Multi-level Modelling and Spatial Inference for Large-Scale Neuroimaging Data

Thomas J. Maullin-Sapey

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctorate of Philosophy

Hilary 2022

Abstract

Recent developments in data sharing and availability provide a vast new window of opportunity for large-sample fMRI analysis. The increased statistical power resulting from larger sample sizes allows researchers to model previously inaccessible phenomena such as the complex grouping structures present in large multi-level datasets (e.g. genetic, familial, geographical) and the spatial variation present in shape, size and locale of excursion sets (identified regions of activation). This thesis is comprised of two halves, which investigate each of these features in turn.

In the first half, we employ the Linear Mixed Model (LMM) to detect and account for grouping and covariance structure present in large experimental designs. At present, the existing fMRI LMM tools are neither scalable nor fully exploit the computational speed-ups afforded by vectorisation of operations over voxels. For this reason, such tools cannot be employed in the large-sample setting. To address this issue, we derive new expressions for LMM estimation and inference and illustrate their usage to perform computationally efficient large-scale fMRI LMM analyses. Our methods are validated empirically via simulations and illustrated using data drawn from the UK Biobank (Allen et al. (2012)).

In the second half, we extend a method proposed by Sommerfeld et al. (2018) to provide confidence bounds for the intersection or 'conjunction' of excursion sets that have been acquired across the same spatial region but under different study conditions. Such 'conjunctions' are of natural interest as they correspond to the question "Where was activation observed under all study conditions?". Our method provides, with a desired confidence, sub- and super-sets of the 'conjunction set' without any assumptions on the dependence between the different conditions. Extensive simulations are provided assessing the method's performance, and a demonstration of the method is given using data drawn from the Human Connectome Project dataset (Van Essen et al. (2013)).

Declarations

I declare that this thesis is entirely my own work and, except where otherwise stated, describes my own research. The content of this thesis is original and has not been submitted in whole or in part for consideration for any other degree or qualification at any other university or institution. Material that is the outcome of work done in collaboration is indicated as such in the text.

- The work presented in Chapter 3 has been published in Statistics and Computing under the title "Fisher Scoring for crossed factor linear mixed models" (c.f. Maullin-Sapey and Nichols (2021)).
- The work presented in Chapter 4 shall be submitted for publication shortly. Aspects of this work were previously submitted as a poster for the Organisation for Human Brain Mapping (OHBM) 2020 conference.
- The work presented in Chapter 5 is currently in submission. Aspects of this work have also been submitted as a poster for the Organisation for Human Brain Mapping (OHBM) 2022 conference.

Thomas J. Maullin-Sapey

January 2022

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Thomas E. Nichols, whose continual guidance and support throughout the PhD process cannot be understated. I also wish to express strong gratitude towards Professor Armin Schwartzman, whose involvement and feedback was crucial to this work's fifth chapter. It is in no small part due to the input of both Tom and Armin that this work is something that I personally feel I can be proud of.

I would also like to thank Professor Jonathan Emberson and Dr Jerome Kelleher for the invaluable feedback they both provided during the transfer of status and confirmation of status assessment process. Thanks must also be given to several members, past and present, of Tom's research group. In particular, I would like to give special thanks to Simon Schwab, Soroosh Afyouni, Habib Ganjgahi, Sam Davenport, Petya Kindalova, Han Bossier, Jessie Liu, Marco Palma, Yifan Yu and Alex Bowing. I would also like to extend this gratitude to Fabian J.E. Telschow and Darren Scott.

Finally, I must thank my family; my mum, my dad and, reluctantly, Ed. I would also like to thank my friends back home, particularly Brad Field, Liam Darkins and Jamie Hanner. Last but not least, I would like to thank Danielle Gilbert for both her personal support and her time in proofreading much of this thesis. Dani, no one has shown more investment in me, personally, academically or otherwise. Thank you.

This work was partially supported by the NIH under Grant [R01EB026859] (Chapters 3-5: TMS, TN, Chapter 5: AS) and the Wellcome Trust award [100309/Z/12/Z] (Chapters 3-4: TN).

Contents

List of Figures

Chapter 1

Introduction

The field of functional Magnetic Resonance Imaging (fMRI) has recently seen a drastic improvement in terms of the volume of data collected and shared among the research community. For many years, sample sizes in fMRI were typically limited to fewer than 60 subjects (Yeung (2018), Szucs and Ioannidis (2020)). However, with recent technological and practical advancements, researchers now have access to datasets containing sample sizes on the scale of the tens of thousands. The substantial increase in data accumulation and availability in fMRI can be attributed to a number of factors including the availability of large-scale cohorts (e.g. the Human Connectome Project, Van Essen et al. (2013), and UK Biobank, Allen et al. (2012)), open-access data platforms (e.g. OpenfMRI, Poldrack et al. (2013), and Neurovault, Gorgolewski et al. (2015)) and tools for distributed data analysis (e.g. the COINSTAC toolbox, Plis et al. (2016)) (see Section 2.1 for further discussion).

The data presented to fMRI researchers is large not only in terms of sample size (e.g. number of subjects) but also in terms of the amount of data collected per subject. At present, in a standard fMRI analysis, a single image typically houses around 10^6 voxels (3D resolution elements, analogous to pixels), approximately a third of which are occupied by brain tissue. This means that the average fMRI image consumes approximately 3-4MB of storage and researchers must analyze data corresponding to approximately 3.3×10^5 'spatial locations' in the brain. With the continuous

improvement of spatial resolution in fMRI data, many researchers have suggested this number is likely to increase (Feinberg and Yacoub (2012), T. Vu et al. (2017)).

It is safe to say that the field of fMRI neuroimaging has firmly entered the "era of *big data*" (Li et al. (2019)). For many researchers, the newfound plenitude of fMRI data is being viewed as a "revolution" (Landhuis (2017)) which promises to provide "greater knowledge about fundamental brain processes [suggesting] new and testable hypotheses that lead to novel experimentation" (Van Horn and Toga (2014)). However, with this deluge of data, many researchers have expressed concern that whilst data is becoming more readily available, it is also becoming more multi-faceted, complex and computationally burdensome to analyze (Bzdok and Yeo (2017)). One recently published opinion piece describes the field of neuroscience as "drowning in a flood of information" in which "all sense of global understanding is in acute danger of getting washed away" (Frégnac (2017)). Alternatively, Webb-Vargas et al. (2017) describe the present state of fMRI analysis as facing an "emerging critical need for Big Data tools and methods"[1].

Such claims are not without foundation, as analysis of large-sample fMRI data is challenging, both in theory and practice. In a typical fMRI analysis, every voxel in an image is first analyzed independently, conventionally using standard regression techniques (c.f. Sections 2.3.1 and 2.3.2). Following this, results are synthesized and presented as images representing brain function. This process involves constructing model matrices and estimating parameters separately for every voxel in the analysis. In all but the simplest settings, adopting such an approach constitutes a significant practical obstacle in terms of both storage and computation time. Previously, this has not been regarded as a particularly significant issue for fMRI research, as although hundreds of thousands of models must be constructed, each model individually tended to be quite small. However, as sample sizes grow, such models increase in size, and the accumulated computational costs become substantial (Li et al. (2019), Bzdok et al. (2019)).

[1]Note the reverential capitalization.

In addition, it is becoming increasingly common that the large-sample datasets available for fMRI analysis have been aggregated from several different sources (i.e. subjects, locations, study protocols, family units) (Horien et al. (2020)). A consequence of this multiplicity of sources is that such datasets are often imbued with complex dependencies, grouping factors and covariance structures. For example, the Adolescent Brain Cognitive Development (ABCD) and UK Biobank (UKB) datasets, which contain imaging data from $10,000$ and $30,000$ subjects respectively, both possess a multi-level covariance structure induced through a repeated measures experimental design where each subject was scanned multiple times (Casey et al. (2018), Allen et al. (2012)). Similarly, the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) cohort, which contains imaging data from tens of thousands of subjects, also possesses a multi-level covariance structure due to its pooling of data sourced from multiple institutions (Bearden and Thompson (2017)).

Many of the methods conventionally employed for drawing inference on large-sample fMRI data (e.g. Ordinary Least Squares regression, t-tests and F-tests) do not account for the presence of covariance structure of this kind and, consequently, can produce erroneous inference. One of the standard, most widely-adopted approaches to modelling such structure is the Linear Mixed Model (LMM), which is exceptionally adept at handling the analysis of longitudinal, heterogeneous or unbalanced clustered datasets (Laird and Ware (1982), Pinheiro and Bates (2009), Demidenko (2013)). While the LMM is widely agreed to be one of the most powerful and flexible solutions to modelling covariance structure, it is also known to be one of the most computationally burdensome. No closed-form solutions are known for the parameters of the LMM, and, as a result, intensive optimization techniques must often be employed to perform numerical estimation (Laird et al. (1987), Bates et al. (2015)). In addition, there is no widely agreed consensus on how to perform null-hypothesis inference on the LMM, and many of the favoured approaches to LMM inference make strong or restrictive assumptions and often require further time-consuming calculation (Verbeke and Molenberghs (2001), West et al. (2014)).

The conventional tools available for LMM analysis of fMRI data are not purposed for big-data applications and, as a result, accrue heavy computational costs as sample sizes increase into the hundreds. Furthermore, such tools also do not account for variability in the patterns of missing data near the edge of the brain, which can cause shrinkage of the final analysis mask if not appropriately accounted for (c.f. Section 2.3.3 and Gebregziabher et al. (2017)). To counter these issues, the first half of this thesis focuses on the development of new methodology for LMM parameter estimation and inference. An implementation of the methodology developed is provided as the Big Linear Mixed Modelling (BLMM) toolbox: a Python package for large-sample fMRI analysis that accounts for missing data near cortical boundaries.

Following parameter estimation and inference, results must be synthesized across voxels and regions of activation must be identified. To combine the results across voxels, standard fMRI analyses employ a multiple comparisons correction to account for the number of tests being performed (c.f. Section 2.3.4). Many common approaches to performing such corrections achieve this goal by leveraging topological and geometric properties of the results obtained (e.g. peak-wise inference, cluster-wise inference). While such inference methods are a mainstay of fMRI analysis and capitalize on spatial characteristics of the statistical results, the established tools can only provide p-values and no notion of spatial uncertainty of cluster location. However, with the increased statistical power afforded by large samples, new methods for assessing the reliability of spatial localization are becoming viable for fMRI analysis.

One such method is that of confidence regions. Proposed by Sommerfeld et al. (2018), confidence regions were developed to provide upper and lower bounds on excursion sets of random functions (e.g. regions of activation in fMRI analysis, c.f. Section 2.3.5). The original text, alongside the subsequent work of Bowring et al. (2019) and Bowring et al. (2021), provided substantive evidence that for sample sizes as small as 80 subjects, the confidence regions method may be employed to obtain, with a desired confidence, sub- and super-sets of regions of activation in an fMRI analysis. While such sample sizes were previously unattainable in practice, the current trends in big data fMRI acquisition open the door to the usage and development of

the theory of confidence regions for fMRI analysis. Such theory provides a powerful tool for assessing how reliably the statistically significant regions identified in an fMRI analysis have been localized to a particular anatomical location.

However, the theory of confidence regions does not currently offer any insight into how analysis results may be compared across study conditions. In fact, despite growing literature, in general, there is little published concerning the comparison of processes that have been sampled across the same spatial region but which reflect different study conditions. As a result, it is currently common practice for researchers to assess the 'overlap' of images subjectively via direct rudimentary visual inspection. In the second half of this thesis, we aim to address this issue by extending the theory of confidence regions to provide quantitative assessment of questions regarding 'overlapping' excursion sets. To be specific, given a set of asymptotically Gaussian random fields, each corresponding to a sample acquired for a different study condition, we aim to provide confidence statements about the intersection of the excursion sets across all fields. Such regions are of strong empirical value as they may be interpreted as the regions of the brain that are active under *all* of the study conditions. The method is verified by extensive simulations, extended to situations involving other combinations of excursion sets (e.g. unions, complements) and demonstrated using a task-fMRI dataset to identify brain regions with activation common to four variants of a working memory task.

In summary, the primary contributions provided by this thesis are:

1. We present and derive new expressions for the extension of an algorithm classically used for single-factor LMM parameter estimation, Fisher Scoring, to multiple, crossed-factor designs. We compare and contrast several variants of the proposed method via simulation and illustrate its extension to models containing arbitrary covariance structures.

2. We provide a new method for LMM Satterthwaite degrees of freedom estimation based on analytical results, which does not require iterative gradient estimation.

Our proposed method is compared to the existing approaches to establish correctness and demonstrate time efficiency.

3. We demonstrate how our proposed methodology may be employed in the large-sample fMRI analysis setting by developing BLMM, a freely available Python tool for large-sample fMRI LMM analysis. In developing BLMM, we also illustrate how voxel-wise missingness near cortical boundaries may be handled in a computationally efficient manner for large-sample analyses.

4. We develop and illustrate a theory of confidence regions for the comparison of multiple excursion sets. Given a set of spatially varying linear models with summary statistics corresponding to asymptotically Gaussian random fields, the proposed theory provides confidence statements about intersections, unions, and complements of the contrast excursion sets taken across all fields.

The remainder of this thesis is organized as follows:

- **Chapter 2: Background**.
  This chapter is partitioned into three sections. The first section, Section 2.1, provides background and context on the study of task-based fMRI. In the second section, Section 2.2, a primer on the statistical analysis methods which will be relevant to this work is provided. The third and final section, Section 2.3, combines the content of the previous two sections to outline how the aforementioned statistical analysis methods are employed in the context of fMRI.

- **Chapter 3: Fisher Scoring for Crossed Factor Linear Mixed Models**.
  In this chapter, we shift attention from fMRI analysis and instead focus solely on the LMM. Here, we derive and provide several new expressions which may be utilized for LMM parameter estimation and inference. The methodology proposed is verified extensively via simulation and univariate real-data examples based on the Longitudinal Evaluation of School Change and Performance (LESCP) dataset (Turnbull et al. (1999)) and the Wu-Minn Human Connectome Project (HCP) cohort (Van Essen et al. (2013)).

- **Chapter 4: Parallelised Computing for Big Linear Mixed Models**.

  In this chapter, we apply the methodology developed in Chapter 3 to perform computationally efficient large-sample fMRI LMM analysis. Our proposed approach is implemented as a toolbox in Python, assessed empirically via simulation and demonstrated using repeated measures data drawn from the UK Biobank (Allen et al. (2012)).

- **Chapter 5: Spatial Confidence Regions for Combinations of Excursion Sets in Image Analysis**.

  In this chapter, we develop a theory of spatial confidence regions for intersections and unions of excursion sets. We first develop a general theory relating confidence statements for excursion sets to an exceedance statement for a well-defined random variable. Following this, we illustrate the application of our results using a Wild t-bootstrapping procedure for a set of spatially varying linear regression models. The material of this chapter is presented in a general context but employs neuroimaging data drawn from the HCP cohort to demonstrate the method.

- **Chapter 6: Concluding Remarks.**

  In this chapter, the contributions of this work are summarized and concluding remarks are provided.

Chapter 2

Background

In this chapter, we provide background material necessary to follow the text of Chapters 3-5. Next, Section 2.1 gives a brief introduction to the history, motivation and science behind fMRI. Following this, Section 2.2 details the univariate statistical models we shall employ throughout this document. Finally, Section 2.3 combines elements of both of the preceding sections to outline the spatial statistics used in fMRI, as well as the topical issues in fMRI analysis that motivate the content of this thesis.

This chapter is intended as a brief overview of relevant fMRI statistics but is not intended as a comprehensive introduction. Further information on the fMRI analysis discussed in Sections 2.1 and 2.3 may be found, for example, in Poldrack et al. (2011) and Jenkinson and Chappell (2018). More detail on the broader statistical methodology discussed in Sections 2.2 and 2.3 may be found in Adler and Taylor (2009), Sommerfeld et al. (2018), Demidenko (2013) and West et al. (2014).

2.1 Neuroimaging Preliminaries

In this section, we provide a short introduction to the field of fMRI. First, Section 2.1.1 provides contextual background on the field of neuroimaging. Following this, Section 2.1.2 describes the physiological processes that underlie the acquisition of fMRI data. Next, Section 2.1.3 provides a brief overview of the experimental designs

typically employed for task-fMRI studies. Finally, Section 2.1.4 outlines some of the common preprocessing steps applied to fMRI data prior to a statistical analysis.

2.1.1 The Development of Neuroscience

Historically, the field of neuroscience has developed in tandem with the advancement of technology. Much of our present understanding of both neuroscience and technology has only arisen in the last two centuries or so (Lorusso et al. (2018), Bear et al. (2020)), with some authors suggesting the formation of neuroscience is as recent as the last seven decades (Brown (2019), Cowan et al. (2000)).

Prior to the establishment of neuroscience as a scientific discipline, for many millennia, it was commonly believed that human thought and emotion emanated from the heart rather than the brain. This belief was particularly prominent in ancient Egypt, China and Greece (c.f. Santoro et al. (2009), Lee (2010)), and is still reflected in everyday language today (e.g. 'learn by heart', 'take to heart', etc.). It was not until the development of the printing press that, in the 16^{th} century, material regarding human anatomy became widely available (c.f. Fig. 2.1), and the notion of the heart being central to thought and emotion came into dispute (Van Laere (1993)).

Following the 16^{th} century, strong debate arose surrounding the location of cognitive function within the human anatomy (Payne (2005)). In the 18^{th} century, this debate came to a head when another advancement in technology occurred: the discovery of electricity. With this discovery came Galvanism, or the inducement of muscle contractions via electrical current stimulation, which was often seen as public entertainment. A wave of investigations into the effect of electricity on the nervous system and new theoretical analogies likening the nervous system to electrical machines (typically telegraph systems, c.f. Smee (1850), Helmholtz (1875)) soon followed. As knowledge of new surgical and electrical experiments spread, the view of the brain as the seat of the mind became solidified as scientific fact. With this shift in thought, a new question quickly entered the scientific discussion; "Are certain functions localised

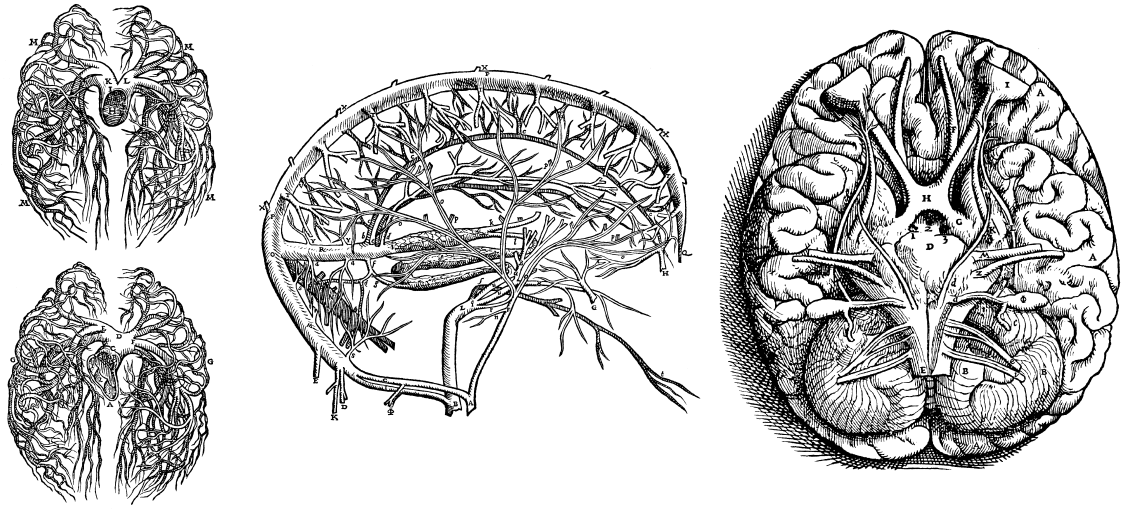

Figure 2.1: Early illustrations of human brain anatomy from 'De Humani Corporis Fabrica' (1543) by Andreas Vesalius, reproduced in Garrison (2015). Modified and reprinted courtesy of Professor Daniel H. Garrison, Department of Classics, Northwestern University.

to specific regions of the brain?" (Swazey (1970)). Although this question was central to the emerging discourse, little could be done to further the debate until new innovations occurred in the realm of technology.

Theoretical breakthroughs came thick and fast with technological advancement throughout the 20^{th} century. As scientists increasingly likened the brain to devices resembling modern computers (Meyer (1911)), concepts such as neurotransmitters and neuronal coupling gained traction (see Section 2.1.2) (Dale (1914)) and the neuron doctrine, the conceptual framing of the nervous system as composed of discrete individual cells, found substantial popularity (Ramón y Cajal (1911))). With the development of electron microscopy in the 1950s, the neuron doctrine was to be established as scientific fact, forming the foundation of modern-day investigation into the sub-atomic structure and properties of neurons (Cowan et al. (2000)). Following these developments, the abstract idea of studying localised brain function directly in a non-invasive, ethical manner shortly became a tangible reality through the advent of medical imaging of the brain.

The concept of "neuroimaging" first appeared in the early 1880s when Italian

physiologist Angelo Mosso attempted to weigh the blood flowing to the brain using a device he referred to as the 'human circulation balance' (Sandrone et al. (2013)). It is unclear how successful Mosso's device was, but his idea paved the way for rapid development of imaging technology. Toward the early 1920s, several methods for imaging the brain were conceived, including Pneumoencephalography (PEG), a highly unethical procedure involving X-rays (Dandy (1919)), and the more successful Electroencephalography (EEG), a procedure still used today which records electrical activity on the scalp (Berger (1929)). EEG provides a powerful window into brain function, but for a long time could not be used to construct the "three-dimensional images" of the brain that we are familiar with today.

In the 1960s, a new form of neuroimaging was developed: the Computerised Axial Tomography (CAT) scan. The transformative development of the CAT scan allowed images of the *brain anatomy* of live subjects to be created in a non-invasive, ethical manner by measuring the x-ray attenuations of different tissue types (Bhattacharyya (2016)). Soon after, in the 1970s, Positron Emission Tomography (PET) was invented, which enabled images of *brain function* to be created by measuring gamma-ray photons indirectly emitted by decaying radioactive tracers. PET imaging fundamentally changed how scientific investigation of brain function was performed but was too invasive, expensive, slow, and low-resolution for widespread adoption at the time (Carlson (2013)).

In May 1991, a revolutionary breakthrough occurred when Dr John Belliveau and colleagues at the Massachusetts General Hospital-NMR Center performed the first functional Magnetic Resonance Imaging (fMRI) experiment (Belliveau et al. (1991)). FMRI measures blood oxygenation in the brain by drawing on the electrochemical properties of neurotransmitters (c.f. Section 2.1.2). Much like PET, fMRI produces images of *brain function* in live subjects in a non-invasive and ethical manner. However, unlike PET, fMRI is widely viewed as a more cost-effective, accessible, and higher resolution option.

Following Belliveau's experiment, the field of fMRI neuroimaging quickly exploded, with an exponential increase in the number of neuroimaging-related publica-

tions observed every year since 1991 (c.f. PubMed statistics (2022)). With fMRI's increased popularity has come many key discoveries, such as the conception of the default mode network (Raichle et al. (2001)), the discovery of bottom-up hierarchical organisation of visual processing (Courtney and Ungerleider (1997)), and the involvement of the cortico-basal ganglia-thalamocortical loop in memory, action-selection and reward (Milardi et al. (2019)). Although many other imaging modalities have been invented since the conception of fMRI (e.g. Arterial Spin Labelling, Koretsky (2012), Vascular-Space-Occupancy, Lu et al. (2013)), none have reached the extensive usage and popularity of fMRI.

At present, the field of fMRI neuroimaging is experiencing some of the fastest growth ever seen, with over $40,000$ papers on fMRI published last year alone (2021, at time of writing, see PubMed statistics (2022)). As happened with the development of the printing press in the 16^{th} century, the discovery of electricity in the 18^{th} century, and the numerous computational advances in the 20^{th} century, the study of brain function is again experiencing tremendous propulsion forward due to technological advances - this time in data storage and availability.

Dubbed the 'big data era' of neuroscience, at present more data is being collected, shared and made accessible to researchers than ever before (Li et al. (2019), Bzdok and Yeo (2017)). While not as groundbreaking or enticing as the invention of electricity or the printing press, recent improvements in data collection, storage and analysis have had an incredible impact on the field of fMRI. Researchers can now access publicly shared data openly using free platforms, such as openfMRI (Poldrack et al. (2013)) and Neurovault (Gorgolewski et al. (2015)). Tools are now available for aggregating data from different sources, which cannot be shared, in a privacy-protected manner (e.g. the COINSTAC toolbox, Plis et al. (2016)). The temporal and spatial resolution of the data being acquired is continually increasing (c.f. Bandettini (2012), Uğurbil et al. (2013)). And, whereas in previous years, the average sample size in a typical fMRI analysis has been around 12-25 subjects (Szucs and Ioannidis (2020)), researchers now have access to massive large-scale datasets containing tens of thousands of images, such as the UK Biobank (Allen et al. (2012)),

the Human Connectome Project (Van Essen et al. (2013)) and the Adolescent Brain Cognitive Development study (Casey et al. (2018)).

These technological developments share a common goal: to make a higher quantity of data easily accessible to more researchers. However, as discussed in Chapter 1, with this overwhelming flood of data often comes increased complexity of analysis, computational overheads and a renewed demand for improved analysis techniques. In an effort to contribute towards meeting such demand, this thesis focuses on developing methodology for 'big data' fMRI analyses. As such, the following sections now provide a brief overview of the specifics of fMRI physiology and data acquisition.

2.1.2 Functional Magnetic Resonance Imaging (fMRI)

To understand what fMRI measures, we must first describe the anatomy of a neuron. The human brain is comprised of approximately 8.6×10^{10} neurons, which come in a wide variety of shapes, sizes and functions (Herculano-Houzel (2009)). Broadly speaking, neurons are comprised of three parts; the cell body (soma), the dendrites and an axon. Neurons are connected to one another via synapses: junctions, or gaps, which lie between the axon terminal (the tip of a small branch at the end of an axon) of one neuron and the dendrite of another. Each neuron shares a synapse with approximately $7,000$ neighbouring neurons, and it is through each of these synaptic connections that the electro-chemical communication, which is believed to represent cognitive function, occurs.

