher power, closely following HOLP. This improvement, suggests that PCA-based approach adapts relatively well to sparse signal configurations. In contrast, more spatially regularised methods like PMS:SPCR-S-Cov and PMS:SPCR-S-Proj show lower power in this scenario. These results highlight the advantage of methods with minimal spatial assumptions, like HOLP and MUA, for dispersed signals and underscore the importance of selecting methods tailored to the spatial characteristics of the data.

Method	$R^2 = 0.5$	$R^2 = 0.9$	$R^2 = 1$
	Median Power (SD)	Median Power (SD)	Median Power (SD)
PMS:PCA	23.67 (1.72)	29.00 (2.23)	28.83 (3.76)
PMS:SPCR	8.67 (1.43)	11.83 (1.14)	11.67 (1.48)
PMS:SPCR-S	31.00 (5.85)	47.00 (5.13)	49.50 (7.64)
PMS:SPCR-S-Cov	41.17 (4.90)	47.50 (4.46)	48.67 (4.11)
PMS:SPCR-S-Proj	43.67 (2.91)	51.00 (4.38)	52.00 (3.52)
HOLP	33.50 (1.99)	41.67 (1.85)	44.67 (1.96)
PMS:Spatial	40.83 (3.80)	53.83 (2.48)	57.00 (2.34)
Ridge	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Ridge-S	29.00 (2.76)	34.50 (1.76)	36.00 (2.33)
MUA	14.83 (1.91)	21.83 (2.18)	24.00 (2.38)

Table 4.5: Aggregated Results for Power across different R^2 values for Scenario 1: A Single Contiguous Region

Method	$R^2 = 0.5$	$R^2 = 0.9$	$R^2 = 1$
	Median Power (SD)	Median Power (SD)	Median Power (SD)
PMS:PCA	23.83 (2.51)	29.00 (2.58)	29.17 (2.46)
PMS:SPCR	8.17 (0.91)	11.50 (1.58)	13.33 (1.74)
PMS:SPCR-S	19.83 (2.88)	26.00 (3.81)	28.50 (2.00)
PMS:SPCR-S-Cov	29.83 (2.83)	40.00 (3.01)	41.67 (2.48)
PMS:SPCR-S-Proj	30.33 (2.69)	39.50 (2.82)	40.83 (2.42)
HOLP	30.67 (1.56)	41.00 (1.80)	41.67 (1.86)
PMS:Spatial	30.50 (2.51)	41.17 (2.60)	42.50 (2.20)
Ridge	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Ridge-S	6.50 (1.21)	8.17 (0.87)	8.33 (0.68)
MUA	14.50 (2.28)	21.17 (2.06)	22.50 (2.69)

Table 4.6: Aggregated Results for Power across different R^2 values for Scenario 2: Double Separated Regions

Method	$R^2 = 0.5$	$R^2 = 0.9$	$R^2 = 1$
	Median Power (SD)	Median Power (SD)	Median Power (SD)
PMS:PCA	23.67 (2.58)	31.00 (2.46)	31.83 (1.51)
PMS:SPCR	7.83 (1.44)	13.33 (2.09)	13.17 (1.33)
PMS:SPCR-S	6.67 (2.45)	9.50 (2.00)	9.50 (1.17)
PMS:SPCR-S-Cov	9.67 (1.47)	15.17 (1.79)	15.67 (1.88)
PMS:SPCR-S-Proj	9.67 (1.77)	15.00 (1.80)	15.50 (2.11)
HOLP	31.50 (2.40)	41.00 (2.16)	42.17 (2.59)
PMS:Spatial	10.33 (1.49)	13.83 (2.12)	14.83 (1.94)
Ridge	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Ridge-S	11.17 (1.60)	12.50 (1.79)	13.00 (1.09)
MUA	14.00 (1.89)	19.67 (2.03)	21.67 (1.22)

Table 4.7: Aggregated Results for Power across different R^2 values for Scenario 3: Sparse Signals

4.5.1.2 Stability

Stability is critical in high-dimensional settings like neuroimaging, where consistency in feature selection across runs greatly influences the reliability of results. To assess stability, we initialise data with different random seeds, run simulations to obtain relevant metrics, and evaluate stability by examining the standard deviation (SD) of these metrics across multiple runs. This process is repeated across varying signal strengths to gauge each method's robustness under different noise levels.

Table 4.5 presents the stability results for Scenario 1 (Single Contiguous Region), showing the SD of median power across R^2 values. Among high-performing methods, HOLP consistently displays the lowest SD, suggesting strong numerical stability and robustness. PMS:Spatial also demonstrates solid stability, with low variability in

power across runs. In contrast, PMS:SPCR-S exhibits greater variability, indicating that its performance fluctuates across different runs. However, the introduction of structural shrinkage with "-Cov" and "-Proj" variants notably improves stability for PMS:SPCR-S, highlighting the role of structural bias in enhancing robustness.

Interestingly, stability does not necessarily improve with higher R^2 values, an initially counterintuitive finding suggesting that noise reduction alone does not guarantee greater numerical consistency. These results indicate that stability gains are more effectively achieved through structural regularisation, as shown by the improved performance of the -Cov and -Proj variants.

4.5.1.3 Screening Consistency

Another powerful analysis for assessing the robustness of variable screening methods is to examine their consistency across different sample sizes. Ideally, an effective data-driven screening method should exhibit a monotonically increasing performance as the sample size grows, capitalising on more signal information and reducing the influence of noise. This concept is analogous to the confidence gained in estimating coefficients β in a linear model: as the sample size increases, the standard error of estimation typically decreases, enhancing reliability.

Figure 4.2 illustrates the screening consistency results, depicting median power for each variable screening method across varying sample sizes for $R^2 = 0.5$ and $R^2 = 0.9$. As expected, most methods display a monotonic increase in median power as sample size grows, reflecting improved performance with greater data availability. Notably, Ridge achieves consistently low power (0%) regardless of sample size, highlighting its limitations in variable screening even as data likelihood improves.

Among the methods, PMS:Spatial and HOLP are the most effective for smaller sample sizes, particularly when the number of subjects equals the number of active voxels (300). In contrast, the PMS:SPCR-S-Proj method performs poorly with limited data but demonstrates competitive performance once the sample size exceeds 800, where it contends or even surpasses PMS:Spatial and HOLP, especially in high-noise conditions ($R^2 = 0.5$).

An anomaly is observed with the PMS:SPCR-S method, which exhibits a sharp drop in median power at specific sample sizes, notably when the number of samples is

800 in Figure 4.2a. This unexpected variation suggests potential instability, possibly due to the impact of random initialisations. Increasing the number of random seeds may stabilise this trend, indicating an area for further investigation.

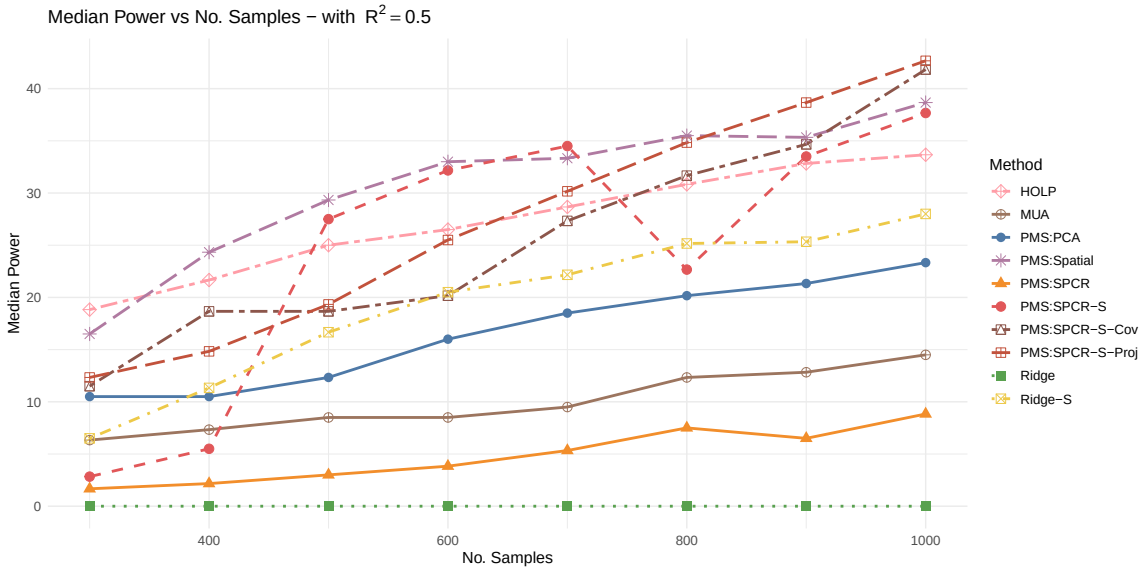
4.6 Real-Data Analysis

4.6.1 Predictive Accuracy

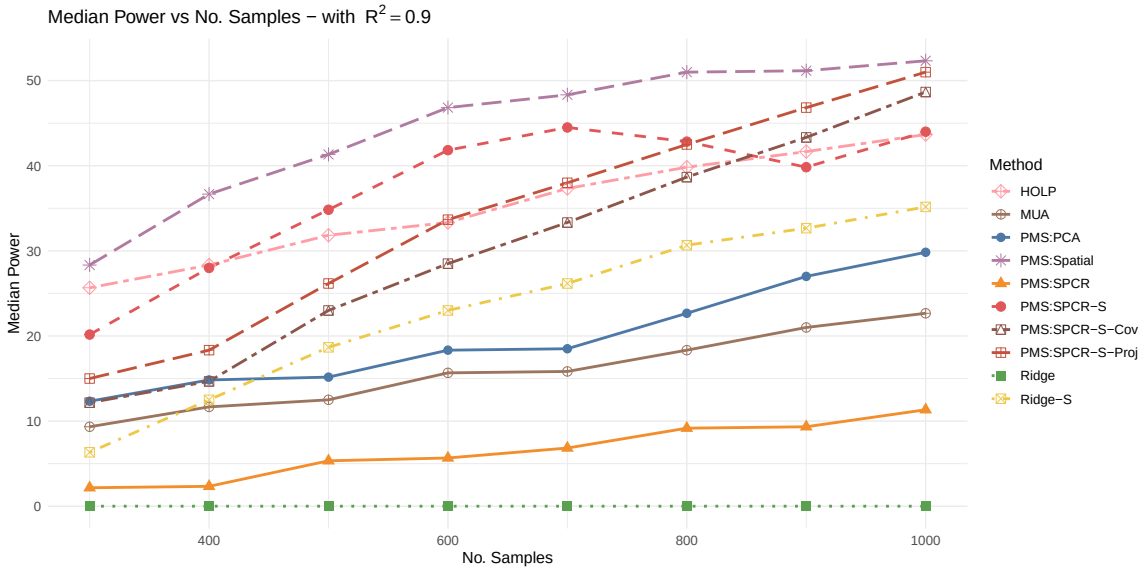
To illustrate the potential of variable screening methods in enhancing predictive accuracy and interpretability in clinical applications, we apply these techniques to the task of age prediction. We begin by estimating the data-driven smoothing parameter σ for our smoothing approach, as outlined in Section 4.3.3.1, using the real dataset. Following this, we perform variable screening to select relevant predictors, and subsequently, employ ridge regression on the screened variables to make age predictions.

Table 4.8 presents the training and test RMSE for age prediction using various variable screening methods across different numbers of selected parameters (p). As expected, we observe a general trend where predictive accuracy improves with an increase in the number of selected variables. This trend is consistent across all methods, demonstrating that larger subsets of informative features tend to enhance model performance by leveraging more relevant signal while reducing noise.

When comparing test RMSE values to the "Overall Ridge" model, which uses all 156,812 voxels without variable screening, it becomes evident that our proposed methods: PMS:PCA, PMS:SPCR-S, and Ridge-S, consistently outperform the full model. This suggests that our proposed variable screening methods can indeed identify a compact subset of predictors (e.g. PMS:PCA with $p = 2,000$, representing 1.275% of the total variables) that maintains or even improves predictive accuracy compared to using the full set of predictors. These results highlight the practical value of variable screening in ultra-high-dimensional neuroimaging data, where reducing dimensionality not only enhances predictive power but also facilitates interpretability in linear predictive models.



(a) Screening Consistency Results for $R^2 = 0.5$



(b) Screening Consistency Results for $R^2 = 0.9$

Figure 4.2: Screening consistency results in the case of Scenario 1: Single Contiguous Region with varying R^2 levels for different methods.

In terms of test RMSE, PMS:PCA emerges as the top-performing method, achieving the lowest error rates in both training and test sets, followed by MUA, PMS:SPCR-S and Ridge-S. The success of PMS:PCA in this context may be attributed to its design, which effectively uses an approximation of the data's covariance structure as a prior, thereby preserving substantial signal information with minimal noise. Both MUA and PMS:SPCR-S-Cov method yield competitive test RMSE, with lower discrepancy between training and test RMSE. This low discrepancy suggests potential for further refinement, possibly optimising the balance between model complexity and overfitting.

Overall, these results underscore the effectiveness of variable screening in high-dimensional predictive models, with methods like PMS:PCA, MUA, PMS:SPCR-S, and Ridge-S enhancing both predictive accuracy and model interpretability by selecting a smaller, more informative subset of features. Notably, our proposed methods: PMS:PCA, PMS:SPCR-S, and Ridge-S, not only outperform the "Overall Ridge" model but also exceed the performance of existing methods such as HOLP and PMS:Spatial in this age prediction task. This highlights the potential of our prior covariance configuration approach, suggesting broader applications and improvements over traditional methods in complex predictive modeling tasks within neuroimaging and beyond.

Method	Training RMSE				Test RMSE			
	p = 2,000	p = 4,000	p = 6,000	p = 10,000	p = 2,000	p = 4,000	p = 6,000	p = 10,000
PMS:PCA	3.023	2.445	2.068	1.628	4.357	4.253	4.220	4.186
PMS:SPCR	4.290	3.707	3.350	2.722	5.334	5.106	4.921	4.743
PMS:SPCR-S	3.679	3.265	3.004	2.673	4.423	4.354	4.335	4.329
PMS:SPCR-S-Cov	3.655	3.182	2.893	2.564	4.435	4.407	4.420	4.458
PMS:SPCR-S-Proj	3.951	3.410	3.090	2.708	4.619	4.573	4.601	4.621
HOLP	2.637	1.948	1.556	1.149	4.769	4.659	4.589	4.574
PMS:Spatial	3.260	2.592	2.180	1.659	4.763	4.671	4.607	4.541
Ridge	3.171	2.553	2.183	1.741	4.784	4.635	4.564	4.486
Ridge-S	4.341	3.965	3.675	3.308	4.730	4.554	4.449	4.402
MUA	4.249	3.724	3.297	2.731	4.772	4.580	4.460	4.327
Data	7.597				7.600			
Overall Ridge	0.911				4.488			

Table 4.8: Training and Testing Errors for Different Methods in Age Prediction

4.6.2 Interpretability

4.6.2.1 Region Importance

Figure 4.3 presents the Relative Importance Index (RII) for the 12-region brain atlas, illustrating the prominence of each region's contribution to the top 10,000 selected voxels across several high-performing variable screening methods, with HOLP and PMS:Spatial included for comparison.

The results show that PMS:PCA, HOLP, and PMS:Spatial share similar regional preferences, predominantly selecting voxels from the Insular Cortex (Region 2), Insular Cortex (Region 3), Basal Ganglia (Region 8), and Thalamus (Region 9), while markedly underrepresenting the Ventral Cerebellum (Region 10), Dorsal Cerebellum Right (Region 11), and Dorsal Cerebellum Left (Region 12). This common focus suggests that, despite each method selecting different specific voxels and achieving distinct predictive accuracies (see Section 4.6.1), they may be capturing similar underlying signals, providing complementary perspectives on feature importance. The selection of these regions aligns well with their clinical relevance in neurodegenerative and cognitive studies, as these areas are often implicated in core brain functions and neurological disorders.

PMS:SPCR-S and PMS:SPCR-S-Cov also prioritise Regions 2, 3, 8, and 9 but with an even stronger emphasis on the Thalamus (Region 9). This concentrated selection implies that these methods may be particularly adept at isolating signals predominantly originating from specific regions, which may offer advantages in studies where spatially focused signals carry clinical relevance. The balanced representation across other regions in these methods suggests that the incorporation of structural shrinkage supports the detection of meaningful regions without overwhelming contributions from less relevant areas.

Similarly, MUA prioritises Regions 2, 3, 8, and 9, along with Regions 6 (Cingulate Cortex), 10, 11, and 12. MUA notably underrepresents the Occipital Lobe (Region 1) and Parietal Lobe (Region 7). This pattern suggests that MUA, despite not explicitly modelling spatial structure, aligns with areas that are functionally significant, capturing both core and peripheral regions across the brain.

Ridge-S, in contrast, displays a distinct selection pattern, showing a strong preference for Regions 2, 8, 9, and 10 while underrepresenting Regions 3 and 7. This unique focus on Region 10 and relative exclusion of Regions 3 and 7 suggests that Ridge-S may be well-suited to identifying dispersed or less localised signals, highlighting spatial patterns that differ from the more concentrated selections seen in other methods.

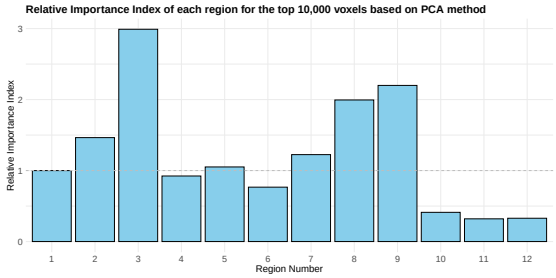
In summary, these RII findings offer a nuanced view of each method's regional priorities, shedding light on the clinical and anatomical relevance of their selections. Methods like PMS:PCA and HOLP focus on well-defined core regions, while PMS:SPCR-S-Cov takes a balanced regional approach, and Ridge-S captures more dispersed patterns. This variability underscores the potential for selecting methods based on the specific clinical or anatomical insights required in neuroimaging analyses.

4.6.2.2 Brain-Mapped Selected Voxels

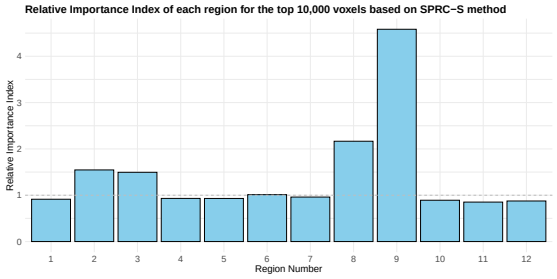
Following the RRI analysis, Figure 4.4 presents the brain-mapped distribution of the top 10,000 selected voxels for each variable screening method. This visualisation offers additional insights into the spatial patterns and selection characteristics of each method, complementing the findings from the RRI results.

Interestingly, MUA yields the smoothest voxel selection patterns among the methods examined, despite not incorporating any inter-covarying data structure or spatial information. This may be due to the inherent smoothness of the VBM data itself, which could impose a natural continuity in the predictor-response relationship. In contrast, PMS:Spatial exhibits strong, localised voxel selections, reflecting the high degree of spatial continuity enforced by its configuration, which emphasises neighbouring correlations. PMS:SPCR-S-Cov also displays a relatively smooth selection pattern, influenced by its integration of both smoothing and covariance structures, which incorporates spatial dependencies across data and response. By comparison, other methods show less spatial continuity, with selected voxels dispersed across broader areas rather than clustering in localised patterns.

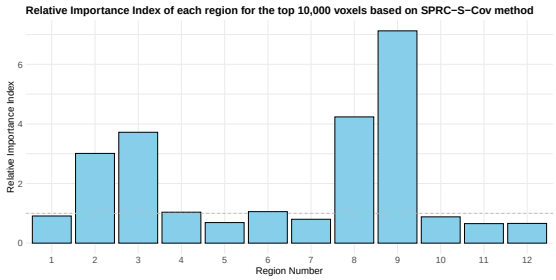
All methods, except Ridge-S, tend to prioritise central brain regions, selecting voxels from these areas in proportions aligned with their respective regional sizes. Ridge-S, however, stands out, favouring voxel selection near the brain's periphery. Edge regions are often susceptible to numerical or computational instability due to



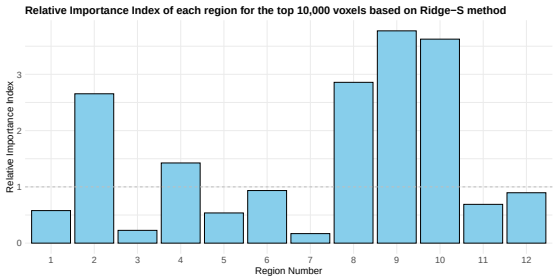
(a) PMS:PCA



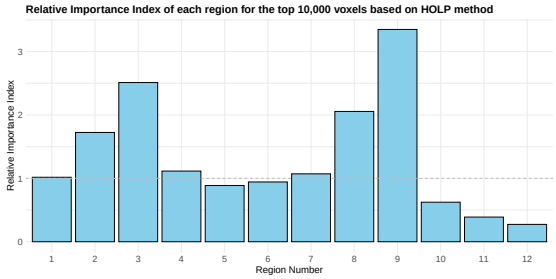
(b) PMS:SPCR-S



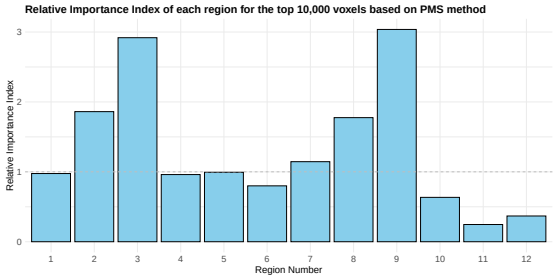
(c) SPCR-S-Cov



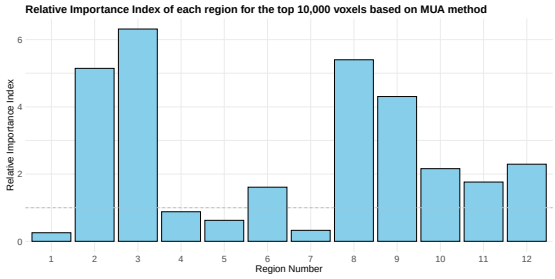
(d) Ridge-S



(e) HOLP



(f) PMS:Spatial

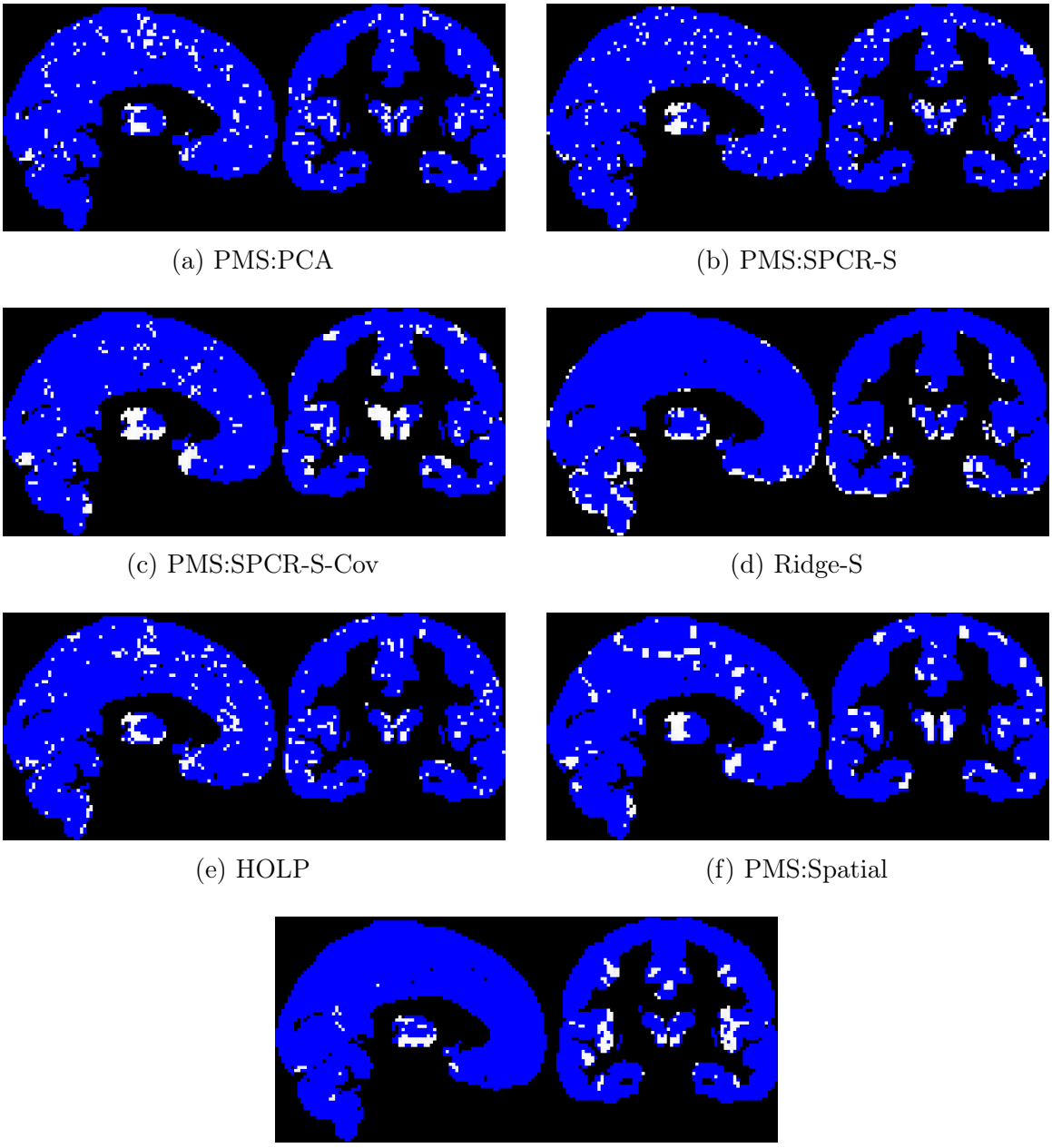


(g) MUA

Figure 4.3: Relative Importance Index of the regions among the top 10,000 variables selected by each variable screening method.

pre-processing artefacts such as motion correction, fewer neighbouring voxels, or slight misalignment during image registration. While our large sample size ($n = 2,000$) should reduce the impact of such artefacts by amplifying genuine signal patterns,

Ridge-S's distinct preference for edge voxels suggests it may be capturing an orthogonal aspect of the data. This unique selection pattern could reveal additional insights into brain age prediction, highlighting the need for further investigation to understand its implications and potential value in clinical or anatomical contexts.



(a) PMS:PCA

(b) PMS:SPCR-S

(c) PMS:SPCR-S-Cov

(d) Ridge-S

(e) HOLP

(f) PMS:Spatial

(g) MUA

Figure 4.4: Top 10,000 variables, in white, selected by each variable screening method, accounting for 6.377% of the total voxels.

4.6.3 Computational Efficiency

Table 4.9 presents the computational times for each variable screening method considered in this study across two scenarios with differing sample and parameter sizes: $n = 1,000$, $p = 15,847$ and $n = 2,000$, $p = 156,812$. It is important to note that certain methods, such as PMS:PCA, PMS:SPCR, Smoothing, Shrinkage Computation, and MUA, represent primitive and un-optimised implementations. In contrast, more established methods like HOLP, PMS, and Ridge have been optimised and are expected to perform efficiently.

For MUA, which is implemented using a for-loop, the observed timing shows a linear increase from 20 seconds for $n = 1,000$, $p = 15,847$ to 274 seconds for $n = 2,000$, $p = 156,812$. This serves as a useful baseline for comparing the scalability of other methods.

HOLP and PMS:Spatial exhibit computational times that are consistent with their lightweight configurations, with HOLP slightly slower than PMS:Spatial (10 seconds vs. 8 seconds for $n = 2,000$, $p = 156,812$), which is an anomaly and likely attributable to implementation-level inefficiencies that can be cautiously overlooked.

The primitive PMS:PCA and PMS:SPCR methods show significant increases in computation time as parameter size grows, from 44–45 seconds at $n = 1,000$, $p = 15,847$ to approximately 1,547–1,560 seconds at $n = 2,000$, $p = 156,812$. These methods, being un-optimised, demonstrate inefficiencies that could be mitigated with proper optimisation techniques or leveraging parallel processing.

Smoothing operations are particularly computationally intensive, requiring 15 minutes (900 seconds) for $n = 1,000$, $p = 15,847$ and 42 minutes (2,520 seconds) for $n = 2,000$, $p = 156,812$. While this process is necessary for applying spatial priors, it could be substantially accelerated using specialised software or efficient matrix operations for smoothing transformations as part of pre-processing.

The shrinkage computation time is 3 seconds for $n = 1,000$, $p = 15,847$, and increases to 183 seconds for $n = 2,000$, $p = 156,812$. This increase reflects the method's computational complexity as the parameter size grows, highlighting the importance of optimising this step for larger datasets.

Overall, the results indicate a clear trade-off between computational efficiency and the sophistication of each method, emphasising the importance of selecting the

appropriate method for specific use cases based on time constraints and scalability requirements.

	Total Time (seconds)	
Method	$n = 1,000$, $p = 15,847$	$n = 2,000$, $p = 156,812$
PMS:PCA	44	1547
PMS:SPCR	45	1570
Smoothing	900	2520
Shrinkage computation	3	183
HOLP	0.27	10
PMS:Spatial	0.43	8
Ridge	0.25	6
MUA	20	274
Smoothing	900	2520

Table 4.9: Computational efficiency of variable screening methods for two different cases of varying total sample and parameter sizes ($n = 1,000$, $p = 15,847$) and ($n = 2,000$, $p = 156,812$). Times are reported in seconds.

4.7 Discussion

In this chapter, we introduced several methodological advancements aimed at enhancing the performance and interpretability of variable screening methods in high-dimensional neuroimaging studies. Our primary contributions include: (1) a new approach for estimating the smoothness parameter σ in the Gaussian smoothing kernel, (2) novel methods for constructing the prior correlation matrix $\mathbf{\Lambda}$, and (3) comprehensive evaluations of these methods in both simulation and real-data analyses.

We proposed a data-driven approach to estimate the smoothness parameter σ, demonstrating its robustness across different scenarios. This parameterisation consistently resulted in stable, high-performing variable screening outcomes, contributing to more reliable and interpretable results.

Additionally, we developed new strategies for constructing a robust prior correlation matrix $\mathbf{\Lambda}$, integrated into the PMS framework. We explored both data-guided (PMS:PCA) and response-guided (PMS:SPCR) approaches, showing that these frameworks can effectively capture essential data structures for projection and covariance estimation. The methods displayed robust performance across simulation scenarios

and in real data, particularly in the age prediction task, where they outperformed traditional methods.

Overall, our proposed methods showed impressive performance metrics in sensitivity (evaluated through power at varying R^2 levels), stability (standard deviation of power across varying R^2), and screening consistency (improvements in power with increasing sample size). For example, PMS:PCA adapted well to sparse signal configurations, showing high power in sensitivity analyses. Their minimal spatial assumptions allowed them to perform comparably to methods like HOLP and MUA, which excel in detecting dispersed signals. In contrast, spatially regularised methods such as PMS:SPCR-S-Cov and PMS:SPCR-S-Proj performed better in scenarios where spatial continuity was essential, underscoring the need to align method choice with the spatial characteristics of the data.

In the real-data age prediction task, our proposed methods (PMS:PCA, PMS:SPCR-S, PMS:SPCR-S-Cov, and Ridge-S) significantly outperformed traditional methods like HOLP and PMS:Spatial, enhancing both predictive accuracy and interpretability by selecting a more informative subset of features. Notably, these methods surpassed the performance of the "Overall Ridge" model, which used all voxels, indicating that effective variable screening can yield a smaller, more interpretable model without sacrificing predictive power. This demonstrates the potential of our prior covariance configurations for broader applications in high-dimensional predictive modeling tasks beyond neuroimaging.

Our regional interpretation analysis, based on the Relative Importance Index (RII), provided additional insights into the anatomical and clinical relevance of selected features. Each method exhibited distinct regional preferences: for example, PMS:PCA and HOLP targeted well-defined core regions like the insular cortex and basal ganglia, while PMS:SPCR-S-Cov offered a balanced regional approach, and Ridge-S captured more distributed patterns across regions. These findings underscore the potential of each method to uncover specific clinical or anatomical insights depending on the focus of the analysis. Additionally, the brain-mapped selected voxels revealed that MUA achieved notably smooth selection patterns, despite not relying on spatial information, which is a phenomenon likely due to the inherent smoothness of VBM data. By contrast, methods like PMS:Spatial and PMS:SPCR-S-Cov exhibited more

localised patterns, with PMS:Spatial demonstrating strong neighbouring correlation-driven selections and PMS:SPCR-S-Cov displaying smoother patterns that integrated covariance structures across the data.

The Data-dependent priors offer the advantage of avoiding rigid, static assumptions and incorporating informative structure, however, their effectiveness relies critically on the representativeness of the data subset used for construction. If this assumption fails, the prior may introduce systematic bias toward subpopulation-specific patterns, impairing generalisability and downstream performance. Moreover, an inherent limitation of approaches such as sample covariance or PCA-based projections is that they prioritise the dominant axes of variation in the data, which may not align with the true directions of clinically meaningful effects. In high-dimensional settings like neuroimaging, important signals may reside in rare, low-variance directions that are poorly captured by unsupervised data-driven structures, leading to underestimation of subtle but clinically relevant effects. These limitation partly explain the occasional underperformance of our methods compared to simpler, assumption-free benchmarks. To mitigate this, we introduced bias correction via covariance shrinkage to regularise the prior.

The data-dependent priors offer the advantage of avoiding rigid, static assumptions and incorporating informative structure; however, their effectiveness relies critically on the representativeness of the data subset used for construction. If this assumption fails, the prior may introduce systematic bias toward subpopulation-specific patterns, impairing generalisability and downstream performance. Moreover, data-driven priors that rely on dominant variance directions can fail when clinically meaningful signals lie in low-variance directions. For instance, in early neurodegeneration studies such as mild cognitive impairment (MCI) or early Alzheimer's disease, critical atrophic changes often occur subtly in small regions like the hippocampus (Schuff et al., 2009; Honea et al., 2010). These effects do not dominate the overall variance structure and thus may be overlooked by variance-maximising methods, resulting in reduced sensitivity to important biological signals. These limitations partly explain the occasional underperformance of our methods compared to simpler, assumption-free benchmarks. To mitigate this, we introduced bias correction via covariance shrinkage to regularise the prior and counteract overfitting to the dominant but potentially

irrelevant variance components.

Moreover, an inherent limitation of approaches such as sample covariance or PCA-based projections is that they prioritise the dominant axes of variation in the data, which may not align with the true directions of clinically meaningful effects. In high-dimensional settings like neuroimaging, important signals may reside in rare, low-variance directions that are poorly captured by unsupervised data-driven structures, leading to underestimation of subtle but clinically relevant effects.

In this study, additional configurations such as PMS:PCA-S, PMS:PCA-S-Proj, and PMS:PCA-S-Cov were developed and tested. These configurations demonstrated additional effects to PMS:PCA highly comparable to when the same augmentations were applied to PMS:SPCR. To maintain clarity, these configurations were excluded from the results section. Similarly, an alternative simulation framework was investigated (see Appendix A.2 Tables A.7 and A.8), wherein, in Section 4.5.0.1, the noise term ζ^2 was held constant while varying the size of $\boldsymbol{\beta}^S$ to achieve a higher *maximum* R^2 across all methods, instead of holding the size of $\boldsymbol{\beta}^S$ constant and varying the noise term. Despite this adjustment, the comparative rankings of the methods remained unchanged relative to the primary setup.

We also investigated the feasibility of using more complex structures for $\boldsymbol{\Lambda}$, such as matrices derived from Robust PCA (Candès et al., 2011) and Partial Least Squares (PLS). Although these methods have potential in theory due to their robust handling of noise and response correlation, their computational demands rendered them impractical for large-scale studies. Our experiments demonstrated that limiting the variance explained in PCA methods yielded performance comparable to that of Robust PCA but with significantly lower computational costs.

There remains considerable potential for further refinement, especially in shrinkage-based methods. For instance, incorporating Ledoit-Wolf shrinkage to adaptively select the optimal shrinkage parameter δ for each application could yield more refined and potentially more interpretable results. Currently, the shrinkage parameter δ is set heuristically; more systematic approaches could improve predictive power and interpretability. Another possible enhancement is to replace the current shrinkage target of $\mathbf{I_p}$ (shrinking towards HOLP) with a structured neighbour correlation matrix, such as that in Eq. (4.11), which would align more closely with the spatial PMS

approach. This configuration would promote local smoothness while retaining global covariance information, resulting in a method that is anatomically coherent.

In conclusion, the advancements proposed in this chapter offer promising directions for variable screening in high-dimensional settings, providing both performance improvements and greater interpretability in neuroimaging analyses. By strategically incorporating data and response-driven structures into the prior covariance matrix, our methods lay a foundation for more adaptable and clinically meaningful models, potentially extending their utility to other domains of complex predictive modeling.

4.8 Code Availability

The code supporting this research, including that used for exploratory data analysis, model developments, simulation study design and results, as well as the real-data study implementation and results, is openly accessible and can be found at:

https://github.com/kan-k/PMS2

Chapter 5

Introduction to Bayesian Gaussian Process Neural Networks

This chapter explores scalar-on-image regression through a novel approach that integrates Gaussian processes with neural networks, advancing the feasibility of applying regression models to large-scale neuroimaging data for predicting medical outcomes. The proposed advancements address key challenges, including scalability, interpretability informed by neuroanatomical and neurofunctional prior knowledge, and uncertainty quantification in machine learning predictions. This work was conducted under the supervision of Prof. Jian Kang, Prof. Timothy D. Johnson, and Prof. Thomas E. Nichols. Additional details related to this chapter are provided in Appendix B.

5.1 Introduction

Dementia is a complex and prevalent health issue of growing importance as the world's population ages. It remains a challenge for scientists to predict and intervene in its development at an early stage. Recent studies (Cole et al., 2019; Franke and Gaser, 2012; Denver and McClean, 2018; Yankner et al., 2008) have highlighted the use of brain age as a meaningful predictor for cognitive decline. Brain age refers to a prediction of a person's chronological age based solely on brain characteristics, and the difference between brain age and chronological age, known as brain age

gap, provides a concise measure of brain health. Various imaging modalities such as T1-weighted Magnetic Resonance Imaging (MRI), functional MRI (fMRI), and Diffusion-Tensor Imaging (DTI) can be used to compute brain age, each offering a distinct perspective on an individual's brain relative to a reference population (Smith et al., 2019). However, the accuracy and generalisability of brain age studies often suffer due to limited subject pools, leading to overfitting and over-optimistic results (Varoquaux and Cheplygina, 2022). This study addresses this concern by utilising the UK Biobank's vast and diverse dataset, encompassing over 50,000 subjects, to enhance the reproducibility and generalisability of our findings.

Brain age is effectively modelled as scalar-on-image regression (SIR), where traditional approaches fall short in capturing complex spatial dependencies and nonlinear associations. To address these complexities, we propose a spatial Bayesian framework. Our framework leverages Gaussian Processes (GPs) to capture the intricate spatial relationships inherent in neuroimaging data, utilising a priori anatomical and functional information from a brain atlas.

While the Bayesian GP framework provides a solid foundation, it lacks the capacity to handle nonlinear associations effectively. This limitation necessitates the integration of neural networks, creating a synergy where neural networks address nonlinearity, and GPs offer spatial coherence and interpretability, enriching our model's analytical power. The input layer of the network corresponds to a whole 3D brain image, allowing for biological plausibility and feeds into a single hidden layer that corresponds to a priori information. Between the input and hidden layers, a non-linear activation function is applied to account for non-linearity in the predictors. Additionally, Gaussian process priors are imposed on the network weights between each layer, acting as a natural enforcer for spatial information in a 3D Euclidean space.

A distinctive element of our methodology is its emphasis on uncertainty quantification, which is often overlooked in traditional machine learning paradigms. This focus is critically relevant in the field of brain age prediction, where accurately quantifying the uncertainty associated with each prediction can essentially enhance the model's utility and reliability. Similar uncertainty quantification for brain age prediction frameworks have recently been proposed by using quantile regression (Palma et al., 2020) and conformal prediction (Ernsting et al., 2023). By integrating this feature, our model

provides users with insights into the credibility of individual predictions, allowing for a more delicate evaluation of the results.

Our approach refines the Bayesian GP-induced neural network model. We incorporate low-rank projections, stochastic learning, and Bayesian conjugate updates to effectively manage the computational challenges typically associated with large-scale neuroimaging data. Interpretable model learning is achieved through the utilisation of prior information from the GPs, coupled with the induction of a biologically plausible network architecture. This allows for a biologically-motivated, brain-mapped model learning representation, potentially promoting the clinical understanding of the model (Scheinost et al., 2019).

Furthermore, by applying Stochastic Gradient Langevin Dynamics, we efficiently sample from posterior distributions, circumventing the computational burdens of traditional MCMC methods (Welling and Teh, 2011; Robert et al., 1999). This framework's ability to make predictions using a posterior mean and to quantify each subject's uncertainty is a key advantage over most existing literature on Brain Age prediction. The derived uncertainty information can be employed in subsequent analyses of brain age and for comparisons within the study.

In this chapter, we demonstrate the efficacy of our model using T1-weighted voxel-based morphometry (VBM) data, containing over 150,000 active voxels with a moderate baseline sample size of 4,000 subjects from the UK Biobank. VBM data captures grey matter intensity and has been found to be the best imaging modality among single-modality age prediction models (Liem et al., 2017a). This application not only underscores the model's capabilities in handling large-scale neuroimaging data but also showcases its potential in providing insightful and reliable predictions in the study of brain age.

5.2 Method

5.2.1 Gaussian Process Neural Network

We develop a Gaussian Process neural network (GPNN) for making predictions on clinical outcomes, e.g. brain age, disease onset, fluid intelligence, etc, using imaging

biomarkers. For each subject i, we have a voxel-level image data $\{X_i(\mathbf{v})\}_{\mathbf{v}\in\mathcal{B}}$ where $\mathcal{B} \subset \mathbb{R}^3$ is a collection of brain imaging voxels. We make use of anchor points $\mathcal{A} \subseteq \mathcal{B}$, corresponding to region centroids in a pre-specified atlas. We also have a single outcome variable of interest y_i. The model is a one-layer spatially-varying neural network that takes in the voxel-level image $\{X_i(\mathbf{v})\}_{\mathbf{v}\in\mathcal{B}}$ with $V = |\mathcal{B}|$ input units, or predictors, one hidden layer consisting of hidden spatially-varying units. The GPNN is given by

$$g[\mathbb{E}(y_i \mid \{\mathbf{X}_i(\mathbf{v}')\}_{\mathbf{v}'\in\mathcal{B}})] = \alpha_0 + \sum_{\mathbf{v}\in\mathcal{A}} \alpha(\mathbf{v}) h\Big\{\eta_0(\mathbf{v}) + \sum_{\mathbf{v}'\in\mathcal{B}} \eta(\mathbf{v},\mathbf{v}')\mathbf{X}_i(\mathbf{v}')\Big\} \tag{5.1}$$

where $g(\cdot)$ is a link function based on the target type, and $h(\cdot)$ is an activation function. In this work, we let $g(\cdot)$ be an identity function and denote $\hat{y}_i := \mathbb{E}(y_i \mid \{\mathbf{X}_i(\mathbf{v}')\}_{\mathbf{v}'\in\mathcal{B}})$, and $h(\cdot)$ be the *Rectified Linear Unit* (ReLU) function: $h(x) = \max\{0, x\}$. The coefficients $\alpha(\mathbf{v})$ and $\eta(\mathbf{v},\mathbf{v}')$ are the parameters of interest where $\eta(\mathbf{v},\mathbf{v}')$ represents the effect of imaging input voxel $\mathbf{v}'$ on the hidden neuron $\mathbf{v}$, and $\alpha(\mathbf{v})$ similarly represents the effect of the hidden neuron $\mathbf{v}$ on the output unit. The coefficients $\eta_0(\mathbf{v})$ and α_0 serve as a constant shift for each hidden neuron and the output unit respectively. We assume the likelihood of the outcome variable follows a Gaussian distribution with a known variance σ^2, i.e. $\mathbf{y} \sim N(f(\mathbf{X}), \sigma^2 I_n)$.

Figure 5.1 depicts the GPNN architecture in an example case of 5 regions of interest (ROI), i.e. $|\mathcal{A}| = 5$. The architecture represent a feedforward network consisting of the input layer, a latent Gaussian Process transformation of the input layer, a dense hidden layer, a Gaussian Process transformation of the hidden layer, and an output layer. Note that the bias terms, i.e. α_0 and $\eta_0(\mathbf{v})$, are omitted in this visualisation. The visual representations of other models investigated in this study is shown in Figure B.1 in the Appendix.

5.2.2 Prior specifications and low-rank representation

In order to capture sparsity and piecewise smoothness of spatial varying weights, we assigns spatially varying properties to our coefficients of interest $\alpha(\mathbf{v})$ and $\eta(\mathbf{v},\mathbf{v}')$ by imposing independent Gaussian Process (GP) priors. Under some mild regularity

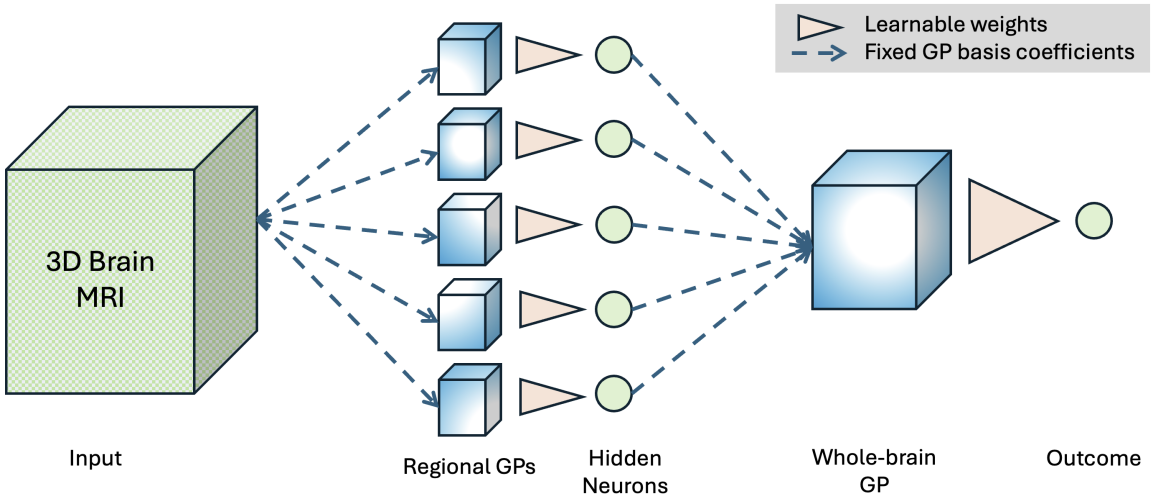


Figure 5.1: Visualisation of the proposed GPNN architecture. The input region is transformed into a low-rank latent representation based using associated groups and anchors. The transformed data is fully connected to a lower-dimensional hidden layer, which is further transformed into another latent representation before producing a scalar output. The example diagram illustrates a GPNN architecture with 5 regions of interest.

conditions, we approximate the GPs through a combination of Mercer's theorem for kernel eigendecomposition (Theorem B.1) and Karhunen-Loève (KL) expansion in (Corollary B.1). The approximation of a GP $f(\mathbf{v})$, can be expressed as a finite sum:

$$f(\mathbf{v}) \approx \mu(\mathbf{v}) + \sum_{\ell=1}^{L} \theta_\ell \sqrt{\lambda_\ell} \psi_\ell(\mathbf{v}). \tag{5.2}$$

where L is a pre-specified highest degree of expansion, $\mu(\mathbf{v})$ is a mean function of the GP evaluated at $\mathbf{v}$, θ_ℓ are unknown such that $\theta_\ell \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ for some variance σ^2. The terms λ_ℓ and $\psi_\ell(\mathbf{v})$ are eigenvalues and orthonormal eigenfunctions of a chosen covariance kernel evaluated at $\mathbf{v}$ respectively, both of which can be pre-computed and effectively fixed during parameter optimisation.

Specifically, the approximation of the bivariate spatially varying coefficient (SVC) $\eta(\mathbf{v}, \mathbf{v}')$ with mean zero can be expressed as

$$\eta(\mathbf{v}, \mathbf{v}') \approx \sum_{\ell=1}^{L(\mathbf{v})} \gamma_\ell(\mathbf{v}', \mathbf{v}) \theta_\ell(\mathbf{v}),$$

where $\gamma_\ell(\mathbf{v}', \mathbf{v})$ capture the deterministic components that encapsulate the spatial dependence of the corresponding neural layer, ensuring smoothness and interpretability. Similarly, for the univariate SVC $\alpha(\mathbf{v})$ with mean zero, we have

$$\alpha(\mathbf{v}) \approx \sum_{k=1}^{K} \rho_k(\mathbf{v})\psi_k,$$

with $\rho_k(\mathbf{v})$ playing a similar role in ensuring smoothness.

The terms $\theta_\ell(\mathbf{v})$ and ψ_k are unknown coefficients assigned Gaussian priors as per Mercer's theorem. Specifically,

$$\theta_\ell(\mathbf{v}) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2\tau^2(\mathbf{v})) \quad \text{and} \quad \psi_k \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2\omega^2),$$

where σ^2 represents the unknown variance of the population response variable, ω^2 is a network-specific hyperparameter, and $\tau^2(\mathbf{v})$ is a data-dependent hyperparameter.

The Gaussian Process priors simplify the problem by projecting the space into a lower-dimensional latent representation, as stated in Corollary B.1. For example, consider an input space of size $\mathbb{R}^p$. With an eigendecomposition matrix $\mathbb{R}^{p\times L}$, where L is the total number of basis expansions of $\eta(\mathbf{v}, \mathbf{v}')$ and $L \ll p$, the corresponding latent representation space becomes $\mathbb{R}^L$. This reduction in dimensionality significantly reduces computational complexity, particularly in the input layer, making parameter optimisation more efficient.

5.2.3 Posterior computation

We assume the distribution of the outcome variable follows a Gaussian distribution with a known variance σ^2, i.e. $\mathbf{y} \sim N(f(\mathbf{X}), \sigma^2 I_n)$, where $f(\cdot)$ is the GPNN function, and $\mathbf{X}$ is the imaging data. Let $\boldsymbol{\Theta} := \left\{\alpha_0, \{\boldsymbol{\psi}\}, \{\boldsymbol{\eta_0}\}, \{\boldsymbol{\theta}\}, \sigma^2\right\}$ denote the set of the unknown parameters which have Gaussian prior distributions as specified in Section 5.2.2. Consequently, the posterior density of the parameters can be expressed as

$$P(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X}) \propto P(\mathbf{y}|\mathbf{X}; \boldsymbol{\Theta})P(\boldsymbol{\Theta}). \tag{5.3}$$

Due to the architecture of the neural network, there is no a closed form solution or guarantees of identifiability. We resort to finding the Maximum-a-Posteriori (MAP) via gradient descent method by setting the loss function $C(\boldsymbol{\theta})$ in terms of negative log posterior density, i.e.

$$C(\boldsymbol{\Theta}) = -\log\big\{P(\mathbf{y}|\mathbf{X};\boldsymbol{\Theta})\big\} - \log\big\{P(\boldsymbol{\Theta})\big\}. \tag{5.4}$$

This gives rise to an optimisation problem, characterised as

$$\arg\min_{\boldsymbol{\Theta}} \ \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(y_i - \hat{y}(X_i)\Big)^2 \tag{5.5}$$

$$+ \frac{K}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\frac{1}{\omega^2}\sum_{k=1}^{K}\psi_k^2 + \frac{1}{2}\alpha_0^2 \tag{5.6}$$

$$+ \sum_{\mathbf{v}\in\mathcal{A}}\frac{L_{\mathbf{v}}}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\mathbf{v}\in\mathcal{A}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\ell=1}^{L(\mathbf{v})}\theta_\ell(\mathbf{v})^2 + \frac{1}{2}\sum_{\mathbf{v}\in\mathcal{A}}\eta_0(\mathbf{v})^2, \tag{5.7}$$

where $\hat{y}(X_i)$ is the prediction function, defined by

$$\hat{y}(X_i) = \alpha_0 + \sum_{k=1}^{K}\psi_k\sum_{\mathbf{v}\in\mathcal{A}}\rho_k(\mathbf{v})h\Big(\eta_0(\mathbf{v}) + \sum_{\ell=1}^{L(\mathbf{v})}\theta_\ell(\mathbf{v})\sum_{\mathbf{v}'\in\mathcal{B}}X_i(\mathbf{v}')\gamma_\ell(\mathbf{v}',\mathbf{v})\Big). \tag{5.8}$$

This specific architecture is referred to as "GPNN-GP-GP", denoting a Gaussian Process Neural Network with GP transformations applied between the input layer and the hidden layer, and an additional whole-brain GP transformation applied between the (activated) hidden layer and the output layer.

5.2.3.1 Architecture variants

Depending on the complexity of the relationship between the imaging features and the outcome variable, utilising Gaussian Processes in both layers can lead to under- or overfitting. We make the following alternative proposals to address such issues:

1. Applying GP transformations in the input layer, and a direct (linear) connection between the (activated) hidden layer and the output layer, "GPNN-GP-Linear", i.e. set-

ting $\alpha(\mathbf{v}) = \psi(\mathbf{v})$, and $\eta(\mathbf{v}, \mathbf{v}') \approx \sum_{\ell=1}^{L(\mathbf{v})} \gamma_\ell(\mathbf{v}', \mathbf{v})\theta_\ell(\mathbf{v})$, such that $\theta_\ell(\mathbf{v}) \sim N(0, \sigma^2\tau^2(\mathbf{v}))$ and $\eta_0(\mathbf{v}) \sim N(0, 1)$.

$$\begin{aligned}\arg\min_{\boldsymbol{\Theta}} \ & \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(y_i - \hat{y}(X_i)\Big)^2 \\ & + \sum_{\mathbf{v}\in\mathcal{A}}\frac{L(\mathbf{v})}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\mathbf{v}\in\mathcal{A}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\ell=1}^{L(\mathbf{v})}\theta_\ell(\mathbf{v})^2 + \frac{1}{2}\sum_{\mathbf{v}\in\mathcal{A}}\eta_0(\mathbf{v})^2, \end{aligned} \tag{5.9}$$

where $\hat{y}(X_i)$ is the prediction function, defined by

$$\hat{y}(X_i) = \alpha_0 + \sum_{\mathbf{v}\in\mathcal{A}}\psi(\mathbf{v})\Big(\eta_0(\mathbf{v}) + \sum_{\ell=1}^{L(\mathbf{v})}\theta_\ell(\mathbf{v})\sum_{\mathbf{v}'\in\mathcal{B}}X_i(\mathbf{v}')\gamma_\ell(\mathbf{v}', \mathbf{v})\Big). \tag{5.10}$$

2. Directly connecting the input layer to the hidden layer (then perform a non-linear activation, e.g. ReLU, post hoc), and a apply a whole-brain GP transformation between the (activated) hidden layer and the output layer, "GPNN-Linear-GP", i.e. Setting $\alpha(\mathbf{v}) \approx \sum_{k=1}^{K}\rho_k(\mathbf{v})\psi_k$, and $\eta(\mathbf{v}, \mathbf{v}') = \theta(\mathbf{v}, \mathbf{v}')$, such that $\psi_k \sim N(0, \sigma^2\omega^2)$, $\alpha_0 \sim N(0, 1)$, $\theta(\mathbf{v}, \mathbf{v}') \sim N(0, \sigma^2\tau^2(\mathbf{v}))$ and $\eta_0(\mathbf{v}) \sim N(0, 1)$.

$$\begin{aligned}\arg\min_{\boldsymbol{\Theta}} \ & \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(y_i - \hat{y}(X_i)\Big)^2 \\ & + \frac{K}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\frac{1}{\omega^2}\sum_{k=1}^{K}\psi_k^2 + \frac{1}{2}\alpha_0^2 \\ & + \frac{|\mathcal{A}||\mathcal{B}|}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\mathbf{v}\in\mathcal{A}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\mathbf{v}'\in\mathcal{B}}\theta(\mathbf{v}, \mathbf{v}')^2 + \frac{1}{2}\sum_{\mathbf{v}\in\mathcal{A}}\eta_0(\mathbf{v})^2, \end{aligned} \tag{5.11}$$

where $\hat{y}(X_i)$ is the prediction function, defined by

$$\hat{y}(X_i) = \alpha_0 + \sum_{k=1}^{K}\psi_k\sum_{\mathbf{v}\in\mathcal{A}}\rho_k(\mathbf{v})h\Big(\eta_0(\mathbf{v}) + \sum_{\mathbf{v}'\in\mathcal{B}}X_i(\mathbf{v}')\theta(\mathbf{v}, \mathbf{v}')\Big). \tag{5.12}$$

5.2.3.2 Minibatch training and Learning rate specification

When working large data in a neural networking setting, updating parameters using the entire training dataset is computationally infeasible. Instead, it is a common

practice to employ minibatches to approximate an unbiased estimator of the learning gradient.

The unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\theta}, \boldsymbol{\eta}_0)$ w.r.t. $\theta(\mathbf{v}, \mathbf{v}')$ is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial \theta(\mathbf{v}, \mathbf{v}')} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\alpha(\mathbf{v})h'(G_i(\mathbf{v}))X_i(\mathbf{v}')\big\}\Big] + \frac{1}{\sigma^2\tau_{\mathbf{v}}^2}\theta(\mathbf{v}, \mathbf{v}'). \tag{5.13}$$

where $G_i(\mathbf{v}) := \eta_0(\mathbf{v}) + \sum_{\mathbf{v}'\in\mathcal{B}} X_i(\mathbf{v}')\theta_{\mathbf{v}\mathbf{v}'}$, and m is the minibatch size.

The corresponding gradient update step at iteration k of the algorithm is

$$\theta(\mathbf{v}, \mathbf{v}')^{(k+1)} = \theta(\mathbf{v}, \mathbf{v}')^{(k)} \tag{5.14}$$

$$- \nu\Big\{n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{(\sigma^{(k)})^2}\big\{(y_i - \hat{y}_i^{(k)})\alpha^{(k)}(\mathbf{v})h'(G_i^{(k)}(\mathbf{v}))X_i(\mathbf{v}')\big\}\Big] \tag{5.15}$$

$$+ \frac{1}{(\sigma^{(k)})^2(\tau^{(k)})_{\mathbf{v}}^2}\theta(\mathbf{v}, \mathbf{v}')^{(k)}\Big\} \tag{5.16}$$

with $\nu \in (0, \frac{1}{n})$ so that $\nu n \in (0, 1)$. Similarly, the MAP gradients w.r.t. the parameters η_0 and σ^2 can be derived.

To account for generalisability and complexity of the problem, an adaptive learning rate can be employed to avoid manual tuning of ν. We employ the Barzilai-Borwein (BB) (Tan et al., 2016), which utilises steepest descent to guide the update, is efficient for large-scale optimisation problems. This method, however, does not satisfy the Robbin-Monro condition for SGD, but conveniently, the Smooth extension (BB-Smoothing) of it can be implemented as suggested by the authors.

5.3 Bayesian Inference

5.3.1 Stochastic Bayesian inference

Langevin Dynamics (LD) is a Markov Chain Monte Carlo (MCMC) technique (Neal et al., 2011) that leverages the concept of Hamiltonian dynamics. It involves updating the proposal state based on the gradient of the target probability density function and the addition of independent Gaussian noise to simulate a random walk. To enhance the

mixing and convergence properties of the random walk, an adaptation of LD called the Metropolis-adjusted Langevin algorithm (MALA) incorporates a Metropolis-Hastings accept/reject mechanism. However, like other generic MCMC algorithms, MALA encounters the curse of dimensionality and requires evaluating probabilities over the entire dataset. To address these challenges, (Welling and Teh, 2011) proposed the Stochastic Gradient Langevin Dynamics (SGLD) algorithm, which combines LD with Stochastic Gradient Descent (SGD). SGLD follows the same principles as MALA, utilising gradient-based updates and random walks, but it iterates through small batches of data, allowing for averaging over time. Moreover, SGLD avoids the need for the accept/reject step of Metropolis-Hastings by ensuring that the rejection rate asymptotically approaches zero.

The SGLD expression for deriving the gradient of update of a parameter $(\cdot)$ is given by

$$\nabla_{SGLD}(\cdot) = \nu\Big\{\frac{n}{m}\sum_{i=1}^{m}\frac{\partial \log P(\mathbf{y}_i|\mathbf{X}_i;\boldsymbol{\Theta})}{\partial\,\cdot} + \frac{\partial \log P(\boldsymbol{\Theta})}{\partial\,\cdot}\Big\} + \zeta \tag{5.17}$$

where $\zeta \sim N(0, 2\nu)$ is the injected noise, which is designed to balance with the learning rate ν so that the variance of the samples matches that of the posterior.

The SGLD algorithm enables the estimation of not only the MAP estimate of the parameters $\boldsymbol{\Theta}$ but also its full posterior distribution by employing sampling techniques. This provides the opportunity to utilise advanced statistical metrics, including posterior parameter estimation, quantification of parameter uncertainty, and computation of the posterior (predictive) mean. These additional metrics allow for a more comprehensive analysis and interpretation of the model's results and facilitate a deeper understanding of the underlying uncertainty associated with the estimated parameters.

5.3.2 Adaptive Gaussian priors

The Gaussian priors in Section 5.2.1 have their variances characterised by the product $\sigma^2\tau_{\mathbf{v}}^2$ where σ^2 signifies the contribution of the target data noise and $\tau_{\mathbf{v}}^2$ are the hyper-variances directly associated with the priors which are of interest in this section. A common noninformative, conjugate hyperprior distribution is used in this study,

namely, Inverse-Gamma distribution. With its conjugacy, the algorithm is able to precisely compute the closed-form Bayesian updates after each iteration. Inverse-Gamma distribution has 2 free parameters, `shape` and `scale` , and their posterior updates are given by (A.3). Specifically, define the minibatch gradient update to be

$$\nabla_{SGD}(\cdot) := \frac{n}{m}\sum_{i=1}^{m}\frac{\partial \log P(\mathbf{y}_i|\mathbf{X}_i;\boldsymbol{\Theta})}{\partial \cdot} + \frac{\partial \log P(\boldsymbol{\Theta})}{\partial \cdot}, \tag{5.18}$$

which corresponds to the MAP estimate from a minibatch with respect to a paramter $(\cdot)$.

Algorithm (1) outlines the parameter update mechanism for the parameter the optimisation.

Algorithm 1: Gaussian Process Neural Network (GPNN) parameter and hyperparameter updates

Input: Data $= \{\mathbf{y}, \mathbf{X}\}$, $\texttt{shape}_{0,\mathbf{v}} > 4, \texttt{scale}_{0,\mathbf{v}} > 0$, learning rate ν.
Result: $\boldsymbol{\theta}, \boldsymbol{\eta}_0, \boldsymbol{\psi}, \boldsymbol{\alpha}, \sigma^2$
Initialisation:
Set $\tau^2(\mathbf{v}) = \frac{\texttt{scale}_{0,\mathbf{v}}}{\texttt{shape}_{0,\mathbf{v}}-1}$
Draw $\theta_\ell(\mathbf{v}) \sim N(0, \sigma^2\tau^2(\mathbf{v})), \eta_0(\mathbf{v}) \sim N(0,1)$
for $e = 1, ..., E$ **do**
  Randomly partition the training set into minibatches $\{1, ..., B_e\}$, each with cardinality $= m$.
  for $b = 1, ..., B_e$ **do**
    Update the parameters
      Obtain hidden-layer parameters ψ_k and α_0 by fitting a linear model
      $\theta_\ell(\mathbf{v}) \leftarrow \theta_\ell(\mathbf{v}) - \nu\nabla_{SGD}(\theta_\ell(\mathbf{v}))$
      $\eta_0(\mathbf{v}) \leftarrow \eta_0(\mathbf{v}) - \nu\nabla_{SGD}(\eta_0(\mathbf{v}))$
      $\sigma^2 \leftarrow \sigma^2 - \nu\nabla_{SGD}(\sigma^2)$
    Update the hyperparameteters
      $\texttt{shape}_{\mathbf{v}} \leftarrow \texttt{shape}_{0,\mathbf{v}} + \frac{L(\mathbf{v})}{2}$
      $\texttt{scale}_{\mathbf{v}} \leftarrow \texttt{scale}_{0,\mathbf{v}} + \frac{\sum_{\ell=1}^{L(\mathbf{v})}\theta_\ell(\mathbf{v})^2}{2\sigma^2}$
      $\tau^2(\mathbf{v}) \sim$ Inverse-Gamma$(\texttt{shape}_{\mathbf{v}}, \texttt{scale}_{\mathbf{v}})$
  end
  Compute ν using BB or BB-Smoothing method
end

5.3.3 Measurements of interest

5.3.3.1 Saliency map

A key tool for the interpretation of neural networks is *saliency maps*, which allow for the investigation of the network's parameter updates during model training. A prediction function $\hat{y}(\cdot)$ is defined with the VBM image $\mathbf{X}_i$ as input for participant i. The gradient of $\hat{y}(\cdot)$ with respect to the input VBM image $\mathbf{X}_i$ is then computed, providing a measure of voxel importance for the neural network's prediction. By visualing these gradients in image form, a saliency map can be generated, highlighting the voxels that are most critical for the neural network's prediction. These maps can help provide insights into the neural network's decision-making process and identify brain regions that are particularly relevant for the task being performed. In partuclar, the prediction function $\hat{y}(\cdot)$ with VBM image $\mathbf{X}_i$ for participant i as input is expressed as:

$$\hat{y}(\mathbf{X}_i) = \alpha_0 + \sum_{\mathbf{v}\in\mathcal{A}} \alpha(\mathbf{v}) h\Big(\eta_0(\mathbf{v}) + \sum_{\mathbf{v}'\in\mathcal{B}} X_i(\mathbf{v}')\eta(\mathbf{v},\mathbf{v}')\Big). \tag{5.19}$$

Then the saliency at each voxel is given by

$$\frac{\partial \hat{y}(\mathbf{X}_i)}{\partial X_i(\mathbf{v}')} = \sum_{\mathbf{v}\in\mathcal{A}} \Big[\alpha(\mathbf{v}) h'\Big(\eta_0(\mathbf{v}) + \sum_{\mathbf{v}''\in\mathcal{B}} X_i(\mathbf{v}'')\eta(\mathbf{v}'',\mathbf{v})\Big)\eta(\mathbf{v},\mathbf{v}')\Big]. \tag{5.20}$$

5.3.3.2 Posterior predictive interval

In Bayesian inference, a posterior predictive interval measure can be constructed to take into account the uncertainty of the parameters $\boldsymbol{\theta}$ calibrated to individual predictions. The α posterior predictive interval can be expressed as:

$$\mathrm{PPI}_i := \bar{y}_i \pm z_\alpha \sqrt{\mathrm{MSE} + \mathrm{PPV}_i}, \tag{5.21}$$

where $\bar{y}_i$ is the posterior predictive mean of subject i, z_α is the standard Gaussian score at level $\left(1 - \frac{(1-\alpha)}{2}\right)$, MSE is the Mean Squared Error of the fitted model which takes into account the overall model error, and PPV_i is the posterior predictive variance of subject i which takes into account parameter uncertainty at the individual level.

With a posterior predictive interval, a comparative statistics can be defined by examining the empirical coverage rate (ECR) of a given Bayesian model across the dataset. ECR is defined as the proportion of times the true value falls within the predicted interval

$$\mathrm{ECR}_{\alpha\%} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\mathrm{PPI}_i}(y_i) \tag{5.22}$$

where $\mathbb{1}_A(\cdot)$ is the indicator function for set A.

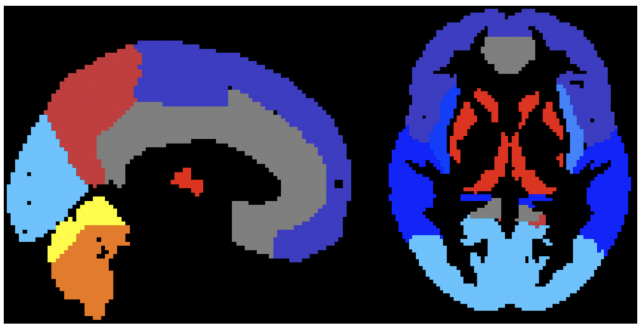
5.3.4 GP construction and parameter tuning

The `BayesGPfit`[1] package, available in R, provides a convenient toolbox for constructing a Gaussian process (GP) using expression (5.2). The GP function employs a Modified Squared Exponential (MSE) kernel with control over the concentration a, smoothness b, and number of eigendecomposition L parameters. Additionally, the central point or anchor of the GP can also be specified, providing four degrees of freedom over the parameterisation of a GP.

We make use of two kinds of GPs, regional and whole-brain. Regional GPs are based on a brain atlas that partitions the brain and also determines the anchor points as the centroid of each region. A single regional GP is determined by one anchor point and is composed of R GPs, one for each region, all sharing a common anchor point. The whole brain GPs uses a single anchor point and consists of a single GP for the entire brain. Thus in total, our model has $R^2 + 1$ GPs.

To generate a GP for each region centroid, an optimal combination of parameters is found by running a grid search on the sets $\{1, 2, 3, 4, 5\}$ and $\{30, 40, 50, 60, 70, 80, 90, 100\}$ for the concentration and smoothness parameters, respectively. Let $c_1, ..., c_R$ denote the set of region centroids and $\{\mathcal{P}_1, ..., \mathcal{P}_R\}$ denote the set of combinations of GP parameters, with each $\mathcal{P}i = a_i, b_i, L_i$ being the combination of the GP's concentration parameter a_i, the smoothness parameter b_i, and the number of K-L expansion terms L_i. Then, a GP in the neural network is defined by $GP(0, \kappa MSE; c_i, \mathcal{P}_j)$, where $i, j = 1, ..., R$. Index i corresponds to the neuron i in the hidden layer, and index j corresponds to the region j to which each input belongs. Figure 5.2 illustrates the

[1] github.com/kangjian2016/BayesGPfit



(a) An example mask, where the light-blue region is Region 1.



(b) The first K-L basis expansion at each voxel $\mathbf{v}'$ of $\eta(R_1, \mathbf{v}')$

Figure 5.2: An example of a bivariate GP; the leading term of the K-L basis expansion at each voxel based with GPs concentrating around the centroid of Region 1.

leading K-L basis expansion term at each voxel according to different regions in the atlas with a concentration around the centroid of Region 1.

5.4 Dataset and Pre-Processing

In this experimental study, the data from the UKBiobank study is used, which currently contains a variety of imaging data of over 40,000 subjects. A total of 4,000 subjects with T1-weighted MRI scans and the corresponding Voxel-Based Morphometry (VBM) data are randomly selected for analysis, with the goal of predicting each participant's age based on their scan and subsequently acquiring information on the Brain Age Gap (BAG). Each VBM image is a 3-dimensional image of size 91 x 109 x 91 with 2mm^3 voxel unit. For prior knowledge, a Gaussian process (GP) kernel is used, which incorporates local spatial structure including the 3D coordinates and brain regions of an atlas. One derived atlas from Tziortzi et al. (2011) is selected, with 12 Regions of

Interests (ROI), in order to provide insight into which grouped voxels should behave similarly based on anatomical regions or brain functions.

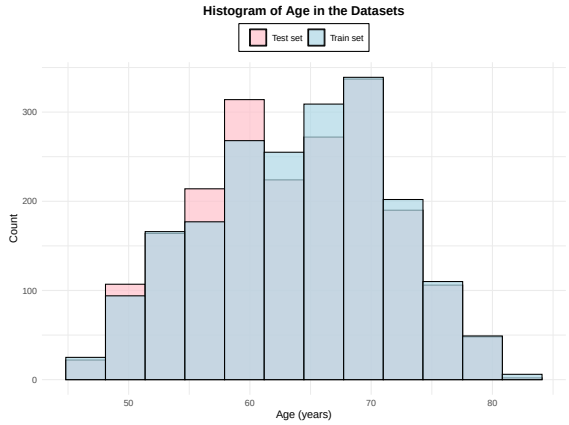
A brain mask is computed by averaging all subject's VBM images and thresholding at 0.1, removing voxels with little or no grey matter. This mask is then applied to the atlas, ensuring that each atlas region consists only of grey matter voxels. The 4,000 subjects are randomly split into a train set of 2,000 and a test set of 2,000. Equally balanced target distributions of the train and test sets are shown for real-data and simulation experiments in Figure 5.3. Additionally, statistical summaries of voxel-level and region-level mean and variance of the VBM images in both sets demonstrate that the train-test splits also have balanced domain distributions. Furthermore, an additional non-overlapping set of 4,000 subjects is also considered to extend the analysis of the models in this study.

While data standardisation, such as centering and scaling to unit variance, is commonly used in high-dimensional machine learning techniques, raw VBM data is used in this study where the lowest variance voxels correspond to low volume of gray matter while high variance voxels are typically are in core cortical areas. Thus, we considered our analysis both with and without data standardisation.