Communication between neurons transpires via electrical impulses known as action potentials (c.f. Fig 2.2). Triggered by complex processes in the soma, an action potential travels along the axon until it reaches an axon terminal, where it then encounters a synapse. At the synapse, the electrical impulse is converted into chemical form via the use of chemical 'neurotransmitters' and other cellular processes, secreted across the synapse and received by a dendrite in the neighbouring neuron. Once the neurotransmitters arrive at the dendrite, the signal is propagated to the neighbouring cell's soma. Inside the soma of the neighbouring cell, the signal may then stimulate

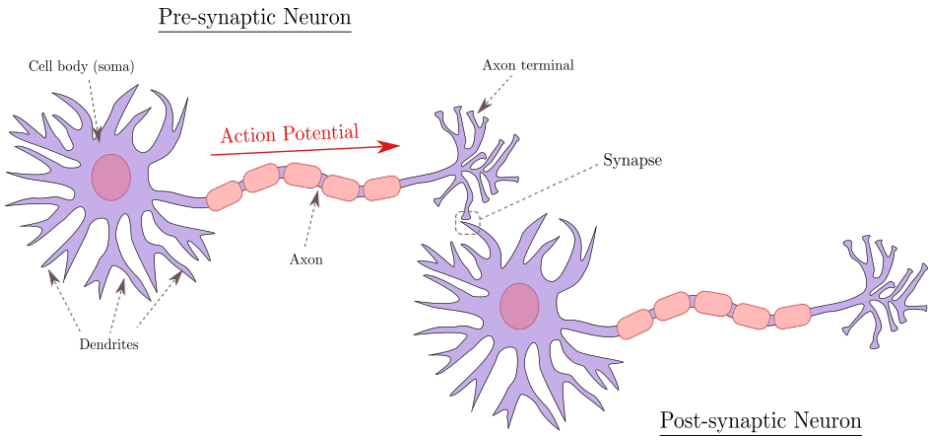

Figure 2.2: Communication between neurons. Information is transported along the axon of the pre-synaptic neuron as an action impulse. At the synapse, the electrical action impulse is converted into chemical neurotransmitters, which are then secreted into the synapse and detected by the dendrites of the post-synaptic neuron.

or inhibit the firing of action potentials, thus allowing the process to continue across neurons.

The transmission of chemicals across synapses is a process known as neuronal coupling and requires large amounts of energy, which must be provided in the form of oxygen and glucose transported in the blood. Via a vast network of arteries (c.f. Fig. 2.1), oxygen is provided to active neurons in the form of haemoglobin in red blood cells. The magnetic properties of haemoglobin differ depending on whether the haemoglobin is carrying oxygen or not, and it is these differences in magnetism that may be detected by an fMRI scanner. To be specific, deoxygenated haemoglobin is more magnetic than oxygenated haemoglobin and induces inhomogeneities in the local magnetic field created by an fMRI scanner. These inhomogeneities are measured by the fMRI scanner and recorded as the Blood Oxygen Level Dependent (BOLD) response. The expected BOLD response over time is governed by the Haemodynamic Response Function (HRF), which is illustrated in the bottom left display of Fig. 2.3.

There are many nuances to the interpretation of the BOLD response, as the signal detected by BOLD is mediated by an array of secondary factors related to localized changes in blood flow and blood oxygenation (Hall et al. (2016)). These factors include; the spiking activity of individual neurons, Cerebral Blood Volume (CBV;

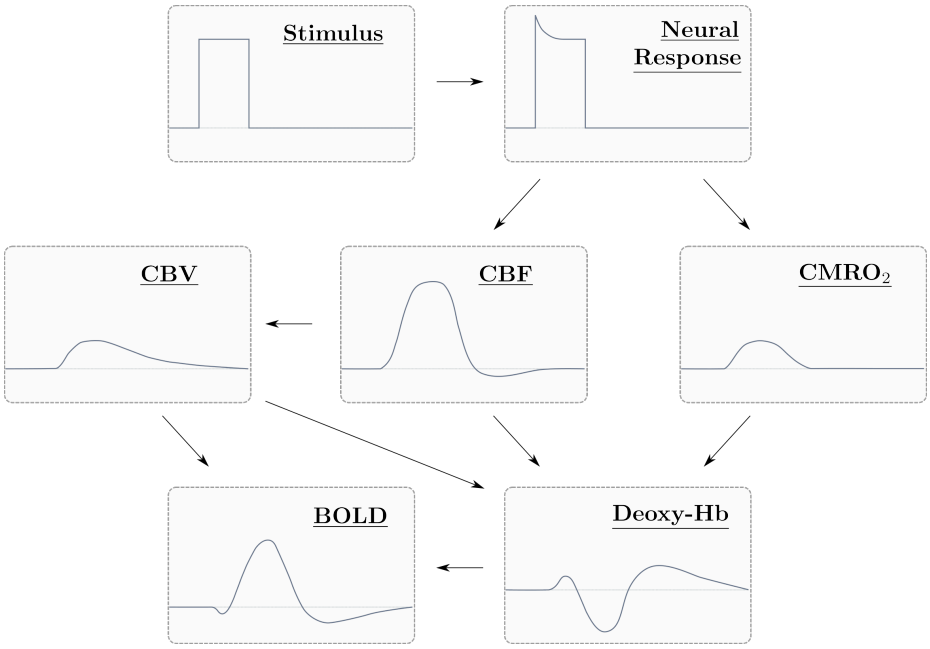

Figure 2.3: The physiological factors which are currently believed to determine the BOLD response. In the presence of a stimulus, BOLD response is influenced a range of factors including CBF, CBV and $CMRO_2$. The above image is based on the models presented in Buxton et al. (2004) and Buxton (2012).

the volume of blood present in a given amount of brain tissue), Cerebral Blood Flow (CBF; the rate at which blood is supplied) and the Cerebral Metabolic Rate of Oxygen ($CMRO_2$; the rate at which oxygen is consumed by the brain). However, it is generally agreed that the BOLD response serves as a reasonable proxy for neuronal activity.

2.1.3 Experimental Design and Task-based fMRI

FMRI provides a non-invasive method for studying changes in BOLD signal across the brain over time. Typically, the aim of an fMRI study is to use the BOLD signal as an indication of how the brain responds to various conditions which the researcher has deemed to be of academic interest. In general, every fMRI study falls into one of two categories; resting-state or task-based. Resting-state fMRI investigates the intrinsic function of the brain when a subject is at rest, whilst task-based fMRI focuses on a subject's response to a stimulus or task. The content of this thesis predominantly concerns task-based fMRI (although it should be noted that the methodology

proposed may also be employed for the analysis of resting-state data). For this reason, here we provide a brief overview of the analysis designs typically employed for a task-based fMRI study.

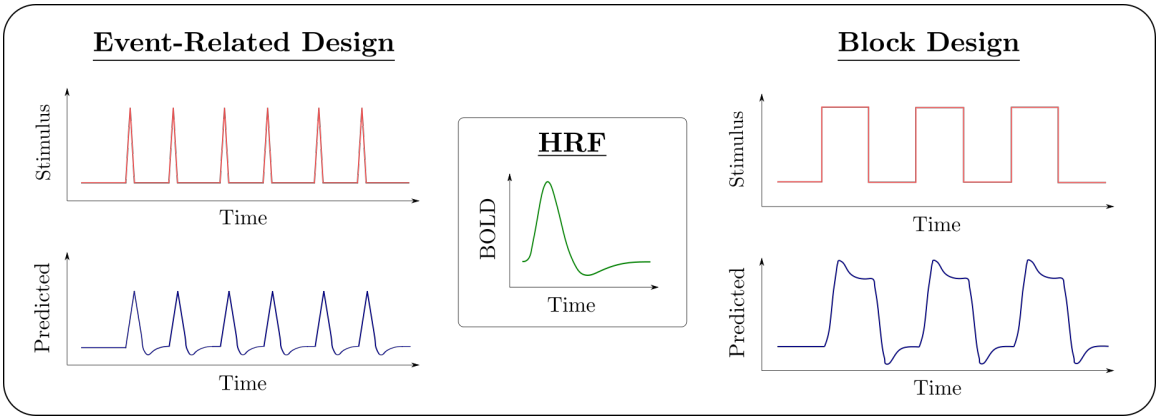

Figure 2.4: Event-related (left) and block (right) designs for task-based fMRI. For each experimental design, the predicted neuronal response (bottom) may be obtained by convolving an approximation of the HRF (center) with the stimulus onset timings (top).

In a task-based fMRI experiment, while in the scanner, subjects are asked to respond to stimuli. The stimuli may be presented to the subject very briefly (e.g. a flash of an image for 500ms or the sound of a single spoken word) or over an extended time period (e.g. a written word shown for a prolonged period of time as part of an antonym generation task). Brief flashes of stimuli are conventionally referred to as "events", whilst stimuli presented to subjects over prolonged periods of time are referred to as "blocks". An experiment involving events is said to have an "event-related design", whilst an experiment involving blocks is said to have a "block design".

To increase the amount of data available and therefore improve the statistical power later in the analysis, the stimuli are presented to the subject multiple times throughout the imaging session, typically at regular or similarly-spaced intervals (Jenkinson and Chappell (2018)). Images of the subject's BOLD response are acquired every few seconds so that the resultant time series of images may be compared to the experimental design to establish which areas of the brain are responsive to the

task. If a region of the brain is responsive to the task at hand, then the predicted BOLD response for that region is given by the convolution of the HRF with a time series representing the experimental design (see Fig. 2.4). The statistical analysis which is used to compare the predicted BOLD response to the observed fMRI timeseries is discussed in Section 2.3.

During the experiment, a structural MRI (sMRI) scan is also acquired in each imaging session. An sMRI scan is an image generated using the magnetic properties of grey matter and white matter, which represents the anatomy (i.e. the *structure*) of the brain. This is in contrast to an fMRI scan, which uses the magnetic properties of haemoglobin in the blood to represent the *function* of the brain. The sMRI image is acquired for the purposes of preprocessing and is crucial to many of the steps detailed in the following section.

2.1.4 Preprocessing and fMRI

Once an fMRI experiment has been conducted, a series of preprocessing steps must be applied to the data. These steps serve to reduce noise in the acquired images, remove unwanted artefacts and biological features and to 'line up' or 'align' the images with one another. A brief list of preprocessing steps to which fMRI data are conventionally subjected is given below:

- **Distortion Correction:** The magnetic field generated during an MRI scan is theoretically homogeneous and uniform across space. However, in practice, inhomogeneities in the magnetic field are common, often caused by differences in the magnetic susceptibility properties of the tissue types being scanned. These inhomogeneities behave in a predictable manner and are often fixed in a preprocessing stage by "warping" an image.

- **Motion Correction:** Image acquisition can be a lengthy process, and it is difficult for participants to remain completely still for such prolonged periods of time. It is inevitable that there will be some head movement during any image

acquisition. In this stage of the preprocessing pipeline, images are transformed to line up with one another (often, the middle image of the time series is used as a reference).

- **Co-registration:** In order to compare them to something meaningful, during preprocessing, images are transformed to resemble the subject's anatomy. This is achieved by "co-registering" the fMRI timeseries images to the subject's anatomical sMRI image. Following co-registration, all fMRI timeseries images have the same shape, size and orientation as the subject's sMRI image.

- **Registration:** Often, scientific interest lies in comparing the results of analysis "between-subjects". For this reason, when performing a group-level analysis, a second transformation is applied, in which the data is transformed to resemble a standard template image to allow between-subject comparison.

- **Brain Extraction:** Non-brain structure, such as the skull and surrounding tissues, are typically removed from all of the images before analysis. This process is performed using the anatomical sMRI image as a reference and is sometimes known as skull-stripping.

- **Spatial Smoothing:** To combat the spatial noise in the data, the images are often convolved with a Gaussian kernel (i.e. made "smoother"). This is a typical step performed during preprocessing, but it is also common to perform smoothing during the first-level analysis instead (c.f. Section 2.3.2).

- **Prewhitening:** FMRI timeseries exhibit strong correlation across time. Such correlation could be accounted for during the modelling stage of an analysis, but such practices have been shown to cause bias in parameter estimation due to poor estimation of the temporal covariance parameters. For this reason, temporal autocorrelation is removed from the data prior to modelling, using methods that average temporal covariance parameters across subjects or smooth temporal covariance parameters across space to obtain improved estimates.

Preprocessing in fMRI is an imperfect process, and there is much which can go awry. An incorrectly calibrated brain extraction, for example, can easily result in the removal of the entire brain from the image. Unfortunately, such occurrences are ubiquitous in fMRI analysis and, whilst typically not as severe as whole-brain removal, can result in substantial variability in the size and shape of the brain in the preprocessed images. The impact that such spatial variability has on a statistical fMRI analysis is an issue which will be explored further in Sections 2.3.3 and 2.3.5, formally accounted for in Chapter 4 and quantified in Chapter 5.

2.2 Statistical Modelling Preliminaries

In this section, we shall describe the univariate statistical methodology which shall be employed to model the BOLD response for a single voxel. Much of the material outlined in this section shall be used in Chapters 3 and 4, in which a Fisher Scoring method is proposed and verified for estimating the parameters of the Linear Mixed Model. As the scope of Chapter 3 generalizes beyond neuroimaging applications, this section draws on examples from non-imaging contexts for reference. The relevance of the material of this section in relation to conventional fMRI analysis is detailed in Section 2.3.2.

2.2.1 The Linear Model

One of the most commonly adopted techniques for statistical analysis is to employ the Linear Model. The Linear Model (LM) containing n observations and p parameters takes the following form:

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I_n), \tag{2.1}$$

where the known quantities in the model are; the $(n \times 1)$ vector of responses, Y, and the $(n \times p)$ design matrix, X. The unknown model parameters are; the $(p \times 1)$ fixed

effects vector, β, and the fixed effects covariance, σ^2. The $(n \times 1)$ vector ϵ is known as the random error. The elements of ϵ are conventionally assumed to be independent and identically distributed (i.i.d.) (although, in a slight abuse of notation and terminology, in Chapter 5, we shall allow ϵ to take an arbitrary covariance structure). Dropping constant terms, the log-likelihood of the LM is given by:

$$l(\beta, \sigma^2) = -\frac{1}{2}\Big\{n \log(\sigma^2) + \sigma^{-2} e'e\Big\},$$

where $e = Y - X\beta$ is the residual vector. The Ordinary Least Squares (OLS) estimators, which maximize the above log-likelihood, for β and σ^2 are given by the below:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y, \qquad \hat{\sigma}^2_{OLS} = \frac{e'e}{n}. \tag{2.2}$$

2.2.2 Simpson's Paradox

When fitting a regression model to a dataset it is a necessity that grouping structure in the data be accounted for. For example, suppose data had been collected from several subjects where each subject had multiple recordings, taken across several evenly spaced time-points, of their weight and the amount of exercise they had done in the previous 3 months. Suppose further that the researcher in this example wished to know if exercise was a good predictor of weight.

An LM analysis for this example may proceed by including an intercept and the exercise data in the design matrix, X, and the weight measurements in the response vector, Y. The model for subject j at timepoint t in this example may appear as:

$$\text{Weight}_{j,t} = \beta_0 + \text{Exercise}_{j,t} \times \beta_1 + \epsilon_{j,t}, \qquad \epsilon_{j,t} \sim N(0, \sigma^2).$$

where $\text{Weight}_{j,t}$ and $\text{Exercise}_{j,t}$ are the weight and amount of exercise recorded for subject j at time t. Estimating this model may produce a fit resembling that displayed on the left hand side of Fig. 2.5. By the left hand side of Fig. 2.5, it seems that it can

be concluded that higher amounts of exercise predict increased weight. Something is clearly wrong here. The issue here becomes apparent upon considering the right hand side of Fig. 2.5; the problem is that subject-level grouping structure was not accounted for in the model.

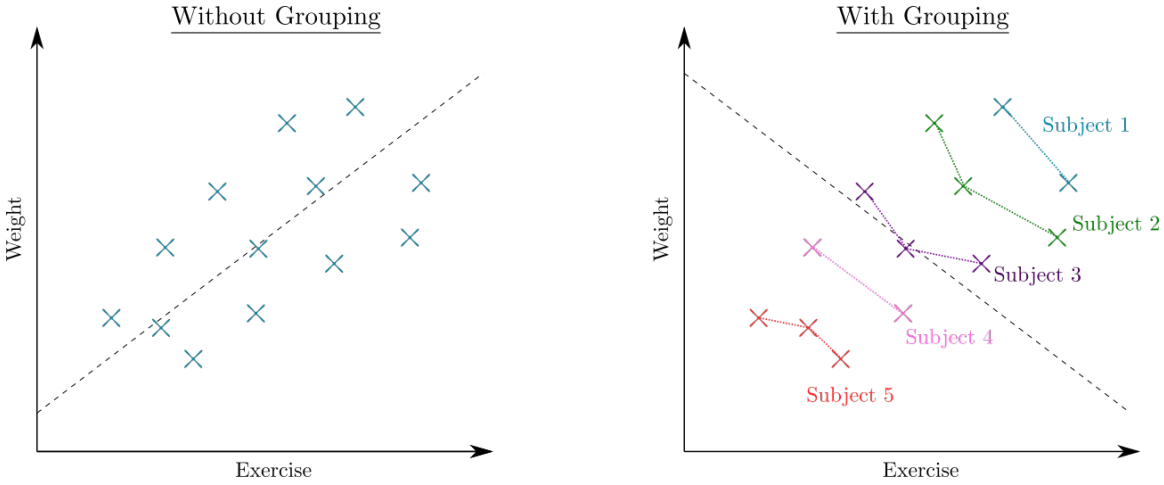

Figure 2.5: Illustration of Simpson's paradox. On the left, grouping structure has not been accounted for, leading to the prediction of an upward trend. On the right, subject-level grouping structure has been accounted for and is indicated with dotted lines and colour. Accounting for grouping structure has reversed the observed trend.

The apparent contradiction that accounting for grouping structure in data can reverse predicted trends is commonly known as Simpson's paradox (Simpson (1951)). A standard approach to accounting for Simpson's paradox is to employ the Linear Mixed Model (LMM). In this example, an LMM may be thought of as first constructing a model for each subject's data separately. The model for subject j may appear as follows;

$$\text{Weight}_{j,t} = \gamma_{j,0} + \text{Exercise}_{j,t} \times \gamma_{j,1} + \epsilon_{j,t}, \qquad \epsilon_{j,t} \sim N(0, \sigma^2).$$

Note how, in the above, the parameters $\{\gamma_{j,i}\}_{i\in\{1,2\}}$ vary with j, as opposed to $\{\beta_i\}_{i\in\{1,2\}}$ in the previous model which were constant across subjects. To combine models across subjects, the LMM assumes that the 'subject-level' intercepts, $\{\gamma_{j,0}\}$, and slopes, $\{\gamma_{j,1}\}$, are distributed around some global mean values, β_0 and β_1, respectively. In other words, in this example an LMM may assume that, for subject j

and each $i \in \{1, 2\}$:

$$\gamma_{j,i} = \beta_i + b_{j,i}, \qquad b_{j,i} \sim N(0, \tau^2).$$

By substituting the above into the previous expression given for the LMM it can be seen that the full model in this case can be written as:

$$\text{Weight}_{j,t} = (\beta_0 + \text{Exercise}_{j,t} \times \beta_1) + (b_{j,0} + \text{Exercise}_{j,t} \times b_{j,1}) + \epsilon_{j,t}.$$

The important distinction between the above model and the LM previously constructed is the inclusion of the random terms $\{b_{j,i}\}_{i \in \{1,2\}}$. These terms are known as 'random effects' and, in this case, model 'within-subject' behaviour. There is one grouping factor in this model; the factor 'subject'.

This example provides a brief insight into the construction and commonplace motivations behind the LMM. One of the most common applications of the LMM is to model datasets containing a single 'grouping factor'. However, the LMM also allows for far more flexible designs that include, for example, multiple grouping factors, and arbitrary 'within-factor' covariance structures. In the following section, we present the general form of the LMM.

2.2.3 The Linear Mixed Model

The LMM, for n observations, takes the following form:

$$\begin{aligned} Y &= X\beta + Zb + \epsilon \\ \epsilon &\sim N(0, \sigma^2 I_n), \quad b \sim N(0, \sigma^2 D). \end{aligned} \tag{2.3}$$

Much like the LM, the known quantities in the model are; Y (the $(n \times 1)$ vector of responses), X (the $(n \times p)$ fixed effects design matrix) and Z (the $(n \times q)$ random effects design matrix). The unknown model parameters are: β (the $(p \times 1)$ fixed effects parameter vector), σ^2 (the scalar fixed effects variance) and D (the $(q \times q)$ random effects covariance matrix). From Equation (2.3) the marginal distribution of

the response vector, Y, can be seen to be $N(X\beta, \sigma^2(I_n + ZDZ'))$. The log-likelihood for the LMM specified by Equation (2.3) is derived from the marginal distribution of Y. Dropping constant terms, this log-likelihood is given by:

$$l(\theta) = -\frac{1}{2}\Big\{n\log(\sigma^2) + \sigma^{-2}e'V^{-1}e + \log|V|\Big\}, \tag{2.4}$$

where θ is shorthand for the parameters (β, σ^2, D), $V = I_n + ZDZ'$ and, as before, $e = Y - X\beta$. Similarly, the restricted log-likelihood is given by:

$$l_R(\theta) = -\frac{1}{2}\Big\{(n-p)\log(\sigma^2) + \sigma^{-2}e'V^{-1}e + \log|V| + \log|X'V^{-1}X|\Big\}. \tag{2.5}$$

Unlike for the LM, no closed-form expression exists for parameter estimation of the LMM. Chapter 3 of this thesis shall focus extensively on numerical parameter estimation performed via Maximum Likelihood (ML) estimation of Equation (2.4). Chapter 4 employs the methods developed in Chapter 3 but adapts them to perform Restricted Maximum Likelihood (ReML) estimation using Equation (2.5). To bridge the gap between the ML methods of Chapter 3 and the ReML methods of Chapter 4, further detail of how ReML estimation may be performed using the methods of Chapter 3 is provided in Appendix A.5.

An LMM may be described as multi-factor or single-factor. The distinction between the multi-factor and single-factor LMM lies in the specification of the random effects in the model. Random effects are often described in terms of factors, categorical variables that group the random effects, and levels, individual instances of such a categorical variable. We highlight here that the term "factor", in this work, refers only to categorical variables which group random effects and does not refer to groupings of fixed effects. We denote the total number of factors in the model as r and denote the k^{th} factor in the model as f_k for $k \in \{1, ..., r\}$. For a given factor f_k, l_k will be used to denote the number of levels possessed by f_k, and q_k the number of random effects which f_k groups. The single-factor LMM corresponds to the case $r = 1$, while the multi-factor setting corresponds to the case $r > 1$. The example

of the previous section was a single-factor LMM. In that example, observations were grouped by the factor 'subject' (e.g. the factor f_1 was 'subject'). There were five subjects (e.g. $l_1 = 5$), and two random effects modelling subject-level behaviour (e.g. $q_1 = 2$).

A more complex example of how this notation may be used in practice is given as follows. Suppose an LMM contains observations that are grouped by 'subject' (i.e. the participant whose observation was recorded) and 'location' (i.e. the place the observation was recorded). Further, suppose that the LMM includes a random intercept and random slope which each model subject-specific behaviour, and a random intercept which models location-specific behaviour. Two factors are present in this design: the factor f_1 is 'subject' and the factor f_2 is 'location'. Therefore, $r = 2$. The number of subjects is l_1 and the number of locations is l_2. The number of covariates grouped by the first factor, q_1, is 2 (i.e. the random intercept and the random slope) and the number grouped by the second factor, q_2, is 1 (i.e. the random intercept).

The values of r, $\{q_k\}_{k\in\{1,\dots,r\}}$ and $\{l_k\}_{k\in\{1,\dots,r\}}$ determine the structure of the random effects design matrix, Z, and random effects covariance matrix, D. To specify Z formally is notationally cumbersome and of little relevance to the aims of this section. For this reason, formal detail of the construction of Z is deferred to Section 4.3.1.1 in which the construction of Z is of practical relevance. Here it suffices to note that, under the assumption that its columns are appropriately ordered, Z is comprised of r horizontally concatenated blocks. The k^{th} block of Z, denoted $Z_{(k)}$, has dimension $(n \times l_k q_k)$ and describes the random effects which are grouped by the k^{th} factor. Additionally, each block, $Z_{(k)}$, can be further partitioned column-wise into l_k blocks of dimension $(n \times q_k)$. The j^{th} block of $Z_{(k)}$, denoted $Z_{(k,j)}$, corresponds to the random effects which belong to the j^{th} level of the k^{th} factor. In summary,

$$Z = [Z_{(1)}, Z_{(2)}, \dots Z_{(r)}], \quad Z_{(k)} = [Z_{(k,1)}, Z_{(k,2)}, \dots Z_{(k,l_k)}] \quad \text{for } k \in \{1, \dots, r\}.$$

An important property of the matrix Z is that for any arbitrary factor f_k, the rows of $Z_{(k)}$ can be permuted in order to obtain a block diagonal matrix. As, for

the single-factor LMM, $Z \equiv Z_{(1)}$, it follows that the observations of the single-factor LMM can be arranged such that Z is block diagonal. This feature of the single-factor LMM greatly simplifies the numerical optimization required for parameter estimation of single-factor designs. However, this simplification cannot be generalized to the multi-factor LMM. In general, it is not true that the rows of Z can be permuted in such a way that the resultant matrix is block-diagonal. As a result of this, many of the results derived for the single-factor LMM have not been generalized to the multi-factor setting (see Section 3.1 for further discussion).