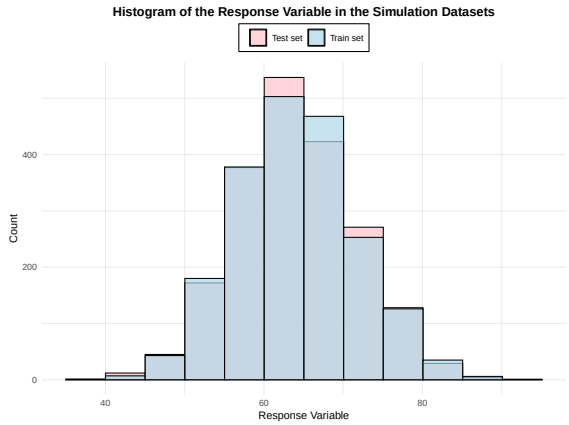
5.5 Results

The accuracy metrics used predominantly in this study is *Root Mean Squared Error* (RMSE) and *Coefficient of Determination* (R^2). The metrics are computed based on the MAP for GPNN-algorithms, and based on the posterior mean for GPNN-SGLD algorithm.

The neural networks are trained on multiple successive epochs. During each epoch, the training data is randomly split into partitioning minibatches, each containing 500 training samples. All first-layer network parameters are randomly initialised from the *a priori* Gaussian distributions. The target variance parameter in the likelihood σ^2 is set to be the variance of the training data. The initial prior variances are set to $\tau^2(\mathbf{v}) = 0.05$. The initial Inverse-Gamma parameters are $(\texttt{shape}_\mathbf{v} = 11, \texttt{scale}_\mathbf{v} = 0.5)$. The learning rate is calculated using BB adaptive learning rate method to search for a good candidate set of initial network parameters prior to running the BB-smoothing



(a) Target distribution (true age) of real-data



(b) Target distribution of simulated response data

Figure 5.3: Target distribution of train and test datasets in real-data and simulation studies. The overlapping areas indicate the overlap between the train and test distributions.

method to ensure the Robbins-Monro condition. The first-layer GPs are constructed according to Section 5.3.4 with 1,771 basis elements. However, the second-layer GP is constructed based on 283 basis elements.

5.5.1 Simulation study

5.5.1.1 Simulation set-up

For the simulation study, each image was sub-sampled by taking every 10th masked voxel. This sub-sampling approach effectively reduces the dimensionality of the data, thereby accelerating the analysis process while preserving spatial information across the three-dimensional space, which is a critical focus of this study.

To ensure spatial dependence in the simulated data, a whole-brain Gaussian Process regression (GPR) was fitted to real data, yielding the fitted intercept $\hat{\alpha}$ and the coefficients $\hat{\beta}$ corresponding to the input voxels. The `BayesGPfit` and `fastBayesReg`[2] libraries were used to fit an approximated GP regression model with a modified squared exponential kernel. The model parameters were set as follows: concentration $a = 2$, smoothness $b = 40$, and number of basis expansions $L = 286$.

Given the fitted model $\widehat{GPR}(\cdot)$, the in-sample predictions for the training set $(i = 1, \ldots, 2000)$ were obtained using

$$\hat{y}_i = \hat{\alpha} + \mathbf{X}_i\hat{\beta}. \tag{5.23}$$

The in-sample mean squared error (MSE), denoted as $\hat{\sigma}^2_*$, was computed using the formula

$$\hat{\sigma}^2_* = \frac{1}{2000}\sum_{i=1}^{2000}(\hat{y}_i - y_i)^2. \tag{5.24}$$

Out-of-sample predictions were then generated using $\widehat{GPR}(\cdot)$ for indices $i = 2001, \ldots, 4000$. Synthetic response variables $\tilde{y}_i$ were created for each subject i by sampling from the Gaussian distribution

$$\tilde{y}_i \sim N(\hat{\alpha} + \mathbf{X}_i\hat{\beta}, \hat{\sigma}^2_*), \tag{5.25}$$

[2] `https://github.com/kangjian2016/fastBayesReg`

or equivalently,

$$\tilde{y}_i \sim N(\hat{y}_i, \hat{\sigma}_*^2), \tag{5.26}$$

for $i = 1, \ldots, 4000$. This construction ensures that the true spatially-induced coefficients, the target distribution, and the noise are known, providing a benchmark for evaluating the methods under investigation.

The optimisation neural networks were trained over 150 epochs, with each epoch comprising four minibatches of 500 training samples. The target variance parameter in the likelihood, σ^2, was initialised to the variance of the training data. Prior variances were initialised to $\tau_{\mathbf{v}}^2 = 0.05$, with inverse-gamma (IG) prior parameters set to $\alpha_{\mathbf{v}} = 11$ and $\beta_{\mathbf{v}} = 0.5$. The first-layer GPs were constructed using 286 bases as described in Section 5.3.4, while the second-layer GP was constructed using 120 bases.

The methods evaluated in this simulation study included Bayesian models: (1) the Horseshoe Prior (HSP) model, a sparsity-inducing Bayesian approach; (2) Gaussian Process Regression (GPR); (3) GPNN-Linear-GP with 12 Regions (GPNN-Linear-GP-12R); and (4) GPNN-Linear-GP with 52 Regions (GPNN-Linear-GP-52R).

5.5.1.2 Performance

The results of the simulation study are summarised in Table 5.1, presenting the RMSE and R^2 for training and test datasets across the evaluated methods. The best-performing model, based on both test RMSE and R^2, is the GPR model. GPR achieves a test RMSE of 4.80 (0.003) and a test R^2 of 60.1% (0.047%), outperforming the other methods in terms of predictive accuracy.

The HSP model demonstrates superior training performance with a training RMSE of 2.97 (0.129) and a training R^2 of 84.6% (1.343%). However, its test RMSE of 5.34 (0.032) and test R^2 of 50.7% (0.582%) reveal a substantial drop in performance on the test set, suggesting overfitting. This observation highlights the limitations of HSP in capturing generalisable patterns in this spatially induced simulation setting.

The GPNN-Linear-GP models, constructed with 12 and 52 regions, show comparable test performances. The GPNN-Linear-GP-12R model achieves a test RMSE of 5.07 (0.051) and test R^2 of 55.5% (0.893%), while the GPNN-Linear-GP-52R model yields similar metrics with a test RMSE of 5.07 (0.054) and test R^2 of 55.5% (0.944%).

Both models demonstrate robustness compared to the HSP model, with less overfitting and better generalisation. However, their performance does not surpass that of the GPR model.

The observed results suggest that the simulation study exhibits a spatial-information-induced setting where the input data lack complex non-linear relationships with the response variable. This characteristic likely favours the GPR model, which efficiently captures the spatially structured relationships through its Gaussian Process framework.

In contrast, the GPNN models may underperform due to the absence of substantial non-linear interactions in the data, limiting the advantages of their hierarchical design. Nevertheless, the GPNN models potentially retain spatial information better than the HSP model, which is a regularised linear approach.

Method	RMSE training, test	R^2 training, test
True GPR model	4.58	64.6%
HSP	2.97 (0.129), 5.34 (0.032)	84.6% (1.343%), 50.7% (0.582%)
GPR	4.52 (0.005), 4.80 (0.003)	64.4% (0.074%), 60.1% (0.047%)
GPNN-Linear-GP-12R	4.60 (0.093), 5.07 (0.051)	63.0% (1.523%), 55.5% (0.893%)
GPNN-Linear-GP-52R	4.47 (0.044), 5.07 (0.054)	65.1% (0.697%), 55.5% (0.944%)

Table 5.1: Simulation: Optimisation Performances Mean (SD)

5.5.2 Real-data study

5.5.2.1 Performance

We examine the effect of data standardisation on traditional linear models, namely Ridge and LASSO regressions, vs. spatial models, namely Gaussian Process Regression (GPR) and GPNN models. Table 5.2 shows that input standardisation has a positive impact on the predictive performance of Ridge and LASSO regressions, while for the spatial models, standardisation of input can have an adverse effect. Specifically, input standardisation remarkably improves the predictive performance of GPR-10, but conversely, it leads to lower training errors but higher test errors in GPR-40, which may indicate poor fitting for highly-parameterised models. This adverse effect is also observed in all of the GPNN models.

Method	RMSE (train, test)	R^2 (train, test)	Total variables (non-zero)	Run time
LASSO	3.34, 4.92	81.2%, 59.4%	628	7 m
Ridge	0.90, 4.59	98.6%, 64.6%	156,812	5 m
LASSO-S	2.46, 4.41	89.7%, 67.3%	859	7 m
Ridge-S	0.30, 4.34	99.8%, 68.4%	156,812	5 m
GPR-10	6.72, 6.79	23.7%, 22.6%	286	1 s
GPR-10-S	4.14, 4.99	71.0%, 58.2%	286	1 s
GPR-20	4.78, 5.63	61.4%, 46.8%	1,771	3 s
GPR-40	4.27, 5.47	69.2%, 49.9%	12,341	20 s
GPR-40-S	0.05, 9.45	100%, -50.7%	12,341	20 s
GPNN-GP-GP	2.92, 5.20	85.6%, 54.5%	21,548	1.5 h
GPNN-GP-Linear	3.39, 5.03	80.6%, 57.4%	21,277	1.3 h
GPNN-Linear-GP	0.00, 4.58	100%, 64.7%	157,096	29 h
GPNN-GP-GP-S	2.24, 5.70	91.4%, 35.3%	21,548	1.5 h
GPNN-GP-Linear-S	0.89, 6.20	98.7%, 45.2%	21,277	1.3 h
GPNN-Linear-GP-S	0.02, 4.61	100%, 64.2%	157,096	29 h

Table 5.2: Real-data: Optimisation Performances (median). The suffix "-S" denotes the cases where the input is standardised. The suffix "-10", "-20", "-40" after GPR models denotes the highest degree of Hermite polynomials used to estimate the eigen functions of the K-L basis.

In cases where the input is not standardised, GPR-40 demonstrates superior performance compared to GPR-10 due to the greater number of variables it incorporates, allowing for higher complexity to be captured by the regression model. However, in this study, both GPR models were outperformed by classical models, namely LASSO and Ridge regression. While Ridge regression may suffer from over-fitting, as indicated by its training RMSE almost equaling zero as compared to the data standard deviation of 5.90, it still achieves the lowest RMSE and highest R^2 on the test datasets among all linear models considered in this study.

The performance metrics of GPNN on real-data train and test datasets are consistent with other brain age literature (Cole et al., 2019). Specifically, the GPNN-GP-GP and GPNN-GP-Linear methods exhibit superior performance compared to GPR-40. While a gradual decrease in training errors is observed for both methods, predictive errors on the test dataset increase over time, with GPNN-GP-GP yielding both lower training and higher test RMSE than GPNN-GP-Linear throughout iterations. Despite this, the results for both GPNN methods are unsatisfactory, as they fail to outperform LASSO, which only employs 0.4% of the variables. On the contrary, the performance of

GPNN-Linear-GP on the test dataset is superior to that of Ridge regression, suggesting that incorporating non-linearity and spatial architecture could be advantageous in modeling such data.

5.5.2.2 Larger sample in real-data study

Method	RMSE (train, test)	R^2 (train, test)	Total variables (non-zero)	Run time
Ridge, 6k	1.91, 4.25	93.8%, 69.7%	156,812	20 m
GPR-10, 6k	5.12, 5.19	55.7%, 54.6%	286	1 s
GPR-40, 6k	4.07, 4.72	72.0%, 62.5%	12,341	10 m
GPNN-Linear-GP, 2k	0.00, 4.58	100%, 64.7%	157,096	26 h
GPNN-Linear-GP, 4k	0.58, 4.36	99.4%, 68.0%	157,096	9 h
GPNN-Linear-GP, 6k	0.12, 4.19	100%, 70.4%	157,096	10 h

Table 5.3: Real-data: Increasing number of training. The suffixes "2k", "4k", and "6k" denotes the doubling and tripling size of training samples respectively. The total number of iterations of 4k and 6k are limited to 600 in contrast to 2,000 iterations in the 2k case.

To address the issue of overfitting in GPNN-Linear-GP, additional samples can be conveniently included in the model training process. A further study was conducted on the model by doubling and tripling the number of training samples to 4,000 and 6,000 samples respectively. The results in Table 5.3 demonstrate that doubling and tripling the training samples, denoted by "GPNN-Linear-GP, 4k" and "GPNN-Linear-GP, 6k" respectively, leads to a monotonic improvement in the predictive performance of the algorithm. Furthermore, even when the training data size is tripled, the training RMSE still hits 0, indicating that there is still room for improving the model through an increase in the training sample size. Conversely, for linear models such as GPR and Ridge, tripling the training sample size has an essential impact on the predictive performance on the test data, yet they become more outperformed by GPNN-Linear-GP.

5.5.2.3 Bayesian sampling

Table 5.4 presents the Bayesian inference performance of the GPR-40, GPNN-MALA, and GPNN-SGLD methods on a reduced training dataset consisting of 200 samples

Method	RMSE (train, test)	R^2 (train, test)	$\text{ECR}_{95\%}$ (train, test)	Avg PPI Width (train, test)	Run time
GPR-40	0.23, 8.64	99.9%, -25.7%	100%, 98.5%	4.9, 43.9	5 m
GPNN-MALA	6.63, 7.51	14.4%, 5.0%	97.5%, 95.7%	27.7, 27.9	26 h
GPNN-SGLD	4.15, 6.43	66.4%, 30.42%	99.0%, 94.8%	22.5, 24.7	15 h
GPNN-SGLD, 2k	5.53, 5.97	48.3%, 40.0%	94.9%, 93.7%	22.1, 22.3	29 h

Table 5.4: Real-data: Posterior Sampling Performances (median) with 200 training subjects and 2,000 test subjects. The suffix "2k" denotes the case of training the model on 2,000 training samples. GPNN-MALA and GPNN-SGLD algorithms are applied upon a GPNN-Linear-GP model.

and a separate test dataset of 2,000 samples, while GPNN-SGLD, 2k method is trained on 2,000 samples as before.

For the GPR model, it is observed that the training RMSE is nearly 0, indicating an overfitting issue. This is further supported by the highest test RMSE and a negative test R^2, indicating poor predictive performance on the test data. The empirical coverage rate (ECR) for a 95% posterior predictive interval (PPI) shows overcoverage, suggesting that the model's uncertainty estimates may be too conservative. However, it is important to note that the substantial discrepancy in PPI width between the training and test datasets (4.9 vs. 43.9) raises concerns about the model's ability to provide reliable Bayesian inference.

In contrast, both the GPNN-MALA and GPNN-SGLD methods exhibit overcoverage in the training ECR for the 95% interval, with GPNN-SGLD achieving near-perfect coverage at 99%. As for the test data, both methods demonstrate good calibration, achieving an average coverage of approximately 95% based on the theoretical 95% interval. However, GPNN-SGLD outperforms GPNN-MALA in terms of accuracy metrics such as RMSE and R^2, indicating superior predictive performance. Additionally, GPNN-SGLD yields a narrower average PPI width, indicating higher confidence in its predictions compared to GPNN-MALA. Notably, GPNN-SGLD also offers significantly faster computational time, making it an attractive choice for practical applications.

Furthermore, upon increasing the number of training samples for GPNN-SGLD from 200 to 2,000, we observe a reduced discrepancy between the training and test outcomes across all metrics. This finding suggests that the GPNN-SGLD method exhibits robustness when provided with a sufficient amount of training data, while also demonstrating better scalability compared to traditional MCMC method such as

MALA.

5.5.2.4 Brain age gap diagnostics

The box plots in Figure 5.4 display the residuals of BAG against the predictions made by the evaluated models. Large, positive values of BAG for individual subjects suggest that their brain ages are considerably advanced relative to the training data. The medians of BAG around 0 suggest that residuals are unbiased towards certain age ranges and are independent of predictions. The BAG plots of GPNN-GP-GP and GPR-40 appear to slightly violate this desirable property. However, GPNN-GP-Linear, GPNN-Linear-GP, and Ridge show good adherence to this property, except for the extreme ends of the prediction ranges. It is worth noting that large BAG values do not necessarily imply a flawed prediction since brain age is an abstract construct, unlike chronological age.

5.5.2.5 Saliency

The absolute-value saliency maps of the GPNN algorithms trained on real data is depicted in Figure 6.4. All methods except LASSO detect importance in the middle region (thalamus) of the brain. GPNN-GP-GP and GPNN-GP-Linear focus on learning through contiguous voxels, similar to the GPR coefficients. The smoothness observed in the saliency maps suggests that the model's learning process is motivated by spatial considerations, and that it accounts for distinct regions, as evidenced by the disassociated colors. Unlike the GPR model, the region-wise smoothness of the GPNN models can facilitate further post-hoc analysis. However, the model representation for GPNN-Linear-GP and Ridge appears almost identical and noisy in native space. This suggests that they have similar learning and that the current setting of smoothness in GPNN-Linear-GP may not be strong enough, leaving room for further adjustment. Lastly, the model representation for LASSO indicates an aggressive variable selection approach, which may not be useful for practical interpretation.

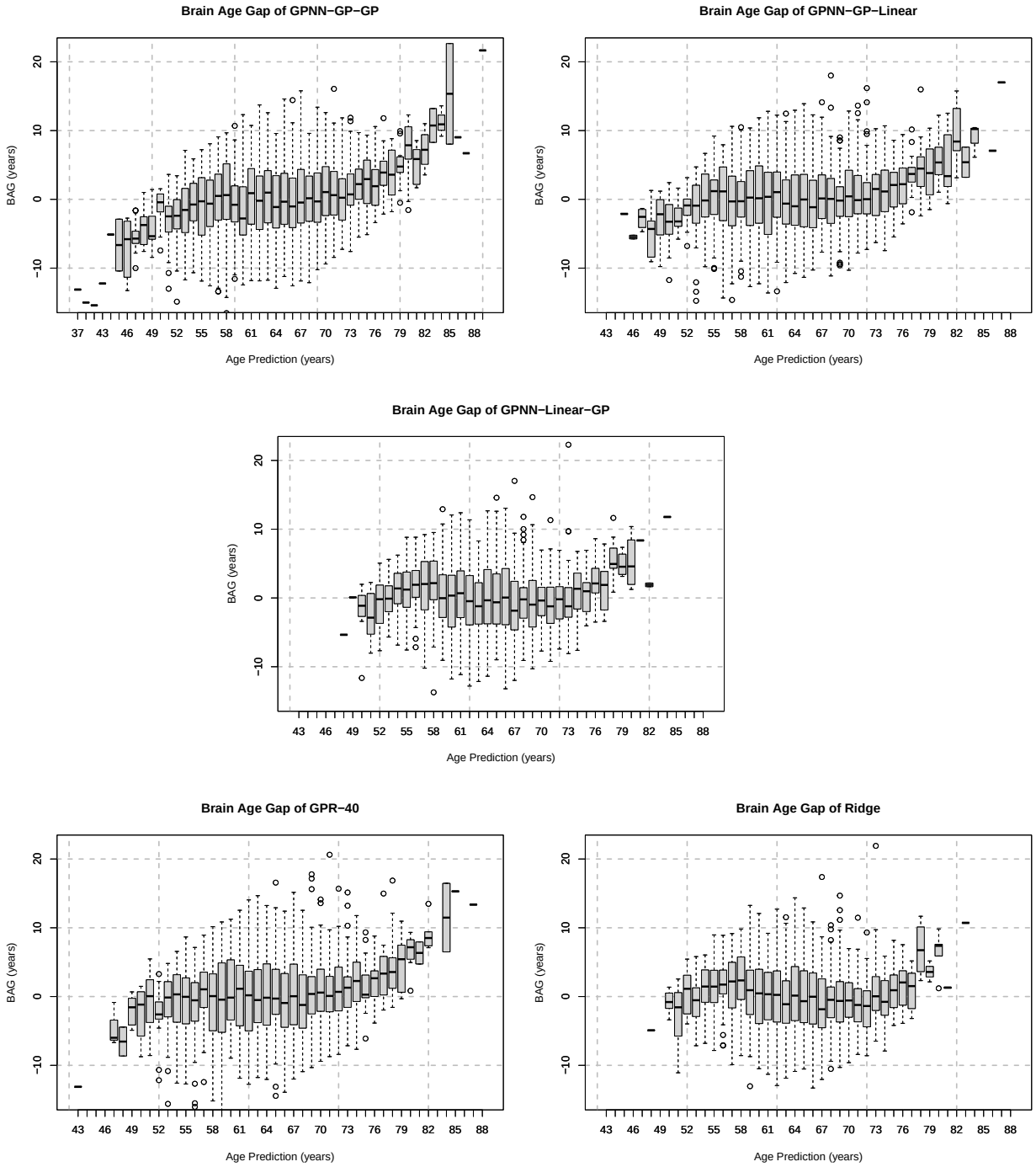
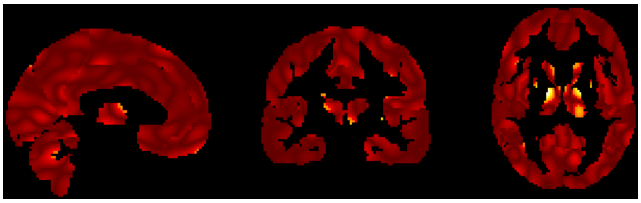
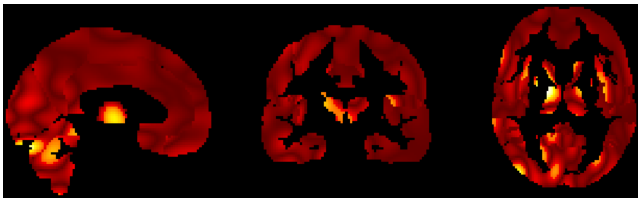


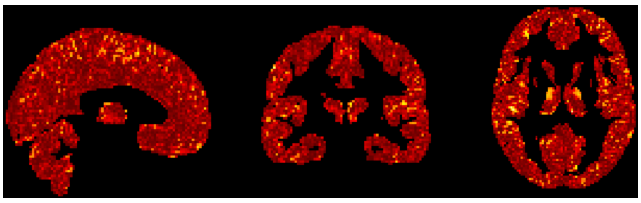
Figure 5.4: Box plots of Brain Age Gap (BAG) of GPNN-GP-GP, GPNN-GP-Linear, GPNN-Linear-GP, GPR-40, and Ridge Regression models, divided into bins of a single year of age. A non-zero average BAG indicates a potential bias of the predictive model for the particular age.



(a) GPNN-GP-GP Saliency



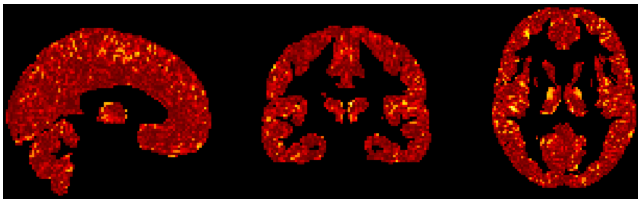
(b) GPNN-GP-Linear Saliency



(c) GPNN-Linear-GP Saliency



(d) GPR-40 coefficients



(e) Ridge coefficients



(f) LASSO coefficients

Figure 5.5: Absolute values of saliency (GPNN) and magnitude of fitted coefficients (regressions).

5.6 Discussion

In this work we develop a Bayesian spatial neural network model to provide flexible predictions while ensuring an interpretable model fit. Although the proposed method does not demonstrate clear superiority over the Gaussian Process Regression (GPR) model in the simulation study, where the data were generated using a GPR framework, the real-data analysis reveals that the GPNN models outperform traditional Gaussian Process models. Specifically, the GPNNs model achieved lower training and test RMSE values, indicating its superior accuracy in predicting the target variable. Additionally, the GPNN model showed higher R^2 values, indicating its better ability to explain the variance in the data. Particularly, based on these performance metrics, the proposed GPNN-Linear-GP model can be considered as the best-performing model among the models considered in this study.

The saliency analysis indicated that the GPNN-GP-GP and GPNN-GP-Linear models may be capable of capturing important spatial information and smoothness in the data, which is crucial for many applications, such as image or signal processing. The biologically smooth saliency scores obtained from the two GPNN models suggest that it can effectively identify and learn the spatial structures in the data. This ability to capture spatial information and smoothness is a key advantage of the GPNN model over traditional Gaussian process models, which may struggle to capture the high complexity of these features. However, they not illustrated its ability to model complex relationship within the data due to the nature of the non-complex problem considered in the study of brain age with VBM data. This is shown in (Smith et al., 2019; Franke et al., 2010; Liem et al., 2017a) which directly use Singular Value Decomposition with a linear model to effectively estimate brain age. Nevertheless, in a large-scale setting (with $n \gg 10,000$ and $p \geq 100,000$), methods such as linear regressions cannot scale up well. The GPNN method allows stochastic learning which alleviate the problem of having large n. Moreover, the GPNN-Linear-GP method, which has the best performance metrics, has non-smooth model representation which resembles that of Ridge regression. This may not be desirable in practical interpretation, and thus the extent to which the model can be smoothed while maintaining the high performance has to be investigated further.

In comparison to the traditional MALA model, the proposed GPNN-SGLD model showed significant improvements in terms of both computational time and accuracy. The SGLD model's ability to leverage the stochastic gradient descent method allows for faster Markov Chain evaluation and more efficient sampling of the posterior distribution, resulting in significant improvements in computational time. Additionally, the GPNN-SGLD model exhibited better accuracy in terms of predicting the target variable, as evidenced by its lower RMSE values. Therefore, the proposed GPNN-SGLD model can be considered as a superior alternative to the traditional MALA model.

The incorporation of hyperpriors through the use of an Inverse-Gamma distribution in the GPNN framework alleviates the need for extensive grid searches. The hyperpriors provide a probabilistic and adaptive approach to update the variances associated with the priors, which are crucial in modeling the uncertainties in the network weights. The conjugate property of the Inverse-Gamma distribution enables closed-form Bayesian updates, simplifying parameter estimation and facilitating effective learning and adaptation without compromising accuracy or computational efficiency.

Although the proposed models have shown promising results, there is still room for improvement. One potential avenue for improvement could be to incorporate more complex and diverse kernel functions in the GPNN model. Additionally, the incorporation of more advanced optimization techniques could further improve the accuracy and efficiency of the model. Furthermore, the development of hybrid models that combine the strengths of multiple approaches could potentially yield even better results. These are exciting areas for future research in the field of Gaussian process modeling.

We plan to further investigate imaging problems that are known to require topological information to aid predictions such as tumour- or cancer-related clinical outcomes to validate the spatial property of our method. Additionally, we aim to extend our methodology to incorporate multimodal imaging data and non-imaging data. Integrating multiple modalities should enhance our understanding of interaction effects and better capture the complex underlying mechanisms in neuroimaging studies.

5.7 Code Availability

The code supporting this research, including that used for exploratory data analysis, model developments, simulation study design and results, as well as the real-data study implementation and results, is openly accessible and can be found at:

https://github.com/kan-k/BSVCNN

Chapter 6

Extensions to Gaussian Process Neural Network

This chapter presents advancements to the GPNN framework introduced in Chapter 5, expanding its capabilities to integrate multiple data modalities simultaneously. These enhancements include the incorporation of non-imaging covariates, such as biological and socio-demographic factors, and the identification of latent subgroups among subjects to uncover sub-population brain-behaviour interactions within a unified framework. Furthermore, the framework is extended to infer interaction effects between different MRI modalities, such as structural and functional data, within a single cohesive model. This work was conducted under the supervision of Prof. Jian Kang, Prof. Timothy D. Johnson, and Prof. Thomas E. Nichols. Additional details regarding this chapter are provided in Appendix C.

6.1 Introduction

Scalar-on-imaging regression has gained considerable attention in the neuroimaging field for its capacity to link complex brain imaging data to scalar outcomes, such as cognitive scores, health metrics, or disease susceptibility. This approach allows for the creation of personalised predictions based on neural structure and function, providing critical insights into individual differences in neurobiological processes. By leveraging

high-dimensional brain imaging data, scalar-on-imaging models can capture subtle, subject-specific variations, which are invaluable for advancing personalised medicine. Our previous Gaussian Process Neural Network (GPNN) framework in Chapter 5 further extended this approach by incorporating uncertainty quantification, enhancing interpretability and robustness in scalar predictions derived from brain images.

Understanding the interaction between brain structure and socio-demographic factors, such as age, socio-economic deprivation, and education level, is critical for elucidating the determinants of cognitive function and health outcomes. Socio-demographic variables, often associated with systemic inequalities, are known to influence brain structure and function through mechanisms involving chronic stress, healthcare access, and lifestyle factors. For instance, socio-economic deprivation has been linked to accelerated brain aging and heightened risk of neurodegeneration due to cumulative exposure to adverse conditions (Klee et al., 2023; Cohen et al., 2024; Röhr et al., 2024). Lifestyle behaviours associated with deprivation, such as substance abuse or lower access to preventive care, further exacerbate these effects, with alcohol consumption linked to reduced grey matter volume and cognitive disorders (Spindler et al., 2021; Daviet et al., 2022), and smoking associated with cortical thinning and impaired cognition (Karama et al., 2015; Khlif et al., 2022). These factors also shape brain development during critical life stages, such as childhood and adolescence, and may modulate neuroplasticity in older adults, where educational attainment and social engagement can slow neurodegeneration (Cai et al., 2014; Marzola et al., 2023). Despite their importance, these covariates are often treated as static variables, limiting our ability to capture their complex and dynamic interactions with neural processes. A unified modelling framework that integrates socio-demographic variables with neuroimaging data is essential to advance our understanding of these multidimensional relationships.

Our study is motivated by the limitations of traditional approaches, which often treat non-imaging covariates as independent modifiers without adequately capturing their interactions with brain structure or behaviour. These methods frequently overlook the latent subgroup structures that may exist within populations, potentially masking medically relevant subgroups that could reveal nuanced relationships between brain and other factors. Addressing this gap, latent subclass analysis provides a powerful

framework to uncover hidden population heterogeneity, enabling the identification of subgroups with distinct neurobiological profiles that may underpin diverse cognitive and health outcomes.

The need for latent subclass analysis is further underscored by the prevalence of Simpson's paradox in neuroscience, where population-level associations can obscure or even reverse within subgroup-level patterns (Kievit et al., 2013). Ignoring this phenomenon can lead to misleading interpretations, potentially resulting in inaccurate or harmful conclusions. For example, previous studies applying latent class analysis (LCA) have successfully identified subgroups in neurobiological data that are meaningfully related to behavioural outcomes, such as ADHD symptoms (Lecei et al., 2019). Similarly, latent profile analysis (LPA) has been used to categorise subgroups based on brain activity patterns associated with risk-taking behaviours, revealing heterogeneity linked to demographic and substance use factors (Martz et al., 2021). These examples illustrate the potential of latent variable modelling to illuminate individual differences and offer more tailored insights into brain-behaviour relationships.

Building on these concepts, we adopt the principles similar to that of Latent Subgroup Identification in Image-On-Scalar Regression (LASIR) (Lin et al., 2024), which optimises latent class allocation alongside outcome prediction in a Bayesian framework. However, unlike LASIR, which operates in an image-on-scalar regression context, our model extends these ideas to scalar-on-image regression, requiring the integration of imaging and non-imaging inputs in a joint framework. This design not only accommodates heterogeneous data sources but also enables uncertainty quantification by assessing expected outcomes across latent classes, rather than relying on rigid categorical assignments. By leveraging the probabilistic nature of Bayesian sampling, our methodology avoids overfitting to specific class configurations and provides robust predictions while capturing complex subgroup effects.

To this end, we propose a novel framework, K-GPNN, which diverges from traditional methods by aggregating the complex effects of biological, socio-demographic, and behavioural factors into a unified model. Instead of separate analyses for subgroup detection and outcome prediction, our approach performs both tasks simultaneously, using Bayesian optimisation and latent class modelling. A unique advantage of K-GPNN lies in its spatially varying coefficients, allowing us to assess the effects of

non-imaging latent classes across distinct brain regions, capturing region-specific associations with socio-demographic factors. Furthermore, our model's focus on uncertainty quantification enables the assessment of prediction confidence for each subgroup, offering clinically relevant insights into the reliability of predictions across different demographic profiles.

Our study investigates cognitive performance in a pair-matching task. The primary focus of this study is to model cognitive ability using Voxel-Based Morphometry (VBM) data, along with socio-demographic covariates such as age, deprivation, and gender. Using data from the UK Biobank, a large-scale biomedical resource encompassing imaging and non-imaging data from over 50,000 participants, we aim to provide a robust foundation for examining brain-cognition relationships in relation to socio-demographic factors.

Additionally, we extend our GPNN model with a bimodality variant, termed "Bimodality-GPNN", which incorporates multiple imaging modalities. Using both VBM data and resting-state fMRI-derived Default Mode Network (DMN) data allows us to leverage complementary information from these imaging types. Multimodal imaging has been shown to improve prediction accuracy in tasks such as brain age estimation (Liem et al., 2017b; Niu et al., 2020; Jirsaraie et al., 2023) and cognitive function prediction (Young et al., 2013; Ajith et al., 2024), as each modality captures distinct aspects of brain structure and function. VBM reflects grey matter density, which is pertinent to structural integrity, while DMN connectivity from resting-state fMRI captures intrinsic functional networks involved in cognitive and self-referential processes. By integrating these modalities, our Bimodal-GPNN model captures modality-specific contributions and their interactions, providing a more comprehensive perspective on brain health. This extension also allows for brain-mapped saliency analysis, revealing how different imaging modalities interact during model training, thus offering valuable insights into multi-modal relationships within the brain.

In summary, this study introduces a novel framework for scalar-on-image regression that not only integrates multi-modal imaging and socio-demographic covariates but also incorporates uncertainty quantification and subgroup identification. Through the K-GPNN and its extensions, we offer a robust, interpretable approach for examining the combined effects of diverse factors on brain and cognitive outcomes. This framework

represents a substantial advancement in neuroimaging analysis, with implications for personalised medicine and a deeper understanding of how socio-demographic and biological factors collectively impact brain heal

In this chapter, we introduce a novel multi-class framework grounded in Gaussian Process Neural Networks (GPNN), incorporating essential methodological advancements to address complex relationships between imaging and non-imaging data. We establish the foundational components of this framework, including interaction effects, scenario-specific activation functions, and their seamless integration into the GPNN architecture. Building on this foundation, we propose a series of evolutionary advancements, such as the integration of single non-imaging covariates, followed by multi-latent-class variants that enable the discovery of latent subgroup structures and their interactions with imaging variables. This innovation allows us to explore intricate relationships influenced by biological and socio-demographic factors. Furthermore, we extend the framework to an imaging bimodality variant, which models the individual and interaction effects of distinct imaging modalities, offering a unified approach to understanding structural and functional neuroimaging data. Alongside detailing these theoretical advancements, we describe the datasets used, including the preprocessing and transformations applied. The efficacy of our methodologies is demonstrated through a comprehensive simulation study, which highlights their capacity to model latent effects and complex interactions. These methods are further validated on real-world cognitive task data, underscoring their practical utility in uncovering meaningful patterns in neuroimaging and behavioural datasets. The chapter concludes with a discussion of the broader implications of these findings, potential limitations, and future directions to extend the proposed models for tackling even more complex neuroimaging questions.

6.2 Background

6.2.1 Interaction Effects in Models

In the context of linear models, an *interaction effect* occurs when the effect of one predictor variable on the outcome depends on the value of another predictor. This is

captured by including a product term of the interacting variables in the model. For example, for two variables X_1 and X_2, the interaction is represented as $X_1 \times X_2$, and the linear model becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \epsilon,$$

where β_3 quantifies the interaction effect, describing how the effect of X_1 on Y changes with X_2, and vice versa. Interaction terms are essential for capturing relationships that go beyond additive effects, particularly in domains like neuroimaging, where interactions between brain regions may significantly impact clinical outcomes. However, the inclusion of these terms increases model complexity and necessitates careful consideration of model fit and interpretability.

In neural networks, interaction effects are captured implicitly through non-linear activation functions and hierarchical layers, which approximate complex dependencies between variables. For instance, in the GPNN architecture, standard non-linear activation functions are applied to the input layer, allowing the model to capture activation patterns across brain regions and their non-linear interactions. While this approach effectively captures high-dimensional dependencies, it often sacrifices interpretability, making it difficult to assess specific interaction effects. Moreover, relying solely on implicit modeling may result in suboptimal performance when interactions are critical to the outcome.

To address these limitations, our extended GPNN architecture incorporates explicit interaction terms in the final layer. By fitting a linear model on the final layer neurons with respect to the response, the design directly models specific interactions between key variables. This enhances the network's capacity to capture important feature dependencies while improving interpretability and reducing underfitting. These explicit terms are particularly advantageous in neuroimaging tasks, where understanding interaction effects can provide deeper insights into the relationships between imaging and non-imaging covariates.

By combining non-linear transformations in the hidden layers with explicit interaction terms in the final layer, this hybrid approach mitigates the limitations of standard neural networks. It offers improved predictive performance and interpretability, making

it an effective solution for complex neuroimaging analyses.

6.2.2 Activation Function

In multi-class classification tasks, our objective is to categorise a given sample $\mathbf{x} \in \mathbb{R}^p$, consisting of p covariates, into one of C distinct classes, where $C \in \mathbb{Z}^+$. This classification is particularly relevant in the context of latent class identification within our GPNN framework, where the model seeks to probabilistically assign data points to latent subgroups based on non-imaging covariates. To achieve this, we employ the **Softmax** function, a widely used activation function in machine learning for multi-class problems.

The Softmax function maps real-valued input vectors into a probability distribution over C classes, ensuring that the outputs are non-negative and sum to one. Let $\mathbf{W} \in \mathbb{R}^{C \times p}$ represent the weight matrix, and $b \in \mathbb{R}^C$ the bias vector. The Softmax function for each class $j = 1, \ldots, C$ is defined as:

$$SM_x(j) = \frac{\exp(\mathbf{W}_j^\top \mathbf{x} + b_j)}{\sum_{c=1}^{C} \exp(\mathbf{W}_c^\top \mathbf{x} + b_c)}. \tag{6.1}$$

This formulation assigns a probability $SM_x(j)$ to class j, where the numerator captures the contribution of the j-th class, and the denominator normalises over all classes. Such probabilistic assignments are essential in our model for tasks like subgroup detection, as they allow the representation of uncertainty in latent class membership.

For optimisation and backpropagation, the gradient of the Softmax function with respect to the weight matrix $\mathbf{W}_j$ is required. The derivative of the full Softmax function is given by:

$$\begin{aligned} \frac{\partial SM_x(\cdot)}{\partial \mathbf{W}_j} &= \sum_{c=1}^{C} \frac{\partial SM_x(c)}{\partial \mathbf{W}_j} \\ &= \mathbf{x} \sum_{c=1}^{C} SM_x(c)(\delta_{jc} - SM_x(j)), \end{aligned} \tag{6.2}$$

where δ_{jc} is the Kronecker delta, which equals 1 if $j = c$ and 0 otherwise.

The Softmax function is integral to the GPNN framework, enabling probabilistic interpretation and robust assignment of latent class probabilities. This supports downstream analyses, such as identifying class-specific imaging effects and assessing group-level uncertainty. Its probabilistic nature aligns with our Bayesian methodology, which emphasises posterior distributions over point estimates, unlike conventional MLE approaches. By integrating uncertainty, the Softmax function enhances the model's capacity to capture latent structures and relationships between imaging and non-imaging features.

6.2.3 Gaussian Process Neural Network (GPNN)

The Gaussian Process Neural Network (GPNN), described in Chapter 5 is designed to integrate raw imaging data with spatially-informed Gaussian Process transformations and non-linear neural network connections. This innovative structure allows the GPNN to capture both individual voxel-level details and broader regional patterns, making it highly suited for spatial data like neuroimaging. The general prediction function in GPNN is given by:

$$g\big[\mathbb{E}(y_i \mid \{\mathbf{X}_i(\mathbf{v}')\}_{\mathbf{v}'\in\mathcal{B}})\big] = \alpha_0 + \sum_{\mathbf{v}\in\mathcal{A}} \alpha(\mathbf{v}) h\Big\{\eta_0(\mathbf{v}) + \sum_{\mathbf{v}'\in\mathcal{B}} \eta(\mathbf{v},\mathbf{v}')\mathbf{X}_i(\mathbf{v}')\Big\} \tag{6.3}$$

Here, $g[\cdot]$ is the link function, $h(\cdot)$ is an activation function, and $\mathbf{X}_i(\mathbf{v}')$ represents the input voxel data, spatially transformed through GP functions. This framework provides a flexible, layered approach to mapping spatial brain data to scalar predictions.

The GPNN framework can be divided into three main architectures, each with distinct characteristics, as outlined in Table 6.1:

- **GPNN-Linear-GP** – This is the most complex model, where raw imaging data is connected to a hidden layer and then transformed by a whole-brain GP before producing a scalar output. It is designed to capture global brain patterns as well as fine-grained voxel effects. However, its computational demand is high, and it tends to overfit due to the extensive model smoothing.

- **GPNN-GP-Linear** – This is the simplest model in the framework. It performs a GP transformation on a regional basis before connecting to a hidden layer, followed by direct output. This architecture is computationally efficient, offering fast processing and capturing localised effects, though its expressiveness is limited compared to more complex GPNN models.

- **GPNN-GP-GP** – This architecture balances regional and global representations by applying regional GP transformations followed by a whole-brain GP transformation before outputting a scalar value. It captures both smooth, local effects and larger patterns, making it effective for nuanced spatial modelling, but it can become overly convoluted.

Architecture	**Description**	**Pros**	**Cons**	**Run time**
GPNN-Linear-GP	Connects raw imaging data to a hidden layer, followed by a whole-brain GP transformation, and links to a scalar output layer.	Captures global brain patterns along with voxel-specific effects	High computational demand, may overfit with less smooth representations	Long
GPNN-GP-Linear	Applies regional GP transformations first, connecting to a hidden layer, and directly to output.	Fast and efficient; captures localised, parsimonious effects	Limited expressiveness	Short
GPNN-GP-GP	Uses region-based GP transformations followed by a whole-brain GP transformation before connecting to output.	Models both regional and global patterns with smooth, localised effects	Can be overly convoluted	Short

Table 6.1: Comparison of each GPNN architecture: GPNN-Linear-GP, GPNN-GP-Linear, and GPNN-GP-GP

Each of these GPNN architectures offers distinct benefits for different neuroimaging

applications, with trade-offs between computational efficiency and the model's ability to capture complex spatial relationships.

6.3 Methodology

6.3.1 Combining Imaging and Non-Imaging Covariates at the Final Layer

Introducing interaction effects into neural networks, particularly at the final layer, is unconventional but necessary in certain applications to capture complex relationships between imaging data and non-imaging covariates. In many neuroimaging problems, non-imaging covariates, such as demographic or biological factors, play a crucial role in modulating the relationship between brain structure or function and the outcome.

We introduce a set of non-imaging covariates $\mathbf{Z}_i = \{Z_{i,1}, \ldots, Z_{i,p}\}$ at the final layer of the model, assuming interaction effects between each non-imaging covariate and the latent representations of the imaging network. The outcome prediction model is defined as:

$$\hat{y}(X_i) = \alpha_0 + \sum_{\mathbf{v}\in\mathcal{A}} \alpha(\mathbf{v}) D_{i,\mathbf{v}} + \sum_{m=1}^{p} \beta_m Z_{i,m} + \sum_{\mathbf{v}\in\mathcal{A}} \sum_{m=1}^{p} \phi_{\mathbf{v},m} Z_{i,m} D_{i,\mathbf{v}}, \tag{6.4}$$

where $D_{i,\mathbf{v}} := h\Big(\eta_0(\mathbf{v}) + \sum_{\mathbf{v}'\in\mathcal{B}} X_i(\mathbf{v}')\theta(\mathbf{v},\mathbf{v}')\Big)$ represents the post-activation of the hidden layer from the imaging data. The term $\phi_{\mathbf{v},m}$ captures the interaction effect between the non-imaging variable $Z_{i,m}$ and the latent imaging variable $D_{i,\mathbf{v}}$. This interaction allows the model to capture how specific non-imaging factors, such as demographic variables, modulate the influence of brain imaging patterns on the outcome.

6.3.1.1 Example: Incorporating Sex as a Non-Imaging Covariate

As an illustrative example, suppose our non-imaging covariate set consists of a single biological factor, *sex*. In this case, Z_i is defined as:

$$Z_i = \begin{cases} -1, & \text{Subject } i \text{ is Female} \\ 1, & \text{Subject } i \text{ is Male} \end{cases}$$

We incorporate this covariate into the *GPNN-Linear-GP* architecture. A vanilla GPNN-Linear-GP follows the general characterisation of Equation 6.3, with additional conditions on the spatially-varying covariates, i.e. setting $\alpha(\mathbf{v}) \approx \sum_{k=1}^{K} \rho_k(\mathbf{v})\psi_k$, and $\eta(\mathbf{v}, \mathbf{v}') = \theta(\mathbf{v}, \mathbf{v}')$, such that $\psi_k \sim N(0, \sigma^2\omega^2)$, $\alpha_0 \sim N(0, 1)$, $\theta(\mathbf{v}, \mathbf{v}') \sim N(0, \sigma^2\tau^2_{\mathbf{v}})$ and $\eta_0(\mathbf{v}) \sim N(0, 1)$. This corresponds to having a direct connection in the input layer followed by a whole-brain GP transformation, along with linearly incorporating the non-imaging covariate with interactions into the hidden layer of the GPNN-Linear-GP. The depiction of this architecture is given in the Appendix Figure C.1a. This gives us the following problem:

$$\begin{aligned} \arg\min_{\boldsymbol{\Theta}} \; & \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big[y_i - \Big\{\alpha_0 + \sum_{k=1}^{K}\psi_k R_{i,k} + \beta_0 Z_i + \sum_{k=1}^{K}\phi_k Z_i R_{i,k}\Big\}\Big]^2 \\ & + \frac{K}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\frac{1}{\omega^2}\sum_{k=1}^{K}\psi_k^2 + \frac{1}{2}\alpha_0^2 \\ & + \frac{|\mathcal{A}||\mathcal{B}|}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\mathbf{v}\in\mathcal{A}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\mathbf{v}'\in\mathcal{B}}\theta(\mathbf{v}, \mathbf{v}')^2 + \frac{1}{2}\sum_{\mathbf{v}\in\mathcal{A}}\eta_0(\mathbf{v})^2, \end{aligned} \tag{6.5}$$

where $R_{i,k} := \sum_{\mathbf{v}\in\mathcal{A}}\rho_k(\mathbf{v})h\Big(\eta_0(\mathbf{v}) + \sum_{\mathbf{v}'\in\mathcal{B}}X_i(\mathbf{v}')\theta(\mathbf{v}, \mathbf{v}')\Big)$ is the latent representation of the hidden layer, and ϕ_k is the interaction effect between sex and the latent imaging variable $R_{i,k}$.

The optimisation problem for the model parameters $\theta(\mathbf{v}, \mathbf{v}')$ can be formulated using maximum a posteriori (MAP) estimation. The unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\theta}, \boldsymbol{\eta}_0)$ with respect to $\theta(\mathbf{v}, \mathbf{v}')$ is given by:

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i,y_i)}{\partial\theta(\mathbf{v},\mathbf{v}')} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i-\hat{y}_i)\sum_{k=1}^{K}\rho_k(\mathbf{v})(\psi_k+\phi_k Z_i)h'(G_i(\mathbf{v}))X_i(\mathbf{v}')\big\}\Big] + \frac{1}{\sigma^2\tau^2_{\mathbf{v}}}\theta(\mathbf{v},\mathbf{v}'), \tag{6.6}$$

where $G_i(\mathbf{v}) := \eta_0(\mathbf{v}) + \sum_{\mathbf{v}'\in\mathcal{B}} X_i(\mathbf{v}')\theta(\mathbf{v},\mathbf{v}')$, and m is the minibatch size. Similarly, the gradient with respect to $\eta_0(\mathbf{v})$ is:

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i,y_i)}{\partial\eta_0(\mathbf{v})} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i-\hat{y}_i)\sum_{k=1}^{K}\rho_k(\mathbf{v})(\psi_k+\phi_k Z_i)h'(G_i(\mathbf{v}))\big\}\Big] + \eta_0(\mathbf{v}). \tag{6.7}$$

Similarly, the unbiased estimator with respect to σ^2 is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i,y_i)}{\partial\sigma^2} = n\frac{1}{m}\sum_{i=1}^{m}\Big\{\frac{1}{2\sigma^2}-\frac{(y_i-\hat{y}_i)^2}{2\sigma^4}\Big\} - \frac{1}{2\sigma^4}\sum_{\mathbf{v}\in\mathcal{A}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\mathbf{v}'\in\mathcal{B}}\theta(\mathbf{v},\mathbf{v}')^2 + \frac{|\mathcal{A}||\mathcal{B}|}{2\sigma^2} \tag{6.8}$$

The coefficients in the hidden layer, including the non-imaging covariate coefficient β_0 and the interaction terms ϕ_k, are computed via MAP estimation, providing interpretable estimates of how non-imaging covariates interact with brain imaging data. This formulation enables a direct and intuitive interpretation of interaction effects between non-imaging covariates and imaging features. Additionally, saliency maps can be used to visualise these effects, with the saliency at each voxel given by:

$$\frac{\partial\hat{y}(X_i)}{\partial X_i(\mathbf{v}')} = \sum_{k=1}^{K}(\psi_k+\phi_k Z_i)\sum_{\mathbf{v}\in\mathcal{A}}\Big[\rho_k(\mathbf{v})h'\Big(\eta_0(\mathbf{v})+\sum_{\mathbf{v}''\in\mathcal{B}}X_i(\mathbf{v}'')\theta_{\mathbf{v}\mathbf{v}''}\Big)\theta(\mathbf{v},\mathbf{v}')\Big]. \tag{6.9}$$

The saliency function demonstrates how the model's predictions are influenced by a linear combination of both the main imaging effect ψ_k and the interaction effect

$\phi_k Z_i$, allowing us to assess how brain regions interact with non-imaging factors like sex.

6.3.2 Multi-class K-GPNN

In healthcare applications, models often need to handle a high-dimensional set of non-imaging covariates, such as demographic or clinical data derived from electronic health records. When these covariates are incorporated directly, as in the specification of Equation (6.4), the model can become overly complex, introducing potential mathematical and computational challenges, including overfitting and interpretability issues. To address this, we introduce the concept of latent group identification as a way to manage this complexity while enhancing the interpretive power of the model.

This approach allows the model to group subjects into categories that simplify the data structure without sacrificing important information. A practical benefit of this latent subgroup approach is that it maintains the interpretability of the model, even with a high number of covariates. Additionally, it facilitates the discovery of new, potentially medically meaningful subgroups in an unsupervised manner, which can lead to more precise and clinically relevant predictions.

The introduction of latent subgroups for non-imaging covariates offers several advantages:

- **Simplified Model Complexity**: By clustering covariates into subgroups, we reduce the dimensionality and complexity of the model, allowing it to generalise more effectively while preserving essential information.
- **Improved Interpretability**: The use of latent subgroups provides a clearer understanding of how non-imaging covariates influence predictions, particularly valuable in clinical contexts where interpretability is crucial.
- **Potential Discovery of New Subgroups**: In an unsupervised setting, this approach enables the identification of previously unrecognised subgroups, which may reveal clinically important patterns and improve prediction accuracy.

In summary, latent subgroups enhance the model's ability to handle high-dimensional non-imaging data by simplifying structure and uncovering potentially valuable covariate interactions. This approach also enables a more nuanced understanding of how demographic and clinical factors interact with brain imaging data, ultimately supporting more precise and interpretable modelling in neuroimaging analyses.

Building on this framework, we extend our architecture to incorporate multiple latent groups using a softmax function. This generalisation enables probablistic—and in the Bayesian sampling sense, distributional—inference for each subject. By moving beyond binary latent groups, the multi-class K-GPNN can identify more granular subpopulations, providing a richer representation of the variability in non-imaging covariates and their impact on neuroimaging outcomes.

It is worth noting that a binary-class version of the K-GPNN, where $K = 2$, was also derived as part of our development process. This binary model utilises the sigmoid function, which can be considered a special case of the softmax function limited to two classes. For brevity, we omit the full derivation of the binary-class model, as it follows naturally from the multi-class framework presented here. The multi-class K-GPNN, however, expands the flexibility of the model, supporting more complex subgroup structures and capturing additional variability in the non-imaging covariates, which is valuable for uncovering deeper insights in clinical data.

6.3.2.1 *K*-GPNN-Linear-GP

The K-GPNN-Linear-GP architecture integrates imaging and non-imaging covariates through a two-stream approach that captures complex cross-modal interactions. The architecture employs two sub-networks: one dedicated to processing imaging data and the other to non-imaging covariates, such as socio-demographic information. Each sub-network processes its respective input independently through its own hidden layer, allowing the model to capture distinct features specific to each data type. After the hidden layers, the outputs from both sub-networks are combined and jointly subjected to a GP transformation, mapping the features into a shared latent space. The depiction of this architecture is given in the Appendix Figure C.1b.

A distinctive aspect of the K-GPNN-Linear-GP architecture is its ability to transform latent classes derived from non-imaging covariates alongside imaging features

within the same GP-transformed space. By applying the GP transformation to both types of features, the model aligns these diverse data modalities within a smooth, probabilistically interpretable latent space, where both local and global patterns can be more effectively identified. This approach is particularly beneficial in capturing cross-modal relationships, as it enables non-imaging covariates to interact meaningfully with imaging-derived features within a unified framework. Such interactions provide a richer context for understanding how structural brain features and socio-demographic factors jointly influence outcomes in a healthcare setting.

To achieve this, we introduce a learnable parameter $\phi_{j,k}$, a latent coefficient that transforms the output of the softmax function $SM(\cdot)$ for class j into a latent representation compatible with the final layer Gaussian coefficient ψ_k. We also specify the following priors: $\alpha(\mathbf{v}) \approx \sum_{k=1}^{K} \psi_k \rho_k(\mathbf{v})$, and $\eta(\mathbf{v}, \mathbf{v}') = \theta(\mathbf{v}, \mathbf{v}')$, such that $\psi_k \sim N(0, \sigma^2\omega^2)$, $\alpha_0 \sim N(0,1)$, $\theta(\mathbf{v}, \mathbf{v}') \sim N(0, \sigma^2\tau^2(\mathbf{v}))$ and $\eta_0(\mathbf{v}) \sim N(0,1)$.

The prediction function $\hat{y}(X_i, Z_i)$ is defined by

$$\hat{y}(X_i, Z_i) = \alpha_0 + \sum_{k=1}^{K} \Big\{ \psi_k \Big[\sum_{\mathbf{v} \in \mathcal{A}} \rho_k(\mathbf{v}) h\Big(G_i(\mathbf{v})\Big) + \sum_{j=1}^{C} \phi_{j,k} SM\Big(F_i(j)\Big) \Big] \Big\} \tag{6.10}$$

where

$$G_i(\mathbf{v}) = \eta_0(\mathbf{v}) + \sum_{\mathbf{v}' \in \mathcal{B}} X_i(\mathbf{v}') \theta(\mathbf{v}, \mathbf{v}') \tag{6.11}$$

is the pre-activated hidden neuron of *imaging* variable $\mathbf{v}$, and

$$F_i(j) = b_j + \sum_{p=1}^{P} w_{p,j} Z_{i,p} \tag{6.12}$$

is the pre-transformed hidden neuron of *non-imaging* latent class j.

The optimisation problem for the K-GPNN-Linear-GP architecture is then formulated as follows:

$$\begin{aligned}\arg\min_{\boldsymbol{\Theta}} \ & \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(y_i - \hat{y}(X_i, Z_i)\Big)^2 \\ & + \frac{K}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\frac{1}{\omega^2}\sum_{k=1}^{K}\psi_k^2 + \frac{1}{2}\alpha_0^2 \\ & + \frac{|\mathcal{A}||\mathcal{B}|}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\mathbf{v}\in\mathcal{A}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\mathbf{v}'\in\mathcal{B}}\theta(\mathbf{v},\mathbf{v}')^2 + \frac{1}{2}\sum_{\mathbf{v}\in\mathcal{A}}\eta_0(\mathbf{v})^2.\end{aligned} \tag{6.13}$$

Suppose that our stochastic gradient descent runs with $m < n$ number of samples. For the first-layer imaging covariates, the unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)$ with respect to $\theta(\mathbf{v}, \mathbf{v}')$ is

$$\begin{aligned}\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial\theta(\mathbf{v},\mathbf{v}')} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k\rho_k(\mathbf{v})h'(G_i(\mathbf{v}))X_i(\mathbf{v}')\big\}\Big] \\ + \frac{1}{\sigma^2\tau_{\mathbf{v}}^2}\theta(\mathbf{v},\mathbf{v}'),\end{aligned} \tag{6.14}$$

Similarly, the gradient with respect to $\eta_0(\mathbf{v})$ is:

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial\eta_0(\mathbf{v})} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k\rho_k(\mathbf{v})h'(G_i(\mathbf{v}))\big\}\Big] + \eta_0(\mathbf{v}). \tag{6.15}$$

For the non-imaging covariate subnetwork, the unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)$ with respect to $\phi_{j,k}$ is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)}{\partial\phi_{j,k}} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k SM(F_i(j))\big\}\Big]. \tag{6.16}$$

The unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)$ with respect to $w_{p,j}$ is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)}{\partial w_{p,j}} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k\phi_{j,k}SM'(F_i(j))Z_{i,j}\big\}\Big]. \tag{6.17}$$

The unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)$ with respect to

b_j is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)}{\partial b_j} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k\phi_{j,k}SM'(F_i(j))\big\}\Big]. \tag{6.18}$$

Finally, the unbiased estimator with respect to σ^2 follows Eq. (6.8).

The saliency at each voxel given by:

$$\frac{\partial\hat{y}(X_i)}{\partial X_i(\mathbf{v}')} = \sum_{k=1}^{K}\psi_k\sum_{\mathbf{v}\in\mathcal{A}}\Big[\rho_k(\mathbf{v})h'\Big(\eta_0(\mathbf{v}) + \sum_{\mathbf{v}''\in\mathcal{B}}X_i(\mathbf{v}'')\theta(\mathbf{v},\mathbf{v}''),\Big)\theta(\mathbf{v},\mathbf{v}')\Big]. \tag{6.19}$$

While interactions are not explicitly defined for this model, the saliency is still influenced by the parameter ψ_k which accounts for both the latent imaging and the latent classes.

6.3.2.2 *K*-GPNN-GP-Linear

The K-GPNN-GP-Linear architecture extends the latent-class GPNN framework by integrating GP transformations at the input level of the imaging data. Multiple GP transformations are applied to the raw imaging data, projecting these onto the hidden layer to capture intricate voxel-level spatial patterns and provide a spatially-aware representation. Simultaneously, non-imaging covariates are processed in a separate subnetwork to generate latent-class representations through a softmax function.

The model combines these latent-class outputs with the GP-transformed imaging features in the hidden layer, capturing both high-dimensional anatomical information and complex socio-demographic patterns. Additionally, linear interaction terms, as outlined in Section 6.2.1, between latent classes and imaging features are incorporated, enabling the model to represent conditional effects where specific brain regions interact with subgroup characteristics. This integration ensures the K-GPNN-GP-Linear model is capable of capturing nuanced relationships between imaging and non-imaging data, offering a robust framework for spatially and demographically-informed neuroimaging analysis. The depiction of this architecture is given in the Appendix Figure C.1c.

To implement this architecture, we recall the model assumptions, which involve a linear connection in the hidden layer and GPs applied in the input layer. Specifically,

we set $\alpha(\mathbf{v}) = \psi_{\mathbf{v}}$ and approximate $\eta(\mathbf{v}, \mathbf{v}') \approx \sum_{\ell=1}^{L_{\mathbf{v}}} \gamma_\ell(\mathbf{v}', \mathbf{v})\theta_\ell(\mathbf{v})$, where $\theta_\ell(\mathbf{v}) \sim N(0, \sigma^2\tau_{\mathbf{v}}^2)$ and $\eta_0(\mathbf{v}) \sim N(0, 1)$. With these assumptions, the resulting prediction function is:

$$\begin{aligned}\hat{y}(X_i, Z_i) = \alpha_0 + \sum_{\mathbf{v}\in\mathcal{A}} \psi_{\mathbf{v}} h\Big(G_i(\mathbf{v})\Big) + \sum_{j=1}^{C} \xi_j SM\Big(F_i(j)\Big) \\ + \sum_{\mathbf{v}\in\mathcal{A}} \sum_{j=1}^{C} \phi_{j,\mathbf{v}} h\Big(G_i(\mathbf{v})\Big) SM\Big(F_i(j)\Big)\end{aligned} \tag{6.20}$$

where

$$G_i(\mathbf{v}) = \eta_0(\mathbf{v}) + \sum_{\ell=1}^{L_{\mathbf{v}}} \theta_\ell(\mathbf{v}) \sum_{\mathbf{v}'\in\mathcal{B}} X_i(\mathbf{v}')\gamma_\ell(\mathbf{v}', \mathbf{v}) \tag{6.21}$$

is the pre-activated hidden neuron of *GP-transformed imaging* variable $\mathbf{v}$, and $F_i(j)$ is as characterised in Eq. (6.12).

The optimisation problem is formulated as follows:

$$\begin{aligned}\arg\min_{\boldsymbol{\Theta}} \ & \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big[y_i - \Big\{\alpha_0 + \sum_{\mathbf{v}\in\mathcal{A}} \psi_{\mathbf{v}} h\Big(G_i(\mathbf{v})\Big) \\ & + \sum_{j=1}^{C} \xi_j SM\Big(F_i(j)\Big) \\ & + \sum_{\mathbf{v}\in\mathcal{A}} \sum_{j=1}^{C} \phi_{j,\mathbf{v}} h\Big(G_i(\mathbf{v})\Big) SM\Big(F_i(j)\Big)\Big\}\Big]^2 \\ & + \sum_{\mathbf{v}\in\mathcal{A}} \frac{L_{\mathbf{v}}}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\mathbf{v}\in\mathcal{A}} \frac{1}{\tau^2(\mathbf{v})} \sum_{\ell=1}^{L_{\mathbf{v}}} \theta_\ell(\mathbf{v})^2 + \frac{1}{2}\sum_{\mathbf{v}\in\mathcal{A}} \eta_0(\mathbf{v})^2.\end{aligned} \tag{6.22}$$

This formulation enables a direct and intuitive interpretation of interaction effects between non-imaging covariates and imaging features. Additionally, saliency maps can be used to visualise these effects, with the saliency at each voxel given by:

$$\begin{aligned}\frac{\partial\hat{y}(X_i)}{\partial X_i(\mathbf{v}')} = \sum_{\mathbf{v}\in\mathcal{A}} \Big[\psi_{\mathbf{v}} + \sum_{\mathbf{v}\in\mathcal{A}} \sum_{j=1}^{C} \phi_{j,\mathbf{v}} SM\Big(F_i(j)\Big)\Big] \times \\ \Big[h'\Big(\eta_0(\mathbf{v}) + \sum_{\ell=1}^{L_{\mathbf{v}}} \theta_\ell(\mathbf{v}) \sum_{\mathbf{v}''\in\mathcal{B}} X_i(\mathbf{v}'')\gamma_{\ell\mathbf{v}''\mathbf{v}}\Big) \sum_{\ell=1}^{L_{\mathbf{v}}} \theta_\ell(\mathbf{v})\gamma_\ell(\mathbf{v}', \mathbf{v})\Big].\end{aligned} \tag{6.23}$$

The saliency function demonstrates how the model's predictions are influenced by a linear combination of both the main imaging effect $\psi_{\mathbf{v}}$ and the interaction effects $\phi_{j,\mathbf{v}} SM\Big(F_i(j)\Big)$, allowing us to assess how brain regions interact with each latent class of non-imaging factors.

The partial derivatives of parameters specific to this architecture are derived similarly to those of the K-GPNN-Linear-GP. Complete derivations are provided in Appendix C.1.

6.3.2.3 *K*-GPNN-GP-GP

The final variation in our latent-class GPNN framework is the K-GPNN-GP-GP architecture, which incorporates GP transformations at both the imaging input layer and the hidden layer. In this model, GP transformations are applied directly to the raw imaging data at the input level, capturing spatial dependencies and local anatomical patterns early in the network. These transformed inputs are then processed through the hidden layer, where an additional GP transformation integrates both imaging and non-imaging covariates. The depiction of this architecture is given in the Appendix Figure C.1d.

This dual-GP configuration enables the model to capture hierarchical spatial relationships, with local patterns learned at the input layer and broader, class-specific interactions integrated at the hidden layer. By allowing GP transformations in both layers, the model can detect complex, multi-scale dependencies across modalities, making it especially suited for datasets with rich spatial structure and diverse covariate information. This setup enhances the model's ability to reveal subtle connections between brain structure and socio-demographic factors, providing a powerful framework for comprehensive analysis in neuroimaging and healthcare applications.

We also specify the following priors: $\alpha(\mathbf{v}) \approx \sum_{k=1}^{K} \psi_k \rho_k(\mathbf{v})$, and $\eta(\mathbf{v}, \mathbf{v}') \approx \sum_{\ell=1}^{L_{\mathbf{v}}} \gamma_\ell(\mathbf{v}', \mathbf{v}) \theta_\ell(\mathbf{v})$, such that $\psi_k \sim N(0, \sigma^2\omega^2)$, $\alpha_0 \sim N(0,1)$, $\theta_\ell(\mathbf{v}) \sim N(0, \sigma^2\tau_{\mathbf{v}}^2)$ and $\eta_0(\mathbf{v}) \sim N(0,1)$.

The prediction function is given by:

$$\hat{y}(X_i, Z_i) = \alpha_0 + \sum_{k=1}^{K} \Big\{ \psi_k \Big[\sum_{\mathbf{v} \in \mathcal{A}} \rho_k(\mathbf{v}) h\Big(G_i(\mathbf{v})\Big) + \sum_{j=1}^{C} \phi_{j,k} SM\Big(F_i(j)\Big) \Big] \Big\} \tag{6.24}$$

where $G_i(\mathbf{v})$ is given by Eq. (6.21) and $F_i(j)$ is given by Eq. (6.12).

The optimisation problem for the K-GPNN-GP-GP architecture is then formulated as follows:

$$\begin{aligned}\arg\min_{\boldsymbol{\Theta}} \; & \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(y_i - \hat{y}(X_i, Z_i)\Big)^2 \\ & + \frac{K}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\frac{1}{\omega^2}\sum_{k=1}^{K}\psi_k^2 + \frac{1}{2}\alpha_0^2 \\ & + \sum_{\mathbf{v}\in\mathcal{A}}\frac{L_{\mathbf{v}}}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\mathbf{v}\in\mathcal{A}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\ell=1}^{L_{\mathbf{v}}}\theta_\ell(\mathbf{v})^2 + \frac{1}{2}\sum_{\mathbf{v}\in\mathcal{A}}\eta_0(\mathbf{v})^2.\end{aligned} \tag{6.25}$$

The complete partial derivatives of parameters specific to this architecture are provided in Appendix C.2.

6.3.3 Imaging Bimodality GPNN

In recent years, neuroimaging research has increasingly embraced multimodal approaches to improve the robustness and precision of predictive models. The integration of multiple imaging modalities, such as structural MRI, functional MRI (fMRI), and diffusion MRI, allows for a more comprehensive understanding of the brain, as each modality captures distinct yet complementary information about brain anatomy and function.

For example, T1-weighted structural MRI provides high-resolution images of brain anatomy, offering insights into grey matter volume, cortical thickness, and potential neurodegenerative changes. VBM images highlights differences in brain structure that may be associated with various neurological conditions. In contrast, functional MRI captures dynamic brain activity, either at rest (resting-state fMRI) or during task performance (task fMRI), revealing patterns of neural activation and connectivity. Diffusion MRI, on the other hand, maps white matter tracts, providing crucial information about the brain's structural connectivity by tracking the diffusion of water molecules along fiber pathways. These imaging modalities, when combined, offer a more holistic view of brain function and structure.

The importance of multimodality lies in the fact that many neuroimaging outcomes

of interest, such as cognitive decline, psychiatric disorders, or treatment response, are influenced by both structural and functional alterations in the brain. Relying on a single modality may provide only partial insight into the underlying neural mechanisms. For instance, structural MRI may reveal atrophy in a specific brain region, but without functional MRI, the model might miss crucial information about abnormal activity patterns or disrupted functional connectivity. Similarly, diffusion MRI adds value by characterising the integrity of white matter tracts, which may not be evident in other modalities. Multimodal neuroimaging thus enhances the ability to capture complex brain-behavior relationships by integrating multiple layers of complementary information.

In this context, we propose an enhanced GPNN architecture, the *Imaging Bimodality GPNN*, designed to effectively integrate data from two imaging modalities and model their interactions. By incorporating interaction terms between modalities, our approach enables the model to capture the synergistic effects that may arise when structural and functional alterations jointly contribute to the outcome. This is particularly important when the outcome is driven by complex, multimodal neural processes that cannot be adequately captured by a single modality alone.

Incorporating multimodal data into the GPNN framework not only improves predictive performance but also enhances interpretability by allowing us to disentangle the contributions of each modality and their interactions. This is crucial for understanding how different aspects of brain structure and function combine to influence clinical outcomes. For instance, our model may reveal how grey matter atrophy interacts with disrupted functional connectivity to predict cognitive decline, providing insights into disease mechanisms that could inform future interventions.

6.3.3.1 Bimodality GPNN-Linear-GP

The Bimodality GPNN-Linear-GP model is designed to process two distinct imaging modalities through separate sub-networks, each connected to its own hidden layer. These hidden neurons undergo a GP transformation, with the resulting basis coefficients linearly connected to the output. The model incorporates interaction terms, enabling a joint analysis of the modalities while maintaining a structure slightly different from

the configuration of the K-GPNN-Linear-GP. The depiction of this architecture is given in the Appendix Figure C.2a.

A key feature of this architecture is the selective interaction mechanism, where interactions are restricted to corresponding eigenbasis terms. Specifically, the leading term from one modality interacts only with the leading term from the other. This constraint ensures that the model captures shared structural patterns across imaging modalities without introducing unnecessary computational complexity, preserving efficiency.

For the Bimodality GPNN-Linear-GP model, priors for Modality 1 are defined as $\alpha^{(1)}(\mathbf{v}) \approx \sum_{k=1}^{K} \psi_k \rho_k^{(1)}(\mathbf{v})$ and $\eta^{(1)}(\mathbf{v}, \mathbf{v}') = \theta(\mathbf{v}, \mathbf{v}')$, where $\psi_k \sim \mathcal{N}(0, \sigma^2\omega^2)$, $\theta(\mathbf{v}, \mathbf{v}') \sim \mathcal{N}(0, \sigma^2\tau^2(\mathbf{v}))$, $\eta_0^{(1)}(\mathbf{v}) \sim \mathcal{N}(0,1)$, and $\alpha_0 \sim \mathcal{N}(0,1)$. For Modality 2, $\alpha^{(2)}(\boldsymbol{u}) \approx \sum_{j=1}^{J} \xi_j \rho_j^{(2)}(\boldsymbol{u})$ and $\eta^{(2)}(\boldsymbol{u}, \boldsymbol{u}') = \mu(\boldsymbol{u}, \boldsymbol{u}')$, where $\xi_j \sim \mathcal{N}(0, \sigma^2\upsilon^2)$, $\mu(\boldsymbol{u}, \boldsymbol{u}') \sim \mathcal{N}(0, \sigma^2\iota^2(\boldsymbol{u}))$, and $\eta_0^{(2)}(\boldsymbol{u}) \sim \mathcal{N}(0,1)$. To ensure symmetry, we assume $J = K$ in this development.

Mathematically, the prediction function $\hat{y}(X_i^{(1)}, X_i^{(2)})$ is defined by

$$\hat{y}(X_i^{(1)}, X_i^{(2)}) = \alpha_0 + \sum_{k=1}^{K} \psi_k D_{i,k} + \sum_{j=1}^{J} \xi_j Q_{i,j} + \sum_{k=j=1}^{K} \phi_{k,j} D_{i,k} Q_{i,j} \tag{6.26}$$

where

$$D_{i,k} = \sum_{\mathbf{v} \in \mathcal{A}^{(1)}} \rho_k^{(1)}(\mathbf{v}) h\Big(\eta_0^{(1)}(\mathbf{v}) + \sum_{\mathbf{v}' \in \mathcal{B}^{(1)}} X_i^{(1)}(\mathbf{v}') \theta(\mathbf{v}, \mathbf{v}')\Big) \tag{6.27}$$

is the k-th basis post-GP-transformed hidden layer of the *first* imaging mode of subject i.

$$Q_{i,j} = \sum_{\boldsymbol{u} \in \mathcal{A}^{(2)}} \rho_j^{(2)}(\boldsymbol{u}) h\Big(\eta_0^{(2)}(\boldsymbol{u}) + \sum_{\boldsymbol{u}' \in \mathcal{B}^{(2)}} X_i^{(2)}(\boldsymbol{u}') \mu(\boldsymbol{u}, \boldsymbol{u}')\Big) \tag{6.28}$$

is the j-th basis post-GP-transformed hidden layer of the *second* imaging mode of subject i.

The corresponding square-loss optimisation problem then becomes

$$\arg\min_{\boldsymbol{\Theta}} \ \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - \left\{\alpha_0 + \sum_{k=1}^{K}\psi_k D_{i,k} + \sum_{j=1}^{J}\xi_j Q_{i,j} + \sum_{k=j=1}^{K}\phi_{k,j} D_{i,k} Q_{i,j}\right\}\right]^2 \tag{6.29}$$

$$\begin{aligned}
&+ \frac{K}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\frac{1}{\omega^2}\sum_{k=1}^{K}\psi_k^2 \\
&+ \frac{L}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\frac{1}{\upsilon^2}\sum_{j=1}^{J}\xi_j^2 \\
&+ \frac{1}{2}\alpha_0^2 \\
&+ \frac{|\mathcal{A}^{(1)}||\mathcal{B}^{(1)}|}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\mathbf{v}\in\mathcal{A}^{(1)}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\mathbf{v}'\in\mathcal{B}^{(1)}}\theta(\mathbf{v},\mathbf{v}')^2 + \frac{1}{2}\sum_{\mathbf{v}\in\mathcal{A}^{(1)}}\eta_0^{(1)}(\mathbf{v})^2 \\
&+ \frac{|\mathcal{A}^{(2)}||\mathcal{B}^{(2)}|}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\frac{1}{\iota^2(\boldsymbol{u})}\sum_{\boldsymbol{u}'\in\mathcal{B}}\mu(\boldsymbol{u},\boldsymbol{u}')^2 + \frac{1}{2}\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\eta_0^{(2)}(\boldsymbol{u})^2.
\end{aligned}$$

The complete partial derivatives of parameters specific to this architecture are provided in Appendix C.3.