To describe the random effects covariance matrix, D, it is assumed that factors are independent from one another and that for each factor, factor f_k, there is a $(q_k \times q_k)$ unknown covariance matrix, D_k, representing the "within-factor" covariance for the random effects grouped by f_k. The random effects covariance matrix, D, appearing in Equation (2.3), can now be given as $D = \bigoplus_{k=1}^{r}(I_{l_k} \otimes D_k)$ where $\oplus$ represents the direct sum, and $\otimes$ represents the Kronecker product. Note that, whilst D is large in dimension (having dimension $(q \times q)$ where $q = \sum_k q_k l_k$), D contains $q_u = \sum_k \frac{1}{2} q_k(q_k + 1)$ unique elements. Typically, it is true that $q_u << q^2$. As a result, the task of describing the random effects covariance matrix D reduces in practice to specifying only a small number of parameters.

2.2.4 Null-Hypothesis Inference

For both the LM and LMM, several methods exist for drawing inference on the fixed effects parameter vector, β. Often, research questions can be expressed as null hypothesis tests of the below form:

$$H_0 : L\beta = 0, \qquad H_1 : L\beta \neq 0,$$

where L is a fixed and known $(1 \times p)$-sized contrast vector specifying a hypothesis, or prior belief, about linear relationships between the elements of β, upon which an inference is to be made. A commonly employed statistic for testing hypotheses of this

form is the T-statistic, given by:

$$T = \frac{L\hat{\beta}}{\sqrt{\widehat{\text{Var}}(L\hat{\beta})}}, \tag{2.6}$$

where $\hat{\beta}$ is typically computed via a maximum likelihood estimation procedure. For the LM, if the variance estimate in the denominator of the above statistic is given by $\hat{\sigma}^2 L(X'X)^{-1}L'$, T is known to follow a students $t-$distribution with $n-p$ degrees of freedom. For the LMM however, the situation is more complex as no equivalent result exists. As a consequence, when performing a $T-$test for the LMM, assumptions must be made regarding the approximate distribution of the T-statistic for a given estimator of $\text{Var}(L\hat{\beta})$. The estimation of the degrees of freedom employed in such 'approximate $T-$tests' is an issue which shall be explored further in Section 3.3.5 of Chapter 3.

Closely related to the $T-$statistic is the $F-$statistic, which can be used to assess whether the inclusion of fixed effects in a model improves the model's goodness of fit. The $F-$statistic takes the following form:

$$F = \frac{\hat{\beta}'L'[L\widehat{\text{Var}}(\hat{\beta})L']^{-1}L\hat{\beta}}{\text{rank}(L)}, \tag{2.7}$$

where, in this instance, L is a contrast matrix consisting of k contrast vectors of dimension $(1 \times p)$ stacked on top of one another. Each of the k contrast vectors represents a covariate in the model, and it is these k covariates whose impact on the model fit is assessed by the $F-$test. In the case $k = 1$, the $F-$statistic is equal the square of the T-statistic. Because of the close relation between the two statistics, much of the above discussion, as well as the material of the following chapters, that concerns the $T-$statistic also applies to the $F-$statistic. The link between the material of Chapter 3 and the $F-$statistic is made explicit in Appendix A.9.2.

In the context of the LMM, it is often of practical interest to make inferences

about the variance parameters contained in D. Such tests take the following form:

$$H_0 : D_{\{J\}} = 0 \quad H_1 : D_{\{J\}} \neq 0$$

where J represents a predetermined set of elements of D which correspond to the removal of $\tilde{q}$ random effects from the study design. Such hypotheses may be tested using a Likelihood Ratio Test (LRT). The LRT statistic is given as:

$$\lambda_{LR} = -2\ln\left(\frac{l(\hat{\theta}_0)}{l(\hat{\theta})}\right), \tag{2.8}$$

where $\hat{\theta}_0$ and $\hat{\theta}$ are the maximum likelihood parameter estimates obtained when the random effects $D_{\{J\}}$ are and are not included in the model specification, respectively. As with the $T-$statistic, an exact distribution for the LRT statistic is not known and thus approximation methods must be employed. One common approach is to assume that λ_{LR} follows a even mixture of χ^2 distributions. This approach is further discussed in Sections 4.3.1.4 and 4.3.3 of Chapter 4, in which likelihood ratio tests of the above form are employed for LMM analysis of large-sample fMRI datasets.

2.3 Neuroimaging Statistics

In this section, we describe how the statistical methods of Section 2.2 are applied to analyse preprocessed fMRI data. First, in Section 2.3.1, we introduce some relevant terminology. Following this, in Section 2.3.2, we outline the conventional two-stage LM and LMM analyses employed to model BOLD response for a single voxel. In Section 2.3.3, we then describe an issue which is particularly prominent in large-sample fMRI analysis concerning missing data near cortical boundaries. Next, in Section 2.3.4, we survey common approaches to accounting for the multiple comparisons problem in fMRI analysis. Lastly, in Section 2.3.5, we outline the theory of confidence regions, which may be used to assess the reliability and spatial uncertainty of identified regions of activation in fMRI results.

2.3.1 The Mass-Univariate Approach

Before discussing how the models outlined in Section 2.2 are applied to the data acquired using the methods of Section 2.1, we now introduce some practical terminology for working with fMRI data. The fMRI data employed in this work is typically stored as a Neuroimaging Informatics Technology Initiative (NIfTI) file. A NIfTI file contains a 3D image, or 'map', of the brain alongside metadata describing the data's provenance. The image is composed of voxels ("3D pixels", resolution elements) arranged in a lattice formation. Typical image dimensions used in this work are approximately $100 \times 100 \times 100$ voxels. In this work we shall often refer to 'brain masks'. A brain mask is a binary (1-0) image with voxels lying within the brain being encoded as 1 and voxels lying outside of the brain being encoded[1] as 0. For further detail, see Section 2.3.3 and Fig. 2.9, which discuss and illustrate issues surrounding 'mask-variability'.

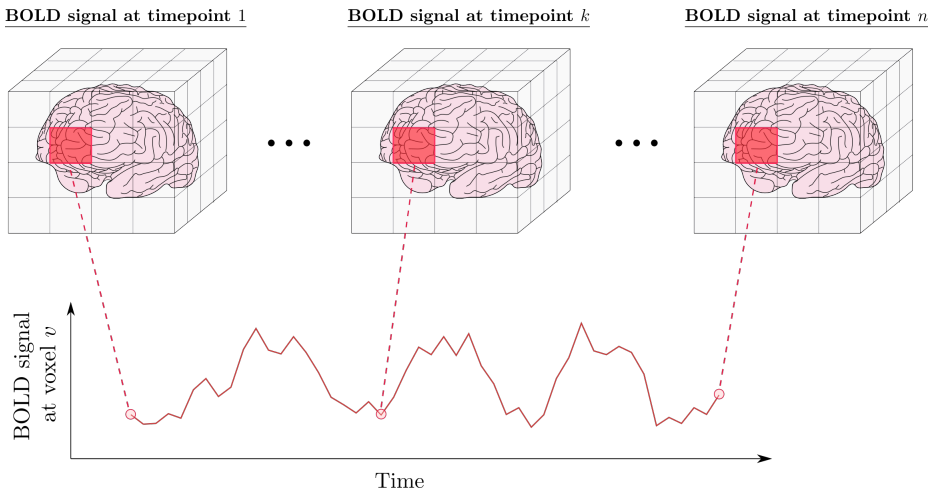

Figure 2.6: Extracting the BOLD activation for a given voxel from an fMRI timeseries. FMRI images are displayed for timepoints 0, k and n for some k such that $0 < k < n$ (top), with the voxel v highlighted in red, and the BOLD activation observed at voxel v is plotted against time (bottom).

In the previous sections, we alluded to the fact that we would be comparing the predicted BOLD response (c.f. Fig. 2.4) to 'BOLD activation in the region of the

[1]Some software packages adopt the convention of encoding voxels outside the brain with NaN values. This convention is not adopted in this work.

brain' but did not explicitly state what was meant by 'in a region'. FMRI commonly adopts a "mass-univariate" approach, in which the BOLD activation at each voxel is modelled separately against the predicted BOLD activation. The BOLD activation at a given voxel may be obtained by considering the value recorded at that voxel in each fMRI image in the timeseries (see Fig. 2.6).

Throughout this work, the set of voxels in an image shall be denoted V and whenever we wish to highlight that a variable belongs to a specific *voxel in the image* a subscript v shall be used to represent said voxel (e.g. Y_v, T_v, etc). To reduce the notation in this work, we shall only employ the subscript v when it is necessary to distinguish between voxels. The subscript v notation features predominantly in Chapter 4. In addition, each image employed in this work shall be assumed to represent some (potentially random) spatially varying function which maps coordinates from a closed spatial domain $S \subseteq \mathbb{R}^3$ to $\mathbb{R}$. When we wish to highlight that a variable belongs to a specific *location in continuous space*, the location shall be denoted as s and the variable shall be denoted as a function of s (e.g. $Y(s), T(s)$, etc). This notation features heavily in Chapter 5.

2.3.2 Multi-Level Analysis

In fMRI analysis, a "two-stage" approach is conventionally adopted in which group-level BOLD response at each voxel is modelled using either two separate LMs or one single-factor LMM. To illustrate this process for a single voxel, we shall first outline the two-stage LM approach. We shall then explain how this approach may be adapted to instead employ the LMM and illustrate why the LMM may be favoured for this task.

For the two-stage LM approach, in the 'first-level' of analysis, also known as the 'subject-level', the below linear model is fit for each subject, subject j:

$$Y_j = X_j\beta_j + \epsilon_j, \qquad \epsilon_j \sim N(0, \tau_j^2 I_{t_j}).$$

where t_j is the number of timepoints (fMRI images) belonging to subject j. Y_j contains the observed BOLD signal across time for subject j (see Fig. 2.6) and X_j contains the predicted BOLD signal for subject j (see Fig. 2.4) alongside other covariates of secondary interest. Common examples of other covariates contained in X_j include; an intercept, representing the signal mean, and low-frequency confound terms, representing drifts caused by scanner noise or physiological factors. An illustration of a first-level analysis is provided by Fig. 2.7.

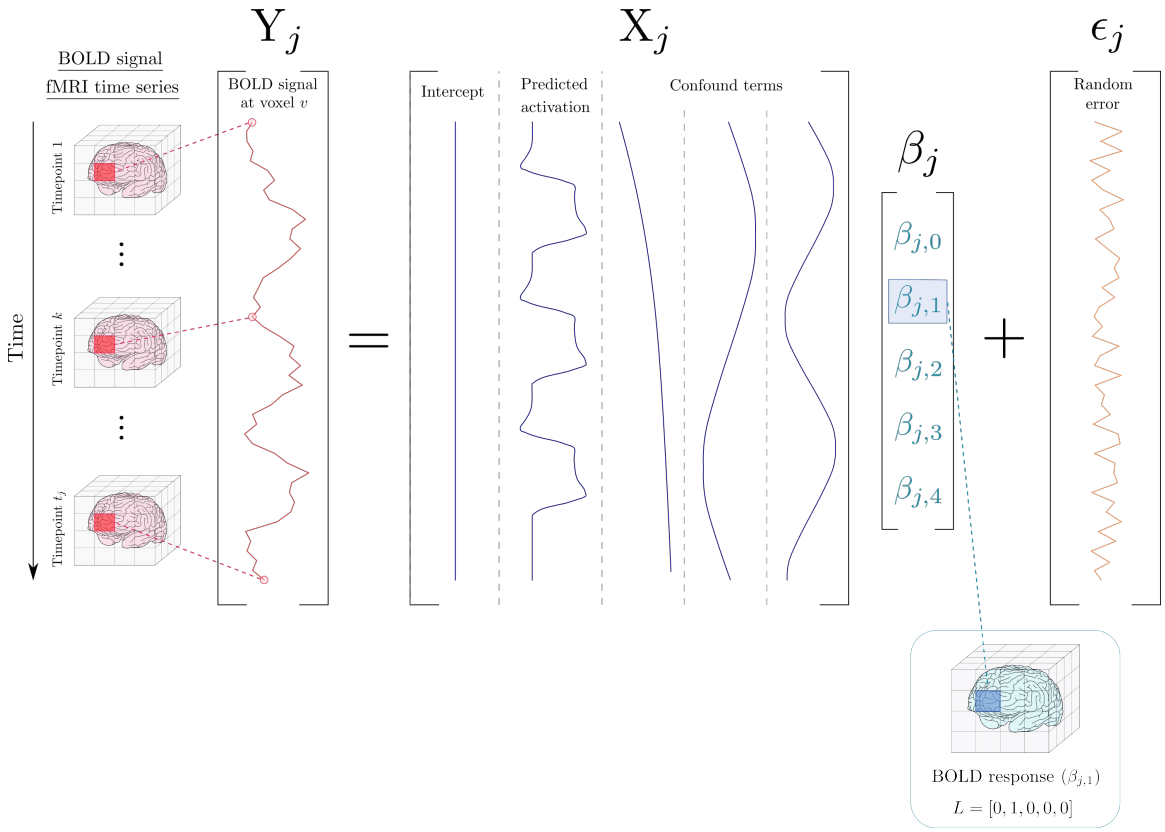

Figure 2.7: An illustration of the first-level analysis for subject j employed for a single voxel. On the left hand side is an fMRI timeseries, which is used to construct the response vector Y_j. The design matrix X_j contains an intercept, the predicted BOLD activation (see Fig. 2.4) and confound terms representing low-frequency drift. The error term, ϵ_j, represents random noise. Numerical values in X_j, Y_j and ϵ_j are represented as functions plotted over time. The output of the analysis is an estimate of the predicted BOLD response coefficient, $\beta_{j,1}$ (bottom).

The model is fitted using Equation (2.2) and an estimate, $\hat{\beta}_j$, is obtained for the fixed effects parameter vector β_j. This estimate is then multiplied by a $(1 \times p)$-shaped contrast vector, L, representing the research question of interest (e.g. in

Fig. 2.7, interest lies in the parameter corresponding to the task-related activation; $\beta_{j,1} = L\beta_j$). As this process is performed for each voxel in the image, the result of a first-level analysis is an image of $L\hat{\beta}_j$ values, known as a 'contrast map', representing task-related BOLD response. As a first-level analysis is performed for each subject, the output following this stage of analysis is a collection of contrast maps, each representing a subject's BOLD response to the contrast of interest.

Following first-level analysis, results must be combined across subjects. To combine the first-level analysis results for l_1 subjects, a 'second-level', or 'group-level', analysis is conducted. Typically, the second-level analysis involves constructing an LM in which the first-level estimates $\{L\hat{\beta}_j\}_{j\in\{1,...l_1\}}$ are used to build the response vector and subject characteristics such as age, sex and BMI are included in the design matrix. For a given voxel, a typical second level analysis takes the following form:

$$\hat{Y}_G = X_G\beta_G + \epsilon^*,$$

where $\hat{Y}_G = [L\hat{\beta}_1, ..., L\hat{\beta}_{l_1}]'$, X_G is the group-level design matrix and β_G is the group-level parameter vector. The aim of a second-level analysis is to estimate the parameters vector β_G to make inference about group averages or between-group differences in BOLD response. For example, in Fig 2.8, we illustrate how a second-level analysis may be employed to investigate the difference in activation between two groups of subjects (e.g. "did group B exhibit a higher level of BOLD activation than group A?").

As, in the above model, $\hat{Y}_G$ contains estimators $\{\hat{\beta}_j\}_{j\in\{1,...l_1\}}$, rather than the true parameters $\{\beta_i\}_{i\in\{1,...l_1\}}$ the assumption that $\epsilon^* \sim N(0, \sigma^2 I_{l_1})$ may not be reasonable. To account for this, a standard approach is to define $\epsilon = Y_G - X_G\beta_G$ where $Y_G = [L\beta_1, ..., L\beta_{l_1}]'$ and assume that $\epsilon \sim N(0, \sigma^2 I_{l_1})$. From the definition of ϵ, it can be seen that;

$$\epsilon^* = \epsilon - Y_G + \hat{Y}_G.$$

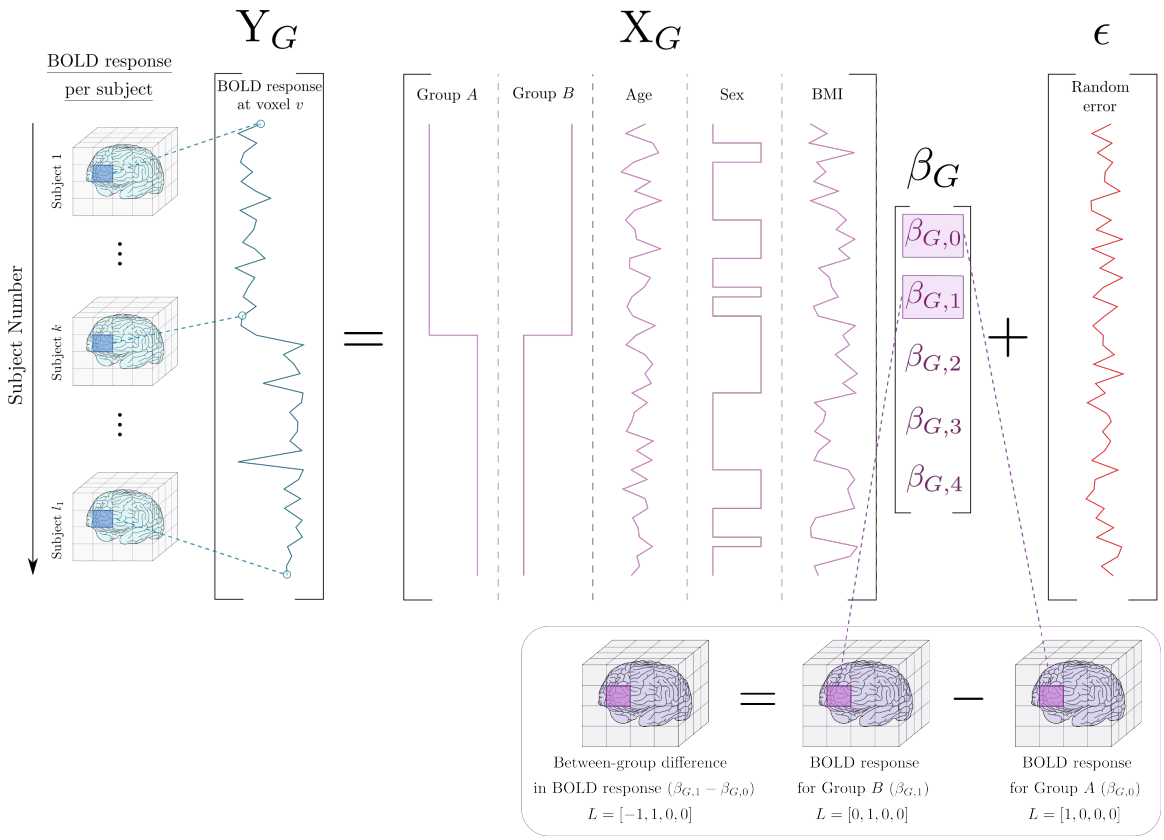

Figure 2.8: An illustration of the second-level model employed for a single voxel. On the left hand side are the $\{L\beta_{G,j}\}$ estimates obtained during the first level analysis (see Fig. 2.7), which are used to construct the response vector Y_G. The design matrix X_G contains two indicator variables representing groups of subjects, subject's age, sex and BMI. The error term, ϵ_G, represents random noise. Numerical values in X_G, Y_G and ϵ_G are represented as functions plotted against subject number. The output of the analysis is an estimate of the between-group difference in BOLD response, $\beta_{G,1} - \beta_{G,0}$ (bottom).

From the above it can be seen that the $\mathrm{Var}(\epsilon^*)$ equals $\mathrm{Var}(\epsilon)$+$\mathrm{Var}(\hat{Y}_G)$ and therefore;

$$\mathrm{Var}(\epsilon^*) = \mathrm{Var}(\epsilon) + \mathrm{Var}(\hat{Y}_G) = \sigma^2 I_{l_1} + \bigoplus_{j=1}^{l_1} \tau_j^2 L(X_j' X_j)^{-1} L'$$

where $\oplus$ is the direct sum. At this stage, various complexities arise in estimating the variance components σ^2 and $\{\tau_j^2\}_{j\in\{1,\ldots l_1\}}$ with multiple software packages adopting different approaches and assumptions.

Retroactively estimating the variance components during two-stage analysis is both theoretically and computationally cumbersome and becomes substantially more

formidable when the experimental design increases in complexity (Smith et al. (2005), Mumford and Nichols (2006)). In particular, this approach runs into difficulties when (1) multiple grouping factors are present in the data, or (2) strong correlation exists between the subject-level fixed effects parameters $\{\beta_{j,i}\}_{i\in\{0,...p-1\}}$ for each subject, subject j (Chen et al. (2013)). Common examples of such situations include (1) longitudinal multi-session designs, in which subjects have attended multiple sessions and observations are now grouped by both 'subject' and 'session', and (2) situations in which the subject-level parameter estimates exhibit heavy within-subject correlation (e.g. due to subject-specific adjustments in the modelling of covariates). In such situations, employing the LMM instead of the two-stage LM can improve the accuracy of variance component estimation and increase the statistical power, especially for unbalanced designs (i.e. designs in which some subjects have many more timepoints than others) (Mumford and Nichols (2009), Lindquist et al. (2012)). Such issues have left many researchers to conclude that the "proper model for group fMRI data is the two-stage summary statistics approach of the [linear] mixed model" instead of linear modelling (Mumford and Poldrack (2007)).

For this reason, the LMM is a popular alternative to the classic two-stage LM approach. An LMM approach to two-stage fMRI analysis begins by concatenating the first-level models described above across subjects as follows:

$$Y = Z\gamma + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I_n)$$

where $Y = [Y_1', ..., Y_{l_1}']'$, $\gamma = [\beta_1', ...\beta_{l_1}']'$, Z =diag$(X_1, ...X_{l_1})$ where 'diag' represents diagonal matrix concatenation, $n = \sum_{j=1}^{l_1} t_j$ and $\epsilon = [\epsilon_1', ...\epsilon_{l_1}']'$. At the group-level, it is again assumed that the parameters contained in γ are distributed about some group-level parameter vector β_G as follows:

$$\gamma = X_G\beta_G + b, \qquad b \sim N(0, \sigma^2 D)$$

where D is an unknown matrix of variance parameters. The key difference between

the 'two-stage' LM approach and the LMM approach is that at the group-level the LMM's response vector, γ, consists of the parameters $\{\beta_j\}_{j\in\{1,\ldots,l_1\}}$, whereas the LM's response vector $\hat{Y}_G$, consists of the estimates $\{L\hat{\beta}_j\}_{j\in\{1,\ldots,l_1\}}$. By combining the above expressions, the LMM in this context can be seen to be given by:

$$Y = ZX_G\beta_G + Zb + \epsilon, \qquad Y \sim N\big(0, \sigma^2(I_n + ZDZ')\big)$$

By defining $X = ZX_G$, it can be seen that the above model fits the specification given by Equation (2.3). A strength of the LMM approach is that the covariance matrix D contains information about the correlation between random effects, and via similar constructions to the above, the model can be extended to include multiple grouping factors[2]. As argued in Chapter 1, accounting for complex multi-level covariance structure is a necessity for large-sample fMRI analyses, for which many more grouping factors may be present in the data (e.g. data may be grouped by 'session' for longitudinal analyses, by 'study protocol' data pooled from different sites,... etc). For this reason, performing LMM parameter estimation in the large-sample fMRI setting in a computationally efficient manner is one of the central objectives of Chapter 4.

Following parameter estimation and model fitting, null-hypothesis testing of the form described in Section 2.2.4 is conventionally performed. This process involves computing statistics of the form outlined in Section 2.2.4 independently for each voxel in the image. A threshold, c, is then chosen such that a voxel is deemed 'significant', if and only if, the statistic at said voxel exceeds c. The output of this process is a 'thresholded' image; an image of statistic values in which the values at any non-significant voxels have been replaced with zero. Voxels that are correctly and incorrectly declared as significant are referred to as true positives and false positives respectively. Similarly, voxels that are correctly and incorrectly not declared as significant are referred to as true negatives and false negatives, respectively. Section

[2]We emphasize here that the conventional formulation of the 'two-stage' approach in fMRI is, to some degree, arbitrary as other grouping factors may be, and often are, included in the analysis design. For example, in a longitudinal analysis it is not usually the case that the *group*-level analysis is the *second*-level, but rather the *third*.

2.3.4 provides a brief overview of the common approaches for choosing the value of the threshold c.

2.3.3 Mask Variability

Another important practical issue that must be addressed prior to or during the parameter estimation stage of an fMRI analysis is the missing data observed on and around cortical boundaries. Such missingness is ubiquitous in whole-brain fMRI analyses and can be attributed to several commonplace sources of between-image spatial variability. Such sources include magnetic susceptibility artefacts, imperfections in the registration process, improper brain extraction, (c.f. Section 2.1.4) differing image acquisition parameters and, indirectly, between-subject biological variation. An illustration of such missingness is provided by Fig. 2.9.

Conventionally, standard fMRI analysis tools address this missing data problem by omitting voxels with incomplete observations from the analysis. As detailed by Vaden et al. (2012), and later by Gebregziabher et al. (2017), adopting this approach can negatively impact both the specificity (true negative rate) and sensitivity (true positive rate) of the analysis results, especially when clusterwise thresholding is employed for multiple comparisons correction (see Section 2.3.4). In terms of specificity, an inflated false negative rate may be observed when the removal of voxels with incomplete data causes brain regions that are near cortical boundaries to be excluded from the analysis. In terms of sensitivity, an inflated false negative rate may result from the smaller number of tests being performed, and consequently, the use of a less conservative multiple comparisons correction.