6.3.3.2 Bimodality GPNN-GP-Linear

The Bimodality GPNN-GP-Linear model adapts the K-GPNN-GP-Linear architecture by replacing non-imaging covariates and latent classes with two subnetworks for imaging data, allowing interactions in the hidden layer. This design uses linear connections between the hidden layer and output, offering a parsimonious framework for analysing hidden-layer mechanisms. The depiction of this architecture is given in the Appendix Figure C.2b. The priors for Modality 1 are defined as $\alpha^{(1)}(\mathbf{v}) = \psi_{\mathbf{v}}$ and $\eta^{(1)}(\mathbf{v},\mathbf{v}') \approx \sum_{\ell=1}^{L_{\mathbf{v}}}\gamma^{(1)}_{\ell\mathbf{v}'\mathbf{v}}\theta^{(1)}_{\ell\mathbf{v}}$, where $\theta^{(1)}_{\ell\mathbf{v}} \sim \mathcal{N}(0,\sigma^2\tau^2(\mathbf{v}))$ and $\eta_0^{(1)}(\mathbf{v}) \sim \mathcal{N}(0,1)$. For Modality 2, the priors are $\alpha^{(2)}(\boldsymbol{u}) = \xi_{\boldsymbol{u}}$ and $\eta^{(2)}(\boldsymbol{u},\boldsymbol{u}') \approx \sum_{k=1}^{K_{\boldsymbol{u}}}\gamma^{(2)}_{k\boldsymbol{u}'\boldsymbol{u}}\mu_k(\boldsymbol{u})$, where $\mu_k(\boldsymbol{u}) \sim \mathcal{N}(0,\sigma^2\iota^2(\boldsymbol{u}))$ and $\eta_0^{(2)}(\boldsymbol{u}) \sim \mathcal{N}(0,1)$.

Mathematically, the prediction function $\hat{y}(X_i^{(1)}, X_i^{(2)})$ is defined by

$$\hat{y}(X_i^{(1)}, X_i^{(2)}) = \alpha_0 + \sum_{\mathbf{v}\in\mathcal{A}^{(1)}}\psi_{\mathbf{v}} D_{i,\mathbf{v}} + \sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\xi_{\boldsymbol{u}} Q_{i,\boldsymbol{u}} + \sum_{\mathbf{v}\in\mathcal{A}^{(1)}}\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\phi_{\mathbf{v},\boldsymbol{u}} D_{i,\mathbf{v}} Q_{i,\boldsymbol{u}} \tag{6.30}$$

where

$$D_{i,\mathbf{v}} = h\Big(\eta_0^{(1)}(\mathbf{v}) + \sum_{\ell=1}^{L_{\mathbf{v}}} \theta_\ell(\mathbf{v}) \sum_{\mathbf{v}' \in \mathcal{B}^{(1)}} X_i(\mathbf{v}')\gamma_\ell(\mathbf{v}', \mathbf{v})^{(1)}\Big) \tag{6.31}$$

is the $\mathbf{v}$-th hidden regional neuron of the *first* imaging mode of subject i and

$$Q_{i,\boldsymbol{u}} = h\Big(\eta_0^{(2)}(\boldsymbol{u}) + \sum_{k=1}^{K_{\boldsymbol{u}}} \mu_k(\boldsymbol{u}) \sum_{\boldsymbol{u}' \in \mathcal{B}^{(2)}} X_i(\boldsymbol{u}')\gamma^{(2)}_{k\boldsymbol{u}'\boldsymbol{u}}\Big) \tag{6.32}$$

is the $\boldsymbol{u}$-th hidden regional neuron of the *second* imaging mode of subject i.

The corresponding square-loss optimisation problem then becomes

$$\begin{aligned} \arg\min_{\boldsymbol{\Theta}} \ & \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(y_i - \hat{y}(X_i^{(1)}, X_i^{(2)})\Big)^2 \\ & + \sum_{\mathbf{v}\in\mathcal{A}^{(1)}} \frac{L_{\mathbf{v}}}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\mathbf{v}\in\mathcal{A}^{(1)}} \frac{1}{\tau^2(\mathbf{v})}\sum_{\ell=1}^{L_{\mathbf{v}}}\theta_\ell(\mathbf{v})^2 + \frac{1}{2}\sum_{\mathbf{v}\in\mathcal{A}^{(1)}}\eta_0^{(1)}(\mathbf{v})^2 \\ & + \sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}} \frac{K_{\boldsymbol{u}}}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}} \frac{1}{\iota^2(\boldsymbol{u})}\sum_{k=1}^{K_{\boldsymbol{u}}}\mu_k(\boldsymbol{u})^2 + \frac{1}{2}\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\eta_0^{(2)}(\boldsymbol{u})^2. \end{aligned} \tag{6.33}$$

The complete partial derivatives of parameters specific to this architecture are provided in Appendix C.4.

6.3.4 Posterior Sampling

A key contribution of this chapter lies in the application of uncertainty quantification in real-data analysis. We employ a Bayesian sampling technique called Stochastic Gradient Langevin Dynamics (SGLD) to approximate the posterior distribution, which provides a probabilistic framework for capturing uncertainty in model parameters.

The SGLD update for a parameter $(\cdot)$ is defined as:

$$\nabla_{\text{SGLD}}(\cdot) = \nu\left\{\frac{n}{m}\sum_{i=1}^{m}\frac{\partial \log P(\mathbf{y}_i|X_i;\boldsymbol{\Theta})}{\partial(\cdot)} + \frac{\partial \log P(\boldsymbol{\Theta})}{\partial(\cdot)}\right\} + \zeta, \tag{6.34}$$

where $\zeta \sim \mathcal{N}(0, 2\nu)$ is the injected noise, specifically designed to balance with the learning rate ν so that the variance of the samples aligns with that of the posterior

distribution.

The parameters subjected to SGLD computations in this chapter include:

$$\theta(\mathbf{v}, \mathbf{v}'), \theta_\ell(\mathbf{v}), \eta_0(\mathbf{v}), b_j, w_{p,j}, \sigma^2, \phi_{j,k}.$$

The SGLD algorithm not only approximates the Maximum A Posteriori (MAP) estimate of the parameters $\mathbf{\Theta}$ but also samples from the full posterior distribution. This allows for a richer statistical analysis, enabling:

- Posterior Parameter Estimation: Deriving point estimates of parameters based on the posterior distribution.
- Uncertainty Quantification: Evaluating the variability and confidence associated with each parameter estimate.
- Posterior Predictive Mean: Computing predictions that account for the full distribution of model parameters.

By incorporating these advanced statistical metrics, SGLD facilitates a deeper understanding of the model's behaviour and the underlying uncertainty associated with parameter estimates. This capability is particularly valuable in real-data scenarios, where accounting for uncertainty enhances both interpretability and robustness of the results.

6.4 Data

In this experimental study, data were taken from the UK Biobank, a comprehensive biomedical resource containing extensive imaging and non-imaging data from over 50,000 participants. From this dataset, we selected a subset of 4,000 participants, including MRI scans and associated imaging-derived measures. This study primarily investigates predicting cognitive performance, specifically the time taken to complete a pair matching task, as a measure of cognitive processing speed.

The 4,000 randomly selected subjects were split into a training set of 2,000 and a test set of 2,000, with balanced underlying non-imaging covariate distributions.

Additionally, an independent, non-overlapping set of 1,205 subjects was further used to assess the robustness of the analysis.

6.4.1 Voxel-Based Morphometry (VBM) Data

For structural analysis, we utilise Voxel-Based Morphometry (VBM) data derived from T1-weighted MRI scans. VBM provides a measure of regional grey matter concentration, with each 3-dimensional image represented by a grid of 91 x 109 x 91 voxels at a 2mm resolution. To incorporate spatial priors, a Gaussian process (GP) kernel is applied, allowing the model to account for local anatomical structures based on 3D coordinates and regional delineations. An atlas from (Tziortzi et al., 2011) was used to define 12 Regions of Interest (ROIs), which group voxels into anatomically or functionally similar clusters.

Preprocessing of VBM data involved creating a grey matter mask by averaging all 4,000 subjects' VBM images and applying a threshold of 0.1, thereby removing voxels with little or no grey matter. This mask was applied to each atlas-defined region to ensure consistency across participants.

6.4.2 Default Mode Network (DMN) Data

In addition to VBM data, the UK Biobank provides 3D spatial maps of the Default Mode Network (DMN) for certain participants, derived through Independent Component Analysis (ICA) on resting-state fMRI data. The DMN images were processed using the FSL MELODIC tool for ICA decomposition, which isolates functional connectivity patterns across brain regions. The DMN itself is a well-documented network associated with internally focused cognition, such as self-referential thought, mind-wandering, and baseline brain activity, and has been shown to exhibit altered connectivity in various neurological and psychiatric conditions.

Voxel intensities in the DMN images represent the strength of temporal correlation between each voxel's time series and the DMN component, providing a spatial map of the network's influence across the brain. This is particularly valuable for studying cognitive function, as the DMN's activity is known to decrease during task-focused behaviour and increase during rest. To maintain consistency, we use the same voxel

resolution, dimensions, brain mask, and GP kernel for the DMN data as used for the VBM images, allowing for direct integration and comparison between the two types of imaging data.

6.4.3 Non-Imaging Covariates

To capture additional socio-demographic influences on neuroimaging outcomes, several non-imaging covariates are included. The biological factor *sex* is incorporated due to its well-established role as a determinant of medical health outcomes. Although prior studies have not identified a strong effect of sex on the neuroimaging measures relevant to the specific health outcomes in this study, its inclusion enables the model to account for potential subtle or unanticipated interactions.

Additionally, *age* is considered as a key covariate. Age is a well-documented biological determinant of cognitive function, given its direct association with neurodegeneration and the progressive decline in cognitive abilities (Deary et al., 2009; Gonzales et al., 2022). By integrating age into the analysis, the model can better capture its pervasive impact on brain structure and function, as well as its influence on the health outcomes under investigation.

Socio-economic status is represented by deprivation indices from England (code 26410), Scotland (code 26427), and Wales (code 26426), derived from common socio-economic domains, including income, employment, health, education, and housing. These indices provide relative rankings of deprivation at small-area levels within each country, with values ranging from 0 to 90. The distribution is heavily right-skewed, indicating that a significant portion of the sample represents more deprived areas, which aligns with UK population demographics. For simplicity, we assume these indices are comparable across countries and treat them on a unified scale, acknowledging that these are relative measures with potential differences in interpretation across regions.

6.4.4 Covariate Categorisation

To simplify analysis and better capture the non-linear effects often seen in healthcare data, we transformed the continuous covariates, age and deprivation index, into discrete categories. In healthcare, significant effects often emerge between distinct subgroups

rather than as gradual, continuous changes. Therefore, both age and deprivation index were divided into three equal quantiles, and these quantile-based groups were used as covariates in place of the original continuous variables.

6.4.5 Response Variable Transformation

Our outcome of interest is the time taken by participants to complete a *pair matching* task, measured in seconds. This cognitive task provides a dynamic measure of cognitive performance. However, the distribution of response times is heavily skewed and displays irregular density patterns, which violate the normality assumption required for traditional regression methods.

To address this, we apply a Gaussian inverse cumulative distribution function (CDF) transformation, which normalises the response variable, making it more suitable for linear regression by aligning it with the assumption of normally distributed errors.

Let y_i denote the raw response time for subject i, and let $\hat{F}(y)$ represent the empirical CDF based on the training data. The transformed response variable, z_i, is calculated as:

$$z_i = \Phi^{-1}(\hat{F}(y_i)),$$

where Φ^{-1} is the inverse of the standard Gaussian CDF. For new data points y_{new}, the transformed value is computed as:

$$z_{new} = \Phi^{-1}(\hat{F}(y_{new})),$$

using the empirical CDF fitted to the training data. This transformation effectively addresses skewness and variability in response times, ensuring the transformed variable meets the assumptions required for robust regression analysis.

6.5 Results

6.5.1 Parameters set-ups

In our proposed GPNN architecture, a hyper-prior is specified on the first-layer imaging parameters to enhance adaptability to the data. For this, we define the hypervariance $\tau^2(\mathbf{v})$ with an Inverse-Gamma prior distribution, using parameters $\texttt{shape}_{\mathbf{v}} = 11$ and $\texttt{scale}_{\mathbf{v}} = 0.5$, representing a relatively uninformative prior to allow flexibility in model fitting.

Parameters associated with non-imaging covariates are initially drawn from a normal distribution $N(0, 0.0001)$. To assess the effect of initialisation, we compared random initialisation against a guided initialisation approach. Specifically, we pre-fitted a Ridge regression model to identify distinct error magnitudes, then fine-tuned initial weights to better separate the groups in the softmax function. Results, however, indicated no significant differences between the guided and random initialisations, suggesting that the proposed architecture is robust to variations in initial non-imaging parameter values.

As outlined in Section 3.4 of Chapter 5, the atlas used as a motivating example in this study is a 12-region bilateral grey-matter brain atlas derived in (Tziortzi et al., 2011). GP parameters within the GPNN are optimised through a region-wise grid search, allowing for fine-tuning across spatial structures. The number of bases in the Karhunen–Loève (K-L) expansions is set to 1,771 for each GP in the first layer, and to 5,456 for the GP in the hidden layer when applied.

6.5.2 Metrics

The predictive accuracy metrics used in this study are *Root Mean Squared Error* (RMSE) and *Coefficient of Determination* (R^2).

For Bayesian inference metrics, the posterior predictive interval (PPI) is used, which accounts for parameter uncertainty in individual predictions. The α posterior predictive interval is given by:

$$\text{PPI}_i := \bar{y}_i \pm z_\alpha \sqrt{\text{MSE} + \text{PPV}_i}, \tag{6.35}$$

where $\bar{y}_i$ is the posterior predictive mean, z_α is the Gaussian score at level $\left(1 - \frac{(1-\alpha)}{2}\right)$, MSE is the model's Mean Squared Error, and PPV_i is the posterior predictive variance for subject i.

The empirical coverage rate (ECR) assesses model performance by measuring how often the true value falls within the predicted interval:

$$\text{ECR}_{\alpha\%} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\text{PPI}_i}(y_i), \tag{6.36}$$

where $\mathbb{1}_A(\cdot)$ is the indicator function.

Saliency map $\frac{\partial \hat{y}(X_i)}{\partial X_i(\mathbf{v}')}$ for neuroimaging-based interpretation.

The Adjusted Rand Index (ARI) is used in this study to measure the similarity between the clusters produced by our model and the ground truth classes, accounting for label permutations. The ARI evaluates the similarity of two partitions by assessing how pairs of samples are grouped into the same or different clusters, while adjusting for chance agreement.

Mathematically, the ARI is defined as:

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}},$$

where n_{ij} is the number of samples shared between cluster i and true class j, a_i is the size of cluster i, b_j is the size of class j, and n is the total number of samples. The ARI ranges from -1 (no similarity) to 1 (perfect similarity), with 0 indicating similarity no better than random chance.

6.5.3 Simulation study

6.5.3.1 Simulation Set-up

In our simulation study, we leverage real data for the predictor variables to closely mimic the inherent noise and structural dependencies present in neuroimaging data. This approach provides a significant advantage over conventional simulations, which often rely on synthetic data generated from predefined signals, potentially missing

realistic data complexities. For this study, we use real VBM imaging data, augmented with non-imaging covariates, to illustrate the added predictive value of biological and socio-demographic factors. Specifically, we incorporate sex and deprivation index as non-imaging covariates, given that deprivation is known to be associated with neurodevelopment and neurodegeneration.

We begin with 4,000 subjects, selecting only those with deprivation indices below the 15th percentile or above the 85th percentile. This yields a set of 1,200 subjects, with a slight imbalance of 16 additional females. To equalise the gender representation, we randomly remove 16 female subjects, resulting in a final sample of 1,184 subjects. From this sample, we create an 80% training set (947 subjects) and a 20% test set, stratified by gender and approximately by age. For the non-imaging covariates, $\mathbf{Z}_i$, we encode sex as a binary variable (-1 for female and 1 for male) and use a one-hot encoding scheme to represent the deprivation index group. For instance, $\mathbf{Z}_j = (1, 0, 1)$ would denote that subject j is male and falls within the top 15th percentile of the deprivation index in this population.

To induce spatial dependence in the simulated response, we construct a synthetic outcome by fitting a Gaussian Process Regression (GPR) model to real imaging data, capturing spatial structure through the fitted coefficients. Denoting the fitted intercept by $\hat{\alpha}$, the voxel-based coefficients by $\hat{\beta}^{\text{img}}$, and the non-imaging coefficients by $\hat{\beta}^{\text{non-img}}$, we fit an approximate GPR model on the training imaging data, following the kernel structure in expression (5.2) with parameters: concentration $a = 0.5$, smoothness $b = 40$, and number of expansion terms $L = 12,341$. This model incorporates non-imaging covariates as linear terms and is trained against the true age of each subject, as prior analyses Chapter 5 indicated apparent VBM signals relevant to age. We denote the fitted model as $\widehat{GPR}(\cdot)$, with performance metrics: training RMSE of 4.202, test RMSE of 6.042, training R^2 of 71.514, and test R^2 of 41.227.

Using the fitted coefficients, we generate predictions for all 1,184 subjects with the expression:

$$\begin{aligned} \hat{y}_i &= \lfloor \widehat{GPR}(\mathbf{X}_i, \mathbf{Z}_i) + 0.5 \rfloor \\ &= \lfloor \hat{\alpha} + \mathbf{X}_i \hat{\beta}^{\text{img}} + \mathbf{Z}_i \hat{\beta}^{\text{non-img}} + 0.5 \rfloor \end{aligned} \tag{6.37}$$

for each $i = 1, \ldots, 1,184$.

To introduce class-specific characteristics, we define three synthetic classes, C_i, for each subject i, based on the following criteria:

1. Class 1: Subject i is female, i.e. $\mathbf{Z}_i = (-1, \cdot, \cdot)$.

2. Class 2: Subject i is male and in the top 15th percentile of deprivation index, i.e. $\mathbf{Z}_i = (1, 0, 1)$.

3. Class 3: Subject i is male and in the bottom 15th percentile of deprivation index, i.e. $\mathbf{Z}_i = (1, 1, 0)$.

For each class, we then adjust the synthetic response, $\hat{y}'_i$, as follows:

$$\hat{y}'_i = \begin{cases} \hat{y}_i, & \text{if } C_i = 1, \\ \hat{y}_i \times 3, & \text{if } C_i = 2, \\ \hat{y}_i - 40, & \text{if } C_i = 3. \end{cases}$$

The distribution of this transformed synthetic response is shown in Figure 6.1, where the class-specific distinctions are evident. This highlights the influence of sex and deprivation index on the synthetic response variable, resulting in distinct, non-overlapping distributions across the three classes.

To evaluate the simulation analysis, we re-randomised the training and test sets, ensuring a balanced distribution with 888 subjects in the training set and 296 in the test set. The simulation was conducted using the following proposed models: K-GPNN-Linear-GP, K-GPNN-GP-Linear, and K-GPNN-GP-GP. In each model, we set k to be higher than the true number of latent classes to examine how the model assigns these classes and assess its success in detecting latent structures.

6.5.3.2 Simulation Results

Table 6.2 presents the results of the simulation study, detailing training and test RMSE, R-square, and Adjusted Rand Index (ARI) values across various methods. These findings reveal notable differences in model performance, particularly when apparent

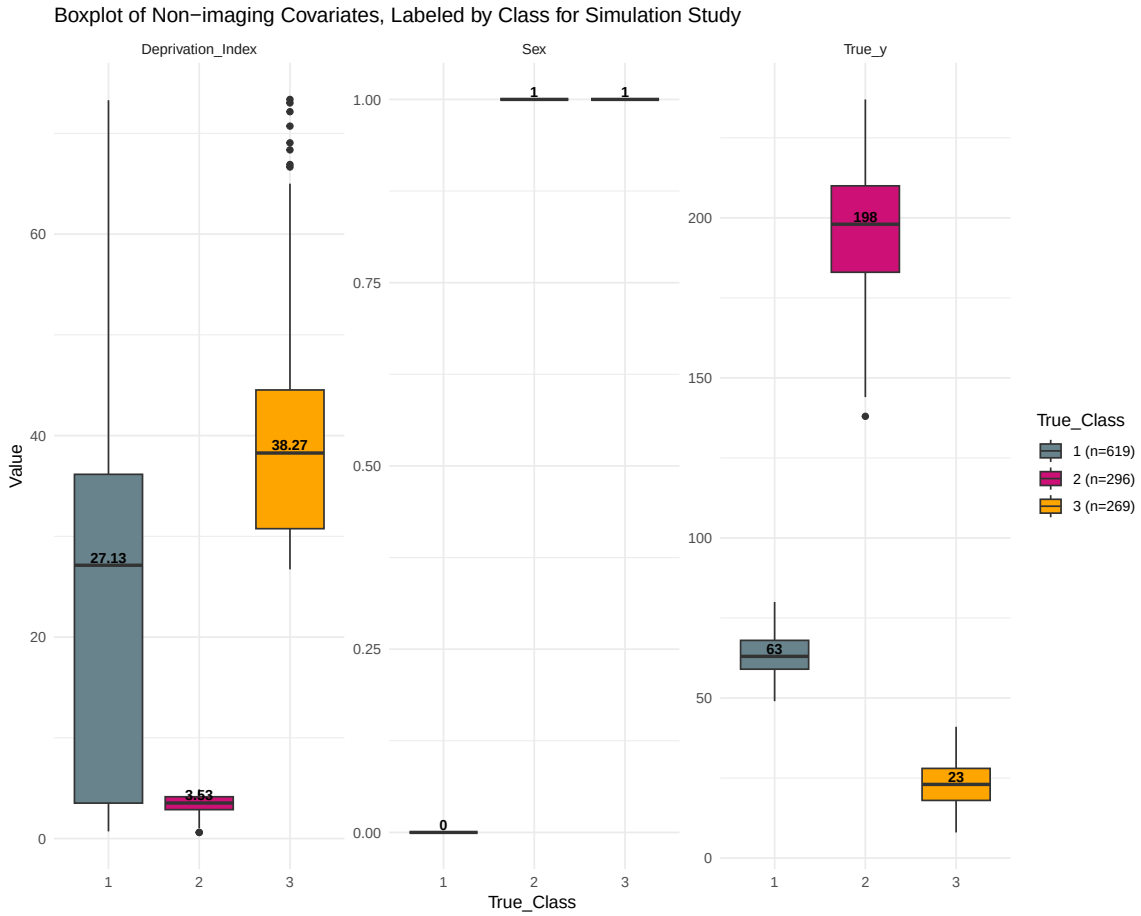


Figure 6.1: Boxplots of the three *true* classes from the simulation study, defined by two factors: Deprivation Index and Sex. Class 1 comprises females regardless of deprivation index, Class 2 includes males with low deprivation index, and Class 3 includes males with high deprivation index. The response variable ($True_y$) shows clear distinctions between these classes.

Method	Training RMSE	Test RMSE	Training R-squared	Test R-squared	ARI
Data	65.670	65.603	-	-	-
LASSO	40.637	44.845	61.71	53.00	-
Ridge	17.459	59.663	92.93	16.81	-
GPR	0.063	113.661	100.00	-201.91	-
GPNN-Linear-GP imaging-only	49.839	59.606	38.61	16.91	-
sex-GPNN-Linear-GP	41.666	56.255	57.01	25.99	-
K-GPNN-Linear-GP K=4	3.228	7.100	99.77	98.82	0.772
K-GPNN-GP-Linear K=4	5.640	6.717	99.19	98.95	0.692
K-GPNN-GP-GP K=4	8.800	9.186	98.09	98.03	0.065

Table 6.2: Simulation study results: training and test RMSE, R-squared and ARI values for various methods.

latent class effects are introduced, highlighting the limitations of traditional models and the advantages of our proposed approaches in handling such complexity. Note that ARI values are computed across all subjects (training and test sets combined).

The linear models, LASSO and Ridge, exhibit significant shortcomings in this context. LASSO achieves a training RMSE of 40.637 with an R-square of 61.706 but performs poorly on the test set, with an RMSE of 44.845 and a test R-square of 53.002. Ridge regression, while demonstrating strong training performance with an RMSE of 17.459 and an R-square of 92.93, fails to generalise effectively, yielding a test RMSE of 59.663 and a low test R-square of 16.812. These results suggest that traditional linear models struggle to capture the underlying non-linear latent class effects embedded in the data, which are necessary for accurate prediction in this setting.

The GPR model, used here as the underlying data-generating model, demonstrates extreme overfitting, achieving a near-perfect fit on the training data with an RMSE of 0.063 and an R-square of 100%. However, its test performance deteriorates drastically, yielding an RMSE of 113.661 and a negative test R-square of -201.905. This result highlights that while GPR captures intricate patterns in training data, it fails to generalise effectively when latent effects are present, rendering it unsuitable for out-of-sample prediction in this context.

The GPNN-Linear-GP img-only model, which incorporates GP transformation solely on imaging data, also struggles to represent the complex latent structure. It achieves a respectable training RMSE of 49.839 and R-square of 38.61, yet its test performance remains weak, with an RMSE of 59.606 and a low test R-square of 16.91.

This result underscores the limitation of omitting latent classes and non-imaging covariates, which leads to a failure to capture critical information required for robust prediction.

The sex-GPNN-Linear-GP model, which introduces a non-imaging covariate (in this case, binary sex) into the hidden layer, aims to enhance the model's learning and inference by interacting with the GP-transformed imaging layer. This approach demonstrates some improvement over the GPNN-Linear-GP img-only model, with lower test RMSE and higher R^2. However, this benefit comes at the expense of a high training RMSE, the highest among all models considered, suggesting it may not be fitting the data effectively during training. Interestingly, the model's results differ markedly from the K-GPNN-Linear-GP K=4 model, which essentially uses a latent structure that could capture binary sex effects (discussed in the next paragraph). This disparity indicates that the sex-GPNN-Linear-GP model may be limited in its configuration or lack sufficient flexibility, despite attempting to encode similar information to that realised in the K-GPNN-Linear-GP K=4 model.

Our proposed methods, designed to accommodate both imaging and non-imaging covariates alongside latent class information, exhibit markedly improved performance. The K-GPNN-Linear-GP K=4 model achieves the highest ARI of 0.772, showing strong latent class recovery, though it tends to split classes by sex rather than capturing all intended latent effects. Despite this limitation, the K-GPNN-Linear-GP K=4 model achieves excellent predictive accuracy, with a training RMSE of 3.228 and R-square of 99.771, and a test RMSE of 7.1 and test R-square of 98.821. These results indicate that even though this model does not always identify the true latent structure, it still generalises effectively, suggesting that a partial recovery of latent class structure is sufficient for accurate prediction.

The K-GPNN-GP-Linear K=4 model, with an ARI of 0.692, suggests less alignment with the true latent structure, as it mainly splits subjects into high deprivation females versus others. However, it achieves the lowest test RMSE of 6.717 and an impressive test R-square of 98.945. This finding implies that the K-GPNN-GP-Linear model may offer an alternative perspective by identifying latent groupings orthogonal to the original classes, providing additional insights into the data structure that contribute to high predictive accuracy.

The K-GPNN-GP-GP K=4 model, with an ARI of 0.065, either oversplits the latent characteristics or achieves near-perfect recovery. It achieves a training RMSE of 8.8 and a training R-square of 98.088, along with a test RMSE of 9.186 and a test R-square of 98.027. This model's ability to overfit or recover latent structure accurately reflects its sensitivity to the presence of strong class effects, making it a potential choice for real-world applications where the latent structure may not be well-defined. Its tendency to oversplit or recover subtle variations indicates a greater capacity to adapt to complex latent structures when strong effects are present.

In conclusion, this simulation study underscores the limitations of traditional linear models and standard GPR in addressing complex latent class effects. While linear models and GPR show poor generalisability, our proposed approaches successfully incorporate imaging and non-imaging covariates, leveraging latent class information for robust and interpretable predictions. Among these, the K-GPNN-GP-Linear K=4 and K-GPNN-Linear-GP K=4 models demonstrate the lowest RMSE, highlighting their effectiveness in capturing intricate data patterns. The K-GPNN-GP-GP K=4 model also shows promise, adapting flexibly to latent class structures, which may be advantageous in applications requiring nuanced subgroup identification. Interestingly, the sex-GPNN-Linear-GP model, while incorporating sex as a binary non-imaging covariate, does not achieve comparable predictive accuracy to the K-GPNN-Linear-GP K=4 model, suggesting that explicit latent class modelling may provide a more flexible and accurate approach. Overall, the K-GPNN-GP-Linear K=4 and K-GPNN-Linear-GP K=4 models emerge as the most generalisable and interpretable options for complex data landscapes involving latent class effects.

6.5.4 Real-data: Cognitive task

6.5.4.1 Selecting number of latent classes "K"

Figure 6.2 illustrates three random initialisations of the K-GPNN-GP-Linear with K=2 model, showing the distributions of non-imaging covariates (Age, Deprivation Index, Sex, Pair Matching score, and Predictions of Pair Matching score) along with squared errors. Additionally, residual vs. fitted scatterplots are presented, separated by the latent classes discovered by our algorithm.

The figure shows that latent class discovery is consistent across random initialisations, as evidenced by the similar distribution patterns in each non-imaging variable and prediction output across runs. Each latent class exhibits distinctive characteristics in the distribution of covariates. Group 1 is associated with individuals who tend to be younger or have a broader range of deprivation index values, independent of sex, and generally exhibit lower pair matching scores. In contrast, Group 2 is characterised by individuals who are typically older or have lower deprivation index values, again independent of sex, and tend to show higher pair matching scores. This differentiation is meaningful, as older age or lower socio-economic deprivation is often linked to variations in cognitive processing speed, likely due to factors like age-related neurodegeneration.

The residual vs. fitted scatterplots reveal distinct class-based patterns in the fitted values. Overall, residuals cluster closely around zero, indicating a good fit with homoscedasticity across classes. The consistency in class-specific patterns suggests that the model achieves a similar level of accuracy for each class, effectively capturing individual class characteristics with comparable precision.

In contrast, Figure 6.3 displays the same set of plots for the K-GPNN-GP-Linear with K=3 model, where the number of latent classes is increased to three. With this configuration, we observe inconsistency in latent class discovery across different random initialisations. While Run #5 yields an interpretable separation of latent classes, primarily grouping individuals by age and deprivation index, other initialisations, such as Runs #4 and #6, produce less distinct subgroups. These alternative runs show considerable overlap in the distributions of characteristics or introduce splits based on gender without clear alignment with meaningful cognitive or demographic subgroups. Additionally, the residual vs fitted scatterplots exhibit poor separation between clusters, suggesting that the algorithm struggles to delineate these groups in a linear manner with the increased number of classes.

Further experimentation with larger values of K (e.g. $K = 8$, $K = 10$, $K = 20$) reveals that increasing the maximum number of latent classes does not necessarily yield additional well-defined subgroups. In many cases, the algorithm fails to utilise all pre-specified classes, and the discovered latent groups lack consistency in their underlying distributions across different initialisations. Moreover, as K increases,

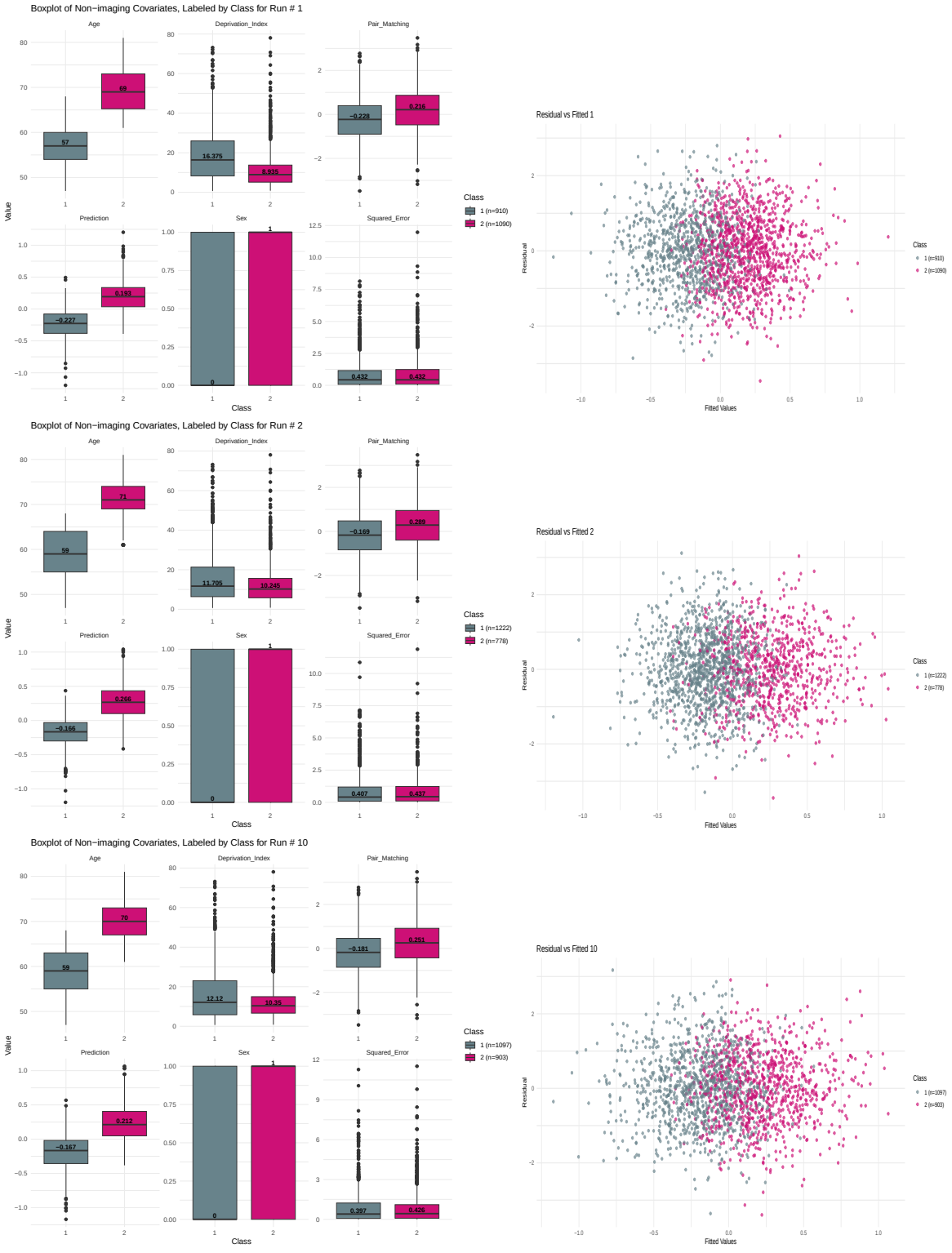


Figure 6.2: Boxplots and Residual vs. Fitted plots of three random initialisations of the K-GPNN-GP-Linear with K=2 model, showing distributions of non-imaging covariates (Age, Deprivation Index, Sex, Pair Matching score, and Predictions) and squared errors, along with class-specific residual vs. fitted scatterplots. The figure highlights consistent latent class discovery across initialisations, with Group 1 (younger or broader deprivation range, lower pair matching scores) and Group 2 (older or lower deprivation, higher pair matching scores) exhibiting distinct covariate distributions. Residuals cluster around zero with homoscedasticity across classes, indicating accurate fits and consistent precision in modelling class-specific characteristics.

overfitting becomes a more prominent issue. We observe that while the training RMSE decreases, indicative of the model fitting the training data closely, the test RMSE increases, reflecting a decline in generalisation performance.

In conclusion, setting $K = 2$ appears to be the most effective choice for this task, as it provides interpretable latent classes that align with meaningful subgroups while maintaining robust predictive performance on unseen data.

6.5.4.2 Predictive Accuracy

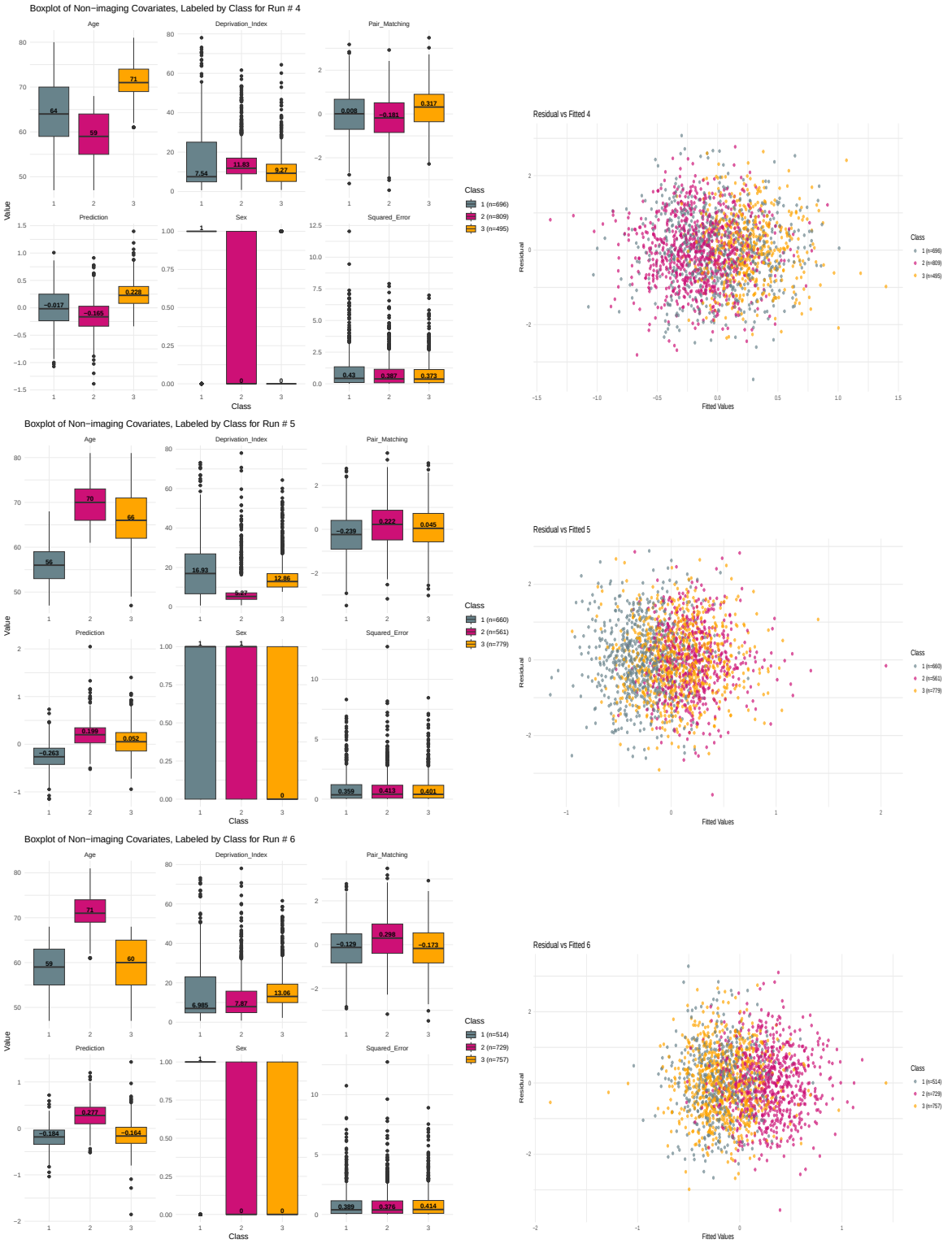


Figure 6.3: Boxplots and Residual vs. Fitted plots of three random initialisations for the K-GPNN-GP-Linear K=3 model, highlighting the impact of increasing the number of latent classes to three. Run #5 shows interpretable latent class separation by age and deprivation index, while Runs #4 and #6 reveal less distinct or overlapping subgroups, often introducing splits based on gender without clear cognitive or demographic alignment. Residual vs fitted scatterplots further reflect poor cluster separation.

Method	Training RMSE	Validation RMSE	Test RMSE	Training R-squared	Validation R-squared	Test R-squared	# Parameters	Run time
Data	1	1.062	1.031	-	-	-	156,812 VBM voxels + 2 genders + 3 Age groups + 3 Deprivation index groups	-
LASSO	0.967	1.038	1.002	6.48	4.73	5.56	17	5s
Ridge	0.905	1.042	1.007	18.13	4.11	4.47	156,791	6.5s
GPR	0.936	1.047	1.011	12.30	3.08	3.78	10,020	15s
***K*-GPNN-GP-Linear K=2**	0.952	1.033	0.996	9.87	5.41	6.63	21,319	7s/epoch
***K*-GPNN-GP-Linear K=2 at convergence**	0.934	1.044	1.018	9.35	3.46	2.46	21,319	7s/epoch
***K*-GPNN-GP-Linear K=3**	0.890	1.034	0.997	11.23	5.19	6.46	21,340	7s/epoch
***K*-GPNN-GP-Linear K=8**	0.938	1.037	1.002	10.63	4.69	5.51	21,445	7s/epoch
***K*-GPNN-Linear-GP K=2**	0.933	1.039	0.998	12.90	4.23	6.29	1,898,141	70s/epoch
***K*-GPNN-Linear-GP K=3**	0.967	1.047	1.015	6.26	2.02	2.99	1,903,605	70s/epoch
***K*-GPNN-GP-GP K=2**	0.986	1.044	1.019	2.68	3.40	2.26	37,649	16s/epoch
***K*-GPNN-GP-GP K=3**	0.982	1.044	1.009	4.89	3.33	4.23	43,113	16s/epoch
GPNN-GP-Linear imaging-only	0.984	1.045	1.002	3.14	3.23	5.57	21,277	7s/epoch
sex-GPNN-Linear-GP	0.940	1.040	0.998	11.60	4.10	6.17	1,892,669	70s/epoch

Table 6.3: Real-data study results: Performance metrics for various methods.

All results presented in Table 6.3 were obtained using model parameters at the minimum validation RMSE, which serves as the main criterion for model selection. Three primary metrics: RMSE and R^2 values for the training, validation, and test sets are used to assess model performance, with a particular focus on test set results as an indicator of generalisation capacity on unseen data. The standard deviation of the response variable within each data set (shown in the "Data" row) provides a baseline level of noise for the prediction task, with models achieving a test RMSE below this baseline considered successful in capturing meaningful predictive patterns. R^2 values complement the RMSE by indicating the proportion of variance explained, providing insight into each model's explanatory power.

The models are divided into two main groups: (1) benchmark linear models (LASSO, Ridge, and Gaussian Process Regression, or GPR) that do not incorporate spatial or anatomical priors, and (2) the proposed models, which leverage Gaussian Process (GP) transformations to incorporate spatial and anatomical information. The linear models serve as baseline comparisons; while they provide competitive RMSE and moderate R^2 values, they are limited by their lack of spatial priors. Among the linear models, Ridge achieves the lowest training RMSE and relatively high R^2 scores, suggesting a robust fit within the training data; however, its test RMSE and R^2 values indicate limited generalisation compared to models that incorporate prior spatial knowledge.

In contrast, the proposed GP-based models demonstrate the impact of integrating spatial and anatomical information through GP transformations. These models generally outperform the linear benchmarks in both validation RMSE and R^2, underscoring the advantage of spatial priors in capturing the underlying structure of neuroimaging data.

The sex-GPNN-Linear-GP model, which explicitly incorporates gender as a binary non-imaging covariate, demonstrates comparable performance to the K-GPNN-Linear-GP K=2 model. This suggests that including gender and its interaction effect with GP-transformed imaging features adds some value to model fitting and prediction of pair matching scores. Notably, this model performs better than the linear benchmarks (LASSO, Ridge, and GPR), highlighting that even a single non-imaging covariate can enhance performance in scalar-on-image regression tasks. However, the sex-

GPNN-Linear-GP model does not significantly outperform K-GPNN-Linear-GP K=2, indicating that the latent classes in the latter may inherently capture gender-related effects alongside other subgroup characteristics.

Among the latent-class models, K-GPNN-GP-Linear with $K = 2$ latent classes achieves the lowest validation and test RMSE, along with the highest test R^2, indicating it captures essential patterns that generalise well to unseen data. By incorporating two latent classes, this model effectively captures distinct subgroups within the data, reflecting nuanced variations that improve prediction accuracy and variance explanation.

The training RMSE across models provides insight into model fit, with lower values indicating a closer capture of training patterns. However, excessively low training RMSE, if not matched by high validation R^2, may suggest potential overfitting. For example, while K-GPNN-GP-Linear with $K = 3$ achieves a competitive test RMSE, its training RMSE is slightly higher than K-GPNN-GP-Linear with $K = 2$, indicating a balance between capturing additional latent structure and avoiding overfitting. This trade-off illustrates how the latent class parameter K influences model complexity, with too many latent classes potentially adding unnecessary detail, while too few (or none) may underfit the data.

The R^2 scores further confirm these findings, particularly in the test set, where the K-GPNN-GP-Linear models consistently achieve higher values than linear benchmarks, validating their effectiveness in predictive accuracy and generalisation. Specifically, K-GPNN-GP-Linear with $K = 2$ demonstrates the highest test R^2 among all models, suggesting that this configuration strikes the best balance across training, validation, and test metrics. This stable fit and robust generalisation likely arise from the optimal use of spatial and latent class information.

Finally, runtime per epoch reflects the computational demands of each model. While linear models such as LASSO and Ridge are computationally efficient, the GP-based methods require moderate additional processing time. For instance, GPNN-Linear-GP models, which include more complex architectures, have longer runtimes due to their additional GP transformation layers and the integration of spatial information. Despite this, the increased runtime of K-GPNN-GP-Linear models is manageable, especially given the performance gains in predictive accuracy and generalisation.

Overall, these results indicate that the GP-based methods, especially K-GPNN-GP-Linear with $K = 2$, provide superior predictive performance in this neuroimaging context by effectively leveraging both spatial and latent class information.

6.5.4.3 Bayesian Sampling

Table 6.4 summarises the Bayesian sampling performance for different models in predicting pair matching scores. Key metrics include the empirical coverage rate (ECR) at a 95% Gaussian significance level, which measures the proportion of subjects for whom the 95% posterior predictive interval (PPI) successfully covers the observed outcome. Ideally, both training and test ECR values should be close to 95%, as this reflects alignment with the theoretical confidence level. An ECR significantly above 95% in training may suggest overfitting, while a substantial drop in test ECR indicates weak generalisation.

Across models, we observe that the training $\text{ECR}_{95\%}$ values range from 95.55 to 95.8, all of which are close to the ideal 95% threshold. This result suggests that all models are achieving reasonably accurate coverage during training without excessive overfitting. The test $\text{ECR}_{95\%}$ values are slightly lower but generally consistent, ranging between 94.35 and 94.75, showing acceptable levels of generalisation across models. The GPR model, however, has the lowest test ECR at 94.35, suggesting it may generalise less effectively compared to other models. This finding points to the benefits of more sophisticated GP-based architectures, such as the proposed GPNN variants, in achieving more reliable predictive coverage.

The average posterior predictive interval width (PPI Width) provides insight into model uncertainty, with lower PPI widths reflecting more confident predictions. Both training and test PPI widths are compared to the 95% interpercentile range ($\text{IPR}_{95\%}$), a baseline measure of data spread. The $\text{IPR}_{95\%}$ for training and test data is 3.912 and 3.956, respectively. Here, the K-GPNN-GP-Linear K=2 model demonstrates the most efficient interval estimates, with a training PPI width of 3.804 and a test PPI width of 3.809, both of which are slightly below the $\text{IPR}_{95\%}$ threshold. This indicates a robust balance of predictive accuracy and model certainty, with the narrowest interval widths across both training and test sets.

Other models, such as K-GPNN-GP-GP K=2 and GPNN-GP-Linear imaging-only, also exhibit competitive PPI widths, though the GPNN-GP-Linear imaging-only model has a slightly higher average test PPI width of 3.845, compared to K-GPNN-GP-Linear K=2. This suggests that while GPNN-GP-Linear imaging-only is effective, incorporating latent classes, as in K-GPNN-GP-Linear K=2, enhances the model's certainty by reducing the PPI width. This difference highlights the advantage of latent class structures in capturing more granular patterns within the data, leading to more confident interval estimates and better alignment with actual variability in cognitive scores.

The K-GPNN-Linear-GP K=2 model shows a slightly broader PPI width in both training (3.886) and test (3.897) sets compared to K-GPNN-GP-Linear K=2. However, these widths remain near the IPR baseline, indicating reasonable model confidence, albeit slightly more conservative than the other GP-based methods. The GPR model lacks PPI width information in this table, limiting our direct comparison but further emphasizing the advantages of K-GPNN-GP-Linear configurations, particularly those with latent classes, for achieving lower PPI widths and high coverage rates.

In summary, the K-GPNN-GP-Linear K=2 model demonstrates the most balanced performance, with a close alignment to the target $\text{ECR}_{95\%}$ in both training and test sets and PPI widths that fall below the IPR threshold, indicating high model certainty. The incorporation of latent classes appears to strengthen the model's capacity for confident predictions, as seen in the K-GPNN-GP-Linear K=2 results compared to the GPNN-GP-Linear imaging-only. These findings underscore the effectiveness of GP-based models, particularly those with latent classes, in Bayesian sampling for neuroimaging tasks where balanced coverage and model certainty are crucial.

Overall Model Bayesian Sampling Performance						
Method	**Training $\text{ECR}_{95\%}$**	**Test $\text{ECR}_{95\%}$**	**Training Avg PPI Width**	**Training $\text{IPR}_{95\%}$**	**Test Avg PPI Width**	**Test $\text{IPR}_{95\%}$**
GPR	95.70	94.35	3.803	3.912	3.807	3.956
GPNN-GP-Linear imaging-only	95.70	94.65	3.844	3.912	3.845	3.956
K-GPNN-GP-Linear K=2	95.80	94.45	3.804	3.912	3.809	3.956
K-GPNN-Linear-GP K=2	95.55	94.75	3.886	3.912	3.897	3.956
K-GPNN-GP-GP K=2	95.55	94.55	3.841	3.912	3.869	3.956

Table 6.4: Real-data study results: Bayesian sampling performance summary of overall training and test coverage, PPI width, and IPR for various methods.

With the incorporation of latent class detection, our models enable a finer exami-

nation of Bayesian sampling performance within distinct subgroups, as compared to analyses without class separation. Table 6.5 and Table 6.6 display results for Group 1 and Group 2, respectively, allowing for comparison of Empirical Coverage Rate ($ECR_{95\%}$), Average Posterior Predictive Interval (PPI) Width, and Interpercentile Range ($IPR_{95\%}$) across methods. The underlying distribution of these subgroups reflects the characteristics identified in Section 6.5.4.1, with each group showing distinct demographic and cognitive profiles:

- **Group 1:** Comprises individuals who are typically younger or have a broader range of deprivation index values, independent of sex, and generally exhibit lower pair matching scores.
- **Group 2:** Comprises individuals who are typically older or have lower deprivation index values, again independent of sex, and generally exhibit higher pair matching scores.

It is important to note that due to noise introduced by SGLD in estimating the latent class parameters, the separation between groups is less distinct. This noise may blur the boundaries between classes, potentially introducing variability and overlap between results within each group. Comparability across methods for each group should therefore be approached with caution.

In Group 1 (Table 6.5), the K-GPNN-GP-Linear K=2 model achieves a training $ECR_{95\%}$ of 95.78% and a test $ECR_{95\%}$ of 94.48%. These values are close to the ideal 95% target, indicating reliable model coverage without signs of overfitting. The narrow PPI widths for K-GPNN-GP-Linear K=2 in both training (3.804) and test (3.808) align closely with the IPR values of 3.806 and 3.929, respectively. This alignment suggests that the model maintains stable uncertainty bounds in both training and testing, with prediction intervals that are suitably narrow for effective inference within this group. The consistency in $ECR_{95\%}$ values and narrow PPI widths makes K-GPNN-GP-Linear K=2 a well-suited model for the heterogeneous socio-demographic composition of Group 1, effectively balancing precision and generalisation.

The K-GPNN-Linear-GP K=2 model, also in Table 6.5, demonstrates a slightly higher training $ECR_{95\%}$ of 96.11%, accompanied by a test $ECR_{95\%}$ of 94.53%, indicating that its coverage rates remain close to the target but with slightly wider PPI widths in

training (3.884) and test (3.895). These widths exceed the training IPR of 3.751 and test IPR of 4.000, reflecting a slight increase in model uncertainty. This result suggests that K-GPNN-Linear-GP K=2 captures additional variability within Group 1, likely due to the group's socio-economic diversity, but may exhibit higher uncertainty than K-GPNN-GP-Linear K=2 when making inferences.

The K-GPNN-GP-GP K=2 model shows a training $\text{ECR}_{95\%}$ of 95.35% and a test $\text{ECR}_{95\%}$ of 94.22%, with PPI widths of 3.840 for training and 3.872 for test, as seen in Table 6.5. These values are close to the IPRs (3.923 for training and 4.039 for test) and indicate that this model has stable, though slightly broader, prediction intervals. The PPI width, only slightly above that of K-GPNN-GP-Linear K=2, suggests that K-GPNN-GP-GP K=2 captures more complex interactions but does so with a manageable increase in uncertainty, suitable for a group with diverse cognitive variability.

In Group 2 (Table 6.6), the K-GPNN-GP-Linear K=2 model achieves a training $\text{ECR}_{95\%}$ of 95.82% and a test $\text{ECR}_{95\%}$ of 94.42%, with PPI widths of 3.804 and 3.809, respectively. These values are close to the IPRs for this group (3.845 for training and 3.998 for test), indicating that K-GPNN-GP-Linear K=2 provides robust predictive certainty and appropriate interval estimates for an older population with distinct cognitive profiles. This model's ability to maintain narrow PPI widths across both training and test sets, alongside consistent ECR values, makes it well-suited for Group 2, highlighting its effectiveness in maintaining precise intervals.

The K-GPNN-Linear-GP K=2 model for Group 2 shows a training $\text{ECR}_{95\%}$ of 95.07% and a test $\text{ECR}_{95\%}$ of 94.95%, with PPI widths of 3.887 for training and 3.899 for test, as presented in Table 6.6. These widths are slightly above the training IPR (3.980) and test IPR (3.930), suggesting that this model has somewhat higher predictive uncertainty, but retains reliable generalisation. The model's slightly higher uncertainty may reflect increased sensitivity to complex socio-demographic patterns in this older group, capturing subtle variations without substantial deviation from ideal coverage.

The K-GPNN-GP-GP K=2 model for Group 2 shows an elevated training $\text{ECR}_{95\%}$ of 96.59% and a test $\text{ECR}_{95\%}$ of 96.26%, as shown in Table 6.6, suggesting overfitting with slightly higher ECRs than the 95% target. The PPI widths of 3.849 for training

and 3.851 for test are close to the training IPR of 3.798 but are slightly above the test IPR of 3.679, indicating tight intervals with limited generalisability. This model's overconfidence in Group 2 implies that while it captures highly specific interactions, its prediction intervals may be excessively narrow, particularly for out-of-sample data. It is important to note, however, that Group 2 has relatively low training ($n = 323$) and test ($n = 321$) sample sizes, which may affect the reliability of these results. The limited number of subjects in this group increases the likelihood that the observed high ECR values and narrow PPI widths are driven by sample-specific patterns rather than generalisable trends, highlighting the need for careful interpretation of these results in contexts with smaller datasets.

Aggregating the information from both groups, we find that the K-GPNN-GP-Linear K=2 model consistently achieves reliable $\text{ECR}_{95\%}$ values close to the target, with relatively lower PPI widths across groups, marking it as the most robust and generalisable model overall. The K-GPNN-Linear-GP K=2 model demonstrates slightly broader PPI widths, with modestly increased uncertainty, particularly in Group 2, which may indicate a nuanced sensitivity to socio-demographic characteristics. Conversely, the K-GPNN-GP-GP K=2 model, while effective at capturing complex non-linear relationships, shows signs of overfitting, especially in Group 2, as reflected in its elevated $\text{ECR}_{95\%}$ values and narrower prediction intervals.

Comparing these group-specific results to the overall model performance, similar trends emerge: Group 1 generally exhibits lower PPI widths, suggesting greater predictive certainty within a younger, socioeconomically diverse cohort. By contrast, Group 2 displays slightly higher PPI widths, indicating increased uncertainty, likely due to cognitive variability associated with an older population. This difference underscores the model's varying degrees of predictive confidence across demographic subpopulations. The K-GPNN-GP-Linear K=2 model consistently stands out as the most effective approach, delivering stable, interpretable predictions across both diverse and homogeneous subgroups. Given its reliable coverage rates and low uncertainty in both the overall and group-specific performances, the K-GPNN-GP-Linear K=2 model is well-suited for applications requiring robust predictions across varied populations.

Group 1 Bayesian Sampling Performance								
Method	**Training $ECR_{95\%}$**	**Test $ECR_{95\%}$**	**Training Avg PPI Width**	**Training $IPR_{95\%}$**	**Test Avg PPI Width**	**Test $IPR_{95\%}$**	**Training n**	**Test n**
K-GPNN-GP-Linear K=2	95.78	94.477	3.804	3.806	3.808	3.929	1019	1014
K-GPNN-Linear-GP K=2	96.112	94.532	3.884	3.751	3.895	4.000	926	951
K-GPNN-GP-GP K=2	95.349	94.223	3.840	3.923	3.872	4.039	1677	1679

Table 6.5: Group 1 Bayesian sampling performance: Training and test $ECR_{95\%}$, average PPI width, and $IPR_{95\%}$ values.

Group 2 Bayesian Sampling Performance								
Method	**Training $ECR_{95\%}$**	**Test $ECR_{95\%}$**	**Training Avg PPI Width**	**Training $IPR_{95\%}$**	**Test Avg PPI Width**	**Test $IPR_{95\%}$**	**Training n**	**Test n**
K-GPNN-GP-Linear K=2	95.821	94.422	3.804	3.845	3.809	3.998	981	986
K-GPNN-Linear-GP K=2	95.065	94.948	3.887	3.980	3.899	3.930	1074	1049
K-GPNN-GP-GP K=2	96.594	96.262	3.849	3.798	3.851	3.679	323	321

Table 6.6: Group 2 Bayesian sampling performance: Training and test $ECR_{95\%}$, average PPI width, and $IPR_{95\%}$ values.

6.5.4.4 Model Representation

Figure 6.4 presents brain-mapped model representations, showing the average magnitude of saliency or fitted coefficients across different random initialisations for various models. In particular, the main effect plot of the K-GPNN-GP-Linear K=2 model illustrates that the model's learning leverages information from several contiguous voxel regions across the brain, indicating that it utilises regional structural information to predict the effect of VBM on cognitive performance, specifically on the time taken to complete the Pair Matching task. This pattern suggests that the model actively incorporates anatomical context when assessing the main structural impact on the cognitive outcome. Similarly, the main effect of K-GPNN-GP-GP K=2 model also shares the same structure of learning, indicating that the first-layer GP transformation may be crucial in guiding the spatial discovery of the framework.

Furthermore, the inclusion of interaction terms in the K-GPNN-GP linear K = 2 model allows us to examine class-specific interaction effects in addition to the main effect for each identified group. The consistency of the latent classes, verified for $K = 2$ (see Section 6.5.4.1), allows for a reliable comparison of these effects. Group 1 (characterised by younger individuals or those with a broader range of deprivation index values, irrespective of sex) shows additional effects primarily localised in the visual association cortex. This region is essential for processing and interpreting visual information, which is expected to be highly active in a healthy brain engaged in a pair matching task, a task that relies on visual and cognitive coordination.

Conversely, Group 2 (typically older individuals or those with lower deprivation index values, also independent of sex, and generally taking longer time to complete the pair matching task) exhibits additional effects localised in the frontal lobe and areas related to emotion. These regions are associated with movement, memory, and complex cognitive functions. This pattern suggests that individuals in Group 2 may rely more heavily on frontal and emotional processing areas, potentially indicating a cognitive strategy focused on task management and memory rather than purely visual processing, as seen in Group 1. Such findings highlight the nuanced ways different brain regions may contribute to cognitive task performance depending on age and socio-economic background.

The GPR model also shows a similar pattern of regionally smooth learning, with voxel saliency distributed across various contiguous brain regions. This alignment with the K-GPNN-GP-Linear K=2 model's main effect indicates that the GPR model also leverages spatially coherent brain regions to predict the cognitive outcome, suggesting that both models are sensitive to the anatomical layout of brain structures. The smoothness observed in these saliency maps underscores that both models incorporate spatial dependencies, potentially allowing for future analyses to link specific structural regions to cognitive outcomes in a spatially meaningful way.

In contrast, both the K-GPNN-Linear-GP K=2 and Ridge models exhibit a less spatially coherent pattern, with non-smooth, dispersed voxel importance across the brain, lacking any concentrated regional effects. This lack of spatial focus implies that the models may not utilise structural or anatomical information effectively in predicting pair matching performance. Instead, it relies on individual voxel weights without accounting for regional coherence, which may limit its interpretative value in neuroimaging analyses. Additionally, the LASSO model, omitted here due to its extremely sparse selection of voxels, produces an overly sparse representation, making it impractical for interpreting spatially distributed effects across brain regions, similar to the limitations observed with Ridge.

Overall, these findings highlight the advantage of models like K-GPNN-GP-Linear K=2 and GPR that incorporate spatial priors, as they reveal interpretable region-wise patterns in voxel saliency. This spatial coherence demonstrates the potential of these

models to elucidate links between structural brain features and cognitive outcomes, offering a promising avenue for further exploration in neuroimaging studies.

6.5.4.5 Bimodality GPNN

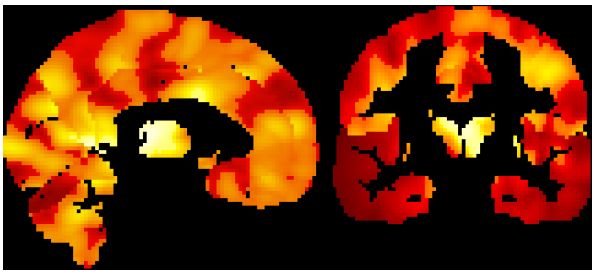
The data used in this section differ slightly from previous analyses, as not all subjects had resting-state fMRI data available. This bimodal dataset combines VBM and DMN imaging data, resulting in a massive input space with 156,812 voxels per modality, totalling 313,624 features.

Table 6.7 summarises the performance of our proposed bimodality interaction architecture compared to linear models and single-modality benchmarks.

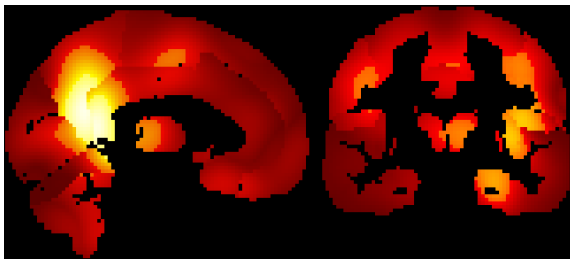
Among the linear models, LASSO achieves the best performance, likely due to its aggressive variable selection, which reduces the risk of overfitting. In contrast, Ridge performs poorly, effectively fitting only the intercept, as evidenced by its near-zero R^2 values for both training and test sets. This suggests Ridge fails to capture any meaningful relationships within this high-dimensional dataset. Similarly, the GPR model demonstrates severe overfitting, achieving strong training performance ($R^2 = 21.70$) but failing to generalise, with a negative test R^2 (-0.67) and a test RMSE exceeding the noise level in the data.

Our proposed GPNN-Linear-GP Bimodality model outperforms the linear models, achieving a training RMSE of 0.884 and a test RMSE of 0.986, with corresponding R^2 values of 25.57 (training) and 1.00 (test). This improvement highlights the model's capacity to capture complex cross-modal interactions between VBM and DMN data, which linear approaches cannot adequately address. However, despite its superior performance, the increased training R^2 suggests the model may be incorporating unnecessary complexity when combining both modalities.

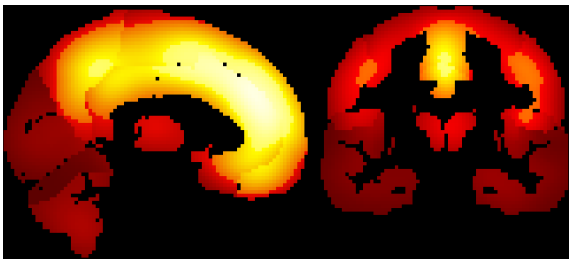
Interestingly, the GPNN-Linear-GP VBM-only model outperforms the bimodality model in terms of test performance ($R^2 = 4.30$ vs. 1.00). In contrast, the GPNN-Linear-GP DMN-only model shows limited predictive power, with a test R^2 of only 0.15. This discrepancy indicates that VBM provides more informative features for predicting pair matching scores, while DMN contributes marginally. Combining the two modalities may have introduced excessive complexity, hampering the model's ability to generalise effectively.



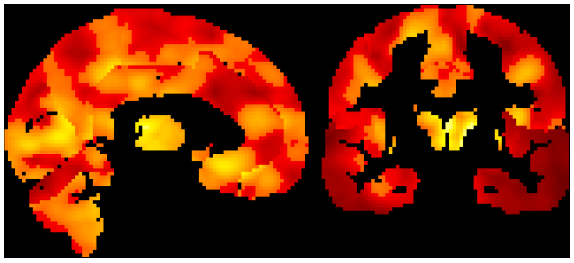
(a) K-GPNN-GP-Linear K=2 Main Effect



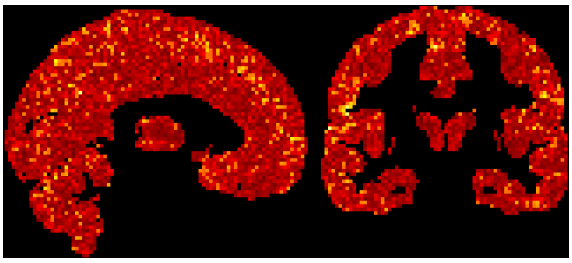
(b) K-GPNN-GP-Linear K=2 Interaction Group 1



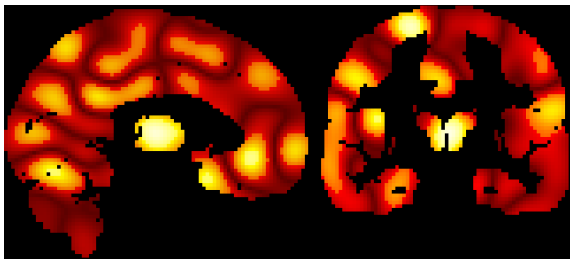
(c) K-GPNN-GP-Linear K=2 Interaction Group 2



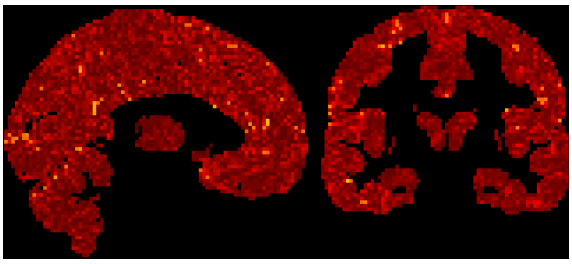
(d) K-GPNN-GP-GP K=2 Main Effect



(e) K-GPNN-Linear-GP K=2 Main Effect



(f) GPR



(g) Ridge

Figure 6.4: Brain-mapped model representations showing the average magnitude of saliency or fitted coefficients across random initialisations for various models. Both the K-GPNN-GP-Linear K=2 and the K-GPNN-GP-GP K=2 model highlight contiguous voxel regions actively utilised for predicting cognitive performance, with class-specific effects localised in the visual association cortex for Group 1 and the frontal lobe for Group 2, reflecting distinct cognitive strategies for K-GPNN-GP-Linear K=2. The GPR model exhibits regionally smooth saliency maps, aligning with spatial anatomical patterns. In contrast, the K-GPNN-Linear-GP K=2 and Ridge models show dispersed voxel importance with trivial spatial coherence.

This result is somewhat counterintuitive, as DMN data, being functional, is typically more closely associated with cognitive outcomes than VBM, which captures structural information. The findings suggest that while functional connectivity data like DMN might offer complementary insights, the structural data in VBM may dominate the predictive signal for this specific task. Further investigations into the weighting of modalities and the interaction architecture may help clarify these observations.

It is important to note that results for the GPNN-GP-Linear Bimodality architecture are not presented in Table 6.7 due to numerical instability encountered during the experiment. This issue likely stems from the computational complexity of simultaneously applying GP transformations across both imaging modalities, which significantly increases the parameter space and amplifies sensitivity to small numerical errors. Addressing this instability may require additional regularisation techniques or alternative optimisation strategies, which could be explored in future work.

Method	**Training RMSE**	**Test RMSE**	**Training R-squared**	**Test R-squared**
Data	1.000	1.000	-	-
LASSO	0.980	0.998	3.99	0.49
Ridge	1.000	1.000	0.05	0.00
GPR	0.885	1.003	21.70	-0.67
GPNN-Linear-GP VBM-only	0.975	0.978	1.99	4.30
GPNN-Linear-GP DMN-only	0.986	0.999	2.07	0.15
GPNN-Linear-GP Bimodality	0.884	0.986	25.57	1.00

Table 6.7: Real-data study results: Training and test RMSE and R-squared values for various methods using bimodality imaging input.

6.6 Discussion

This chapter introduced a suite of methodological advancements extending the Gaussian Process Neural Network framework from Chapter 5. Central to these developments is the ability to integrate non-imaging covariates, such as socio-demographic and biological factors, into predictive models. This integration enables the proposed methods to simultaneously account for the effects of non-imaging covariates and their interactions with imaging features. Additionally, a bimodality approach was developed, allowing two distinct imaging modalities to interact and inform predictions. These methodological contributions are underpinned by an innovative approach to latent

class modelling, which includes a strategy for determining the optimal number of latent classes based on class consistency.

In the simulation study, the limitations of traditional linear models and GPR were highlighted, particularly in their inability to handle complex latent class effects. Linear models, such as LASSO and Ridge, exhibited poor generalisability, while GPR severely overfitted the training data, resulting in negative test R^2 values. Conversely, our proposed methods, specifically K-GPNN-GP-Linear K=4 and K-GPNN-Linear-GP K=4, achieved the lowest RMSE values, demonstrating their robustness and generalisability. These models effectively captured intricate data patterns by incorporating both imaging and non-imaging covariates. The K-GPNN-GP-GP K=4 model also showed promise, particularly in its flexible adaptation to latent class structures, making it a potential tool for applications requiring nuanced subgroup identification. Interestingly, the sex-GPNN-Linear-GP model, which explicitly incorporated sex as a binary covariate, failed to match the predictive accuracy of the K-GPNN-Linear-GP K=4 model, underscoring the superiority of explicit latent class modelling for capturing complex interactions.

In the real-data cognitive task, K-GPNN-GP-Linear with K=2 emerged as the most robust and accurate model. Its ability to maintain consistent underlying latent class structures underscores its stability and reliability. The method achieved superior fit and prediction accuracy compared to both benchmark linear models and other proposed methods, solidifying its value for modelling neuroimaging outcomes.

Bayesian sampling results further demonstrated the advantages of incorporating latent classes. K-GPNN-GP-Linear K=2 exhibited narrower posterior predictive intervals and lower uncertainty compared to imaging-only models, highlighting the benefit of including non-imaging covariates and their aggregated interaction effects. Importantly, the latent class framework facilitated subgroup-specific Bayesian analysis, offering deeper insights into performance variations across subgroups. This capability enables targeted analyses that can inform personalised modelling and subsequent investigations.

In terms of interpretability, the K-GPNN-GP-Linear model provided brain-mapped class-specific effects, which appeared smooth and interpretable. This feature allows for a nuanced understanding of how subclasses differentially affect brain regions, offering a

valuable tool for understanding complex neuroimaging data. Similarly, the bimodality models can also enable the visualisation of interactions between different imaging modalities, further enriching interpretability and application potential.

To enhance the robustness of latent group identification, future work may incorporate statistical tests, such as multi-logistic regression analysis or chi-square tests, to evaluate the significance of group differentiation based on latent class identification similar to Lichenstein et al. (2022). Such tests could provide stronger evidence for the validity of class splits and offer practical guidelines for their use in real-world applications.

Preliminary experiments on increasing the number of GPs in the atlas from 12 to 52 showed improvements in model fit and interpretation by enabling finer anatomical granularity. However, the associated computational burden limited rigorous evaluation in this chapter. Future work could explore this extension with optimised computational strategies to balance granularity and efficiency.

While joint GP transformations in the K-GPNN-Linear-GP and K-GPNN-GP-GP models implicitly account for non-linear interaction effects, the addition of explicit interaction terms could further enhance the models' explanatory power. However, the high dimensionality resulting from projecting K-L bases poses significant computational challenges. Addressing these challenges in future work could unlock the full potential of this framework.

Finally, extending the proposed methods to other outcomes and covariates could provide additional insights into their performance under diverse conditions. Preliminary efforts to model complex cognitive tasks from the UK Biobank dataset, such as Fluid Intelligence (code 20016), Duration to complete numeric path (code 6348), Duration to complete alphanumeric path (code 6350), and Number of puzzles correctly solved (code 6373), revealed limited success, even with benchmark models. These findings suggest room for improvement in the current framework to better accommodate tasks with greater complexity.