Whilst the removal of missing-data voxels in the small-sample setting typically has a negligible effect, in the large-sample setting omitting voxels can profoundly influence results of an analysis, often 'deleting' large chunks of the final images produced. The reason that the severity of this issue becomes notably more pronounced in the large-sample setting is that the probability of a given voxel being missing in at least one image increases with the number of images in the analysis. To address this issue,

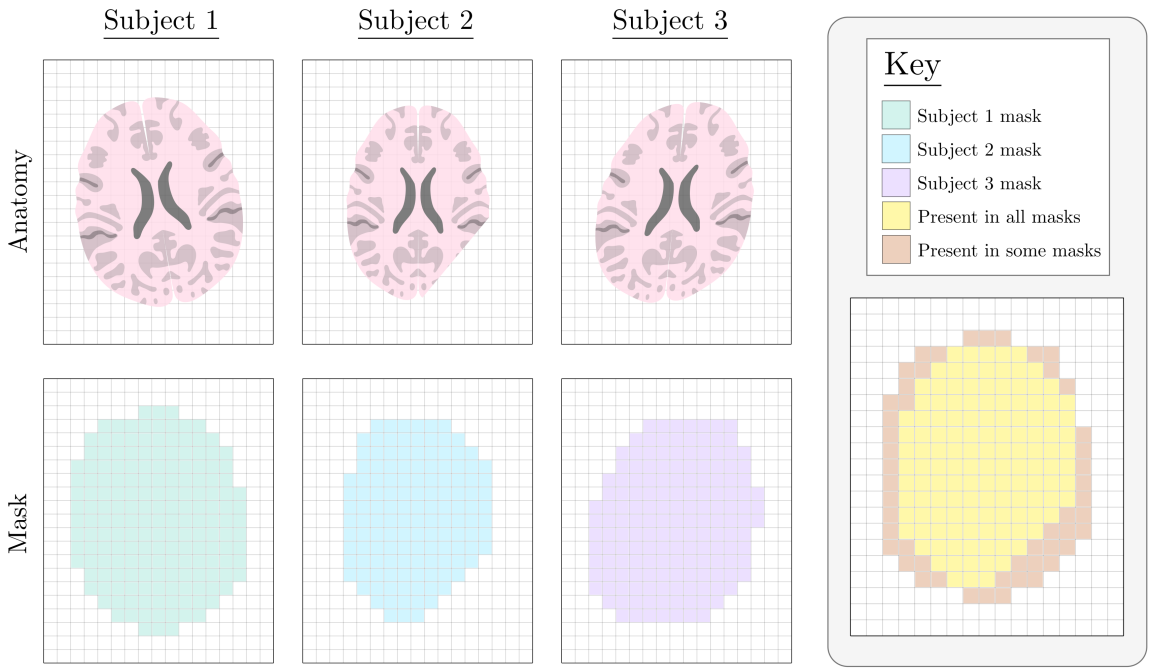

Figure 2.9: Illustration of subject-level mask variability. Displayed on the left are representations of three subject's anatomical images following preprocessing, with the corresponding analysis masks below. On the right hand side is an illustration of the overlap of the subject masks. Using standard fMRI tools, only voxels displayed in yellow would be retained for analysis, and the brown voxels would be discarded. In this example, the spatial variability observed could have arisen from poor image co-registration for Subject 1 and Subject 3 and conservative brain extraction for Subject 2.

the patterns of missingness observed for each voxel must be carefully considered and accounted for in large-sample analyses. In Chapter 4, we shall explore this issue further by explicitly accounting for such missingness in the setting of mass-univariate LMM fMRI analysis.

2.3.4 The Multiple Comparisons Problem

As mentioned in Chapter 1, in a typical fMRI analysis approximately 3.3×10^5 independent statistical tests are conducted (one per voxel). If the null hypothesis were true for all voxels and a statistic image was thresholded using a standard 5% uncorrected significance level, it would be expected that 1.7×10^4 voxels ($\approx 0.05 \times 3.3 \times 10^5$ voxels) in the image would be incorrectly declared significant. In general, the more

voxels that are in an image, the more false positives will be observed. This issue is known as the multiple comparisons problem.

To control for the multiple comparisons problem, the choice of threshold, c must be adjusted in order to reflect the number of tests being conducted. Two conventional approaches exist for performing such an adjustment. The first approach is to control the False Discovery Rate (FDR) and the second is to control the Family Wise Error (FWE). Formally, the FWE and FDR are defined as $FDR = \mathbb{E}[V/R]$ and $FWE = \mathbb{P}[V \geq 1]$, where V is the number of false positives and R is the number of tests for which the null hypothesis was rejected. To say that an analysis "controls for multiple comparisons at the α significance level" is equivalent to saying that either $FWE \leq \alpha$ or $FDR \leq \alpha$.

Multiple comparison corrections which attempt to control the number of voxels which are falsely declared significant (via either FDR or FWE) are known as voxel-wise methods. One popular voxel-wise FWE approach is the Bonferonni method which divides the significance level, α, by the number of voxels, $|V|$, to obtain that:

$$\mathbb{P}\left[\max_{v \in V}(p_v) \leq \frac{\alpha}{|V|}\right] \leq \sum_{v \in V} \mathbb{P}\left[p_v \leq \frac{\alpha}{|V|}\right] = \alpha$$

under the assumption that the null hypothesis is true for all voxels, where p_v is the p-value derived from the statistic at voxel v.

A second popular approach to controlling the voxel-wise FWE is to employ Random Field Theory (RFT). A standard RFT approach to controlling the voxel-wise FWE is to employ topological properties of the $T-$statistic to make the below approximation;

$$\mathbb{P}\left[\max_{v \in V}(T_v) > c\right] \approx \mathbb{E}[\chi(\mathcal{T}_c)]$$

where $\mathcal{T}_c = \{v \in V : T_v \geq c\}$ and χ is Euler Characteristic (a measure of topological structure derived from the number of 'holes', 'handles' and 'connected components' in the set $\mathcal{T}_c$). Another approach to controlling the voxel-wise FWE is to employ permutation-based methodology. As such methodology is not employed in this work,

we shall not detail it here but refer to Nichols and Hayasaka (2003) and Good (2013) for further detail. For voxel-wise FDR, one of the most popular procedures is the Benjamini-Hochberg process. For brevity, again this shall not be detailed fully here (c.f. Benjamini and Hochberg (1995) for further detail).

An alternative to voxel-wise procedure is to adopt a cluster-wise procedure. Conceptually, the idea behind a cluster-wise procedure this is that small, noisier regions of activation are likely significant by chance, whereas larger contiguous regions of space are more likely to reflect true activation. For this reason, given an excursion set (the set of voxels, or spatial locations, over which the function of interest exceeds a threshold c), a cluster-wise approach treats small 'clusters' (disconnected components in the excursion set) as false positives. Again by employing RFT, the conventional cluster-wise FWE procedure controls the maximum cluster size in the image as follows:

$$\mathbb{P}\left[\max_{i\in I}(C_i) \geq k\right] = 1 - e^{-\mathbb{E}[|I|]\mathbb{P}[C\geq k]}$$

where I is the set of clusters in the excursion set, C_i is the size of the i^{th} cluster and $\mathbb{P}[C \geq k]$ is the probability of an arbitrary cluster being larger than k voxels in size (Friston et al. (1994)). Typically, $\mathbb{E}[|I|]$ is estimated using smoothness properties of the image and $\mathbb{P}[C \geq k]$ is approximated using an exponential distribution for $C^{2/D}$. Despite having been a source of contention in recent years (Eklund et al. (2016), Flandin and Friston (2019)) and suffering from the arbitrary requirement of a "cluster-forming" threshold (the value of c used to define the excursion set is arbitrary), RFT cluster-wise FWE is one of the most popular approaches currently employed for multiple comparison correction.

Many other noteworthy methods exist for controlling for multiple comparisons. Such methods include; peakwise inference, cluster mass inference, all resolutions inference and threshold free cluster enhancement. Peakwise inference employs properties of local maxima to control the 'peakwise-FWE' (Schwartzman et al. (2011)). Cluster mass inference controls the 'cluster-mass-FWE' by utilizing information about excursion set intensity volumes (Zhang et al. (2009)). All resolutions inference recur-

sively controls the voxel-wise FWE across different resolutions of inference (Rosenblatt et al. (2018)). And threshold free cluster enhancement employs an altogether different approach involving integrals of weighted cluster sizes taken across a range of cluster-forming thresholds (Smith and Nichols (2009)).

The above survey of multiple comparisons corrections serves to highlight how spatial and topological properties of fMRI data are incorporated into statistical analysis. It is worth emphasizing that until this stage of the statistical analysis, every voxel in the image has been treated as independent of the rest, with no notion of spatial uncertainty being incorporated in the analysis. In general, despite many of the above methods drawing on the spatial properties of excursion sets, it is rare for the conclusions of a neuroimaging analysis to discuss the variability in the size, shape and locale of such regions. To account for this issue, the theory of confidence regions was developed.

2.3.5 Confidence Regions

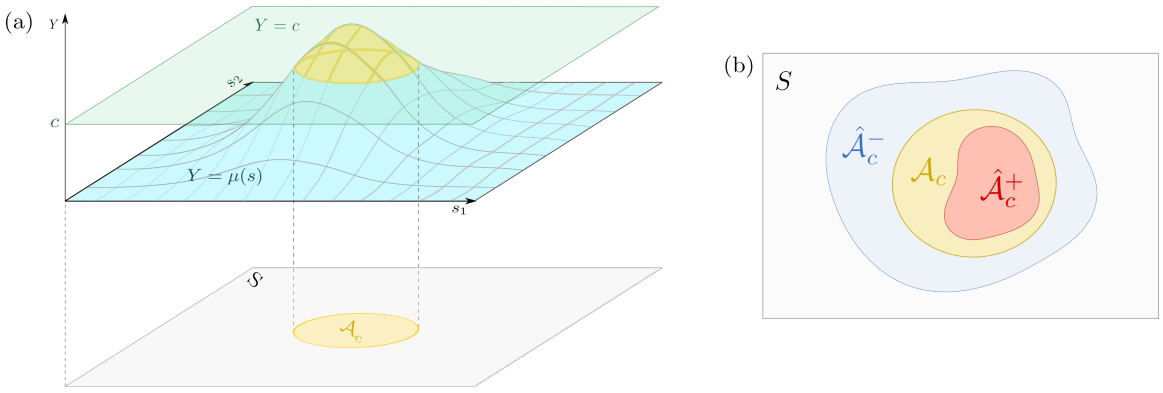

Figure 2.10: A two-dimensional illustration of the set $\mathcal{A}_c$. Shown in (a) is the spatially-varying target function (blue), $\mu(s)$, thresholded at the level c. In (b), the spatial region over which $\mu(s) \geq c$, $\mathcal{A}_c$, is illustrated in yellow with potential confidence regions, $\hat{\mathcal{A}}_c^+$ and $\hat{\mathcal{A}}_c^-$, illustrated in red and blue, respectively. In an fMRI context, S may be thought of as the brain mask used for analysis, the function $\mu(s)$ may be thought of as group-level BOLD response across the brain, and $\mathcal{A}_c$ may be thought of as the regions for which significant BOLD response values were observed.

The theory of confidence regions was first proposed in Sommerfeld et al. (2018) and later applied to neuroimaging by Bowring et al. (2019). To describe the theory,

we first assume that interest lies in a spatially varying target function $\mu(s) : S \to \mathbb{R}$ for which we have an estimate, $\hat{\mu}_n(s) : S \to \mathbb{R}$, drawn from n observations. In the notation of the previous sections, $\mu(s)$ may be a contrast estimate of the form $L\beta(s)$. The theory of confidence regions aims to provide sets which bound the excursion set $\mathcal{A}_c = \{s \in S : \mu(s) \geq c\}$ with a desired confidence. More specifically, for a given confidence level α (e.g. $\alpha = 0.05$), the theory aims use the estimate $\hat{\mu}_n$ to define sets $\hat{\mathcal{A}}_c^+$ and $\hat{\mathcal{A}}_c^-$ such that:

$$\lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{A}}_c^+ \subseteq \mathcal{A}_c \subseteq \hat{\mathcal{A}}_c^-] = 1 - \alpha, \tag{2.9}$$

An illustration of how such sets may appear is given by Fig. 2.10. For an arbitrary spatial set $B \subseteq S$, we now denote its complement, boundary and interior as $B^c, \partial B$ and B°, respectively. To define sets $\hat{\mathcal{A}}_c^+$ and $\hat{\mathcal{A}}_c^-$ satisfying the above, Sommerfeld et al. (2018) assumed the following:

Assumption 2.1. Every open ball around a point in $\partial\mathcal{A}_c$ has a nonempty intersection with $(\mathcal{A}_c)^\circ$ and with $(\mathcal{A}_c)^c$. Further, there exists $\eta_0 > 0$ such that the target function $\mu(s)$ is continuous on the set $U = \{s \in S : |\mu(s) - c| \leq \eta_0\}$.

Assumption 2.2. For $n \in \mathbb{N}$ large enough, the estimators $\hat{\mu}_n(s)$ are continuous on U. Further, there exist a sequence of numbers $\tau_n \to 0$ and a continuous bounded function $\sigma : U \to \mathbb{R}^+$ such that;

$$\left\{ \frac{\hat{\mu}_n(s) - \mu(s)}{\tau_n \sigma(s)} \right\}_{s\in U} \xrightarrow{d} \{G(s)\}_{s\in U}$$

as $n \to \infty$, where $\xrightarrow{d}$ represents convergence in distribution and G is a Gaussian field on U with mean zero, unit variance and continuous sample paths with probability one.

Assumption 2.3. There exists a constant $C > 0$ such that for all $s \notin U$ the below holds;

$$(\hat{\mu}_n(s) - c) \cdot \operatorname{sgn}(\mu(s) - c) \geq \sigma(s)(C + Z_n(s))$$

almost surely, where $\{Z_n(s)\}_{n\in\mathbb{N}}$ is a sequence of stochastic processes on S such that $\sup_{s\in S}|Z_n(s)| = \mathcal{O}_p(\tau_n)$.

Under the above assumptions, Sommerfeld et al. (2018) defined the sets $\hat{\mathcal{A}}_c^-$ and $\hat{\mathcal{A}}_c^+$ as follows;

$$\hat{\mathcal{A}}_c^+ := \left\{ s \in S : \frac{\hat{\mu}_n(s) - c}{\tau_n \sigma(s)} \geq +a \right\} \quad \text{and} \quad \hat{\mathcal{A}}_c^- := \left\{ s \in S : \frac{\hat{\mu}_n(s) - c}{\tau_n \sigma(s)} \geq -a \right\},$$

and provided proofs of the following result:

$$\lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{A}}_c^+ \subseteq \mathcal{A}_c \subseteq \hat{\mathcal{A}}_c^-] = \mathbb{P}\left[\sup_{s\in\partial\mathcal{A}_c} |G(s)| \leq a \right].$$

The above result is valuable as it demonstrates that confidence regions which satisfy Equation (2.9) may be generated provided the distribution of $\sup_{s\in\partial\mathcal{F}_c}|G(s)|$ is known. Via the use of a Gaussian multiplier bootstrap, Sommerfeld et al. (2018) demonstrated that this distribution could be evaluated empirically in an efficient manner when data were sampled from a spatially-varying LM. This procedure was later refined by Bowring et al. (2019) to instead employ a wild t-bootstrap for improved asymptotic performance. Both Sommerfeld et al. (2018) and Bowring et al. (2019) report evidence of convergence of the method for approximately 80 subjects under a conventional confidence level of $\alpha = 0.05$.

Such confidence regions possess several features which make them desirable for consideration in fMRI analysis. Firstly, there is the guarantee that across many samples they will, on average, accurately capture the set $\mathcal{A}_c$ with a desired probability. And secondly, both of the confidence regions will increasingly resemble the estimated set $\hat{\mathcal{A}}_c := \{s \in S : \hat{\mu}_n(s) \geq c\}$ as the sample size increases (this can be seen by considering the role of τ_n in the definitions of $\hat{\mathcal{A}}_c^-$ and $\hat{\mathcal{A}}_c^+$). It follows from these observations that confidence regions provide information about the reliability of $\hat{\mathcal{A}}_c$ as an estimate of $\mathcal{A}_c$. To be specific, the strength of resemblance between the confidence regions and $\hat{\mathcal{A}}_c$ indicates the reliability of $\hat{\mathcal{A}}_c$ as an estimate of $\mathcal{A}_c$.

In Chapter 5, we generalize the theory of confidence regions to ask questions not just about a single excursion set, but rather about the intersection and union of multiple excursion sets. Such questions are of natural interest as they often arise when samples acquired under different study conditions must be compared to one another. To meet this aim, we shall adjust the assumptions given above (and drop, for ease, the notation U, which was initially motivated by requirements specific to the field of climatology) to develop an extension of the above theory which may be employed to compare multiple excursion sets. We shall also provide, via simulation, evidence that the coverage results provided by the method converge for sample sizes of approximately 80 subjects or higher for realistic practical scenarios, thus confirming the findings of Sommerfeld et al. (2018) and Bowring et al. (2019).

2.4 Summary

In the previous sections, we have introduced several concepts from fMRI analysis and statistical modelling upon which the following chapters shall rely. To summarize, key points outlined in the above sections are given as follows:

- The LMM is a powerful tool commonly applied to model grouping factors and complex covariance structure in fMRI data. As no closed-form expressions exist for the parameters of the LMM, estimating the LMM parameters can be computationally intensive and performing inference for the LMM typically requires approximate distributional assumptions.

- FMRI data is extremely large. Not only does computation scale with the number of images in the analysis, n, but also with the number of voxels in an image, v. The values of n and v we are interested in are typically on the scale of 10^3 and 10^6, respectively. For this reason, when we speak of computational time efficiency in the following chapters, our interest lies not only in improving computation speeds for a single model (voxel), but streamlining computation across hundreds of thousands of models.

- FMRI data collection and preprocessing is an imperfect process. As a result, strong spatial variability is often present in preprocessed fMRI data. One consequence of this variability is missingness around the edge of the brain and cortical boundaries which can severely impact large-sample analyses.

- Another consequence of spatial variation in fMRI data is that regions of activation identified in an analysis may be subject to variation in shape, size and locale. Current fMRI statistical practices sometimes exploit spatial properties of such regions to account for multiple comparisons. However, the impact of spatial variability on an analysis is rarely considered when assessing the reliability of statistical results.

- A new approach for modelling spatial variability is provided by the theory of confidence regions. Developed by Sommerfeld et al. (2018), confidence regions have been demonstrated to provide reasonable confidence bounds on excursion sets for sample sizes as low as $n = 80$.

The following chapters will focus on addressing key issues surrounding LMM computation, large-sample fMRI analysis and spatial variability. Chapter 3 shall focus extensively on the LMM and provide new expressions which may be employed for LMM parameter estimation and inference. By utilizing the expressions provided in Chapter 3, Chapter 4 shall exploit design commonality between voxels to improve the computational performance of large-sample fMRI LMM analyses, whilst simultaneously accounting for the missingness caused by mask variability. Finally, Chapter 5 shall extend the theory of confidence regions to facilitate the comparison of multiple excursion sets acquired across the same spatial domain but under different study conditions.

Chapter 3

Fisher Scoring for Crossed Factor Linear Mixed Models

This chapter focuses on parameter estimation and inference methodology for a single univariate LMM. The content of this chapter was developed under the supervision of Professor Thomas E. Nichols and has been published in the journal 'Statistics and Computing' under the title *'Fisher Scoring for Crossed Factor Linear Mixed Models'* (c.f. Maullin-Sapey and Nichols (2021)). Proofs and further theoretical detail associated with this chapter are provided in Appendix A. Supplementary results may be found in the 'Supplementary Material A' document submitted alongside this thesis.

3.1 Background

Since its conception in the seminal work of Laird and Ware (1982), the literature on Linear Mixed Model (LMM) estimation and inference has evolved rapidly. At present, many software packages exist which are capable of performing LMM estimation and inference for large and complex LMMs in an incredibly quick and memory-efficient manner. For some packages, this exceptional speed and efficiency arises from simplifying model assumptions, whilst for others, complex mathematical operations such as

sparse matrix methodology and sweep operators are utilized to improve performance (Wolfinger et al. (1994), Bates et al. (2015)).

However, due to practical implementation concerns, the current methodology cannot be applied in certain situations. For example, in the mass-univariate analyses used in fMRI applications (c.f. Section 2.3.2), standard practice involves estimating hundreds of thousands of models concurrently. To efficiently perform a mass-univariate analysis within a practical time-frame, the use of vectorized computation which exploits the repetitive nature of simplistic operations to streamline calculation must be employed (Smith and Nichols (2018), Li et al. (2019)). Unfortunately, many existing LMM tools utilize complex operations, for which vectorized support does not currently exist. As a result, alternative methodology, using more conceptually simplistic mathematical operations for which vectorized support exists, is required.

The complex nature of LMM computation has partly arisen from the gradual expansion of the definition of "Linear Mixed Model". Previously, the term was primarily used to refer to an extension of the linear regression model containing random effects grouped by a single random factor. Examples of this definition can be seen in Laird and Ware (1982), in which "Linear Mixed Model" refers only to "single-factor" longitudinal models, and in Lindstrom and Bates (1988), where more complex, multi-factor models are described as an *"extension"* of the "Linear Mixed Model". Consequently, throughout the late 1970s and 1980s, one of the main focuses of the LMM literature was to provide parameter estimation methods such as Fisher Scoring, Newton-Raphson and Expectation Maximization for the single-factor LMM (e.g. Dempster et al. (1977), Jennrich and Schluchter (1986) and Laird et al. (1987)). By exploiting structural features of the single-factor model, implementation of these methods required only conceptually simplistic mathematical operations. For instance, the Fisher Scoring algorithm proposed by Jennrich and Schluchter (1986) relies only upon vector addition and matrix multiplication, inversion, and reshaping operations.

More recently, usage of the term "Linear Mixed Models" has grown substantially to include models which contain random effects grouped by multiple random factors. Examples of this more general definition are found in Pinheiro and Bates (2009) and

Tibaldi et al. (2007). For this more general "multi-factor" LMM definition, models can be described as exhibiting either a hierarchical factor structure (i.e. factor groupings are nested inside one another) or crossed factor structure (i.e. factor groupings are not nested inside one another). For instance, a study involving students grouped by the factors "school" and "class" contains hierarchical factors (as every class belongs to a specific school). In contrast, a study involving observations grouped by "subject" and "location," where subjects change location between recorded measurements, contains crossed factors (as each subject is not generally associated with a specific location or vice versa). In either case, the computational approaches used in the single-factor setting can not be directly applied to the more complex multi-factor model. For this reason, parameter estimation of the multi-factor LMM has often been viewed as a much more "difficult" problem than its single-factor counterpart (see, for example, the discussion in Chapters 2 and 8 of West et al. (2014)).

Several authors have proposed and implemented methods for multi-factor LMM estimation. However, such methods typically require conceptually complex mathematical operations, which are not naturally amenable to vectorized computation, or restrictive simplifying model assumptions, which prevent some designs from being estimated. For example, the popular R package lme4 performs LMM estimation via minimization of a penalized least squares cost function based on a variant of Henderson's mixed model equations (Henderson et al. (1959), Bates et al. (2015)). However, sparse matrix methods are required to achieve fast evaluation of the cost function, and advanced numerical approximation methods are needed for optimization (e.g. the Bound Optimization BY Quadratic Approximation, BOBYQA, algorithm, Powell (2009)). The commonly used SAS and SPSS packages, PROC-MIXED and MIXED, employ a Newton-Raphson algorithm proposed by Wolfinger, Tobias, and Sall (1994). In this approach, though, derivatives and Hessians must be computed using the sweep operator, W-transformation, and Cholesky decomposition updating methodology (SAS Institute Inc. (2015), IBM Corp (2015)). The Hierarchical Linear Models (HLM) package takes an alternative option and restricts its inputs to only LMMs which contain hierarchical factors (Raudenbush and Bryk (2002)); although,

the Hierarchical Cross-classified Models (HCM) sub-module does make allowances for specific use-case crossed LMMs. To perform parameter estimation, HLM employs a range of different methods, each tailored to a particular model design. As a result, there are many models for which HLM does not provide support.

Methods have also been proposed for evaluating the score vector (the derivative of the log-likelihood function) and the Fisher Information Matrix (the expected Hessian of the negative log-likelihood function) required to perform Fisher Scoring for the multi-factor LMM. For example, alongside the Newton-Raphson approach adopted by PROC-MIXED and MIXED, Wolfinger et al. (1994) also describe a Fisher Scoring algorithm. However, evaluation of the expressions they provide requires the use of the sweep operator, W-transformation, and Cholesky decomposition updating methodology, operations for which widespread vectorized support do not yet exist. More recently, expressions for score vectors and Fisher Information matrices were provided by Zhu and Wathen (2018). However, the approach Zhu and Wathen (2018) adopt to derive these expressions produces an algorithm that requires independent computation for each variance parameter in the LMM. This algorithm's serialized nature results in substantial overheads in terms of computation time, thus limiting the method's utility in practical situations where time-efficiency is a crucial consideration.

To our knowledge, no approach has yet been provided for multi-factor LMM parameter estimation, which utilizes only simplistic, universally available operations which are naturally amenable to vectorized computation. In this chapter, we revisit a Fisher Scoring approach suggested for the single-factor LMM, described in Demidenko (2013), extending it to the multi-factor setting, with the intention of revisiting the motivating example of the mass-univariate model for fMRI analysis in Chapter 4.

The novel contribution of this chapter is to provide new derivations and closed-form expressions for the score vector and Fisher Information matrix of the multi-factor LMM. We show how these expressions can be employed for LMM parameter estimation via the Fisher Scoring algorithm and can be further adapted for constrained optimization of the random effects covariance matrix. Additionally, we demonstrate

that such closed-form expressions are also of use in the setting of mixed model inference, where the degrees of freedom of the approximate T-statistic are not known and must be estimated (Verbeke and Molenberghs (2001)). We show how our derived results may be combined with the Satterthwaite-based method for approximating degrees of freedom for the LMM by using an approach based on the work of Kuznetsova et al. (2017).

In this chapter, we first propose five variants of the Fisher Scoring algorithm. Following this, we provide a discussion of initial starting values for the algorithm and methods for improving the algorithm's computational efficiency during implementation. Detail on constrained optimization, allowing for structural assumptions to be placed on the random effects covariance matrix, is then provided. Proceeding this, new expressions for Satterthwaite estimation of the degrees of freedom of the approximate T-statistic for the multi-factor LMM are given. Finally, we verify the correctness of the proposed algorithms and degrees of freedom estimates via simulation and real data examples, benchmarking the performance against the R package lme4.

3.2 Notation

In this section, we introduce notation which will be used throughout the remainder of this chapter. The hat operator is used to denote estimators resulting from likelihood maximisation procedures (e.g. β represents the true fixed effects parameter vector whilst the maximum likelihood estimate of β is denoted $\hat{\beta}$). Subscript notation, $A_{[X,Y]}$, is used to denote the sub-matrix of matrix A, composed of all elements of A with row indices $x \in X$ and column indices $y \in Y$. The replacement of X or Y with a colon, :, represents all rows or all columns of A, respectively. If a scalar, x or y, replaces X or Y, this represents the elements with row indices $x = X$ or column indices $y = Y$, respectively. The notation (k) may also replace X and Y where (k) represents the indices of the columns of Z which correspond to factor f_k (c.f. Section 2.2.3). Similarly, X and Y may be substituted for the ordered pair (k, j) where (k, j)

represents the indices of the columns of Z which correspond to level j of factor f_k. We highlight again the notation presented earlier in Section 2.2.3, $Z_{(k,j)}$, which, due to its frequent occurrence acts as a shorthand for $Z_{[:,(k,j)]}$, i.e. the columns of Z corresponding to level j of factor f_k.

Finally, we shall also adopt the notations 'vec', 'vech', N_k, $K_{m,n}$, $\mathcal{D}_k$ and $\mathcal{L}_k$ as used in Magnus and Neudecker (1980), defined as follows:

- 'vec' represents the mathematical vectorizarion operator which transforms an arbitrary $(k \times k)$ matrix, A, to a $(k^2 \times 1)$ column vector, vec(A), composed of the columns of A stacked into a column vector. (Note: this is different to the concept of computational vectorization discussed in Section 3.1).

- 'vech' represents the half-vectorization operator which transforms an arbitrary square matrix, A, of dimension $(k \times k)$ to a $(k(k+1)/2 \times 1)$ column vector, vech(A), composed by stacking the elements of A which fall on and below the diagonal into a column vector.

- N_k is defined as the unique matrix of dimension $(k^2 \times k^2)$ which implements symmetrization for any arbitrary square matrix A of dimension $(k \times k)$ in vectorized form. I.e. N_k satisfies the following relation:

$$N_k \text{vec}(A) = \text{vec}(A + A')/2.$$

- $K_{m,n}$ is the unique "Commutation" matrix of dimension $(mn \times mn)$, which permutes, for any arbitrary matrix A of dimension $(m \times n)$, the vectorization of A to obtain the vectorization of the transpose of A. I.e. $K_{m,n}$ satisfies the following relation:

$$\text{vec}(A) = K_{m,n} \text{vec}(A').$$

- $\mathcal{D}_k$ is the unique "Duplication" matrix of dimension $(k^2 \times k(k+1)/2)$, which maps the half-vectorization of any arbitrary symmetric matrix A of dimension

$(k \times k)$ to it's vectorization. I.e. $\mathcal{D}_k$ satisfies the following relation:

$$\text{vec}(A) = \mathcal{D}_k \text{vech}(A).$$

- $\mathcal{L}_k$ is the unique $1-0$ "Elimination" matrix of dimension $(k(k+1)/2 \times k^2)$, which maps the vectorization of any arbitrary lower triangular matrix A of dimension $(k \times k)$ to it's half-vectorization. I.e. $\mathcal{L}_k$ satisfies the following relation:

$$\text{vech}(A) = \mathcal{L}_k \text{vec}(A).$$

To help track the notational conventions employed in this work, a Nomenclature is provided at the end of this thesis.

3.3 Methods

3.3.1 Fisher Scoring Algorithms

In this section, we employ the Fisher Scoring algorithm for ML estimation of the parameters (β, σ^2, D). The Fisher Scoring algorithm update rule takes the following form:

$$\theta_{s+1} = \theta_s + \alpha_s \mathcal{I}(\theta_s)^{-1} \frac{dl(\theta_s)}{d\theta}, \tag{3.1}$$

where θ_s is the vector of parameter estimates given at iteration s, α_s is a scalar step size, the score vector of θ_s, $\frac{dl(\theta_s)}{d\theta}$, is the derivative of the log-likelihood with respect to θ evaluated at $\theta = \theta_s$, and $\mathcal{I}(\theta_s)$ is the Fisher Information matrix of θ_s;

$$\mathcal{I}(\theta_s) = \mathbb{E}\left[\left(\frac{dl(\theta)}{d\theta} \right) \left(\frac{dl(\theta)}{d\theta} \right)' \middle| \theta = \theta_s \right].$$

A more general formulation of Fisher Scoring, which allows for low-rank Fisher Information matrices, is given by Rao and Mitra (1972):

$$\theta_{s+1} = \theta_s + \alpha_s \mathcal{I}(\theta_s)^+ \frac{dl(\theta_s)}{d\theta}, \tag{3.2}$$

where superscript plus, $^+$, is the Moore-Penrose (or "Pseudo") inverse. For notational brevity, when discussing algorithms of the form (3.1) and (3.2) in the following sections, the subscript s, representing iteration number, will be suppressed unless its inclusion is necessary for clarity.

For the LMM, several different representations of the parameters of interest, (β, σ^2, D), can be used for numerical optimization and result in different Fisher Scoring iteration schemes. In this section, we consider the following three representations for θ:

$$\theta^h = \begin{bmatrix} \beta \\ \sigma^2 \\ \text{vech}(D_1) \\ \vdots \\ \text{vech}(D_r) \end{bmatrix}, \theta^f = \begin{bmatrix} \beta \\ \sigma^2 \\ \text{vec}(D_1) \\ \vdots \\ \text{vec}(D_r) \end{bmatrix}, \theta^c = \begin{bmatrix} \beta \\ \sigma^2 \\ \text{vech}(\Lambda_1) \\ \vdots \\ \text{vech}(\Lambda_r) \end{bmatrix},$$

where Λ_k represents the lower triangular Cholesky factor of D_k, such that $D_k = \Lambda_k \Lambda_k'$. We will refer to the representations (θ^h, θ^f and θ^c) as the "half", "full" and "Cholesky" representations of (β, σ^2, D), respectively. In a slight abuse of notation, the function l will be allowed to take any representation of θ as input, with the interpretation unchanged (i.e. $l(\theta^f) = l(\theta^h) = l(\theta^c)$). For example, if the full representation is being used, the log-likelihood will be denoted $l(\theta^f)$, but if the half representation is being used, the log likelihood will be denoted $l(\theta^h)$.

In the following sections, the score vectors and Fisher Information matrices required to perform five variants of Fisher Scoring for the multi-factor LMM will be stated with proofs provided in Appendices A.1 and A.3. For notational convenience, we denote the sub-matrix of the Fisher Information matrix of θ^h with row indices corresponding to parameter vector a and column indices corresponding to parameter

vector b as $\mathcal{I}^h_{a,b}$. In other words, $\mathcal{I}^h_{a,b}$ is the sub-matrix of $\mathcal{I}(\theta^h)$, defined by:

$$\mathcal{I}^h_{a,b} = \mathbb{E}\left[\left(\frac{dl(\theta^h)}{da}\right)\left(\frac{dl(\theta^h)}{db}\right)'\middle|\theta = \theta_s\right].$$

For further simplification of notation, when a and b are equal, the second subscript will be dropped and the matrix $\mathcal{I}^h_{a,b} = \mathcal{I}^h_{a,a}$ will be denoted simply as $\mathcal{I}^h_a$. Analogous notation is used for the full and Choleksy representations.

3.3.1.1 Fisher Scoring

The first variant of Fisher Scoring we provide uses the "half"-representation for (β, σ^2, D), θ^h, and is based on the standard form of Fisher Scoring given by Equation (3.1). This may be considered the most natural approach for a Fisher Scoring algorithm as θ^h is an unmodified vector of the unique parameters of the LMM and (3.1) is the standard update rule. For this approach the elements of the score vector are:

$$\frac{dl(\theta^h)}{d\beta} = \sigma^{-2}X'V^{-1}e, \tag{3.3}$$

$$\frac{dl(\theta^h)}{d\sigma^2} = -\frac{n}{2}\sigma^{-2} + \frac{1}{2}\sigma^{-4}e'V^{-1}e. \tag{3.4}$$

For $k \in \{1, \ldots, r\}$:

$$\frac{dl(\theta^h)}{d\text{vech}(D_k)} = \frac{1}{2}\mathcal{D}'_{q_k}\text{vec}\left(\sum_{j=1}^{l_k} Z'_{(k,j)}V^{-1}\left(\frac{ee'}{\sigma^2} - V\right)V^{-1}Z_{(k,j)}\right). \tag{3.5}$$

and the entries of the Fisher Information matrix are given by:

$$\mathcal{I}^h_\beta = \sigma^{-2}X'V^{-1}X, \quad \mathcal{I}^h_{\beta,\sigma^2} = \mathbf{0}_{p,1}, \quad \mathcal{I}^h_{\sigma^2} = \frac{n}{2}\sigma^{-4}. \tag{3.6}$$

For $k \in \{1, \ldots, r\}$:

$$\begin{aligned}
\mathcal{I}^h_{\beta,\text{vech}(D_k)} &= \mathbf{0}_{p,q_k(q_k+1)/2}, \\
\mathcal{I}^h_{\sigma^2,\text{vech}(D_k)} &= \frac{1}{2}\sigma^{-2}\text{vec}'\Bigg(\sum_{j=1}^{l_k} Z'_{(k,j)}V^{-1}Z_{(k,j)}\Bigg)\mathcal{D}_{q_k}.
\end{aligned} \tag{3.7}$$

For $k_1, k_2 \in \{1, \ldots, r\}$:

$$\mathcal{I}^h_{\text{vech}(D_{k_1}),\text{vech}(D_{k_2})} = \frac{1}{2}\mathcal{D}'_{q_{k_1}}\sum_{j=1}^{l_{k_2}}\sum_{i=1}^{l_{k_1}}(Z'_{(k_1,i)}V^{-1}Z_{(k_2,j)} \otimes Z'_{(k_1,i)}V^{-1}Z_{(k_2,j)})\mathcal{D}_{q_{k_2}}. \tag{3.8}$$

where $\mathbf{0}_{n,k}$ denotes the $(n \times k)$ dimensional matrix of zeros. Due to its to natural approach to Fisher Scoring, this algorithm is referred to as FS in the remainder of this text. Pseudocode for the FS algorithm is given in Algorithm 1. Discussion of the initial estimates used in the algorithm is deferred to Section 3.3.2. To ensure D is non-negative definite, a commonly employed Eigendecomposition based approach is used (c.f. Appendix A.6).

Algorithm 1: Fisher Scoring (FS)

1 Assign θ^h to an initial estimate using Equations (2.2) and (3.20).
2 **while** *current* $l(\theta^h)$ *and previous* $l(\theta^h)$ *differ by more than a predefined tolerance* **do**
3 Calculate $\frac{dl(\theta^h)}{d\theta^h}$ using Equations (3.3)-(3.5).
4 Calculate $\mathcal{I}(\theta^h)$ using Equations (3.6)-(3.8).
5 Assign $\theta^h = \theta^h + \alpha\mathcal{I}(\theta^h)^{-1}\frac{dl(\theta^h)}{d\theta^h}$.
6 Project D to be non-negative definite.
7 Assign $\alpha = \frac{\alpha}{2}$ if $l(\theta^h)$ has decreased in value.
8 **end**

3.3.1.2 Full Fisher Scoring

The second variant of Fisher Scoring considered in this chapter uses the "full", θ^f, representation of the model parameters, and shall therefore be referred to as "Full

Fisher Scoring" (FFS). In this approach, for each factor, f_k, the elements of $\text{vec}(D_k)$ are to be treated as distinct from one another with numerical optimization for D_k performed over the space of all $(q_k \times q_k)$ matrices. This approach differs from the FS method proposed in the previous section, in which optimization was performed on the space of only those $(q_k \times q_k)$ matrices that are symmetric. This optimization procedure is realized by treating symmetric matrix elements of D_k as distinct and, for a given element, using the partial derivative with respect to the element during optimization instead of the total derivative with respect to both the element and its symmetric counterpart. This change is reflected by denoting the elements of the score vector which correspond to $\text{vec}(D_k)$ using a partial derivative operator, ∂, rather than the total derivative operator, d. The primary motivation for the inclusion of the FFS approach is that it serves as a basis for which the constrained covariance approaches of Section 3.3.4 can be built upon. However, it should be noted that, as it does not require the construction or use of duplication matrices, FFS also provides simplified expressions and potential improvement in terms of computation speed. As a result, FFS is of some theoretical and practical interest and is detailed in full here.

An immediate obstacle to this approach is that the Fisher Information matrix of θ^f is rank-deficient, and, therefore, cannot be inverted. Intuitively, this is to be expected, as repeated entries in θ^f result in repeated rows in $\mathcal{I}(\theta^f)$. Mathematically, this can be seen by noting that $\mathcal{I}^f_{\text{vec}(D_k)}$ can be expressed as a product containing the matrix N_{q_k} (defined in Section 3.2), which is low-rank by construction. Consequently, the Fisher Scoring update rule for θ^f must be based on the Pseudo-inverse formulation of Fisher Scoring given in Equation (3.2).

As the derivatives of the log-likelihood with respect to β and σ^2 do not depend upon the parameterisation of D, both FFS and FS employ the same expressions for the elements of the score vector which correspond to β and σ^2, given by Equations (3.3) and (3.4) respectively. The score vector for $\{\text{vec}(D_k)\}_{k\in\{1,\dots r\}}$ used by FFS is given by:

$$\frac{\partial l(\theta^f)}{\partial \text{vec}(D_k)} = \frac{1}{2}\text{vec}\left(\sum_{j=1}^{l_k} Z'_{(k,j)} V^{-1}\left(\frac{ee'}{\sigma^2} - V\right)V^{-1}Z_{(k,j)}\right). \tag{3.9}$$

The entries of the Fisher Information matrix of θ^f, based on the derivatives given in Equations (3.3), (3.4) and (3.9), are given by:

$$\mathcal{I}^f_{\beta} = \mathcal{I}^h_{\beta}, \qquad \mathcal{I}^f_{\beta,\sigma^2} = \mathcal{I}^h_{\beta,\sigma^2}, \qquad \mathcal{I}^f_{\sigma^2} = \mathcal{I}^h_{\sigma^2}.$$

For $k \in \{1, \ldots, r\}$:

$$\begin{aligned} \mathcal{I}^f_{\beta,\mathrm{vec}(D_k)} &= \mathbf{0}_{p,q_k^2}, \\ \mathcal{I}^f_{\sigma^2,\mathrm{vec}(D_k)} &= \frac{1}{2}\sigma^{-2}\mathrm{vec}'\Bigg(\sum_{j=1}^{l_k} Z'_{(k,j)}V^{-1}Z_{(k,j)}\Bigg). \end{aligned} \tag{3.10}$$

For $k_1, k_2 \in \{1, \ldots, r\}$:

$$\mathcal{I}^f_{\mathrm{vec}(D_{k_1}),\mathrm{vec}(D_{k_2})} = \frac{1}{2}\sum_{j=1}^{l_{k_2}}\sum_{i=1}^{l_{k_1}}(Z'_{(k_1,i)}V^{-1}Z_{(k_2,j)} \otimes Z'_{(k_1,i)}V^{-1}Z_{(k_2,j)})N_{q_k}. \tag{3.11}$$

Derivations for the above can be found in Appendices A.1 and A.3. It can also be seen that the Full Fisher Scoring algorithm can also be expressed in the form:

$$\theta^f_{s+1} = \theta^f_s + \alpha_s F(\theta^f_s)^{-1}\frac{\partial l(\theta^f_s)}{\partial \theta}, \tag{3.12}$$

where, unlike in Equations (3.1) and (3.2), $F(\theta^f)$ is not the Fisher Information matrix. Rather, $F(\theta^f)$ is a matrix of equal dimensions to $\mathcal{I}(\theta^f)$ with all of its elements equal to those of $\mathcal{I}(\theta^f)$, apart from those which were specified by Equation (3.11), which instead are, for $k_1, k_2 \in \{1, \ldots, r\}$:

$$F_{\mathrm{vec}(D_{k_1}),\mathrm{vec}(D_{k_2})} = \frac{1}{2}\sum_{j=1}^{l_{k_2}}\sum_{i=1}^{l_{k_1}}(Z'_{(k_1,i)}V^{-1}Z_{(k_2,j)} \otimes Z'_{(k_1,i)}V^{-1}Z_{(k_2,j)}) \quad , \tag{3.13}$$

where the same subscript notation has been adopted to index $F(\theta^f)$ as was adopted for $\mathcal{I}(\theta^f)$. This alternative representation of the FFS algorithm can be derived di-

rectly using well-known properties of the commutation matrix (c.f. Appendix A.4). Pseudocode for the FFS algorithm using the representation of the update rule given by Equation (3.12), is provided by Algorithm 2.

Algorithm 2: Full Fisher Scoring (FFS)

Assign θ^f to an initial estimate using Equations (2.2) and (3.20).
while *current* $l(\theta^f)$ *and previous* $l(\theta^f)$ *differ by more than a predefined tolerance* **do**
    Calculate $\frac{\partial l(\theta^f)}{\partial \theta^f}$ using Equations (3.3),(3.4) and (3.9).
    Calculate $F(\theta^f)$ using Equations (3.10) and (3.13).
    Assign $\theta^f = \theta^f + \alpha F(\theta^f)^{-1} \frac{\partial l(\theta^f)}{\partial \theta^f}$.
    Project D to be non-negative definite.
    Assign $\alpha = \frac{\alpha}{2}$ if $l(\theta^f)$ has decreased in value.
end

3.3.1.3 Simplified Fisher Scoring

The third Fisher Scoring algorithm proposed in this chapter represents the parameters (β, σ^2, D) using θ^h, the half-representation, and takes an approach, similar to that of coordinate ascent, which is commonly adopted in the single-factor setting (c.f. Demidenko (2013)). In this approach, instead of performing a single update step upon the entire vector θ^h in the form of (3.1), updates for β, σ^2 and $\{D_k\}_{k\in\{1,\dots,r\}}$ are performed individually in turn. For β and σ^2 each iteration uses the standard Generalized Least Squares (GLS) estimators given by:

$$\beta_{s+1} = (X'V_s^{-1}X)^{-1}X'V_s^{-1}Y, \quad \sigma^2_{s+1} = \frac{e'_{s+1}V_s^{-1}e_{s+1}}{n}. \tag{3.14}$$

To update the random effects covariance matrix, D_k, for each factor, f_k, individual Fisher Scoring updates are applied to $\text{vech}(D_k)$. These updates are performed using the block of the Fisher Information matrix corresponding to $\text{vech}(D_k)$, given

by Equation (3.8), and take the following form for $k \in \{1, \ldots, r\}$:

$$\text{vech}(D_{k,s+1}) = \text{vech}(D_{k,s}) + \alpha_s \left(\mathcal{I}^h_{\text{vech}(D_{k,s})}\right)^{-1} \frac{dl(\theta^h_s)}{d\text{vech}(D_{k,s})}. \tag{3.15}$$

In line with the naming convention used in Demidenko (2013), this method shall be referred to as "Simplified" Fisher Scoring (SFS). This is due to the relative simplicity, both in terms of notational and computational complexity, of the updates (3.14) and (3.15) used in the SFS algorithm in comparison to those used in the FS algorithm of Section 3.3.1.1, given by Equations (3.3)-(3.8). Pseudocode for the SFS algorithm is given by Algorithm 3.

Algorithm 3: Simplified Fisher Scoring (SFS)

```
Assign θ^f to an initial estimate using Equations (2.2) and (3.20).
while current l(θ^h) and previous l(θ^h) differ by more than a predefined tolerance do
    Update β and σ^2 using Equation (3.14).
    for k ∈ {1, ...r} do
        Update vech(D_k) using Equation (3.15).
        Project D_k to be non-negative definite.
    end
    Assign α = α/2 if l(θ^h) has decreased in value.
end
```

3.3.1.4 Full Simplified Fisher Scoring

The Full Simplified Fisher Scoring algorithm (FSFS) combines the "Full" and "Simplified" approaches described in Sections 3.3.1.2 and 3.3.1.3. In the FSFS algorithm individual updates are applied to β and σ^2 using Equation (3.14) and to $\{\text{vec}(D_k)\}_{k\in\{1,\ldots,r\}}$ using a Fisher Scoring update step, based on the matrix $F_{\text{vec}(D_k)}$ given by Equation (3.13). The update rule for $\{\text{vec}(D_k)\}_{k\in\{1,\ldots,r\}}$ takes the following form:

$$\text{vec}(D_{k,s+1}) = \text{vec}(D_{k,s}) + \alpha_s F^{-1}_{\text{vec}(D_{k,s})} \frac{\partial l(\theta^f_s)}{\partial \text{vec}(D_{k,s})}. \tag{3.16}$$

We note that the above can be seen to be equivalent to an update rule of the form (3.2), given by:

$$\text{vec}(D_{k,s+1}) = \text{vec}(D_{k,s}) + \alpha_s \big(\mathcal{I}^f_{\text{vec}(D_{k,s})}\big)^+ \frac{\partial l(\theta^f_s)}{\partial \text{vec}(D_{k,s})}.$$

Justification of this claim can be found in Appendix A.4. Pseudocode for the FSFS algorithm is given by Algorithm 4.

Algorithm 4: Full Simplified Fisher Scoring (FSFS)

```
Assign θ^f to an initial estimate using Equations (2.2) and (3.20).
while current l(θ^f) and previous l(θ^f) differ by more than a predefined tolerance do
    Update β and σ^2 using Equation (3.14).
    for k ∈ {1, ...r} do
        Update vec(D_k) using Equation (3.16).
        Project D_k to be non-negative definite.
    end
    Assign α = α/2 if l(θ^f) has decreased in value.
end
```

3.3.1.5 Cholesky Simplified Fisher Scoring

The final variant of the Fisher Scoring algorithm we consider is based on the "simplified" approach described in Section 3.3.1.3 and uses the Cholesky parameterisation of (β, σ^2, D), θ^c. This approach follows directly from the below application of the chain rule of differentiation for vector-valued functions,

$$\frac{dl(\theta^c)}{d\text{vech}(\Lambda_k)} = \frac{\partial \text{vech}(D_k)}{\partial \text{vech}(\Lambda_k)} \frac{\partial l(\theta^c)}{\partial \text{vech}(D_k)}.$$

An expression for the derivative which appears second in the above product was given by Equation (3.5). It therefore follows that in order to evaluate the score vector of $\text{vech}(\Lambda_k)$ (i.e. the derivative of l with respect to $\text{vech}(\Lambda_k)$), only an expression

for the first term of the above product is required. This term can be evaluated to $\mathcal{L}_{q_k}(\Lambda_k' \otimes I_{q_k})(I_{q_k^2} + K_{q_k})\mathcal{D}_{q_k}$, proof of which is provided in Appendix A.8.

Through similar arguments to those used to prove Corollaries A.4-A.6 of Appendix A.3, it can be shown that the Fisher Information matrix for θ^c is given by:

$$\mathcal{I}^c_{\beta} = \mathcal{I}^h_{\beta}, \quad \mathcal{I}^c_{\beta,\sigma^2} = \mathcal{I}^h_{\beta,\sigma^2}, \quad \mathcal{I}^c_{\sigma^2} = \mathcal{I}^h_{\sigma^2}.$$

For $k \in \{1, \ldots, r\}$:

$$\begin{aligned} \mathcal{I}^c_{\beta,\text{vech}(\Lambda_k)} &= \mathbf{0}_{p,q_k(q_k+1)/2}, \\ \mathcal{I}^c_{\sigma^2,\text{vech}(\Lambda_k)} &= \mathcal{I}^h_{\sigma^2,\text{vech}(D_k)} \left(\frac{\partial \text{vech}(D_k)}{\partial \text{vech}(\Lambda_k)} \right)'. \end{aligned} \tag{3.17}$$

For $k_1, k_2 \in \{1, \ldots, r\}$:

$$\mathcal{I}^c_{\text{vech}(\Lambda_{k_1}),\text{vech}(\Lambda_{k_2})} = \left(\frac{\partial \text{vech}(D_{k_1})}{\partial \text{vech}(\Lambda_{k_1})} \right) \mathcal{I}^h_{\text{vech}(D_{k_1}),\text{vech}(D_{k_2})} \left(\frac{\partial \text{vech}(D_{k_2})}{\partial \text{vech}(\Lambda_{k_2})} \right)'. \tag{3.18}$$

From the above, it can be seen that a non-"simplified" Cholesky-based variant of the Fisher Scoring algorithm, akin to the FS and FFS algorithms described in Sections 3.3.1.1 and 3.3.1.2, could be constructed. However, preliminary tests have indicated that the performance of such an approach, in terms of computation time, is significantly worse than the previously proposed algorithms. For this reason, we only consider the "simplified" version of the Cholesky Fisher Scoring algorithm, analogous to the "simplified" approaches described in Sections 3.3.1.3 and 3.3.1.4. The Cholesky Simplified Fisher Scoring (CSFS) algorithm adopts (3.14) as the update rule for β and σ^2, and employs the following update rule for $\{\text{vech}(\Lambda_k)\}_{k \in \{1,\ldots,r\}}$.

$$\text{vech}(\Lambda_{k,s+1}) = \text{vech}(\Lambda_{k,s}) + \alpha_s \big(\mathcal{I}^c_{\text{vech}(\Lambda_{k,s})} \big)^{-1} \frac{dl(\theta^c_s)}{d\text{vech}(\Lambda_{k,s})}. \tag{3.19}$$

Pseudocode summarizing the CSFS approach is given by Algorithm 5.