In conclusion, this chapter demonstrates the versatility and effectiveness of the proposed extensions to the GPNN framework. By integrating imaging and non-imaging covariates, leveraging latent class information, and enabling bimodality interactions, these methods advance the field of neuroimaging analysis. Future developments hold

the promise of further enhancing their applicability and precision in addressing complex biomedical and cognitive research questions.

6.7 Code Availability

The code supporting this research, including that used for exploratory data analysis, model developments, simulation study design and results, as well as the real-data study implementation and results, is openly accessible and can be found at:

https://github.com/kan-k/KGPNN

Chapter 7

Conclusion

This thesis has presented a comprehensive exploration of methodological advancements in neuroimaging analysis, addressing fundamental challenges in the field. Spanning temporal modelling, dimensionality reduction, predictive frameworks, and multi-modal data integration, the research bridges traditional statistical methods and state-of-the-art machine learning approaches. Each chapter tackled a specific aspect of the neuroimaging pipeline, from pre-processing to prediction and interpretability, collectively advancing the precision, robustness, and scalability of neuroimaging analytics. The proposed methods not only address critical gaps in current techniques but also lay the groundwork for novel applications, including personalised modelling and enhanced interpretability in complex biomedical contexts.

- **Chapter 3: Spectral Adjusted Least Square for Autocorrelated Noise**

  Temporal autocorrelation in task-based fMRI data has long been a challenge, undermining the validity of statistical inferences and inflating false positive rates. Chapter 3 introduced **SALSA (Spectral Adjusted Least Squares for Autocorrelation)**, a novel pre-whitening method designed to address this issue. By employing a two-stage approach, first accounting for tissue-specific long-range dependencies globally and then refining voxel-specific autocorrelation estimates, SALSA effectively mitigates biases while preserving statistical integrity. This framework ensures that fMRI analyses remain robust, even in the context of fast-acquisition datasets with high temporal resolution.

When compared to established methods such as FAST, fMRIStat, and 3dREMLfit, SALSA demonstrated superior performance across several benchmarks. It consistently achieved accurate false positive rates and exhibited higher sensitivity in detecting synthetic signals. Its robustness against temporal smoothing and high sampling rates further distinguishes SALSA, making it particularly well-suited for modern neuroimaging applications, including pre-surgical fMRI and group-level analyses. SALSA's ability to adapt to tissue-specific autocorrelation properties, such as the elevated dependencies in grey matter and cerebrospinal fluid, underscores its versatility and relevance.

Despite these advancements, certain limitations remain. Discrepancies observed in group-level comparisons with 3dREMLfit highlight the need for further investigation into method-specific behaviours in diverse datasets. Additionally, extending SALSA to incorporate more complex spatiotemporal dependencies or multimodal imaging contexts presents a promising avenue for future work. Nonetheless, SALSA provides a robust foundation for improving temporal modelling in fMRI, bridging the gap between advanced imaging technology and the analytical precision required for reliable neuroimaging research.

- **Chapter 4: Bayesian High-Dimensional Variable Screening: Extensions to Posterior Mean Screening**

  Dimensionality reduction remains a critical challenge in neuroimaging, where the number of predictors (p) vastly exceeds the number of subjects (n). Chapter 4 addressed this issue by extending the Posterior Mean Screening (PMS) framework with a cohesive suite of innovations, focusing on constructing the prior covariance matrix $\mathbf{\Lambda}$ and estimating the smoothness parameter σ in the Gaussian kernel. The proposed methods included data-guided (PCA) and response-guided (SPCR) approaches to construct $\mathbf{\Lambda}$, alongside a shrinkage method to enhance robustness against noise. These elements were integrated into a unified framework that effectively captured spatial and functional dependencies, offering reliable and interpretable variable screening outcomes.

  The proposed methods outperformed traditional benchmarks, including HOLP, MUA, and standard PMS, particularly in real-data applications like age pre-

diction. By selecting more informative subsets of features, these methods demonstrated superior predictive accuracy and interpretability while achieving better results than full-dataset models. Regional interpretation analyses, using the Relative Importance Index, revealed distinct anatomical patterns unique to each method, further underscoring their ability to uncover clinically meaningful insights.

Despite these advancements, challenges remain. Future work could refine the shrinkage parameter δ through adaptive techniques like Ledoit-Wolf shrinkage or incorporate structured neighbour correlation matrices to better align with anatomical brain organisation. Additionally, exploring more complex covariance structures and integrating multimodal datasets could further enhance the robustness and scalability of the PMS framework. This chapter provides a strong foundation for advancing dimensionality reduction in high-dimensional neuroimaging, bridging the gap between analytical precision and biological insight.

- **Chapter 5: Introduction to Bayesian Gaussian Process Neural Network**

  Scalar-on-image regression tasks in neuroimaging pose significant challenges due to the need to capture non-linear relationships and spatial coherence in high-dimensional data. Chapter 5 addressed these challenges by introducing the Gaussian Process Neural Network (GPNN), a hybrid Bayesian framework that combines the flexibility of neural networks with the spatial modelling capabilities of Gaussian processes. Leveraging stochastic learning and Bayesian updates, GPNN effectively handled the computational demands of large-scale neuroimaging datasets while maintaining interpretability.

  The GPNN framework outperformed traditional Gaussian process and linear models across key metrics, achieving lower RMSE and higher R^2 scores, demonstrating its superior predictive accuracy and explanatory power. The GPNN-Linear-GP variant emerged as the best-performing model, and saliency analyses revealed the framework's ability to capture biologically plausible spatial structures. These findings underscore GPNN's robustness in handling large-scale

datasets ($n \gg 8,000$, $p \geq 100,000$), where traditional methods like ridge regression become computationally infeasible. Moreover, the integration of Stochastic Gradient Langevin Dynamics for posterior sampling reduced computational time significantly compared to traditional MCMC approaches like MALA, enhancing efficiency without compromising accuracy.

Despite these advancements, opportunities remain for refining the framework. Future work could explore more complex kernel functions and advanced optimisation techniques to further improve model performance and interpretability. Additionally, balancing smoothness and predictive accuracy, particularly in models like GPNN-Linear-GP, remains an open question. Extensions to incorporate multimodal imaging and non-imaging data, as well as applications to clinical imaging tasks requiring topological information, represent promising directions. These enhancements will expand GPNN's applicability and solidify its role as a versatile and powerful tool for tackling complex neuroimaging challenges.

- **Chapter 6: Extensions to Gaussian Process Neural Network**

  Chapter 6 addressed the challenge of integrating socio-demographic covariates, latent subgroup structures, and multi-modal imaging data into predictive models, extending the Gaussian Process Neural Network framework introduced in Chapter 5. By incorporating latent class modelling and enabling interactions between imaging and non-imaging factors, the proposed methods provided a unified framework to capture complex brain-behaviour relationships. This innovative approach facilitated class-specific analyses and offered a nuanced understanding of subgroup effects, significantly enhancing model interpretability and adaptability.

  The proposed methods demonstrated clear advantages over traditional approaches like LASSO, Ridge regression, and GPR, which struggled with generalisation and overfitting in the simulation study with induced latent class effects. Notably, even when GPR was used as the generative model, our models, such as K-GPNN-GP-Linear K=4 and K-GPNN-Linear-GP K=4, achieved the lowest RMSE values, outperforming GPR and other benchmarks. These models effectively captured intricate latent class interactions and demonstrated robust

performance, highlighting their ability to handle both imaging and non-imaging covariates in high-dimensional setups.

In real-data applications, K-GPNN-GP-Linear K=2 emerged as the most robust model, achieving superior predictive accuracy for cognitive outcomes while maintaining stable latent class structures. Bayesian sampling analyses highlighted the advantages of incorporating latent classes, with K-GPNN-GP-Linear K=2 exhibiting narrower posterior predictive intervals and reduced uncertainty compared to imaging-only counterpart. Subgroup-specific Bayesian analyses provided deeper insights into variability across subgroups, while brain-mapped class-specific effects illuminated how latent subgroups and imaging features differentially influenced brain regions. Additionally, the bimodality models enriched interpretability by capturing interactions between structural MRI, resting-state fMRI, and socio-demographic factors, offering a comprehensive view of brain-behaviour relationships.

Future work will focus on refining latent class identification using statistical tests like multi-logistic regression or F-tests to validate subgroup differentiation and class splits. Increasing the anatomical granularity of the Gaussian Process atlas could enhance model fit and interpretability, though computational strategies must be optimised to manage the associated demands. Expanding the framework to explicitly model interaction terms and applying it to a broader range of cognitive outcomes will further extend its versatility and impact. These advancements position the extended GPNN framework as a powerful tool for addressing complex biomedical and cognitive research challenges, paving the way for personalised neuroimaging studies.

The innovations presented in this thesis offer transformative frameworks that redefine the capabilities of neuroimaging analysis. By introducing methods that rigorously capture spatial and temporal dependencies, incorporate multimodal and socio-demographic data, and provide robust uncertainty quantification, this work equips researchers with a versatile framework for tackling the complexities of brain-behaviour interactions. These advancements not only tackle longstanding methodological challenges but also open new avenues for studying brain dynamics, enabling

more accurate predictions and population-level insights unifiedly. As the field of neuroscience progresses toward integrating increasingly complex and large datasets, the contributions of this thesis provide a foundation for developing more precise and impactful approaches, paving the way for transformative discoveries in neuroscience and beyond.

Appendix A

Bayesian High-Dimensional Variable Screening: Extensions to Posterior Mean Screening

A.1 Extended results tables

A.1.1 Simulations

A.1.1.1 Smoothing

Method	Median Power			
	$\sigma = 0.25$	$\sigma^* = 0.8$	$\sigma = 1$	$\sigma = 2$
PMS:PCA-S	27.17	31.67	30.33	27.50
PMS:PCA-S-Cov	43.17	43.50	39.67	73.83
PMS:PCA-S-Proj	43.33	46.17	40.83	68.17
PMS:SPCR-S	11.17	54.83	63.33	58.00
PMS:SPCR-S-Cov	43.17	45.67	43.33	72.00
PMS:SPCR-S-Proj	43.33	49.17	45.33	73.67
PMS:Spatial-S	43.50	54.00	53.00	7.67
Ridge-S	0.00	33.67	36.67	26.00

Table A.1: Median Power across different smoothing parameters (σ) for Scenario 1: A Single Contiguous Region. σ^* is derived from averaged correlations as proposed in Section 4.3.3.1.

Method	Median Power				
	$\sigma = 0.25$	$\sigma^* = 0.55$	$\sigma^\dagger = 0.72$	$\sigma = 1$	$\sigma = 2$
PMS:PCA-S	31.00	21.67	31.00	30.00	26.50
PMS:PCA-S-Cov	40.83	40.00	34.83	25.17	44.50
PMS:PCA-S-Proj	40.83	39.67	36.50	27.67	44.50
PMS:SPCR-S	11.33	26.67	33.67	37.83	27.67
PMS:SPCR-S-Cov	40.83	40.17	35.83	25.50	41.67
PMS:SPCR-S-Proj	40.50	40.50	35.33	26.67	42.83
PMS:PMS:Spatial-S	40.83	42.00	40.00	36.33	4.17
Ridge-S	0.00	8.17	24.50	28.33	17.00

Table A.2: Median Power across different smoothing parameters (σ) for Scenario 2: Double Separated Regions. σ^* is derived from averaged correlations as proposed in Section 4.3.3.1, and $\sigma^\dagger$ is derived from averaged absolute correlations.

A.1.1.2 Sensitivity and Stability

Method	$R^2 = 0.5$	$R^2 = 0.9$	$R^2 = 1$
	Median Power (SD)	Median Power (SD)	Median Power (SD)
PMS:PCA	23.67 (1.72)	29.00 (2.23)	28.83 (3.76)
PMS:PCA-S	24.50 (7.98)	31.67 (2.08)	33.00 (3.54)
PMS:PCA-S-Cov	37.17 (3.98)	45.00 (4.35)	45.67 (3.96)
PMS:PCA-S-Proj	40.50 (4.44)	46.50 (4.27)	48.50 (3.95)
PMS:SPCR	8.67 (1.43)	11.83 (1.14)	11.67 (1.48)
PMS:SPCR-S	31.00 (5.85)	47.00 (5.13)	49.50 (7.64)
PMS:SPCR-S-Cov	41.17 (4.90)	47.50 (4.46)	48.67 (4.11)
PMS:SPCR-S-Proj	43.67 (2.91)	51.00 (4.38)	52.00 (3.52)
HOLP	33.50 (1.99)	41.67 (1.85)	44.67 (1.96)
PMS:Spatial	40.83 (3.80)	53.83 (2.48)	57.00 (2.34)
Ridge	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Ridge-S	29.00 (2.76)	34.50 (1.76)	36.00 (2.33)
MUA	14.83 (1.91)	21.83 (2.18)	24.00 (2.38)

Table A.3: Aggregated Results for Power across different R^2 values for Scenario 1: A Single Contiguous Region

Method	$R^2 = 0.5$	$R^2 = 0.9$	$R^2 = 1$
	Median Power (SD)	Median Power (SD)	Median Power (SD)
PMS:PCA	23.83 (2.51)	29.00 (2.58)	29.17 (2.46)
PMS:PCA-S	26.33 (4.81)	24.33 (7.26)	30.50 (6.88)
PMS:PCA-S-Cov	29.50 (2.85)	39.50 (2.67)	40.33 (1.87)
PMS:PCA-S-Proj	29.67 (2.68)	38.67 (2.51)	40.67 (2.56)
PMS:SPCR	8.17 (0.91)	11.50 (1.58)	13.33 (1.74)
PMS:SPCR-S	19.83 (2.88)	26.00 (3.81)	28.50 (2.00)
PMS:SPCR-S-Cov	29.83 (2.83)	40.00 (3.01)	41.67 (2.48)
PMS:SPCR-S-Proj	30.33 (2.69)	39.50 (2.82)	40.83 (2.42)
HOLP	30.67 (1.56)	41.00 (1.80)	41.67 (1.86)
PMS:Spatial	30.50 (2.51)	41.17 (2.60)	42.50 (2.20)
Ridge	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Ridge-S	6.50 (1.21)	8.17 (0.87)	8.33 (0.68)
MUA	14.50 (2.28)	21.17 (2.06)	22.50 (2.69)

Table A.4: Aggregated Results for Power across different R^2 values for Scenario 2: Double Separated Regions

A.1.2 Real-data experiment

A.1.2.1 Predictive accuracy

Method	$R^2 = 0.5$	$R^2 = 0.9$	$R^2 = 1$
	Median Power (SD)	Median Power (SD)	Median Power (SD)
PMS:PCA	23.67 (2.58)	31.00 (2.46)	31.83 (1.51)
PMS:PCA-S	20.00 (7.38)	30.17 (2.39)	31.67 (1.62)
PMS:PCA-S-Cov	10.67 (2.44)	15.33 (1.79)	16.50 (2.05)
PMS:PCA-S-Proj	10.83 (2.22)	15.67 (1.71)	17.17 (2.04)
PMS:SPCR	7.83 (1.44)	13.33 (2.09)	13.17 (1.33)
PMS:SPCR-S	6.67 (2.45)	9.50 (2.00)	9.50 (1.17)
PMS:SPCR-S-Cov	9.67 (1.47)	15.17 (1.79)	15.67 (1.88)
PMS:SPCR-S-Proj	9.67 (1.77)	15.00 (1.80)	15.50 (2.11)
HOLP	31.50 (2.40)	41.00 (2.16)	42.17 (2.59)
PMS:Spatial	10.33 (1.49)	13.83 (2.12)	14.83 (1.94)
Ridge	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Ridge-S	11.17 (1.60)	12.50 (1.79)	13.00 (1.09)
MUA	14.00 (1.89)	19.67 (2.03)	21.67 (1.22)

Table A.5: Aggregated Results for Power across different R^2 values for Scenario 3: Sparse Signals

Method	Training RMSE				Test RMSE			
	p = 2,000	p = 4,000	p = 6,000	p = 10,000	p = 2,000	p = 4,000	p = 6,000	p = 10,000
PMS:PCA	3.023	2.445	2.068	1.628	4.357	4.253	4.220	4.186
PMS:PCA-S	4.314	3.829	3.560	3.381	4.893	4.718	4.661	4.607
PMS:PCA-S-Cov	3.810	3.224	2.843	2.373	4.713	4.641	4.629	4.607
PMS:PCA-S-Proj	3.810	3.224	2.843	2.373	4.713	4.641	4.629	4.607
PMS:SPCR	4.290	3.707	3.350	2.722	5.334	5.106	4.921	4.743
PMS:SPCR-S	3.679	3.265	3.004	2.673	4.423	4.354	4.335	4.329
PMS:SPCR-S-Cov	3.655	3.182	2.893	2.564	4.435	4.407	4.420	4.458
PMS:SPCR-S-Proj	3.951	3.410	3.090	2.708	4.619	4.573	4.601	4.621
HOLP	2.637	1.948	1.556	1.149	4.769	4.659	4.589	4.574
PMS:Spatial	3.260	2.592	2.180	1.659	4.763	4.671	4.607	4.541
Ridge	3.171	2.553	2.183	1.741	4.784	4.635	4.564	4.486
Ridge-S	4.341	3.965	3.675	3.308	4.730	4.554	4.449	4.402
MUA	4.249	3.724	3.297	2.731	4.772	4.580	4.460	4.327
Data	7.597				7.600			
Overall Ridge	0.911				4.488			

Table A.6: Training and Testing Errors for Different Methods in Age Prediction

A.2 Additional experiments

A.2.1 Alternative simulation setup 1

In the simulation setup as outlined in Section 4.5.0.1, we change step 6 to 8 to keep the induced-noise level fixed, while varying the size of the true signal coefficients $\beta^S(\mathbf{v})$, as follows:

1. ...

2. ...

3. ...

4. ...

5. ...

6. For each simulation scenario, select a subset $s \leq n$ of *true effect* voxels within the ROI, define the set of *active* predictors S, and assign the *true* initial coefficient values $\boldsymbol{\beta}^S$. Set $\beta^S(\mathbf{v}) = 0$ for voxels not in S, and $\beta^S(\mathbf{v}) = \hat{\beta}^{\text{HOLP}}(\mathbf{v})$ for $v \in S$.

7. Create a vector η of synthetic Gaussian noises $N(0, \zeta^2)$, where ζ is arbitrarily set as $\sqrt{5}$ for Table A.7.

8. Given a desired R^2 level, α (e.g. 0.5, 0.9) such that $0 < \alpha < 1$, define a scaling factor, γ to be
$$\sqrt{\frac{\alpha\zeta^2}{(1-\alpha)\text{var}(X)\|\beta^S(\mathbf{v})\|_2}}$$

9. Construct synthetic response variable $\boldsymbol{y}$ by computing $\mathbf{X}\boldsymbol{\beta} * \gamma + \eta$, which yields $R^2 = \alpha$.

10. Assess the Power of detection for varying signal strengths.

A.2.2 Alternative simulation setup 2

In the simulation setup as outlined in Section 4.5.0.1, we change step 5 to 8 to keep the induced-noise level fixed, while varying the size of the true signal coefficients $\beta^S(\mathbf{v})$, along with changing the size of the synthetic voxel values in the designated ROI to follow a uniform distribution $U[50, 51]$, as follows:

1. ...

2. ...

3. ...

4. ...

Method	$R^2 = 0.5$	$R^2 = 0.9$
	Median Power (SD)	Median Power (SD)
PMS:PCA	22.83 (1.76)	30.83 (4.10)
PMS:PCA-S	24.67 (9.60)	32.67 (3.05)
PMS:PCA-S-Cov	35.17 (4.80)	45.50 (4.93)
PMS:PCA-S-Proj	38.33 (4.65)	47.83 (4.84)
PMS:SPCR	8.00 (1.89)	12.33 (2.06)
PMS:SPCR-S	31.17 (8.34)	53.17 (7.96)
PMS:SPCR-S-Cov	37.50 (3.67)	50.83 (5.20)
PMS:SPCR-S-Proj	40.67 (4.25)	50.17 (4.93)
HOLP	34.50 (2.75)	43.00 (2.48)
PMS:Spatial	38.67 (4.58)	53.83 (3.25)
Ridge	0.00 (0.00)	0.00 (0.00)
Ridge-S	28.50 (1.51)	33.50 (2.18)
MUA	14.50 (2.09)	21.67 (3.36)

Table A.7: Aggregated Results for Power across different R^2 values for Scenario 1: A Single Contiguous Region of Alternative simulation setup 1

5. Replace voxel values in the ROI for each subject with independent samples drawn from a uniform distribution $U[50, 51]$, expressed as:

$$X_i(v) = w_i(v) \quad \text{where} \quad w_i(v) \sim U[50, 51],$$

for subject i and voxel v within the ROI. This is done to mimic the intensity characteristics within the ROI and to disassociate any structural correlation of the ROI from other parts of the brain.

6. For each simulation scenario, select a subset $s \leq n$ of *true effect* voxels within the ROI, define the set of *active* predictors S, and assign the *true* initial coefficient values $\boldsymbol{\beta}^S$. Set $\beta^S(\mathbf{v}) = 0$ for voxels not in S, and $\beta^S(\mathbf{v}) = \hat{\beta}^{\text{HOLP}}(\mathbf{v})$ for $v \in S$.

7. Create a vector η of synthetic Gaussian noises $N(0, \zeta^2)$, where ζ is arbitrarily set as $\sqrt{1}$ for Table A.8.

8. Given a desired R^2 level, α (e.g. 0.5, 0.9) such that $0 < \alpha < 1$, define a scaling

factor, γ to be

$$\sqrt{\frac{\alpha\zeta^2}{(1-\alpha)\text{var}(X)\|\beta^S(\mathbf{v})\|_2}}$$

9. Construct synthetic response variable $\boldsymbol{y}$ by computing $\mathbf{X}\boldsymbol{\beta} * \gamma + \eta$, which yields $R^2 = \alpha$.

10. Assess the Power of detection for varying signal strengths.

Method	$R^2 = 0.5$	$R^2 = 0.9$
	Median Power (SD)	Median Power (SD)
PMS:PCA	38.67 (2.71)	38.67 (2.25)
PMS:PCA-S	33.17 (2.72)	35.00 (2.71)
PMS:PCA-S-Cov	98.17 (17.89)	98.33 (0.19)
PMS:PCA-S-Proj	98.33 (0.25)	98.33 (0.21)
PMS:SPCR	24.17 (2.06)	25.00 (1.97)
PMS:SPCR-S	85.67 (1.38)	85.33 (1.22)
PMS:SPCR-S-Cov	97.00 (9.76)	96.67 (0.42)
PMS:SPCR-S-Proj	97.83 (0.23)	97.67 (0.16)
HOLP	89.17 (1.54)	97.00 (0.70)
PMS:Spatial	78.17 (2.13)	90.00 (1.03)
Ridge	0.00 (0.00)	0.00 (0.00)
Ridge-S	27.83 (2.06)	33.83 (2.97)
MUA	15.17 (1.87)	22.17 (2.94)

Table A.8: Aggregated Results for Power across different R^2 values for Scenario 1: A Single Contiguous Region Alternative simulation setup 2

Appendix B

Introduction to Gaussian Process Neural Network

B.1 Bayesian Derivations

A. 1. Consider the following linear regression set up, we observe $\mathbf{x}_1, ..., \mathbf{x}_n$ as inputs and $y_1, ..., y_n$ as response variables such that $y \sim N(\mu, \sigma^2)$ with unknown μ and σ^2. We can express $\mu = \mu(\mathbf{x})$ as a function of our data $\mathbf{x}$. Then the log-likelihood can be

written as

$$\begin{aligned}
\log P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) &= \log \prod_{i=1}^{n} P(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log \Big\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_i - \mu(\mathbf{x}_i))^2}{2\sigma^2} \Big) \Big\} \\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{n} \log \Big\{ \exp\Big(- \frac{(y_i - \mu(\mathbf{x}_i))^2}{2\sigma^2} \Big) \Big\} \\
&= n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}_i))^2 \\
&= C_1 + n \log \frac{1}{\sqrt{\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}_i))^2 \\
&\propto n \log \frac{1}{\sqrt{\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}_i))^2
\end{aligned}$$

for some constant C_1 independent of μ and σ^2.

A. 2. Suppose $\boldsymbol{\theta} \sim MVN(\underline{0}, \sigma^2\tau^2 I_p)$, then the corresponding log density is

$$\begin{aligned}
\log P(\boldsymbol{\theta}|\underline{0}, \sigma^2\tau^2 I_p) &= \log \Big(\frac{1}{(\sqrt{2\pi\sigma^2\tau^2})^p} \exp\Big(- \frac{\boldsymbol{\theta}^\top \boldsymbol{\theta}}{2\sigma^2\tau^2} \Big) \Big) \\
&= C_2 + p \log \frac{1}{\sqrt{\sigma^2\tau^2}} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\theta}}{2\sigma^2\tau^2} \\
&\propto p \log \frac{1}{\sqrt{\sigma^2\tau^2}} - \frac{1}{2\sigma^2\tau^2} ||\boldsymbol{\theta}||_2^2
\end{aligned}$$

for some constant C_2 independent of σ^2 and τ^2.

A. 3. Consider a Gaussian likelihood on $\theta \sim N(0, \sigma^2\tau^2)$, with an Inverse-Gamma prior

on τ^2 such that $\tau^2 \sim \Gamma^{-1}(\alpha, \beta)$, then the conjugate posterior is

$$\begin{aligned}
P(\tau^2|\theta, \sigma^2, \alpha, \beta) &\propto \prod_{i=1}^{n} \Big(\frac{1}{\sqrt{2\pi\sigma^2\tau^2}} \exp\Big(-\frac{\theta_i^2}{2\sigma^2\tau^2} \Big) \Big) \Big(\frac{\beta^\alpha}{\Gamma(\alpha)} (\tau^2)^{-\alpha-1} \exp\big(-\frac{\beta}{\tau^2} \big) \Big) \\
&= \Big(\frac{1}{(2\pi\sigma^2)^{n/2}} (\tau^2)^{-n/2} \exp\Big(-\sum_{i=1}^{n} \frac{\theta_i^2}{2\sigma^2\tau^2} \Big) \Big) \Big(\frac{\beta^\alpha}{\Gamma(\alpha)} (\tau^2)^{-\alpha-1} \exp\big(-\frac{\beta}{\tau^2} \big) \Big) \\
&\propto (\tau^2)^{-(\alpha+n/2)-1} \exp\Big(-\frac{1}{\tau^2} \big(\beta + \sum_{i=1}^{n} \frac{\theta_i^2}{2\sigma^2} \big) \Big) \\
&\propto \Gamma^{-1}\Big(\alpha + \frac{n}{2}, \beta + \sum_{i=1}^{n} \frac{\theta_i^2}{2\sigma^2} \Big)
\end{aligned}$$

B.2 Covariance Kernel in Gaussian Processes

Let $\mathcal{GP}(\mu, \kappa)$ represent a Gaussian Process (GP) with mean function $\mu(\mathbf{v})$ and covariance kernel $\kappa(\mathbf{v}, \mathbf{v}')$ for any $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^d$. Hence if

$$f \sim \mathcal{GP}(\mu, \kappa),$$

then

$$\mathrm{E}\{f(\mathbf{v})\} = \mu(\mathbf{v})$$

and

$$\mathrm{Cov}\{f(\mathbf{v}), f(\mathbf{v}')\} = \kappa(\mathbf{v}, \mathbf{v}').$$

The kernel used in this study is called "Modified Squared Exponential" (MSE) kernel, which is defined as

$$\kappa_{MSE}(\mathbf{v}, \mathbf{v}') = \tau^2 \exp\{-a(||\mathbf{v}||^2 + ||\mathbf{v}'||^2) - b||\mathbf{v} - \mathbf{v}'||^2\}, \text{ for } a, b > 0, \tag{B.1}$$

where norm $||\mathbf{v}||^2 = \mathbf{v}^\top \mathbf{v}$. We have the following properties if $f \sim \mathcal{GP}(0, \kappa_{MSE})$:

1. $\mathrm{Var}\{f(0)\} = \tau^2$.
2. $\mathrm{Var}\{f(\mathbf{v})\} = \tau^2 \exp(-a||\mathbf{v}||^2)$.
3. $\mathrm{Cor}\{f(\mathbf{v}), f(\mathbf{v}')\} = \exp\{-b||\mathbf{v} - \mathbf{v}'||^2\}$.

Theorem B.1 (Mercer's theorem for kernel eigendecomposition)**.** *Under some mild regularity conditions, the covariance kernel can be decomposed as*

$$\kappa(\mathbf{v}, \mathbf{v}') = \sum_{\ell=1}^{\infty} \lambda_\ell \psi_\ell(\mathbf{v}) \psi_\ell(\mathbf{v}'),$$

where $\{\lambda_\ell\}_{\ell=1}^{\infty}$ *are eigenvalues and* $\{\lambda_\ell \psi_\ell(\mathbf{v})\}_{\ell=1}^{\infty}$ *are corresponding orthonormal eigenfunctions which satisfy the following properties:*

$$\int_{\mathbb{R}^d} \psi_\ell(\mathbf{v}) dv = 1, \ \textit{for } \ell \geq 1, \tag{B.2}$$

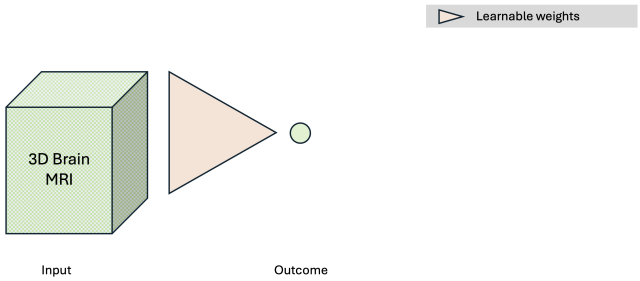
$$\int_{\mathbb{R}^d} \psi_\ell(\mathbf{v}) \psi_{\ell'}(\mathbf{v}) dv = 0, \ \textit{for } \ell \neq \ell'. \tag{B.3}$$

Corollary B.1 (Karhunen-Loeve expansion)**.** *From Mercer's Theorem, we can represent a GP* $f(\mathbf{v})$ *by the following expansion:*

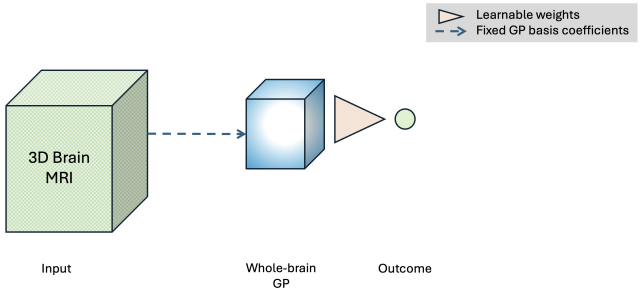
$$f(\mathbf{v}) = \mu(\mathbf{v}) + \sum_{\ell=1}^{\infty} \theta_\ell \sqrt{\lambda_\ell} \psi_\ell(\mathbf{v}), \ \textit{where } \theta_\ell \overset{\text{i.i.d.}}{\sim} N(0, \tau^2).$$

Thus, the scalar parameter τ^2 controls the maximum marginal variance of $f(\mathbf{v})$ and the maximum marginal variance of $f(\mathbf{v})$ is achieved at the origin 0. The parameter a controls the decay rate of $\text{Var}\{f(\mathbf{v})\}$ compared to $\text{Var}\{f(0)\}$. A larger a corresponds to a faster decay rate. The parameter b controls the smoothness of $f(\mathbf{v})$. A larger b corresponds to a smoother $f(\mathbf{v})$.

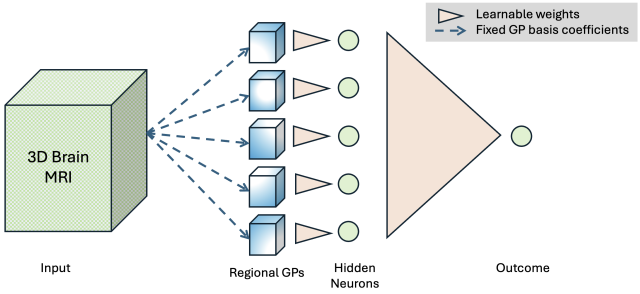
B.3 Additional figures



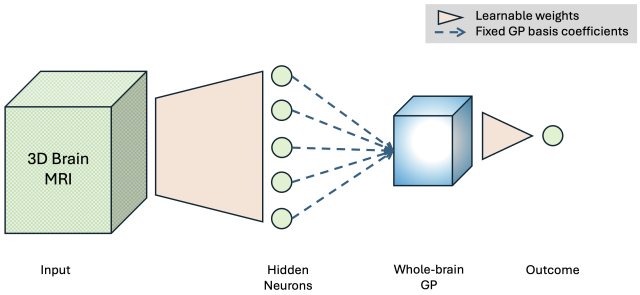
(a) Linear Regression model



(b) Gaussian Processs Regression model



(c) GPNN-GP-Linear model (case of 5 ROIs), given by Problem (6.22)



(d) GPNN-Linear-GP model (case of 5 ROIs), given by Problem (5.11)

Figure B.1: Additional visualisation of model architectures considered in the study.

Appendix C

Extensions to GPNN

C.1 *K*-GPNN-GP-Linear

For the first-layer imaging covariates, the unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)$ with respect to $\theta_\ell(\mathbf{v})$ is

$$\begin{aligned}&\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial \theta_\ell(\mathbf{v})} = \\ &n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\Big\{(y_i-\hat{y}_i)\Big(\psi_{\mathbf{v}}+\sum_{j=1}^{C}\phi_{j,\mathbf{v}}SM\big(F_i(j)\big)\Big)h'(G_i(\mathbf{v}))X_i(\mathbf{v}')\gamma_\ell(\mathbf{v}',\mathbf{v})\Big\}\Big] \\ &+\frac{1}{\sigma^2\tau_{\mathbf{v}}^2}\theta_\ell(\mathbf{v})\end{aligned} \tag{C.1}$$

where $G_i(\mathbf{v})$ is given by Eq. (6.21) and $F_i(j)$ is given by Eq. (6.12).

Similarly, the unbiased gradient estimator with respect to $\eta_0(\mathbf{v})$ is:

$$\begin{aligned}&\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial \eta_0(\mathbf{v})} = \\ &n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\Big\{(y_i-\hat{y}_i)\Big(\psi_{\mathbf{v}}+\sum_{j=1}^{C}\phi_{j,\mathbf{v}}SM\big(F_i(j)\big)\Big)h'(G_i(\mathbf{v}))\Big\}\Big]+\eta_0(\mathbf{v})\end{aligned} \tag{C.2}$$

For the first-layer non-imaging covariates, the unbiased estimator of the MAP

gradient of $C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)$ with respect to $w_{p,j}$ is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)}{\partial w_{p,j}} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\Big(\xi_j + \sum_{\mathbf{v}\in\mathcal{A}}\phi_{j,\mathbf{v}}h\big(G_i(\mathbf{v})\big)\Big)SM'\big(F_i(j)\big)Z_{i,j}\big\}\Big] \tag{C.3}$$

The unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)$ with respect to b_j is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)}{\partial w_{p,j}} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\Big(\xi_j + \sum_{\mathbf{v}\in\mathcal{A}}\phi_{j,\mathbf{v}}h\big(G_i(\mathbf{v})\big)\Big)SM'\big(F_i(j)\big)\big\}\Big] \tag{C.4}$$

Finally, the unbiased gradient estimator with respect to σ^2 is

$$\begin{aligned}&\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial \sigma^2} = \\ &n\frac{1}{m}\sum_{i=1}^{m}\Big\{\frac{1}{2\sigma^2} - \frac{(y_i - \hat{y}_i)^2}{2\sigma^4}\Big\} - \frac{1}{2\sigma^4}\sum_{\mathbf{v}\in\mathcal{A}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\ell=1}^{L_{\mathbf{v}}}\theta_\ell(\mathbf{v})^2 + \sum_{\mathbf{v}\in\mathcal{A}}\frac{L_{\mathbf{v}}}{2\sigma^2}\end{aligned} \tag{C.5}$$

C.2 *K*-GPNN-GP-GP

For the first-layer imaging covariates, the unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)$ with respect to $\theta_\ell(\mathbf{v})$ is

$$\begin{aligned}&\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial \theta_\ell(\mathbf{v})} = \\ &n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k\rho_k(\mathbf{v})h'(G_i(\mathbf{v}))X_i(\mathbf{v}')\gamma_\ell(\mathbf{v}', \mathbf{v})\big\}\Big] + \frac{1}{\sigma^2\tau_{\mathbf{v}}^2}\theta_\ell(\mathbf{v})\end{aligned} \tag{C.6}$$

where $G_i(\mathbf{v})$ is given by Eq. (6.21) and $F_i(j)$ is given by Eq. (6.12).