Algorithm 5: Cholesky Simplified Fisher Scoring (CSFS)

Assign θ^c to an initial estimate using Equations (2.2) and (3.20).
while *current* $l(\theta^c)$ *and previous* $l(\theta^c)$ *differ by more than a predefined tolerance* **do**
    Update β and σ^2 using Equation (3.14).
    for $k \in \{1, ...r\}$ **do**
        Update $\text{vec}(\Lambda_k)$ using Equation (3.19).
    end
    Assign $\alpha = \frac{\alpha}{2}$ if $l(\theta^c)$ has decreased in value.
end

3.3.2 Initial Values

Choosing which initial values of β, σ^2 and D will be used as starting points for optimization is an important consideration for the Fisher Scoring algorithm. Denoting these initial values as β_0, σ_0^2 and D_0, respectively, this work follows the recommendations of Demidenko (2013) and evaluates β_0 and σ_0^2 using the OLS estimates given by Equation (2.2) in Section 2.2.1. For the initial estimate of $\{D_k\}_{k\in\{1,...,r\}}$, an approach similar to that suggested in Demidenko (2013) is also adopted, which substitutes V for I_n in the update rule for $\text{vec}(D_k)$, Equation (3.16). The resulting initial estimate for $\{D_k\}_{k\in\{1,...,r\}}$ is given by

$$\begin{aligned}\text{vec}(D_{k,0}) = \quad & \left(\sum_{j=1}^{l_k} Z'_{(k,j)}Z_{(k,j)} \otimes Z'_{(k,j)}Z_{(k,j)}\right)^{-1} \times \\ & \text{vec}\left(\sum_{j=1}^{l_k} Z'_{(k,j)}\left(\frac{e_0 e'_0}{\sigma_0^2} - I_n\right)Z_{(k,j)}\right). \end{aligned} \tag{3.20}$$

3.3.3 Computational Efficiency

This section provides discussion on the computational efficiency of evaluating the Fisher Information matrices and score vectors of Sections 3.3.1.1-3.3.1.5. This discussion is, in large part, motivated by the mass-univariate setting of fMRI (c.f. Sections 2.3.1 and 3.1), in which not one, but rather hundreds of thousands of models must be estimated concurrently. As a result, this discussion prioritizes both time efficiency

and memory consumption concerns and, further, operates under the assumption that sparse matrix methodology, such as that employed by the R package lme4, cannot be employed. This assumption is motivated by the current lack of support for computationally vectorized sparse matrix operations.

A primary concern, for both memory consumption and time efficiency, stems from the fact that many of the matrices used in the evaluation of the score vectors and Fisher Information matrices possess dimensions which scale with n, the number of observations. For example, V has dimensions $(n \times n)$ and is frequently inverted in the FSFS algorithm. In practice, it is not uncommon for studies to have n ranging into the thousands. As such, inverting V directly may not be computationally feasible. To address this issue, we define the "product forms" as;

$$P = X'X,\ Q = X'Y,\ R = X'Z,\ S = Y'Y,\ T = Y'Z,\ U = Z'Z. \tag{3.21}$$

Working with the product forms is preferable to using the original matrices X, Y and Z, as the dimensions of the product forms do not scale with n, but instead scale with p and q. As an example, consider the expressions below, which appear frequently in Equations (3.9) and (3.13) and have been reformulated in terms of the product forms:

$$Z'_{(k_1,i)}V^{-1}Z_{(k_2,j)} = (U - UD(I_q + DU)^{-1}U)_{[(k_1,i),(k_2,j)]},$$
$$Z'_{(k,j)}V^{-1}e = ((I_q - UD)(I_q + DU)^{-1}(T' - R'\beta))_{[(k,j),:]}.$$

For computational purposes, the right-hand side of the above expressions are much more convenient than the left-hand side. In order to evaluate the left-hand side in both cases, an $(n \times n)$ inversion of the matrix V must be performed. In contrast, the right-hand side is expressible solely in terms of the product forms and (β, σ^2, D), with the only inversion required being that of the $(q \times q)$ matrix $(I_q + DU)$. These examples can be generalized further. In fact, all of the previous expressions (3.3)-(3.20) can be rewritten in terms of only the product forms and (β, σ^2, D). This observation is

important as it implies that an algorithm for mixed model parameter estimation may begin by taking the matrices X, Y and Z as inputs, but discard them entirely once the product forms have been constructed. As a result, both computation time and memory consumption no longer scale with n.

Another potential source of concern regarding computation speed arises from noting that the algorithms we have presented contain many summations of the following two forms:

$$\sum_{i=1}^{c_0} A_i B_i' \quad \text{and} \quad \sum_{i=1}^{c_1}\sum_{i=1}^{c_2} G_{i,j} \otimes H_{i,j}. \tag{3.22}$$

where matrices $\{A_i\}$ and $\{B_i\}$ are of dimension $(m_1 \times m_2)$, and matrices $\{G_{i,j}\}$ and $\{H_{i,j}\}$ are of dimension $(n_1 \times n_2)$. We denote the matrices formed from vertical concatenation of the $\{A_i\}$ and $\{B_i\}$ matrices as A and B, respectively, and G and H the matrices formed from block-wise concatenation of $\{G_{i,j}\}$ and $\{H_{i,j}\}$, respectively. Instances of such summations can be found, for example, in Equations (3.5) and (3.8).

From a computational standpoint, summations of the forms shown in Equation (3.22) are typically realized by 'for' loops. This cumulative approach to computation can cause a potential issue for LMM computation since typically the number of summands corresponds to the number of levels, l_k, of some factor, f_k. In particular applications of the LMM, such as repeated measures and longitudinal studies, some factors may possess large quantities of levels. As this means 'for' loops of this kind could hinder computation and result in slow performance, we provide alternative methods for calculating summations of the forms shown in Equation (3.22).

For the former summation shown in Equation (3.22), we utilize the "generalized vectorisation", or "vec_m" operator, defined by Turkington (2013) as the operator which performs the below mapping for a horizontally partitioned matrix M:

$$M = \begin{bmatrix} M_1 & M_2 & \dots M_{c_0} \end{bmatrix} \rightarrow \text{vec}_m(M) = \begin{bmatrix} M_1' & M_2' & \dots M_{c_0}' \end{bmatrix}'$$

where the partitions $\{M_i\}$ are evenly sized and contain m columns. Using the definition of the generalized vectorisation operator, the former summation in Equation

(3.22) can be reformulated as:

$$\sum_{i=1}^{l} A_i B_i' = \text{vec}_{m_2}(A')'\text{vec}_{m_2}(B').$$

The right-hand side of the expression above is of practical utility as the 'vec_m' operator can be implemented efficiently. The 'vec_m' operation can be performed, for instance, using matrix resizing operations such as the 'reshape' operators commonly available in many programming languages such as MATLAB and Python. The computational costs associated with this approach are significantly lesser than those experienced when evaluating the summation directly using 'for' loops.

For the latter summation in Equation (3.22), we first define the notation $\tilde{M}$ to denote the transformation below, for the block-wise partitioned matrix M:

$$M = \begin{bmatrix} M_{1,1} & M_{1,2} & ... & M_{1,c_2} \\ M_{2,1} & M_{2,2} & ... & M_{2,c_2} \\ \vdots & \vdots & \ddots & \vdots \\ M_{c_1,1} & M_{c_1,2} & ... & M_{c_1,c_2} \end{bmatrix} \rightarrow \tilde{M} = \begin{bmatrix} \text{vec}(M_{1,1})' \\ \vdots \\ \text{vec}(M_{1,c_2})' \\ \text{vec}(M_{2,1})' \\ \vdots \\ \text{vec}(M_{c_1,c_2})' \end{bmatrix}. \tag{3.23}$$

Using a modification of Lemma 3.1 (i) of Neudecker and Wansbeek (1983), we obtain the below identity:

$$\text{vec}\left(\sum_{i,j} G_{i,j} \otimes H_{i,j} \right) = (I_{n_2} \otimes K_{n_1,n_2} \otimes I_{n_1})\text{vec}(\tilde{H}'\tilde{G}), \tag{3.24}$$

where K_{n_1,n_2} is the $(n_1 n_2 \times n_1 n_2)$ Commutation matrix (c.f. Section 3.2). The matrices $\tilde{H}$ and $\tilde{G}$ can be obtained from H and G using resizing operators with little computation time. The matrix $(I_{n_2} \otimes K_{n_1,n_2} \otimes I_{n_1})$ can be calculated via simple means and is a permutation matrix depending only on the dimensions of H and G. As a result, this matrix can be calculated once and stored as a vector. Therefore, the matrix multiplication in the above expression does not need to be performed directly,

but instead can be evaluated by permuting the elements of $\text{vec}(\tilde{H}'\tilde{G})$ according to the permutation vector representing $(I_{n_2} \otimes K_{n_1,n_2} \otimes I_{n_1})$. To summarize, evaluation of expressions of the latter form shown in Equation (3.22) can be performed by using only reshape operations, a matrix multiplication, and a permutation. This method of evaluation provides notably faster computation time than the corresponding 'for' loop evaluated over all values of i and j.

The above method, combined with the product form approach, was used to obtain the results of Sections 3.4.1-3.5.2.

3.3.4 Constrained Covariance Structure

In many applications involving the LMM, it is often desirable to use a constrained parameterisation for D_k. Examples include compound symmetry, first-order auto-regression and a Toeplitz structure. A more comprehensive list of commonly employed covariance structures in LMM analyses can be found, for example, in Wolfinger (1996). In this section, we describe how the Fisher Scoring algorithms of the previous sections can be adjusted to model dependence between covariance elements.

When a constraint is placed on the covariance matrix D_k, it is assumed that the elements of D_k can be defined as continuous, differentiable functions of some smaller parameter vector, $\text{vecu}(D_k)$. Colloquially, $\text{vecu}(D_k)$ may be thought of as the vector of "unique" parameters required to specify the constrained parameterization of D_k. To perform constrained optimization, we adopt a similar approach to the previous sections and define the constrained representation of θ as $\theta^{con} = [\beta', \sigma^2, \text{vecu}(D_1)', ...\text{vecu}(D_k)']'$. Denoting the Jacobian matrix $\partial\text{vec}(D_k)/\partial\text{vecu}(D_k)$ as $\mathcal{C}_k$, the score vector and Fisher Information matrix of θ^{con} can be constructed as follows; for $k \in \{1, \ldots, r\}$,

$$\begin{aligned} &\frac{dl(\theta^{con})}{d\text{vecu}(D_k)} = \mathcal{C}_k \frac{\partial l(\theta^{con})}{\partial \text{vec}(D_k)}, \\ &\mathcal{I}^{con}_{\beta,\text{vecu}(\tilde{D}_k)} = \mathbf{0}_{p,\tilde{q}_k}, \quad \mathcal{I}^{con}_{\sigma^2,\text{vecu}(\tilde{D}_k)} = \mathcal{I}^{f}_{\sigma^2,\text{vec}(D_k)}\mathcal{C}_k', \end{aligned} \tag{3.25}$$

and, for $k_1, k_2 \in \{1, \ldots, r\}$,

$$\mathcal{I}^{con}_{\text{vecu}(\tilde{D}_{k_1}),\text{vecu}(\tilde{D}_{k_2})} = \mathcal{C}_{k_1} \mathcal{I}^{f}_{\text{vec}(D_{k_1}),\text{vec}(D_{k_2})} \mathcal{C}'_{k_2},$$

where $\tilde{q}_k$ is the length of $\text{vecu}(D_k)$ and $\mathcal{I}^f$ is the "full" Fisher Information matrix defined in Section 3.3.1.2. The above expressions can be derived trivially using the definition of the Fisher Information matrix and the chain rule. In the remainder of this chapter, and throughout Appendix A, the matrix $\mathcal{C}_k$ is referred to as a "constraint matrix" due to the fact it "imposes" constraints during optimization. A fuller discussion of constraint matrices, alongside examples, is provided in Appendix A.7.

We note here that this approach can be extended to situations in which all covariance parameters can be expressed in terms of one set of parameters, ρ_D, common to all $\{\text{vec}(D_k)\}_{k\in\{1,\ldots,r\}}$. In such situations, a constraint matrix, $\mathcal{C}$, may be defined as the Jacobian matrix $\partial[\text{vec}(D_1)', \ldots \text{vec}(D_r)']'/\partial\rho_D$ and Fisher Information matrices and score vectors may be derived in a similar manner to the above. Models requiring this type of commonality between factors are rare in the LMM literature, since the covariance matrices $\{D_k\}_{k\in\{1,\ldots r\}}$ are typically defined independently of one another. However, one such example is provided by the ACE model employed for twin studies, in which the elements of $v(D)$ can all be expressed in terms of three variance components; $\rho_D = [\sigma^2_a, \sigma^2_c, \sigma^2_e]'$. Further information on the ACE model is presented in Section 3.5.1.2.

3.3.5 Degrees of Freedom Estimation

In the setting of the LMM, a commonly employed statistic for testing hypotheses of the form given in Section 2.2.4 is the approximate T-statistic, given by:

$$T = \frac{L\hat{\beta}}{\sqrt{\hat{\sigma}^2 L(X'\hat{V}^{-1}X)^{-1}L'}},$$

where $\hat{V} = I_n + Z\hat{D}Z'$. For LMM hypothesis testing, it is assumed that T follows a student's t_v-distribution where, for this chapter only, the subscript v shall represent degrees of freedom instead of voxel index. As noted in Dempster et al. (1981), using an estimate of D in this setting results in an underestimation of the true variability in $\hat{\beta}$. For this reason, the relation $T \sim t_v$ is not exact and is treated only as an approximating distribution (i.e. an "approximate T statistic") where the degrees of freedom, v, must be estimated empirically. A standard method for estimating the degrees of freedom, v, utilizes the Welch-Satterthwaite equation, originally described in Satterthwaite (1946) and Welch (1947), given by:

$$v(\hat{\eta}) = \frac{2(S^2(\hat{\eta}))^2}{\text{Var}(S^2(\hat{\eta}))}, \tag{3.26}$$

where $\hat{\eta}$ represents an estimate of the variance parameters $\eta = (\sigma^2, D_1, \dots D_r)$ and $S^2(\hat{\eta}) = \hat{\sigma}^2 L(X'\hat{V}^{-1}X)^{-1}L'$. Typically, as the variance estimates obtained under ML estimation are biased downwards, ReML estimation is employed to obtain the estimate of η employed in the above expression. If ML were used to estimate η instead, the degrees of freedom, $v(\hat{\eta})$, would be underestimated and, consequently, conservative p-values and a reduction in statistical power for resulting hypothesis tests would be observed.

To obtain an approximation for $v(\hat{\eta})$, a second order Taylor expansion is applied to the unknown variance on the denominator of Equation (3.26):

$$\text{Var}(S^2(\hat{\eta})) \approx \left(\frac{dS^2(\hat{\eta})}{d\hat{\eta}}\right)' \text{Var}(\hat{\eta}) \left(\frac{dS^2(\hat{\eta})}{d\hat{\eta}}\right). \tag{3.27}$$

This approach is well-documented, and has been adopted, most notably, by the R package lmerTest, first presented in Kuznetsova et al. (2017). To obtain the derivative of S^2 and asymptotic variance covariance matrix, lmerTest utilizes numerical estimation methods. As an alternative, we define the "half" representation of $\hat{\eta}$, $\hat{\eta}^h$, by $\hat{\eta}^h = [\hat{\sigma}^2, \text{vech}(\hat{D}_1)', \dots, \text{vech}(\hat{D}_r)']'$ and present the below exact closed-form

expression for the derivative of S^2 in terms of $\hat{\eta}^h$;

$$\frac{dS^2(\hat{\eta}^h)}{d\hat{\sigma}^2} = L(X'\hat{V}^{-1}X)^{-1}L',$$
$$\frac{dS^2(\hat{\eta}^h)}{d\text{vech}(\hat{D}_k)} = \hat{\sigma}^2\mathcal{D}'_{q_k}\left(\sum_{j=1}^{l_k}\hat{B}_{(k,j)}\otimes\hat{B}_{(k,j)}\right),$$

where $\hat{B}_{(k,j)} = Z'_{(k,j)}\hat{V}^{-1}X(X'\hat{V}^{-1}X)^{-1}L'$. We obtain an expression for $\text{var}(\hat{\eta}^h)$ by noting that the asymptotic variance of $\hat{\eta}^h$ is given by $\mathcal{I}(\hat{\eta}^h)^{-1}$ where $\mathcal{I}(\hat{\eta}^h)$ is a sub-matrix of $\mathcal{I}(\hat{\theta}^h)$, given by Equations (3.6)-(3.8).

In summary, we have provided all the closed-form expressions necessary to perform the Welch-Satterthwaite degrees of freedom estimation method for any LMM described by Equation (2.3). For the remainder of this chapter, estimation of v using the above expressions is referred to as the "direct-WS" method. This name reflects the direct approach taken for the evaluation of the right-hand side of Equation (3.27). We note that this approach may also be extended to models using the constrained covariance structures of Section 3.3.4 by employing the Fisher Information matrix given by Equation (3.25) and transforming the derivative of $S^2(\eta)$ appropriately (details of which are provided by Theorem A.9 of Appendix A.9.3). We conclude this section by noting that this method can also be used in a similar manner for estimating the degrees of freedom of an approximate F-statistic based on the multi-factor LMM. Additional material describing how the Welch-Satterthwaite equation may be applied for approximate F-statistics is provided in Appendix A.9.2.

3.4 Simulations

To assess the accuracy and efficiency of each of the proposed LMM parameter estimation methods described in Sections 3.3.1.1-3.3.1.5 and the direct-WS degrees of freedom estimation method described in Section 3.3.5, extensive simulations were conducted. These simulations are described fully in Section 3.4.1 and the results

are presented in Section 3.4.2. All reported results were obtained using an Intel(R) Xeon(R) Gold 6126 2.60GHz processor with 16GB RAM.

3.4.1 Simulation Methods

3.4.1.1 Parameter Estimation

The algorithms of Sections 3.3.1.1-3.3.1.5 have been implemented in the programming language Python. Under three different simulation settings, each with a different design structure, the algorithms are compared against one another, the 'lmer' function from the R package lme4, and the baseline truth used to generate the simulations. All methods are contrasted in terms of output, computation time and, for the methods presented in this chapter, the number of iterations until convergence. Convergence of each method was assessed by verifying whether successive log-likelihood estimates differed by more than a predefined tolerance of 10^{-6}. 1000 individual simulation instances were run for each simulation setting.

All model parameters were held fixed across all runs. In every simulation setting, test data was generated according to model (2.3). Each of the three simulation settings imposed a different structure on the random effects design and covariance matrices, Z and D. The first simulation setting employed a single factor design ($r = 1$) with two random effects (i.e. $q_1 = 2$) and 50 levels (i.e. $l_1 = 50$). The second simulation setting employed two crossed factors ($r = 2$) where the number of random effects and numbers of levels for each factor were given by $q_1 = 3$, $q_2 = 2$, $l_1 = 100$ and $l_2 = 50$, respectively. The third simulation setting used three crossed factors (i.e. $r = 3$) with the number of random effects and levels for each of the factors given by $q_1 = 4$, $q_2 = 3$, $q_3 = 2$, $l_1 = 100$, $l_2 = 50$ and $l_3 = 10$, respectively. In all simulations, the number of observations, n, was held fixed at 1000. In each simulated design, the first column of the fixed effects design matrix, X, and the first random effect in the random effects design matrix, Z, were treated as intercepts. Within each simulation instance, the remaining (non-zero) elements of the variables X and Z, as well as those of the error vector ϵ, were generated at random according to the standard

univariate normal distribution. The random effects vector b was simulated according to a normal distribution with covariance D, where D was predefined, exhibited no particular constrained structure and contained a mixture of both zero and non-zero off-diagonal elements. The assignment of observations to levels for each factor was performed at random with the probability of an observation belonging to any specific level held constant and uniform across all levels.

Fisher Scoring methods for the single-factor design have been well-studied (c.f. Demidenko (2013)) and the inclusion of the first simulation setting is only for comparison purposes. The second and third simulation settings represent more complex crossed factor designs for which the proposed Fisher Scoring based parameter estimation methods did not previously exist. To assess the performance of the proposed algorithms, the Mean Absolute Error (MAE) and Mean Relative Difference (MRD) were used as performance metrics. The methods considered for parameter estimation were those described in Sections 3.3.1.1-3.3.1.5 and the baseline truth used for comparison was either the baseline truth used to generate the simulated data or the lmer computed estimates.

All methods were contrasted in terms of the MAE and MRD of both β and the variance product $\sigma^2 D$. The variance product $\sigma^2 D$ was chosen for comparison instead of the individual components σ^2 and D as the variance product $\sigma^2 D$ is typically of intrinsic interest during the inference stage of conventional statistical analyses and is often employed for further computation. In all simulations, both ML and ReML estimation variants of the methods were assessed. The computation time for each method was also recorded and contrasted against that of lmer. To ensure a fair comparison of computation time, the recorded computation times for lmer were based only on the performance of the 'optimizeLmer' function, which is the module employed for parameter estimation by lmer. The specific values of β, σ^2, and D employed for each simulation setting can be found in Section S1 of Supplementary Material A. Formal definitions of MAE and MRD measures used for comparison are given in Section S2 of Supplementary Material A.

3.4.1.2 Degrees of Freedom Estimation

The accuracy and validity of the direct-WS degrees of freedom estimation method proposed in Section 3.3.5 were assessed through further simulation. To achieve this, as LMM degrees of freedom are assumed to be specific to the experiment design, a single design was chosen at random from each of the three simulation settings described in Section 3.4.1.1. With the fixed effects and random effects matrices, X and Z, now held constant across simulations, 1000 simulations were run for each simulation setting. The random effects vector, b, and random error vector, ϵ, were allowed to vary across simulations according to normal distributions with the appropriate variances. In each simulation, degrees of freedom were estimated via the direct-WS method for a predefined contrast vector, corresponding to a fixed effect that was truly zero.

The direct-WS estimated degrees of freedom were compared to a baseline truth and the degrees of freedom estimates produced by the R package lmerTest. Baseline truth was established in each simulation setting using $1,000,000$ simulations to empirically estimate $\text{Var}(S^2(\hat{\eta}))$, giving a single value for the denominator of $v(\hat{\eta})$ from Equation (3.26). Following this, in each of the 1000 simulation instances described above, the numerator of Equation (3.26) was recorded, giving 1000 estimates of $v(\hat{\eta})$. The final estimate was obtained as an average of these 1000 values. All lmerTest degrees of freedom estimates were obtained using the same simulated data as was used by the direct-WS method and were computed using the 'contest1D' function from the lmerTest package. Whilst all baseline truth measures were computed using parameter estimates obtained using the FSFS algorithm of Section 3.3.1.4, we believe this has not induced bias into the simulations as, as discussed in Section 3.4.2.1, all simulated FSFS-derived parameter estimates considered in this work agreed with those provided by lmer within a tolerance level on the scale of machine error.

3.4.2 Simulation Results

3.4.2.1 Parameter Estimation Results

The results of the parameter estimation simulations of Section 3.4.1.1 were identical. In all three settings, all parameter estimates and maximised likelihood criteria produced by the FS, FFS, SFS and FSFS methods were identical to those produced by lmer up to a machine error level of tolerance. In terms of the maximisation of likelihood criteria, there was no evidence to suggest that lmer outperformed any of the FS, FFS, SFS or FSFS methods, or vice versa. The CSFS method provided worse performance, however, with maximised likelihood criteria which were lower than those reported by lmer in approximately 2.5% of the simulations run for simulation setting 3. The reported maximised likelihood criteria for these simulations were all indicative of convergence failure. For all methods considered during simulation, the observed performance was unaffected by the choice of likelihood estimation criteria (i.e. ML or ReML) employed.

The MRD and MAE values provided evidence for strong agreement between the parameter estimates produced by lmer and the FS, FFS, SFS and FSFS methods. The largest MRD values observed for these methods, taken relative to lmer, across all simulations and likelihood criteria, were 1.03×10^{-3} and 2.12×10^{-3} for β and $\sigma^2 D$, respectively. The largest observed MAE values for these methods, taken relative to lmer, across all simulations, were given by 1.02×10^{-5} and 4.30×10^{-4} for β and $\sigma^2 D$, respectively. Due to the extremely small magnitudes of these values, MAE and MRD values are not reported in detail here. For further detail, see Supplementary Material A Sections S3-S6.

To compare the proposed methods in terms of computational performance, Tables 3.1 and 3.2 present the average time and number of iterations performed for each of the 5 methods, using ML and ReML likelihood criteria, respectively. Corresponding computation times are also provided for lmer. For both ML and ReML likelihood criteria, all Fisher Scoring methods demonstrated considerable efficiency in terms of computation speed. While the improvement in speed is minor for simulation setting

1, for multi-factor simulation settings 2 and 3, the performance gains can be seen to be considerable. Overall, the only method that consistently demonstrated notably worse performance than the others was the CSFS algorithm. In addition to requiring a longer computation time, the CSFS algorithm employed many more iterations.

Method	**FS**	**FFS**	**SFS**	**FSFS**	**CSFS**	**lmer**
Simulation 1						
t (Time/s)	0.041	0.037	0.026	0.055	0.057	0.059
	(0.012)	(0.011)	(0.007)	(0.058)	(0.017)	(0.021)
n_{it} (No. of iterations)	5.243	5.243	6.048	6.048	10.461	-
	(0.613)	(0.613)	(0.256)	(0.256)	(1.851)	-
Simulation 2						
t (Time/s)	0.129	0.112	0.105	0.125	0.220	0.648
	(0.047)	(0.042)	(0.036)	(0.061)	(0.070)	(0.050)
n_{it} (No. of iterations)	6.694	6.694	8.323	8.323	14.281	-
	(0.964)	(0.964)	(0.534)	(0.534)	(1.381)	-
Simulation 3						
t (Time/s)	1.081	0.889	1.872	1.941	3.970	5.979
	(0.434)	(0.429)	(0.915)	(0.869)	(1.094)	(0.299)
n_{it} (No. of iterations)	8.133	8.133	16.054	16.054	33.437	-
	(0.743)	(0.743)	(0.835)	(0.835)	(24.265)	-

Table 3.1: The average time in seconds and the number of iterations reported for maximum likelihood estimation performed using the FS, FFS, SFS, FSFS and CSFS methods. For each simulation setting, results displayed are taken from averaging across 1000 individual simulations. Also given are the average times taken by lmer to perform maximum likelihood parameter estimation on the same simulated data. Standard deviations are given in brackets below each entry in the table.

Within a setting where q (i.e. the number number of columns in the random effects design matrix) is small, the results of these simulations demonstrate strong computational efficiency for the FS, FFS, SFS and FSFS methods. However, we emphasise that no claim is made to suggest that the observed results will generalise to larger values of q or all possible designs. For large sparse LMMs that include a large number of random effects, without further adaptation of the methods presented in this chapter to employ sparse matrix methodology, it is expected that lmer will provide superior performance (c.f. Chapter 4 and Appendix B.3 for further consider-

ation of sparsity patterns in single-factor models). We again emphasise here that our purpose in undertaking this work is not to compete with the existing LMM parameter estimation tools. Instead, the focus of this work is to develop methodology, which provides performance comparable to the existing tools, for use in situations where the existing tools may not be applicable due to practical implementation considerations.