Similarly, the unbiased gradient estimator with respect to $\eta_0(\mathbf{v})$ is:

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial \eta_0(\mathbf{v})} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k\rho_k(\mathbf{v})h'(G_i(\mathbf{v}))\big\}\Big] + \eta_0(\mathbf{v}) \tag{C.7}$$

For the non-imaging covariate subnetwork, the unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)$ with respect to $\phi_{j,k}$ is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)}{\partial \phi_{j,k}} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k SM(F_i(j))\big\}\Big] \tag{C.8}$$

The unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)$ with respect to $w_{p,j}$ is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)}{\partial w_{p,j}} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k\phi_{j,k} SM'(F_i(j))Z_{i,j}\big\}\Big] \tag{C.9}$$

The unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)$ with respect to b_j is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|Z_i, y_i)}{\partial b_j} = n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\psi_k\phi_{j,k} SM'(F_i(j))\big\}\Big] \tag{C.10}$$

Similarly, the unbiased gradient estimator with respect to σ^2 is given be Eq. (C.5).

C.3 GPNN-Linear-GP Bimodality

For the first-layer imaging covariates, the unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)$ with respect to $\theta(\mathbf{v}, \mathbf{v}')$ is

$$\begin{aligned}&\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)}{\partial \theta(\mathbf{v}, \mathbf{v}')} = \\ &n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\big(\psi_k + \phi_{k,k}Q_{i,k}\big)\rho_k^{(1)}(\mathbf{v})h'(G_i^{(1)}(\mathbf{v}))X_i^{(1)}(\mathbf{v}')\big\}\Big] \\ &+ \frac{1}{\sigma^2\tau_{\mathbf{v}}^2}\theta(\mathbf{v}, \mathbf{v}')\end{aligned} \tag{C.11}$$

where $Q_{i,k}$ is given by Eq. (6.28), and

$$G_i^{(1)}(\mathbf{v}) = \eta_0^{(1)}(\mathbf{v}) + \sum_{\mathbf{v}' \in \mathcal{B}^{(1)}} X_i^{(1)}(\mathbf{v}')\theta(\mathbf{v}, \mathbf{v}') \tag{C.12}$$

Similarly, the unbiased gradient estimator with respect to $\eta_0^{(1)}(\mathbf{v})$ is:

$$\begin{aligned} &\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial \eta_0^{(1)}(\mathbf{v})} = \\ &n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{k=1}^{K}\big(\psi_k + \phi_{k,k}Q_{i,k}\big)\rho_k^{(1)}(\mathbf{v})h'(G_i^{(1)}(\mathbf{v}))\big\}\Big] \\ &+ \eta_0^{(1)}(\mathbf{v}) \end{aligned} \tag{C.13}$$

The unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)$ with respect to $\mu(\boldsymbol{u}, \boldsymbol{u}')$ is

$$\begin{aligned} &\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)}{\partial \mu(\boldsymbol{u}, \boldsymbol{u}')} = \\ &n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i - \hat{y}_i)\sum_{j=1}^{J}\big(\xi_j + \phi_{j,j}D_{i,j}\big)\rho_j^{(2)}(\boldsymbol{u})h'(G_i^{(2)}(\boldsymbol{u}))X_i^{(2)}(\boldsymbol{u}')\big\}\Big] \\ &+ \frac{1}{\sigma^2\iota^2(\boldsymbol{u})}\mu(\boldsymbol{u}, \boldsymbol{u}') \end{aligned} \tag{C.14}$$

where $D_{i,j}$ is given by Eq. (6.27), and

$$G_i^{(2)}(\boldsymbol{u}) = \eta_0^{(2)}(\boldsymbol{u}) + \sum_{\boldsymbol{u}' \in \mathcal{B}^{(2)}} X_i^{(2)}(\boldsymbol{u}')\mu(\boldsymbol{u}, \boldsymbol{u}') \tag{C.15}$$

Similarly, the unbiased gradient estimator with respect to $\eta_0^{(2)}(\boldsymbol{u})$ is:

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial \eta_0^{(2)}(\boldsymbol{u})} = \tag{C.16}$$
$$n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\big\{(y_i-\hat{y}_i)\sum_{j=1}^{J}\big(\xi_j+\phi_{j,j}D_{i,j}\big)\rho_j^{(2)}(\boldsymbol{u})h'(G_i^{(2)}(\boldsymbol{u}))\big\}\Big]+\eta_0^{(2)}(\boldsymbol{u})$$

Finally, the unbiased gradient estimator with respect to σ^2 is

$$\begin{aligned}\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i, y_i)}{\partial \sigma^2} &= n\frac{1}{m}\sum_{i=1}^{m}\Big\{\frac{1}{2\sigma^2}-\frac{(y_i-\hat{y}_i)^2}{2\sigma^4}\Big\} \\ &- \frac{1}{2\sigma^4}\sum_{\boldsymbol{u}\in\mathcal{A}^{(1)}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\boldsymbol{u}'\in\mathcal{B}^{(1)}}\theta(\mathbf{v},\mathbf{v}')^2+\frac{|\mathcal{A}^{(1)}||\mathcal{B}^{(1)}|}{2\sigma^2} \\ &- \frac{1}{2\sigma^4}\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\frac{1}{\iota^2(\boldsymbol{u})}\sum_{\boldsymbol{u}'\in\mathcal{B}^{(2)}}\mu(\boldsymbol{u},\boldsymbol{u}')^2+\frac{|\mathcal{A}^{(2)}||\mathcal{B}^{(2)}|}{2\sigma^2}\end{aligned} \tag{C.17}$$

C.4 GPNN-GP-Linear Bimodality

For the first-layer imaging covariates, the unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)$ with respect to $\theta_\ell(\mathbf{v})$ is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)}{\partial \theta_\ell(\mathbf{v})} = \tag{C.18}$$
$$n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\Big\{(y_i-\hat{y}_i)\Big(\psi_{\mathbf{v}}+\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\phi_{\mathbf{v},\boldsymbol{u}}Q_{i,\boldsymbol{u}}\Big)h'(G_i^{(1)}(\mathbf{v}))X_i^{(1)}(\mathbf{v}')\gamma_\ell(\mathbf{v}',\mathbf{v})^{(1)}\Big\}\Big]$$
$$+\frac{1}{\sigma^2\tau_{\mathbf{v}}^2}\theta_\ell(\mathbf{v})$$

where $Q_{i,\boldsymbol{u}}$ is given by Eq. (6.32), and

$$G_i^{(1)}(\mathbf{v}) = \eta_0^{(1)}(\mathbf{v}) + \sum_{\ell=1}^{L_{\mathbf{v}}}\theta_\ell(\mathbf{v})\sum_{\mathbf{v}'\in\mathcal{B}^{(1)}}X_i^{(1)}(\mathbf{v}')\gamma_\ell(\mathbf{v}',\mathbf{v})^{(1)} \tag{C.19}$$

Similarly, the unbiased gradient estimator with respect to $\eta_0^{(1)}(\mathbf{v})$ is:

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)}{\partial \eta_0^{(1)}(\mathbf{v})} = \tag{C.20}$$
$$n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\Big\{(y_i-\hat{y}_i)\Big(\psi_{\mathbf{v}}+\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\phi_{\mathbf{v},\boldsymbol{u}}Q_{i,\boldsymbol{u}}\Big)h'(G_i^{(1)}(\mathbf{v}))\Big\}\Big]+\eta_0^{(1)}(\mathbf{v})$$

The unbiased estimator of the MAP gradient of $C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)$ with respect to $\mu_k(\boldsymbol{u})$ is

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)}{\partial \mu_k(\boldsymbol{u})} = \tag{C.21}$$
$$n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\Big\{(y_i-\hat{y}_i)\Big(\xi_{\boldsymbol{u}}+\sum_{\mathbf{v}\in\mathcal{A}^{(1)}}\phi_{\mathbf{v},\boldsymbol{u}}D_{i,\mathbf{v}}\Big)h'(G_i^{(2)}(\boldsymbol{u}))X_i^{(2)}(\boldsymbol{u}')\gamma^{(2)}_{k\boldsymbol{u}'\boldsymbol{u}}\Big\}\Big]$$
$$+\frac{1}{\sigma^2\iota_{\boldsymbol{u}}^2}\mu_k(\boldsymbol{u}),$$

where $D_{i,\mathbf{v}}$ is given by Eq. (6.31), and

$$G_i^{(2)}(\boldsymbol{u}) = \eta_0^{(2)}(\boldsymbol{u}) + \sum_{k=1}^{K_{\boldsymbol{u}}}\mu_k(\boldsymbol{u})\sum_{\boldsymbol{u}'\in\mathcal{B}^{(2)}}X_i^{(2)}(\boldsymbol{u}')\gamma^{(2)}_{k\boldsymbol{u}'\boldsymbol{u}} \tag{C.22}$$

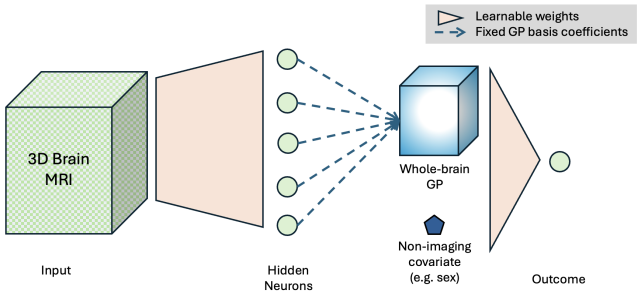
Similarly, the unbiased gradient estimator with respect to $\eta_0^{(2)}(\mathbf{v})$ is:

$$\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)}{\partial \eta_0^{(2)}(\mathbf{v})} = \tag{C.23}$$
$$n\frac{1}{m}\sum_{i=1}^{m}\Big[\frac{-1}{\sigma^2}\Big\{(y_i-\hat{y}_i)\Big(\xi_{\boldsymbol{u}}+\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\phi_{\mathbf{v},\boldsymbol{u}}D_{i,\mathbf{v}}\Big)h'(G_i^{(2)}(\boldsymbol{u}))\Big\}\Big]$$
$$+\eta_0^{(2)}(\boldsymbol{u})$$

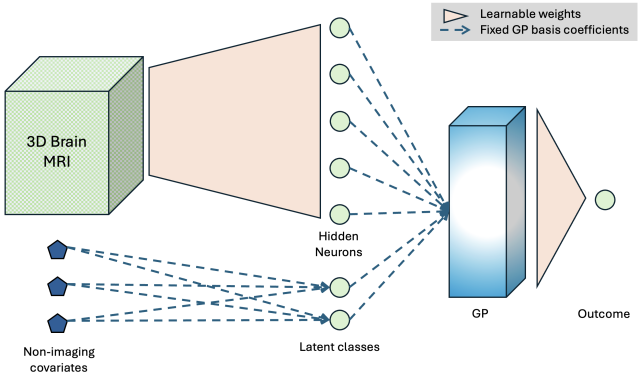
Finally, the unbiased gradient estimator with respect to σ^2 is

$$\begin{aligned}\frac{\partial C_{\arg\min}(\boldsymbol{\Theta}|X_i^{(1)}, X_i^{(2)}, y_i)}{\partial \sigma^2} &= n\frac{1}{m}\sum_{i=1}^{m}\left\{\frac{1}{2\sigma^2} - \frac{(y_i - \hat{y}_i)^2}{2\sigma^4}\right\} \\ &- \frac{1}{2\sigma^4}\sum_{\mathbf{v}\in\mathcal{A}^{(1)}}\frac{1}{\tau^2(\mathbf{v})}\sum_{\ell=1}^{L_{\mathbf{v}}}\theta_\ell(\mathbf{v})^2 + \sum_{\mathbf{v}\in\mathcal{A}}\frac{L_{\mathbf{v}}}{2\sigma^2} \\ &- \frac{1}{2\sigma^4}\sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\frac{1}{\iota_{\boldsymbol{u}}^2}\sum_{k=1}^{K_{\boldsymbol{u}}}\mu_k(\boldsymbol{u})^2 + \sum_{\boldsymbol{u}\in\mathcal{A}^{(2)}}\frac{K_{\boldsymbol{u}}}{2\sigma^2}\end{aligned} \tag{C.24}$$

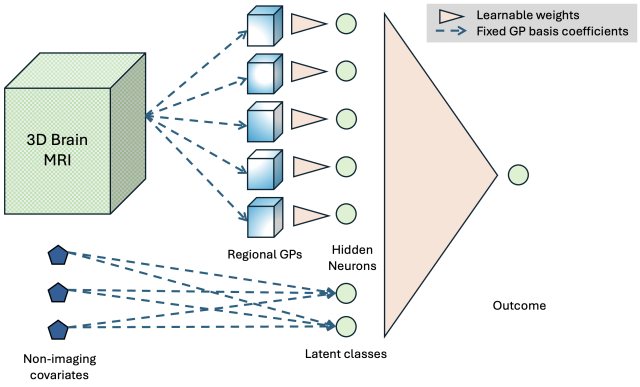
C.5 Additional figures



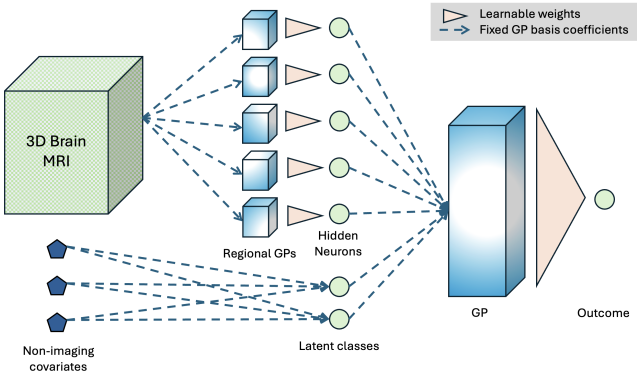
(a) GPNN-Linear-GP with one non-imaging covariate (proposed in Section 6.3.1.1)



(b) GPNN-Linear-GP with latent classes for non-imaging covariates (proposed in Section 6.3.2.1)

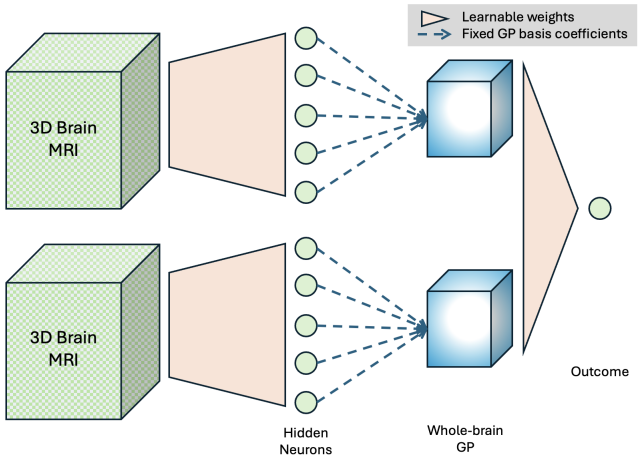


(c) GPNN-GP-Linear with latent classes for non-imaging covariates (proposed in Section 6.3.2.2)

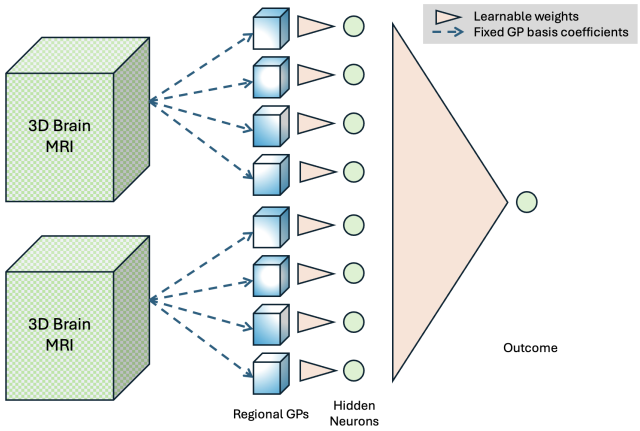


(d) GPNN-GP-GP with latent classes for non-imaging covariates (proposed in Section 6.3.2.3)

Figure C.1: Additional visualisation of model architectures with non-imaging covariates proposed in this study.



(a) GPNN-Linear-GP with bimodality imaging input (proposed in Section 6.3.3.1)



(b) GPNN-GP-Linear with bimodality imaging input (proposed in Section 6.3.3.2)

Figure C.2: Additional visualisation of model architectures with bimodality imaging input considered in the study.

References

Afyouni S, Smith SM, Nichols TE (2019) Effective degrees of freedom of the pearson's correlation coefficient under autocorrelation. NeuroImage 199:609–625

Aguirre GK, Detre JA, Zarahn E, Alsop DC (2002) Experimental design and the relative sensitivity of bold and perfusion fmri. Neuroimage 15(3):488–500

Ajith M, Spence JS, Chapman SB, Calhoun VD (2024) Multimodal predictive modeling: Scalable imaging informed approaches to predict future brain health. bioRxiv

Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, Gallacher J, Green J, Matthews P, Pell J, et al. (2012) Uk biobank: Current status and what it means for epidemiology. Health Policy and Technology 1(3):123–126

Azevedo FA, Carvalho LR, Grinberg LT, Farfel JM, Ferretti RE, Leite RE, Filho WJ, Lent R, Herculano-Houzel S (2009) Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. Journal of Comparative Neurology 513(5):532–541

Barut E, Fan J, Verhasselt A (2016) Conditional sure independence screening. Journal of the American Statistical Association 111(515):1266–1277

Beckmann CF, Jenkinson M, Smith SM (2003) General multilevel linear modeling for group analysis in FMRI. NeuroImage 20(2):1052–1063, doi:10.1016/S1053-8119(03)00435-X, URL http://linkinghub.elsevier.com/retrieve/pii/S105381190300435Xhttp://www.ncbi.nlm.nih.gov/pubmed/14568475https://linkinghub.elsevier.com/retrieve/pii/S105381190300435X

Belliveau JW, Kennedy DN, McKinstry RC, Buchbinder BR, Weisskoff RM, Cohen MS, Vevea J, Brady TJ, Rosen BR (1991) Functional mapping of the human visual cortex by magnetic resonance imaging. Science 254(5032):716–719

Berger H (1933) Über das elektrenkephalogramm des menschen. Archiv für Psychiatrie und Nervenkrankheiten 99(1):555–574

Bhattacharyya KB (2016) Godfrey newbold hounsfield (1919–2004): The man who revolutionized neuroimaging. Annals of Indian Academy of Neurology 19(4):448–450

Bianciardi M, Cerasa A, Patria F, Hagberg GE (2004) Evaluation of mixed effects in event-related fMRI studies: impact of first-level design and filtering. NeuroImage 22(3):1351–70, doi:10.1016/j.neuroimage.2004.02.039, URL `http://www.ncbi.nlm.nih.gov/pubmed/15219607`

Bollmann S, Puckett AM, Cunnington R, Barth M (2018) Serial correlations in single-subject fMRI with sub-second TR. NeuroImage 166(October 2017):152–166, doi:10.1016/j.neuroimage.2017.10.043, URL `https://doi.org/10.1016/j.neuroimage.2017.10.043`

Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. John Wiley & Sons

Bullmore E (2012) The future of functional MRI in clinical medicine. NeuroImage 62(2):1267–1271, doi:10.1016/j.neuroimage.2012.01.026, URL `http://dx.doi.org/10.1016/j.neuroimage.2012.01.026`

Bzdok D, Nichols TE, Smith SM (2019) Towards algorithmic analytics for large-scale datasets. Nature Machine Intelligence 1(7):296–306

Cai L, Chan JS, Yan JH, Peng K (2014) Brain plasticity and motor practice in cognitive aging. Frontiers in aging neuroscience 6:31

Candès EJ, Li X, Ma Y, Wright J (2011) Robust principal component analysis? Journal of the ACM (JACM) 58(3):1–37

Carlson NR, Birkett MA (2021) Physiology of behavior. Pearson Higher Ed

Casey BJ, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, Soules ME, Teslovich T, Dellarco DV, Garavan H, et al. (2018) The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. Developmental cognitive neuroscience 32:43–54

Chatfield C (2016) The analysis of time series: an introduction. CRC press

Cheang WK, Reinsel GC (2000) Bias Reduction of Autoregressive Estimates in Time Series Regression Model through Restricted Maximum Likelihood. Journal of the American Statistical Association 95(452):1173–1184, doi:10.1080/01621459.2000.10474318

Chen G, Saad ZS, Nath AR, Beauchamp MS, Cox RW (2012) FMRI group analysis combining effect estimates and their variances. NeuroImage 60(1):747–65, doi:10.1016/j.neuroimage.2011.12.060, URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3404516{&}tool=pmcentrez{&}rendertype=abstract

Chen JE, Polimeni JR, Bollmann S, Glover GH (2019) On the analysis of rapidly sampled fmri data. Neuroimage 188:807–820

Cleveland WS, Grosse E, Shyu WM (2017) Local regression models. In: Statistical models in S, Routledge, pp 309–376

Cochrane D, Orcutt GH (1949) Application of least squares regression to relationships containing auto-correlated error terms. Journal of the American statistical association 44(245):32–61

Cohen JW, Ramphal B, DeSerisy M, Zhao Y, Pagliaccio D, Colcombe S, Milham MP, Margolis AE (2024) Relative brain age is associated with socioeconomic status and anxiety/depression problems in youth. Developmental psychology 60(1):199

Cole JH, Franke K, Cherbuin N (2019) Quantification of the Biological Age of the Brain Using Neuroimaging, Springer International Publishing, Cham, pp 293–328. doi:10.1007/978-3-030-24970-0_19, URL https://doi.org/10.1007/978-3-030-24970-0_19

Corbin N, Todd N, Friston KJ, Callaghan MF (2018) Accurate modeling of temporal correlations in rapidly sampled fMRI time series. Human Brain Mapping 39(10):3884–3897, doi:10.1002/hbm.24218, URL https://onlinelibrary.wiley.com/doi/10.1002/hbm.24218

Daviet R, Aydogan G, Jagannathan K, Spilka N, Koellinger PD, Kranzler HR, Nave G, Wetherill RR (2022) Associations between alcohol consumption and gray and white matter volumes in the uk biobank. Nature communications 13(1):1175

Deary IJ, Corley J, Gow AJ, Harris SE, Houlihan LM, Marioni RE, Penke L, Rafnsson SB, Starr JM (2009) Age-associated cognitive decline. British medical bulletin 92(1):135–152

Denver P, McClean PL (2018) Distinguishing normal brain aging from the development of alzheimer's disease: inflammation, insulin signaling and cognition. Neural regeneration research 13(10):1719

Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, Anderson JS, Assaf M, Bookheimer SY, Dapretto M, et al. (2014) The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry 19(6):659–667

Edwards AD, Rueckert D, Smith SM, Abo Seada S, Alansary A, Almalbis J, Allsop J, Andersson J, Arichi T, Arulkumaran S, et al. (2022) The developing human connectome project neonatal data release. Frontiers in neuroscience 16:886772

Ernsting J, Winter NR, Leenings R, Sarink K, Barkhau CB, Fisch L, Emden D, Holstein V, Repple J, Grotegerd D, et al. (2023) From group-differences to single-subject probability: Conformal prediction-based uncertainty estimation for brain-age modeling. arXiv preprint arXiv:230205304

Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(5):849–911, doi:https://doi.org/10.1111/j.1467-9868.2008.00674.x, URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00674.x, https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2008.00674.x

Franke K, Gaser C (2012) Longitudinal changes in individual brainage in healthy aging, mild cognitive impairment, and alzheimer's disease. GeroPsych 25(4):235–245, doi:10.1024/1662-9647/a000074, URL https://doi.org/10.1024/1662-9647/a000074, https://doi.org/10.1024/1662-9647/a000074

Franke K, Ziegler G, Klöppel S, Gaser C, Initiative ADN, et al. (2010) Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: exploring the influence of various parameters. Neuroimage 50(3):883–892

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RS (1994) Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping 2(4):189–210

Friston KJ, Glaser DE, Henson RN, Kiebel S, Phillips C, Ashburner J (2002) Classical and bayesian inference in neuroimaging: applications. Neuroimage 16(2):484–512

Glover GH (2011) Overview of functional magnetic resonance imaging. Neurosurgery Clinics 22(2):133–139

Goldsmith J, Huang L, Crainiceanu CM (2014) Smooth scalar-on-image regression via spatial bayesian variable selection. Journal of Computational and Graphical Statistics 23(1):46–64

Gonzales MM, Garbarino VR, Pollet E, Palavicini JP, Kellogg DL, Kraig E, Orr ME, et al. (2022) Biological aging processes underlying cognitive decline and neurodegenerative disease. The Journal of clinical investigation 132(10)

Hannan E, McDougall A (1988) Regression procedures for arma estimation. Journal of the American Statistical Association 83(402):490–498

Hannan EJ, Rissanen J (1982) Recursive estimation of mixed autoregressive-moving average order. Biometrika 69(1):81–94

Harville DA (1974) Bayesian inference for variance components using only error contrasts. Biometrika 61(2):383–385

He J, Kang J (2021) Prior knowledge guided ultra-high dimensional variable screening with application to neuroimaging data. Statistica Sinica (SS-2020-0427), doi:10.5705/ss.202020.0427

He K, Kang J, Hong HG, Zhu J, Li Y, Lin H, Xu H, Li Y (2019) Covariance-insured screening. Computational statistics & data analysis 132:100–114

Heine L, Soddu A, Gómez F, Vanhaudenhuyse A, Tshibanda L, Thonnard M, Charland-Verville V, Kirsch M, Laureys S, Demertzi A (2012) Resting state networks and consciousness: alterations of multiple resting state network connectivity in physiological, pharmacological, and pathological consciousness states. Frontiers in psychology 3:295

Honea R, Swerdlow R, Vidoni E, Goodwin J, Burns J (2010) Reduced gray matter volume in normal adults with a maternal family history of alzheimer disease. Neurology 74(2):113–120

Hong HG, Kang J, Li Y (2018) Conditional screening for ultra-high dimensional covariates with survival outcomes. Lifetime data analysis 24:45–71

Jirsaraie R, Gorelik A, Gatavins M, Engemann D, Bogdan R, Barch D, Sotiras A (2023) A systematic review of multimodal brain age studies: Uncovering a divergence between model accuracy and utility. patterns, 4 (4), 100712

Kang J, Hong HG, Li Y (2017) Partition-based ultrahigh-dimensional variable screening. Biometrika 104(4):785–800

Kang J, Reich BJ, Staicu AM (2018) Scalar-on-image regression via the soft-thresholded gaussian process. Biometrika 105(1):165–184

Karama S, Ducharme S, Corley J, Chouinard-Decorte F, Starr JM, Wardlaw JM, Bastin ME, Deary IJ (2015) Cigarette smoking and thinning of the brain's cortex. Molecular psychiatry 20(6):778–785

Khlif MS, Egorova-Brumley N, Bird LJ, Werden E, Brodtmann A (2022) Cortical thinning 3 years after ischaemic stroke is associated with cognitive impairment and apoe $\varepsilon 4$. NeuroImage: Clinical 36:103200

Kievit RA, Frankenhuis WE, Waldorp LJ, Borsboom D (2013) Simpson's paradox in psychological science: a practical guide. Frontiers in psychology 4:513

Klee M, Leist AK, Veldsman M, Ranson JM, Llewellyn DJ (2023) Socioeconomic deprivation, genetic risk, and incident dementia. American Journal of Preventive Medicine 64(5):621–630

Lecei A, van Hulst BM, de Zeeuw P, van der Pluijm M, Rijks Y, Durston S (2019) Can we use neuroimaging data to differentiate between subgroups of children with adhd symptoms: A proof of concept study using latent class analysis of brain activity. NeuroImage: Clinical 21:101601

Ledoit O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis 88(2):365–411

Lichenstein SD, Roos C, Kohler R, Kiluk B, Carroll KM, Worhunsky PD, Witkiewitz K, Yip SW (2022) Identification and validation of distinct latent neurodevelopmental profiles in the adolescent brain and cognitive development study. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 7(4):352–361

Liem F, Varoquaux G, Kynast J, Beyer F, Kharabian Masouleh S, Huntenburg JM, Lampe L, Rahim M, Abraham A, Craddock RC, Riedel-Heller S, Luck T, Loeffler M, Schroeter ML, Witte AV, Villringer A, Margulies DS (2017a) Predicting brain-age from multimodal imaging data captures cognitive impairment. NeuroImage 148:179–188, doi:https://doi.org/10.1016/j.neuroimage.2016.11.005, URL `https://www.sciencedirect.com/science/article/pii/S1053811916306103`

Liem F, Varoquaux G, Kynast J, Beyer F, Masouleh SK, Huntenburg JM, Lampe L, Rahim M, Abraham A, Craddock RC, et al. (2017b) Predicting brain-age from multimodal imaging data captures cognitive impairment. Neuroimage 148:179–188

Lin Z, Si Y, Kang J (2024) Latent subgroup identification in image-on-scalar regression. The annals of applied statistics 18(1):468

Marchini JL, Ripley BD (2000) A new statistical approach to detecting significant activation in functional mri. NeuroImage 12(4):366–380

Martz ME, Cope LM, Hardee JE, Brislin SJ, Weigard A, Zucker RA, Heitzeg MM (2021) Subtypes of inhibitory and reward activation associated with substance use variation in adolescence: A latent profile analysis of brain imaging data. Cognitive, Affective, & Behavioral Neuroscience 21(5):1101–1114

Marzola P, Melzer T, Pavesi E, Gil-Mohapel J, Brocardo PS (2023) Exploring the role of neuroplasticity in development, aging, and neurodegeneration. Brain Sciences 13(12):1610

Michon KJ, Khammash D, Simmonite M, Hamlin AM, Polk TA (2022) Person-specific and precision neuroimaging: Current methods and future directions. NeuroImage 263(July):119589, doi:10.1016/j.neuroimage.2022.119589, URL `https://doi.org/10.1016/j.neuroimage.2022.119589`

Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JL, et al. (2016) Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nature neuroscience 19(11):1523–1536

Monti MM (2011) Statistical analysis of fmri time-series: a critical review of the glm approach. Frontiers in human neuroscience 5:28

Mumford JA, Nichols TE (2009) Simple group fMRI modeling and inference. NeuroImage 47(4):1469–75, doi:10.1016/j.neuroimage.2009.05.034, URL `http://www.ncbi.nlm.nih.gov/pubmed/19463958`

Neal RM, et al. (2011) Mcmc using hamiltonian dynamics. Handbook of markov chain monte carlo 2(11):2

Niu X, Zhang F, Kounios J, Liang H (2020) Improved prediction of brain age using multimodal neuroimaging data. Human brain mapping 41(6):1626–1643

Nooner KB, Colcombe SJ, Tobe RH, Mennes M, Benedict MM, Moreno AL, Panek LJ, Brown S, Zavitz ST, Li Q, et al. (2012) The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry. Frontiers in neuroscience 6:152

Olszowy W, Aston J, Rua C, Williams GB (2019) Accurate autocorrelation modeling substantially improves fmri reliability. Nature communications 10(1):1220

Palma M, Tavakoli S, Brettschneider J, Nichols TE, Initiative ADN, et al. (2020) Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression. Neuroimage 219:116938

Penny WD, Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE (2011) Statistical parametric mapping: the analysis of functional brain images. Elsevier

Robbins H, Monro S (1951) A stochastic approximation method. The annals of mathematical statistics pp 400–407

Robert CP, Casella G, Casella G (1999) Monte Carlo statistical methods, vol 2. Springer

Röhr S, Gibson RH, Alpass FM (2024) Higher socioeconomic deprivation in areas predicts cognitive decline in new zealanders without cognitive impairment. Scientific Reports 14(1):28314

Saad ZS, Chen G, Reynolds RC, Christidis PP, Hammett KR, Bellgowan PS, Cox RW (2006) Functional imaging analysis contest (fiac) analysis according to afni and suma. Human brain mapping 27(5):417–424

Scheinost D, Noble S, Horien C, Greene AS, Lake EM, Salehi M, Gao S, Shen X, O'Connor D, Barron DS, et al. (2019) Ten simple rules for predictive modeling of individual differences in neuroimaging. NeuroImage 193:35–45

Schuff N, Woerner N, Boreta L, Kornfield T, Shaw L, Trojanowski J, Thompson P, Jack Jr C, Weiner M, Initiative ADN (2009) Mri of hippocampal volume loss in early alzheimer's disease in relation to apoe genotype and biomarkers. Brain 132(4):1067–1077

Smith M, Fahrmeir L (2007) Spatial bayesian variable selection with application to functional magnetic resonance imaging. Journal of the American Statistical Association 102(478):417–431

Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, Filippini N, Watkins KE, Toro R, Laird AR, et al. (2009) Correspondence of the brain's functional architecture during activation and rest. Proceedings of the national academy of sciences 106(31):13040–13045

Smith SM, Vidaurre D, Alfaro-Almagro F, Nichols TE, Miller KL (2019) Estimation of brain age delta from brain imaging. Neuroimage 200:528–539

Spindler C, Trautmann S, Alexander N, Bröning S, Bartscher S, Stuppe M, Muehlhan M (2021) Meta-analysis of grey matter changes and their behavioral characterization in patients with alcohol use disorder. Scientific reports 11(1):5238

Tan C, Ma S, Dai YH, Qian Y (2016) Barzilai-borwein step size for stochastic gradient descent. Advances in neural information processing systems 29

Ten Kate M, Ingala S, Schwarz AJ, Fox NC, Chételat G, van Berckel BN, Ewers M, Foley C, Gispert JD, Hill D, et al. (2018) Secondary prevention of alzheimer's dementia: neuroimaging contributions. Alzheimer's research & therapy 10:1–21

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 58(1):267–288

Tziortzi AC, Searle GE, Tzimopoulou S, Salinas C, Beaver JD, Jenkinson M, Laruelle M, Rabiner EA, Gunn RN (2011) Imaging dopamine receptors in humans with [11c]-(+)-phno: Dissection of d3 signal and anatomy. NeuroImage 54(1):264–277, doi:https://doi.org/10.1016/j.neuroimage.2010.06.044, URL `https://www.sciencedirect.com/science/article/pii/S1053811910008888`

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. NeuroImage 15(1):273–289, doi:https://doi.org/10.1006/nimg.2001.0978, URL `https://www.sciencedirect.com/science/article/pii/S1053811901909784`

Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, Consortium WMH, et al. (2013a) The wu-minn human connectome project: an overview. Neuroimage 80:62–79

Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, Consortium WMH, et al. (2013b) The wu-minn human connectome project: an overview. Neuroimage 80:62–79

Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine 5(1):48

Vasilkovska T (2023) The physiological basis of the fmri bold response. BioRender URL https://app.biorender.com/profile/TamaraVasilkovska/templates/64dcc3ced89f68954a23d7bc

Wang H (2009) Forward regression for ultra-high dimensional variable screening. Journal of the American Statistical Association 104(488):1512–1524

Wang X, Leng C (2016) High dimensional ordinary least squares projection for screening variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78(3):589–611, doi:https://doi.org/10.1111/rssb.12127, URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12127, https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12127

Welling M, Teh YW (2011) Bayesian learning via stochastic gradient langevin dynamics. In: Proceedings of the 28th international conference on machine learning (ICML-11), Citeseer, pp 681–688

Westfall J, Nichols TE, Yarkoni T (2016) Fixing the stimulus-as-fixed-effect fallacy in task fmri. Wellcome open research 1

Woolrich MW, Ripley BD, Brady M, Smith SM (2001) Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage 14(6):1370–86, doi:10.1006/nimg.2001.0931, URL http://www.ncbi.nlm.nih.gov/pubmed/11707093

Woolrich MW, Behrens TE, Beckmann CF, Jenkinson M, Smith SM (2004) Multilevel linear modelling for FMRI group analysis using Bayesian inference. NeuroImage 21(4):1732–47, doi:10.1016/j.neuroimage.2003.12.023, URL http://www.ncbi.nlm.nih.gov/pubmed/15050594

Worsley K, Friston K (1995) Analysis of fMRI Time-Series Revisited—Again. NeuroImage 2(3):173–181, doi:10.1006/nimg.1995.1023, URL https://linkinghub.elsevier.com/retrieve/pii/S1053811985710233

Worsley KJ (2000) Comment. Journal of the American Statistical Association 95(451):711–716, doi:10.1080/01621459.2000.10474258, URL https://doi.org/10.1080/01621459.2000.10474258, https://doi.org/10.1080/01621459.2000.10474258

Worsley KJ (2001) Statistical analysis of activation images. Functional MRI: An introduction to methods 14(1):251–70

Worsley KJ, Liao CH, Aston J, Petre V, Duncan G, Morales F, Evans AC (2002) A general statistical analysis for fmri data. Neuroimage 15(1):1–15

Yankner BA, Lu T, Loerch P (2008) The aging brain. Annual review of pathology 3(1):41–66

Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, Ourselin S, Initiative ADN, et al. (2013) Accurate multimodal probabilistic prediction of conversion to alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clinical 2:735–745

Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American statistical association 101(476):1418–1429

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology 67(2):301–320