Method	**FS**	**FFS**	**SFS**	**FSFS**	**CSFS**	**lmer**
Simulation 1						
t (Time/s)	0.057	0.052	0.043	0.064	0.088	0.071
	(0.014)	(0.011)	(0.008)	(0.042)	(0.021)	(0.021)
n_{it} (No. of iterations)	5.261	5.261	6.053	6.053	10.404	-
	(0.591)	(0.591)	(0.257)	(0.257)	(1.829)	-
Simulation 2						
t (Time/s)	0.175	0.154	0.157	0.169	0.306	0.810
	(0.043)	(0.039)	(0.035)	(0.041)	(0.067)	(0.065)
n_{it} (No. of iterations)	6.758	6.758	8.413	8.413	14.283	-
	(0.984)	(0.984)	(0.561)	(0.561)	(1.404)	-
Simulation 3						
t (Time/s)	1.615	1.378	2.968	3.002	6.604	7.457
	(0.404)	(0.344)	(0.752)	(0.814)	(5.455)	(0.442)
n_{it} (No. of iterations)	8.084	8.084	16.018	16.018	34.611	-
	(0.730)	(0.730)	(0.798)	(0.798)	(25.526)	-

Table 3.2: The average time in seconds and the number of iterations reported for restricted maximum likelihood estimation performed using the FS, FFS, SFS, FSFS and CSFS methods. For each simulation setting, results displayed are taken from averaging across 1000 individual simulations. Also given are the average times taken by lmer to perform restricted maximum likelihood parameter estimation on the same simulated data. Standard deviations are given in brackets below each entry in the table.

It must be stressed that reported computation times do not solely reflect theoretical differences between the proposed methods and those employed by lmer. Practical considerations related to implementation, such as which programming language and software libraries are used, also influence computational performance. Again, we emphasize that the aim of these simulations is not to perform a like-for-like comparison between software packages but instead, to demonstrate the viability and efficacy of

our proposed methods as assessed in relation to the tools currently available.

Two potential causes for the poor performance of the CSFS method have been identified. The first cause is presented in Demidenko (2013), in which it is argued that the Cholesky parameterised algorithm will provide slower convergence than that of the unconstrained alternative methods due to structural differences between the likelihood surfaces over which parameter estimation is performed. This reasoning offers a likely explanation for the higher number of iterations and computation time observed for the CSFS method across all simulations. However, this does not explain the small number of simulations in simulation setting 3 in which evidence of convergence failure was observed.

The second possible cause also provides an explanation for the observed convergence failure. Pinheiro and Bates (1996) note that Cholesky parameterisations are not unique and the number of possible choices for the Cholesky factorisation of a positive definite matrix of dimension $(q \times q)$ increases with the dimension q. This means that, for each covariance matrix D_k, there are multiple Cholesky factors, Λ_k, which satisfy $\Lambda_k \Lambda_k' = D_k$. The larger D_k is in dimension, the greater the number of Λ_k that correspond to D_k there are. Pinheiro and Bates (1996) argue that when optimal solutions are numerous and close together in the parameter space, numerical problems can arise during optimisation. We note that in comparison to the other simulation settings considered here, the design for simulation setting 3 contains the largest number of factors and random effects. Consequently, the covariance matrices, D_k, are large for this simulation setting and numerous Cholesky factors which correspond to the same optimal solution for the covariance matrix, D, exist. For this reason, simulation setting 3 is the most susceptible to numerical problems of the kind described by Pinheiro and Bates (1996). We suggest that this is a likely accounting for the 2.5% of simulations in which convergence failure was observed. In summary, the simulation results offer strong evidence for the correctness and efficiency of the Fisher Scoring methods proposed, with the exception of the CSFS method, which experiences slower performance and convergence failure in rare cases.

Method	Truth	direct-WS	lmerTest
Simulation 1			
Mean	910.93	910.68	906.62
Standard Deviation	0.0	2.36	2.39
Mean Squared Error	0.0	5.64	24.23
Simulation 2			
Mean	844.61	842.17	837.79
Standard Deviation	0.0	7.16	7.26
Mean Squared Error	0.0	57.20	99.21
Simulation 3			
Mean	707.01	700.96	695.08
Standard Deviation	0.0	17.19	17.67
Mean Squared Error	0.0	332.01	454.53

Table 3.3: The mean, standard deviation and mean squared error for $1,000$ degrees of freedom estimates in each simulation setting. Results are displayed for both the lmerTest and direct-WS methods, alongside "true" mean values, which were established using the moment-matching based approach outlined in Section 3.4.1.2 and computed using $1,000,000$ simulation instances.

3.4.2.2 Degrees of Freedom Estimation Results

Across all degrees of freedom simulations, results indicated that the degrees of freedom estimates produced by the direct-WS method possessed both lower bias and lower variance than those produced by lmerTest. The degrees of freedom estimates, for each of the three simulation settings, are summarized in Table 3.3.

It can be seen from Table 3.3 that throughout all simulation settings both direct-WS and lmerTest appear to underestimate the true value of the degrees of freedom. However, the bias observed for the lmerTest estimates is notably more severe that of the direct-WS method, suggesting that the estimates produced by direct-WS have a higher accuracy than those produced by lmerTest. The observed difference in the standard deviation of the degrees of freedom estimates between the lmerTest and direct-WS methods is less pronounced. However, in all simulation settings, lower

standard deviations are reported for direct-WS, suggesting that the estimates produced by direct-WS have a higher precision than those produced by lmerTest.

For both lmerTest and direct-WS, the observed bias and variance increase with simulation complexity. The simulations provided here indicate that the severity of disagreement between direct-WS and lmerTest increases as the complexity of the random effects design increases. This observation matches the expectation that the accuracy of the numerical gradient estimation employed by lmerTest will worsen as the complexity of the parameter space increases. Reported mean squared errors are also provided, indicating further that the direct-WS method outperformed lmerTest in terms of accuracy and precision in all simulation settings.

3.5 Real Data Examples

To illustrate the usage of the methodology presented in Sections 3.3.1-3.3.5 in practical situations, we provide two real data examples. These examples are described fully in Section 3.5.1 and the results are presented in Section 3.5.2. Again, all reported results were obtained using an Intel(R) Xeon(R) Gold 6126 2.60GHz processor with 16GB RAM. For each reported computation time, averages were taken across 50 repeated runs of the given analysis. The real data examples provided in this chapter are univariate, and do not involve imaging data of the form described in Chapter 2.

3.5.1 Real Data Methods

3.5.1.1 The SAT Score Example

The first example presented here is based on data from the Longitudinal Evaluation of School Change and Performance (LESCP) dataset (Turnbull et al. (1999)). This dataset has notably been previously analyzed by Hong and Raudenbush (2008) and was chosen for inclusion in this work because it previously formed the basis for between-software comparisons of LMM software packages by West et al. (2014). The LESCP study was conducted in 67 American schools in which SAT (Student Aptitude

Test) math scores were recorded for randomly selected samples of students. As in West et al. (2014), one of the 67 schools from this dataset was chosen as the focus for this analysis. For each individual SAT score, unique identifiers (ID's) for the student who took the test and for the teacher who prepared the student for the test were recorded. As many students were taught by multiple teachers and all teachers taught multiple students, the grouping of SAT scores by student ID and the grouping of SAT scores by teacher ID constitute two crossed factors. In total, $n = 234$ SAT scores were considered for analysis. The SAT scores were taken from 122 students taught by a total of 12 teachers, with each student sitting between 1 and 3 tests.

The research question in this example concerns how well a student's grade (i.e. year of schooling) predicted their mathematics SAT score (i.e. did students improve in mathematics over the course of their education?). For this question, between-student variance and between-teacher variance must be taken into consideration as different students possess different aptitudes for mathematics exams and different teachers possess different aptitudes for teaching mathematics. In practice, this is achieved by employing an LMM which includes random intercept terms for both the grouping of SAT scores according to student ID and the grouping of SAT scores according to teacher ID. For the k^{th} mathematics SAT test taken by the i^{th} student, under the supervision of the j^{th} teacher, such a model could be stated as follows:

$$\text{MATH}_{i,j,k} = \beta_0 + \beta_1 \times \text{YEAR}_{i,j,k} + s_i + t_j + \epsilon_{i,j,k},$$

where $\text{MATH}_{i,j,k}$ is the SAT score achieved and $\text{YEAR}_{i,j,k}$ is the grade of the student at the time the test was taken. In the above model, β_0 and β_1 are unknown parameters and s_i, t_j and $\epsilon_{i,j,k}$ are independent mean-zero random variables which differ only in terms of their covariance. s_i is the random intercept which models between-student variance, t_j is the random intercept which models between-teacher variance, and $\epsilon_{i,j,k}$ is the random error term. The random variables s_i, t_j and $\epsilon_{i,j,k}$ are assumed to be

mutually independent and follow the below distributions:

$$s_i \sim N(0, \sigma_s^2), \quad t_j \sim N(0, \sigma_t^2), \quad \epsilon_{i,j,k} \sim N(0, \sigma^2),$$

where the parameters σ_s^2, σ_t^2 and σ^2 are the unknown student, teacher and residual variance parameters, respectively.

The random effects in the SAT score model can be described using the notation presented in previous sections as follows; $r = 2$ (i.e. observations are grouped by two factors, student ID and teacher ID), $q_1 = q_2 = 1$ (i.e. one random effect is included for each factor, the random intercepts s_i and t_j, respectively), and $l_1 = 122, l_2 = 12$ (i.e. there are 122 students and 12 teachers). When the model is expressed in the form described by Equation (2.3), the random effects design matrix Z is a $0-1$ matrix. In this setting, the positioning of the non-zero elements in Z indicates the student and teacher associated with each test score. The random effects covariance matrices for the two factors, student ID and teacher ID, are given by $D_0 = [\sigma_s^2]$ and $D_1 = [\sigma_t^2]$, respectively.

For the SAT score model, the estimated parameters obtained using each of the methods detailed in Sections 3.3.1.1-3.3.1.5 are reported. For comparison, the parameter estimates obtained by lmer are also given. As the estimates for the fixed effects parameters, β_0 and β_1, are of primary interest, ML was employed to obtain all reported parameter estimates. For this example, methods are contrasted in terms of output, computation time and the number of iterations performed.

3.5.1.2 The Twin Study Example

The second example presented in this chapter aims to demonstrate the flexibility of the constrained covariance approaches described in Section 3.3.4. This example is based on the ACE model; an LMM commonly employed to analyze the results of twin studies by accounting for the complex covariance structure exhibited between related individuals. The ACE model achieves this by separating between-subject response variation into three categories: variance due to additive genetic effects (σ_a^2), variance

due to common environmental factors (σ_c^2), and residual error (σ_e^2). For this example, we utilize data from the Wu-Minn Human Connectome Project (HCP) (Van Essen et al. (2013)). The HCP dataset contains brain imaging data collected from $1,200$ healthy young adults, aged between 22 and 35, including data from 300 twin pairs and their siblings. We do not make use the imaging data in the HCP dataset but, instead, focus on the baseline variables for cognition and alertness.

The primary research question considered in this example focuses on how well a subject's quality of sleep predicts their English reading ability. The variables of primary interest used to address this question are subject scores in the Pittsburgh Sleep Quality Index (PSQI) questionnaire and an English language recognition test (ENG). Other variables included the subjects' age in years (AGE) and sex (SEX), as well as an age-sex interaction effect. A secondary research question considered asks "How much of the between-subject variance observed in English reading ability test scores can be explained by additive genetic and common environmental factors?". To address this question, the covariance parameters σ_a^2, σ_c^2 and σ_e^2 must be estimated.

To model the covariance components of the ACE model, a simplifying assumption that all family units share common environmental factors was made. Following the work of Winkler et al. (2015), family units were first categorized by their internal structure into what shall be referred to as "family structure types" (i.e. unique combinations of full-siblings, half-siblings and identical twin pairs which form a family unit present in the HCP dataset). In the HCP dataset, 19 such family structure types were identified. In the following text, each family structure type shall be treated as a factor in the model. For the i^{th} observation to belong to the j^{th} level of the k^{th} factor in the model may be interpreted as the i^{th} subject belonging to the j^{th} family exhibiting family structure of type k. The model employed for this example is given by:

$$\begin{aligned}\text{ENG}_{k,j,i} =& \beta_0 + \beta_1 \times \text{AGE}_i + \beta_2 \times \text{SEX}_i + \ldots \\ & \beta_3 \times \text{AGE}_i \times \text{SEX}_i + \beta_4 \times \text{PSQI}_i + \gamma_{k,j,i} + \epsilon_{k,j,i},\end{aligned}$$

where both $\gamma_{k,j,i}$ and $\epsilon_{k,j,i}$ are mean-zero random variables. The random error, $\epsilon_{k,j,i}$, is distributed $N(0, \sigma_e^2)$, and the random term $\gamma_{k,j,i}$ models the within-"family unit" covariance. Further detail on the specification of $\gamma_{k,j,i}$ can be found in Appendix A.10.1.

In the notation of the previous sections, the number of factors in the ACE model, r, is equal to the number of family structure types present in the model (i.e. 19). For each family structure type present in the model, family structure type k, l_k is the number of families who exhibit such structure and q_k is the number of subjects present in any given family unit with such structure. As there is a unique random effect (i.e. a unique random variable, $\gamma_{k,j,i}$) associated with each individual subject, none of which are scaled by any coefficients, the random effects design matrix, Z, is the $(n \times n)$ identity matrix. To describe $\{D_k\}_{k\in\{1,\ldots,r\}}$ requires the known matrices $\mathbf{K}_k^a$ and $\mathbf{K}_k^c$, which specify the kinship (expected genetic material) and environmental effects shared between individuals, respectively (See Appendix A.10.2 for more details). Given $\mathbf{K}_k^a$ and $\mathbf{K}_k^c$, the covariance components σ^2 and $\{D_k\}_{k\in\{1,\ldots,r\}}$ are given as $\sigma^2 = \sigma_e^2$ and $D_k = \sigma_e^{-2}(\sigma_a^2\mathbf{K}_k^a + \sigma_c^2\mathbf{K}_k^c)$ respectively.

As the covariance components σ_a and σ_c are of practical interest in this example, optimization is performed according to the ReML criterion using the adjustments described in Appendix A.5 and covariance structure is constrained via the methods outlined in Section 3.3.4. Further detail on the constrained approach for the ACE model can be found in Appendix A.10.3. Discussion of computational efficiency for the ACE model is also provided in Appendix A.10.4.

Section 3.5.2.2 reports the maximized restricted log-likelihood values and parameter estimates obtained using the Fisher Scoring method. Also given are approximate T-statistics for each fixed effects parameter, alongside corresponding degrees of freedom estimates and p-values obtained via the methods outlined in Section 3.3.5. To verify correctness, the restricted log-likelihood of the ACE model was also maximized numerically using the implementation of Powell's bi-directional search based optimization method (Powell (1964)) provided by the SciPy Python package. The maximized restricted log-likelihood values and parameter estimates produced were

then contrasted against those provided by the Fisher Scoring method. The OLS (Ordinary Least Squares) estimates, which would be obtained had the additive genetic and common environmental variance components not been accounted for in the LMM analysis, are also provided for comparison.

3.5.2 Real Data Results

3.5.2.1 The SAT Score Results

For the SAT score example described in Section 3.5.1.1, the log-likelihood, fixed effect parameters and variance component estimates produced by the Fisher Scoring algorithms were identical to those produced by lmer. Further, the reported computation times for this example suggested little difference between all methods considered in terms of computational efficiency. Of the Fisher Scoring methods considered, the FS and FFS methods took the fewest iterations to converge while the SFS and FSFS methods took the most iterations to converge. The results presented here exhibit strong agreement with those reported in West et al. (2014), in which the same model was used as the basis for a between-software comparison of LMM software packages. For completeness, the full table of results can be found in Supplementary Material A Section S9.

3.5.2.2 The Twin Study Results

The results for the twin study example described in Section 3.5.1.2 are presented in Table 3.4. It can be seen from Table 3.4 that the Powell optimizer and Fisher Scoring method attained extremely similar optimized likelihood values, with Fisher Scoring converging notably faster. This result offers further evidence for the correctness of the parameter estimates produced by the Fisher Scoring method as it is unlikely that both Powell estimation and Fisher Scoring would converge to the same solution if the solution were suboptimal. The parameter estimates produced by Fisher Scoring and Powell optimization can be seen to be smaller in magnitude than those estimated by

Estimation Method		OLS	Powell	FS
Performance				
l (Log-likelihood)		-2133.49	-2005.83	-2005.81
t (Time in seconds)		$< .001$	93.32	2.31
Fixed effects parameters (Standard Errors)				
β_0 (Intercept)		121.46 (3.61)	118.43 (3.40)	118.34 (3.42)
β_1 (Age)		-0.11 (0.12)	-0.04 (0.11)	-0.04 (0.11)
β_2 (Sex)		-10.74 (5.07)	-7.22 (4.64)	-7.61 (4.66)
β_3 (Age:sex)		0.45 (0.18)	0.33 (0.16)	0.34 (0.16)
β_4 (PSQI score)		-0.49 (0.12)	-0.31 (0.10)	-0.30 (0.10)
Covariance parameters				
σ_a^2 (Additive genetic)		0.00	48.20	50.67
σ_c^2 (Common environment)		0.00	27.53	26.89
σ_e^2 (Residual error)		110.04	33.57	32.71
Tests for Fixed Effects				
Intercept	*T-statistic*	33.67	34.83	34.65
	degrees of freedom	1105.00	920.59	921.57
	p-value	$< .001^*$	$< .001^*$	$< .001^*$
Age	*T-statistic*	-0.94	-0.350	-0.320
	degrees of freedom	1105.00	907.44	908.42
	p-value	.350	.726	.749
Sex	*T-statistic*	-2.12	-1.56	-1.63
	degrees of freedom	1105.00	833.01	832.96
	p-value	$.034^*$	.120	.103
Age:sex	*t-statistic*	2.58	2.03	2.10
	degrees of freedom	1105.00	830.53	830.61
	p-value	$.010^*$	$.043^*$	$.036^*$
PSQI score	*T-statistic*	-4.25	-3.19	-3.17
	degrees of freedom	1105.00	933.67	928.87
	p-value	$< .001^*$	$.001^*$	$.001^*$

Table 3.4: Performance metrics, parameter estimates and approximate T-test results for the twin study example. Standard errors for the fixed effects parameter estimates are given in brackets alongside the corresponding estimates. For each model parameter, a T-statistic, direct-WS degrees of freedom estimate, and p-value are provided, corresponding to the approximate t-test for a non-zero effect. P-values that are significant at the 5% level are indicated using a * symbol.

OLS, highlighting how the inclusion of additional variance terms in the model can have a meaningful impact on the conclusion of the analysis.

Also provided in Table 3.4, are approximate t-tests based on the Fisher Scoring and Powell optimizer parameter estimates, with corresponding standard t-tests given based on the OLS estimates. At the 5% significance level, the approximate t-tests conclude that the fixed effects parameters corresponding to the 'Intercept', 'Age and Sex Interaction' and the 'PSQI Score' are non-zero in value. While it may be expected that age should affect reading ability, no significant effect was observed for the 'Age' covariate. This lack of observed effect may be explained by the narrow age range of subjects present in the HCP dataset, with subjects ranging from 22 to 35 years in age, and by the fact that individual observations were recorded in units of years. In general, the OLS-based t-tests produced similar conclusions to those produced by the FS-based approximate t-tests. A notable exception, however, is given by the t-tests for the fixed effect associated with the 'Sex' covariate. While the standard OLS t-test reported at the 5% significance level that the 'Sex' fixed effect was non-zero in value, the FS-based approximate t-test concluded that there was not evidence to support this claim. This result further highlights the importance of modeling all relevant variance terms.

3.6 Discussion

In this chapter, we have presented derivations for and demonstrated potential applications of, score vector and Fisher Information matrix expressions for LMMs containing multiple random factors. While many of the examples presented in this chapter were benchmarked against existing software, it is not the authors' intention to suggest that the proposed methods are superior to existing software packages. Instead, this work aims to complement existing LMM parameter estimation research. This aim is realized through careful exposition of the score vectors and Fisher Information matrices and detailed description of methodology and algorithms.

Modern approaches to LMM parameter estimation typically depend upon conceptually complex mathematical operations which require support from a range of software packages and infrastructure. The work in this chapter has been, in large part, motivated by current deficiencies in vectorized support for such operations. Vectorized computation is a crucial requirement for fMRI applications, in which hundreds of thousands of mixed models must be estimated concurrently. The methods proposed in this work take advantage over the existing tools in such situations, where many LMM's must be run concurrently, as the theoretically simplistic mathematical operations they employ are more naturally amenable to vectorized computation. It is with such applications in mind that this work has been undertaken. The application of mass-univariate computation for fMRI analysis is pursued further in Chapter 4.

Although the methods presented in this chapter are well suited for the desired application, we stress that there are many situations where current software packages will likely provide superior performance. One such situation can be found by observing that the Fisher Scoring method requires the storage and inversion of a matrix of dimensions $(q \times q)$. This is problematic, both in terms of memory and computation accuracy, for designs which involve very large numbers of random effects, grouping factors or factor levels. While such applications have not been considered extensively in this work, we note that many of the expressions provided in this document may benefit from combination with sparse matrix methodology to overcome this issue. We suggest that the potential for improved computation time via the combination of the Fisher Scoring approaches described in this chapter with sparse matrix methodology may also be the subject of future research. In Chapter 4, this idea is further explored via explicit modeling of the sparsity patterns present in the product forms for single-factor LMMs (see Appendix B.3 for further detail).

An active area of LMM research that has not been discussed extensively in this work is hypothesis testing for random effects covariance parameters. In general, development of hypothesis testing procedures for random effects is a more complex task than for that of fixed effects. This additional complexity stems from the fact that the random effects parameters can lie on the boundary of the parameter space

and must be modelled using mixture distributions. Many methods are available which provide approximate testing procedures for random effects covariance parameters (c.f. Scheipl et al. (2008), and Section 4.3.1.4 of Chapter 4). However, it is not immediately apparent to us whether the derivations we have provided may be used in conjunction with such methods to improve the testing procedures for random effects. We suggest this idea may form a potential basis for future investigation.

3.7 Data and Code Availability

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

The longitudinal evaluation of school change and performance (LESCP) dataset employed in Sections 3.5.1.1 and 3.5.2.1 is publicly available and can be found, for example, as one of the example datasets included in the Heirachical Linear Models (HLM) software package (Raudenbush and Bryk (2002)).

All code used to generate the results of the simulations and real data examples presented in Sections 3.4.1-3.5.2.2 is publicly available and can be found in the below GitHub repository:

https://github.com/TomMaullin/LMMPaper

Included in this repository are detailed notebooks demonstrating how the code may be executed in order to reproduce the results presented in this work.

Chapter 4

Parallelised Computing for Big Linear Mixed Models

In this chapter, we develop a computationally efficient framework for large-sample fMRI LMM analysis. This chapter was written under the supervision of Professor Thomas E. Nichols and a corresponding manuscript will be submitted for publication shortly. Additional detail associated with this chapter is provided in Appendix B. Supplementary results and further simulation specifics are given in the 'Supplementary Material B' document provided alongside this thesis.

4.1 Background

Covariance structure in an experimental design often arises from grouping factors present in the data (c.f. Chapter 1 and Sections 2.2.2-2.2.3). Accounting for complex grouping factors during an fMRI analysis is an historically routine practice for small-sample studies. Commonly employed analysis designs in the small-sample setting include longitudinal multi-session analyses (observations grouped by subject), comparative group analyses (subjects grouped by study conditions) and mega-analyses (analysis results grouped according to study protocols). As stated in previous chapters, the widely-accepted, conventional approach to modelling such datasets is to em-

ploy the LMM to account for grouping structure via the inclusion of both fixed effects and random effects during model specification (c.f. Laird and Ware (1982), Friston et al. (2002)). Here, we recall that fixed effects are unknown constant parameters that are associated with covariates in the experimental design. In contrast, random effects are random variables that model the systematic differences between instances of a categorical variable (e.g. between-subject differences, between-site differences).

There are many strong obstacles to the practical execution of LMM analyses for large-sample fMRI datasets. An LMM analysis consists of two stages: parameter estimation and statistical inference, each of which presents unique computational and theoretical challenges in the large-sample fMRI analysis setting. An overview of the challenges specific to each of these stages is provided by Sections 4.1.1 and 4.1.2, respectively. In addition, in the large-sample setting, missing data near cortical boundaries is ubiquitous and must be carefully accounted for in the analysis design (c.f. Section 2.3.3). In the univariate (non-imaging) setting, many tools exist for estimating the parameters of and performing inference upon the LMM (c.f. Section 3.1). However, many of these tools are not designed to (a) be scalable to arbitrarily large datasets and (b) exploit vectorisation speed-ups from processing multiple voxels simultaneously. In Chapter 3, we derived novel closed form expressions for the Fisher Information matrices and gradient vectors of the LMM. Following this work, in this chapter, we shall demonstrate that these expressions may be employed to perform fast and scalable LMM parameter estimation and inference in the context of large-sample fMRI analysis.

In this chapter, we present "Big" Linear Mixed Models (BLMM), a Python-based tool for parameter estimation and statistical inference of mass-univariate LMMs, designed for usage on a High Performance Cluster (HPC) with Sun Grid Engine (SGE). The BLMM tool partitions fMRI analyses to limit memory consumption, while still being able to exploit vectorization speed-ups from working with multiple voxels. In the following sections, we first provide background on LMM parameter estimation and inference in fMRI. Following this, we formally define the mass-univariate voxel-wise LMM employed in the fMRI analysis setting. In the methods section, we then outline

the computational pipeline of BLMM, starting with the input specification, followed by the distributed stages of computation, parameter estimation and finally, inference. Next, the correctness and performance of BLMM are assessed via simulation. We conclude by providing a real data example based on the UK Biobank.

4.1.1 FMRI LMM Parameter Estimation

In Chapter 3, we provided a survey of the software and tools commonly employed for parameter estimation of the 'univariate' (single model) LMM. However, in mass-univariate fMRI, parameter estimation is not performed for a single model, but rather for hundreds of thousands of models, each corresponding to a different voxel in the analysis mask. For a mass-univariate voxel-wise analysis to truly utilise the computational power available, it is a necessity that computation is vectorised across voxels. As noted in Section 3.1, many of the established LMM tools are reliant upon operations that are not naturally amenable to vectorisation. Previously discussed examples of such operations included the sparse Cholesky decomposition employed by lme4 (Bates and DebRoy (2004)) and the sparse sweep operation employed by PROC-MIXED and MIXED (Wolfinger, Tobias, and Sall (1994)). Operations that are not amenable to vectorisation create bottlenecks for mass-univariate computation as they must be executed separately for each voxel in the image. Serial execution of such operations can result in severe computational overheads, especially when modelling large sample sizes. As a result, many of the tools available for univariate LMM analysis cannot be employed in the large-sample fMRI setting.

In the small-sample fMRI setting, several tools exist for mass-univariate LMM parameter estimation. These include SPM's built-in mixed-effects module (Friston et al. (2005)), FSL's FLAME package (Beckmann et al. (2003)), FreeSurfer's longitudinal analysis pipeline (Bernal-Rusiel et al. (2013a)) and AFNI's 3dLME package (Chen et al. (2013)). The tools provided by SPM, FSL and FreeSurfer perform parameter estimation for only LMMs in which observations are grouped by a single categorical factor. Specifically, SPM and FSL allow LMM parameter estimation of second-level

designs (i.e. designs with timepoints grouped by subject, c.f. Section 2.3.2) whilst FreeSurfer allows for parameter estimation of longitudinal LMM designs (i.e. designs with repeated visits grouped by subject) (FIL Methods Group (2020), Woolrich et al. (2004), Bernal-Rusiel et al. (2013b), Madhyastha et al. (2018)). In SPM and FreeSurfer, parameter estimates are obtained via Restricted Maximum Likelihood (ReML) estimation, whilst in FSL a Bayesian approach is adopted. In contrast, AFNI's 3dLME package allows parameter estimation of a much broader range of LMM designs and provides similar options to those offered by standard tools in the univariate setting. 3dLME provides this support by calling directly to the R package lme4 and parallelising computation across all available processors and cluster nodes via the use of the Dask Python package (Rocklin (2015)).

The tools provided by SPM, FSL, FreeSurfer and AFNI are computationally efficient for parameter estimation in the small-sample mass-univariate analysis setting, but were not originally designed for applications involving large sample sizes. In the context of large-sample analyses, SPM, FSL and FreeSurfer quickly encounter memory errors as sample sizes increase into the hundreds whilst AFNI experiences overheads in terms of computation time. For all tools, reduced computational performance in the large-sample setting primarily stems from two issues: (1) construction and storage of the analysis design (i.e. the design matrices and response vector) and (2) bottleneck computations which must be performed independently for each voxel. In addition, none of the aforementioned tools account for the missingness observed near cortical boundaries which was described in Section 2.3.3. Such missingness increases in severity with large sample sizes and must be accounted for during or prior to parameter estimation.

4.1.2 FMRI LMM Inference

In the small-sample setting, fMRI LMM analyses conclude with significance-based hypothesis tests using test statistics of the form described in Section 2.2.4. However, hypothesis testing of this nature has long been a contentious topic in the broader

LMM literature, due to the unknown variability in the estimation of variance components and the lack of exact distribution for the T and F test statistics (Verbeke and Molenberghs (2001)). As a result, whilst many of the popular tools for univariate LMM analysis provide support for calculating T-statistics and F-statistics, there is debate concerning how the corresponding p-values are to be calculated and whether such practices should be widely adopted (c.f. Manor and Zucker (2004), Baayen et al. (2008), Luke (2017)).

The lack of consensus in the LMM literature is reflected by the range of options available for performing LMM inference in the univariate setting. For example, HLM, MIXED and PROC-MIXED each adopt the assumption that the T and F test statistics for the LMM 'approximately' follow student's t- and F-distributions, but employ different techniques for estimating the associated unknown degrees of freedom. HLM approximates the degrees of freedom via closed-form expressions resembling those employed for multi-level linear model analyses (c.f. Raudenbush and Bryk (2002), West et al. (2014)). On the other hand, MIXED and PROC-MIXED utilize the Welch-Satterthwaite equation (c.f. Section 3.3.5) and numerical gradient optimization (Satterthwaite (1946), Welch (1947), Fai and Cornelius (1996)) in order to obtain degrees of freedom estimates for a more general breadth of LMM applications. For brevity, in the remainder of this chapter we shall refer to methods involving degrees of freedom estimation using the Welch-Satterthwaite equation as 'WSDF' (Welch-Satterthwaite Degrees of Freedom) based methods.

By employing this 'approximate distribution' assumption, HLM, MIXED and PROC-MIXED are able to provide support for significance-based hypothesis testing, outputting p-values alongside T and F statistics for a wide variety of LMMs. While lmer rejects any approximation and offers no p-values (Bates (2006)), the supporting lmerTest package (Kuznetsova et al. (2017)) provides p-values by using the same methods as MIXED and PROC-MIXED. Numerous studies have found that the WSDF-based approach employed by MIXED, PROC-MIXED and lmerTest is notably more robust to small sample sizes, unbalanced designs, covariance heterogeneity (when fitting the correct covariance structure) and non-normality than the approach

adopted by HLM (c.f. Keselman et al. (1999), Schaalje et al. (2002), Kuznetsova et al. (2017), Luke (2017), Francq et al. (2019)). However, the computational burden of the numerical gradient estimation required for the WSDF-based approach is substantial and constitutes a significant obstacle to large-sample LMM analysis in the mass-univariate fMRI setting.

Several tools are available in the small-sample fMRI setting for significance-based hypothesis testing via T and F statistics. As with LMM parameter estimation, these include SPM's built-in mixed-effects module, FSL's FLAME package, FreeSurfer's longitudinal analysis pipeline and AFNI's 3dLME package. Due to the low computational costs required, SPM, FreeSurfer and FSL each employ a similar approach to that used by HLM in the univariate setting by using closed-form expressions, which can be found, for example, in Pinheiro and Bates (2009), to approximate the degrees of freedom. AFNI alternately employs the WSDF approach by acting as a wrapper for the lmerTest package. While the former approach is more efficient in terms of computation time and memory, it provides less accurate estimates for the degrees of freedom. Conversely, the latter approach provides more accurate estimation of the approximate sampling distributions of the test statistics, but at the cost of increased computation time that scales with the number of observations. In this chapter, we make use of the novel closed-form expressions developed in Section 3.3.5 of Chapter 3 for evaluating the gradients required by the WSDF approach. These expressions offer a viable and accurate alternative to gradient estimation and may be employed in the large-sample fMRI setting using vectorised computation without sacrificing statistical accuracy.

4.2 Preliminaries

In this section, we provide a brief overview and statement of the general form of the mass-univariate LMM that accounts for missingness due to mask variability. As in the previous chapters, we shall denote θ as shorthand for the parameters (β, σ^2, D), p as the number of fixed effect parameters in the design, $V = I_n + ZDZ'$ as the marginal

variance, and $e = Y - X\beta$ as the residual vector. Throughout this chapter, we assume that θ takes the form $\theta = [\beta', \sigma^2, \text{vec}(D_1)', \ldots, \text{vec}(D_r)']'$, where vec represents the vectorization operator. We highlight that whilst this may seem like a natural representation of θ, it is by no means the only possible representation (c.f. Section 3.3.1 for further discussion).

4.2.1 The Mass Univariate Model

In the mass-univariate setting we fit and infer on hundreds of thousands of LMMs concurrently. In the setting of fMRI, each LMM corresponds to a voxel in the study's analysis mask. Adapting the notation of the previous chapters, this is represented as:

$$\begin{aligned} &Y_v = X\beta_v + Zb_v + \epsilon_v \\ &\epsilon_v \sim N(0, \sigma_v^2 I_n), \quad b_v \sim N(0, \sigma_v^2 D_v) \end{aligned} \tag{4.1}$$

where the subscript v represents voxel number. In Equation (4.1), the fixed effects and random effects design matrices (X and Z) are treated as constant across all voxels, whilst the response vector (Y), design parameters (β, σ^2 and D) and random terms (ϵ and b) vary from voxel to voxel (see Section 2.3.2 for illustration). By extension, this also means that $n, r, \{l_k\}_{k\in\{1,\ldots r\}}$ and $\{q_k\}_{k\in\{1,\ldots r\}}$, defined in Section 2.2.3, are also treated as constant across voxels. Equation (4.1) is the conventional form of the mass-univariate LMM which is typically adopted by the existing fMRI LMM software packages. As alluded to in Section 2.3.3, this model does not account for mask-variability and, as a result, must be adapted to reflect the pattern of missingness observed at each voxel.

To account for such missingness in Y_v, we adopt an MCAR (Missing Completely At Random) assumption and define the $(n \times n)$-dimensional 'missingness matrix', M_v, as a diagonal indicator matrix, where the $(i,i)^{th}$ element is 1 if the i^{th} element of Y_v is not missing and 0 otherwise. We now define $X_v = M_v X$ and $Z_v = M_v Z$ and recall that missingness in Y_v and ϵ_v is encoded as 0 (c.f. Section 2.3.1). The model

specification for a mass-univariate LMM that accounts for missingness induced by mask-variability is now given as follows:

$$\begin{aligned} Y_v &= X_v\beta_v + Z_v b_v + \epsilon_v \\ \epsilon_v &\sim N(0, \sigma_v^2 M_v), \quad b_v \sim N(0, \sigma_v^2 D_v) \end{aligned} \tag{4.2}$$

The inclusion of the missingness matrix, M_v, ensures that rows of the design at which missingness occurred for voxel v are replaced with zeros, or "zero-ed out". This process ensures that the analysis proceeds as though such 'missing-data' rows had not been included in the design at all.

An important ramification of this model construction is that the fixed effects and random effects design, X_v and Z_v, as well as the number of observations, n_v, are now treated as spatially varying. If naively implemented, a model involving a spatially varying design (such as Equation (4.2)) presents a much more formidable computational challenge than its non-spatially varying counterpart (i.e. Equation (4.1)), as additional expenses in memory and computation time arise from accounting for voxel-specific design matrices. However, in this context, it must be noted that as X_v and Z_v are constructed by 'zero-ing out' rows of X and Z, many voxels will employ identical design matrices for analysis (i.e. it is often the case that $X_{v_1} = X_{v_2}$ and $Z_{v_1} = Z_{v_2}$ for two separate voxels v_1 and v_2). In particular, it is expected that inside the brain, far from cortical boundaries, no missingness will be observed, and therefore, for many voxels, $X_v = X$ and $Z_v = Z$. In other words, whilst it is true that the design matrices, X_v and Z_v, vary spatially across the whole brain, it is expected that large contiguous groups of voxels will exist over which X_v and Z_v do not vary at all. Accounting for this "between-voxel design commonality" can result in drastic improvements in computation speed and efficiency, and thus, constitutes a key motivation behind the approach adopted by BLMM.

In the following sections, unless stated otherwise the subscript v, representing voxel number, will be dropped from our notation. For the remainder of this chapter, it is assumed implicitly that any equations provided correspond to a model of the

form (4.2) for some given voxel. In this thesis, we have now defined several conflicting subscript notations. We emphasize here that subscript s represents iteration number (c.f. Section 3.3.1), subscript v represents voxel number (c.f. Section 2.3.1) and, in the case of D, subscript k represents factor (c.f. Section 2.2.3). For example, D_s represents the $(q \times q)$ estimate of the random effects covariance matrix, D, observed during iteration s (for some voxel v, implicitly), D_v represents the $(q \times q)$ random effects covariance explicitly at the voxel v, and D_k represents the $(q_k \times q_k)$ sub-matrix of D corresponding to factor f_k (for some voxel v, implicitly). To prevent confusion, where necessary, in the following sections we shall explicitly state which subscript notation is being employed.

4.3 Methods

In this section, we describe the computational pipeline employed by BLMM to perform mass-univariate LMM analysis of large-sample fMRI data, as well as the simulations and real-data examples for which results are later presented in Section 4.4. To begin, Section 4.3.1 gives an in-depth overview of the stages of the BLMM computational pipeline. Following this, Section 4.3.2 describes simulations that shall be used to assess the correctness and performance of BLMM. Finally, Section 4.3.3 details a real-world example based on repeated-measures data drawn from the UK Biobank, demonstrating BLMM's usage in practice.

4.3.1 The BLMM Pipeline

A visual overview of the BLMM pipeline is provided by Fig. 4.1 in the form of an activity diagram. Highlighted in Fig. 4.1 are the four "stages" of the BLMM pipeline: input specification, product form computation, parameter estimation, and inference and output. Each of these stages is described in turn by Sections 4.3.1.1-4.3.1.4, respectively. Also labelled in Fig. 4.1 are the steps of the pipeline which employ

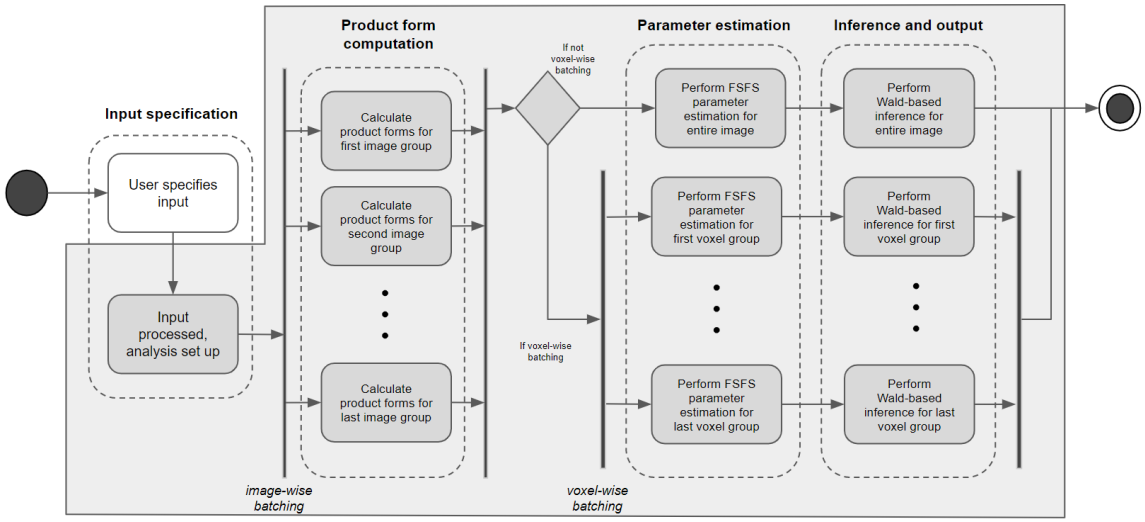

Figure 4.1: Activity diagram detailing the BLMM pipeline. The boundary of the BLMM code is indicated by the gray outline. The start and end nodes of the pipeline are represented by the black circle and nested black and white circles, respectively. Decision nodes are represented by diamonds and parallel stages of computation are represented with vertical bars. Also included are dotted lines indicating the distinct "stages" of the BLMM pipeline.

distributed computation: "Image-wise batching" and "Voxel-wise batching", each of which is further discussed further in Sections 4.3.1.2 and 4.3.1.3-4.3.1.4, respectively.

4.3.1.1 Input Specification

To specify an analysis design in BLMM, the user must provide the fixed effects design matrix, X, the response images, Y, and the random effects design matrix, Z. For specification of the random effects design matrix, Z, a similar approach to that of lmer is adopted (Bates et al. (2015)). For each factor in the design, the user provides a factor vector, g_k, and a "raw" regressor matrix, z_k. The raw regressor matrix, z_k, is a matrix of dimension $(n \times q_k)$ and contains the covariates which correspond to the random effects that are to be grouped by the k^{th} factor in the model. The factor vector, g_k, is a numerical vector of dimension $(n \times 1)$ with entries indicating to which level of the k^{th} factor each observation belongs. From these inputs, the following construction is used to obtain the random effects design matrix Z:

$$Z = \left[J_1' * z_1', \quad \ldots, \quad J_r' * z_r' \right], \quad \text{where} \quad (J_k)_{[i,j]} = \begin{cases} 1 & \text{if } (g_k)_{[i]} = j \\ 0 & \text{otherwise} \end{cases}$$

and $*$ is the Khatri-Rao product. For example, in a longitudinal design, g_1 would be a vector of subject identifiers and z_1 could contain a column of 1's for a random intercept and a column of study times for a random slope for the time effect. See Supplementary Material B Section S1 for a complete worked example.

During input specification, BLMM also provides a range of masking options. By default, the user is required to specify an analysis mask. In addition, the user may specify one mask per input image, as well as a "missingness threshold". The missingness threshold, which may be specified as either a percentage or an integer, indicates how many input images a voxel must have recorded data for in order for it to be retained in the final analysis. This threshold is essential, as while accommodating missingness is a central feature of BLMM, allowing excessive missingness (down to a small faction of the data) is not advised and may result in rank-deficient designs. Implicit masking is also supported by BLMM, with any voxel set to 0 or NaN in an input image being treated as 'missing' from the analysis.

Whilst specifying the analysis model, the user may also opt to perform hypothesis testing using an approximate T-test or F-test. To specify a hypothesis test of this form, the user must provide a contrast vector, L, representing the null hypothesis, and the type of statistic to be used to perform the test (i.e. T or F). Further detail on hypothesis testing via approximate T-statistics and F-statistics may be found in Section 2.2.4.

4.3.1.2 Product Form Computation

Computation within the BLMM pipeline begins by, for each voxel v, computing the "product forms" defined as $P_v = X_v'X_v, Q_v = X_v'Y_v, R_v = X_v'Z_v, S_v = Y_v'Y_v, T_v = Y_v'Z_v$ and $U_v = Z_v'Z_v$ (See Section 3.3.3). Following the computation of the product

forms, the original matrices, X_v, Y_v and Z_v, can be discarded since only the product forms are required for future computation (c.f. Sections 4.3.1.3-4.3.1.4). This approach is adopted by BLMM as the dimensions of the product forms do not scale with n but rather with p and q, the second dimensions of the fixed effects and random effects design matrices, respectively. As p and q are assumed to be much smaller than n, working with the product forms instead of X_v, Y_v and Z_v can provide large reductions in both memory consumption and computation time (See Section 3.3.3 for further discussion).

To compute the product forms efficiently, BLMM employs an image-wise batching approach. In this approach, the input images and model matrices are split into batches and computation is performed in parallel across several cluster nodes. More precisely, given B nodes, X and Z are vertically partitioned into evenly sized blocks $\{X^{(b)}\}_{b\in\{1,...,B\}}$ and $\{Z^{(b)}\}_{b\in\{1,...,B\}}$, respectively. The list of input images, Y, is also partitioned into B lists of $\frac{n}{B}$ input images, $\{Y^{(b)}\}_{b\in\{1,...,B\}}$, each corresponding to a partition of X and Z. Each node is assigned a partition of X, the corresponding partition of Z and the corresponding list of input images. For every voxel in the analysis mask, this means that the b^{th} node now possesses $\frac{n}{B}$ observations (including missing values encoded as zero). We denote the response vector of observations at voxel v, taken from the images $Y^{(b)}$, as $Y_v^{(b)}$.

The b^{th} node now applies voxel-specific masking to $X^{(b)}$ and $Z^{(b)}$ to obtain the spatially-varying designs $\{X_v^{(b)}\}$ and $\{Z_v^{(b)}\}$. For each voxel, v, this is achieved by considering whether v has missing data in the input images listed in $Y^{(b)}$ and "zeroing" out the corresponding rows of $X^{(b)}$ and $Z^{(b)}$ accordingly (c.f. Section 4.2.1). Given the spatially varying designs, $\{X_v^{(b)}\}$ and $\{Z_v^{(b)}\}$, and response vector, $\{Y_v^{(b)}\}$, the product forms for this partition of the model are now computed as: $P_v^{(b)} = X_v^{(b)'}X_v^{(b)}$, $Q_v^{(b)} = X_v^{(b)'}Y_v^{(b)}$,... and so forth.

For each voxel, the product forms for each partition are then sent to a designated central node. For the v^{th} voxel, the central node now computes the product forms for the entire model by summing over those sent from each node (i.e. $P_v = \sum_b P_v^{(b)}$, $Q_v = \sum_b Q_v^{(b)}$,...). This approach can be seen to produce the product forms which

were defined by Equation (3.21) by noting, for arbitrary matrices of appropriate dimension, A and B, with corresponding vertical partitions $\{A^{(b)}\}$ and $\{B^{(b)}\}$, that $A'B = \sum_b A^{(b)'}B^{(b)}$. Pseudocode for the product form computation stage of the BLMM pipeline is provided by Algorithm 6.

To prevent convergence failure due to rank deficiency, following product form calculation, any voxels for which $\text{rank}(P_v) < p$ or $\text{rank}(U_v) < q$ are dropped from the analysis. Removal of such voxels is advised during analysis but can be prevented by setting the "safeMode" option to 0. The approach described in this section was initially motivated by a similar method employed for parameter estimation of the Linear Model. Details of this method are provided in Appendix B.1, alongside a corresponding implementation written in Python. Further notes on the computational efficiency of Algorithm 6 are also provided in Appendix B.2.

Algorithm 6: Product Form Computation Pseudocode

```
On the central node
    Partition X and Z vertically into B batches.
    Partition list of Y images into B batches.
end
On each node, node b
    Read in the b^th batch of Y images as Y^(b).
    Read in the b^th batches of X and Z as X^(b) and Z^(b), respectively.
    for all voxels do
        Compute mask M_v^(b) by considering the pattern of zeros in Y_v^(b).
        Apply masking to X^(b) and Z^(b) to obtain X_v^(b) = M_v^(b) X^(b) and Z_v^(b) = M_v^(b) Z^(b).
        Compute the product forms for this partition: P_v^(b), Q_v^(b), R_v^(b), S_v^(b), T_v^(b) and U_v^(b).
        Send the product forms to the central node.
    end
end
On the central node
    for all voxels do
        Compute the product forms; P_v, Q_v, R_v, S_v, T_v and U_v by summing over results
          from nodes (e.g. P_v = sum_b P_v^(b)).
    end
end
```

4.3.1.3 Parameter Estimation

The most computationally intensive stage of any LMM analysis is the estimation of the unknown model parameters (β, σ^2, D). A common approach to estimating the unknown model parameters of the LMM is to perform Restricted Maximum Likelihood (ReML) estimation using Equation (2.5). BLMM employs the ReML variant of the Full Simplified Fisher Scoring (FSFS) algorithm of Section 3.3.1.4 to perform this task for each voxel in the analysis mask. The FSFS algorithm iteratively performs updates for the fixed effects parameter vector, β, and fixed effects variance, σ^2, using the Generalized Least Squares (GLS) estimators, and for $\{D_k\}_{k\in\{1,...,r\}}$ separately using a Fisher Scoring update step based on the "full" representation of D_k; $\text{vec}(D_k)$. We recall from Chapter 3, that in this context the word "full" refers to the fact that all of the elements of $\text{vec}(D_k)$ are treated as distinct from one another during optimization (c.f. Section 3.3.1).

Formally, during each iteration, iteration s, β and σ^2 are updated according to the GLS update rules given by Equation (3.14) and for $k \in \{1, ..., r\}$, the update rule employed for D_k takes the following form:

$$\text{vec}(D_{k,s+1}) = \text{vec}(D_{k,s}) + \alpha_s (F_{\text{vec}(D_k),s})^{-1} \partial_{k,s}. \tag{4.3}$$

Here, α_s is a scalar step size, which is initialized to $\alpha_0 = 1$ and halved each time a decrease in log-likelihood is observed between iterations, $F_{\text{vec}(D_k),s}$ is the matrix given by Equation (3.13) and $\partial_{k,s}$ is the ReML score vector for $\text{vec}(D_{k,s})$ given by:

$$\partial_{k,s} = \text{vec}\Bigg(\sum_{j=1}^{l_k} (Z'_{(k,j)} V_s^{-1} e_s)(Z'_{(k,j)} V_s^{-1} e_s)' - Z'_{(k,j)} V_s^{-1} Z_{(k,j)} + Z'_{(k,j)} V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1} Z_{(k,j)} \Bigg).$$

In this approach, to ensure that the estimates of $\{D_k\}_{k\in\{1,...r\}}$ are non-negative definite following each evaluation of the above update rules, an eigendecomposition based approach is used to project D_k to the space of non-negative definite matrices. Further

detail on the use of the eigendecomposition in this manner can be found in Appendix A.6. We recall from Chapter 3 that $F_{\text{vec}(D_k)}$ is technically not a Fisher Information matrix, but rather a simplified version of the true Fisher Information matrix for the "full" representation of θ that provides the same results during optimisation (c.f. Section 3.3.1.2).

It is important to note that, in every equation referenced above, the right-hand side can be reformulated to be expressed solely in terms of the product forms and the parameter estimates (β, σ^2, D). This is crucial to the BLMM framework as, as is noted in Section 4.3.1.2, to prevent memory consumption and computation time from scaling with n, only the product forms are retained in memory following product form computation. The GLS update rules for β and σ^2, Equation (3.14), can be written in terms of the product forms and parameter estimates as follows:

$$\begin{aligned}
\beta_{s+1} &= (P - RD_s(I + D_sU)^{-1}R')^{-1}(P - RD_s(I + D_sU)^{-1}R'), \\
\sigma^2_{s+1} &= \frac{1}{n}(S - 2Q'\beta_s + \beta_s'P\beta_s),
\end{aligned}$$

and the expressions for $F_{\text{vec}(D_k),s}$ and $\partial_{k,s}$ can be seen to be composed entirely of sub-matrices of $X'V_s^{-1}X$, $Z'V_s^{-1}Z$ and $Z'V_s^{-1}e_s$, which are given by:

$$\begin{aligned}
X'V_s^{-1}X &= P - RD_s(I + D_sU)^{-1}R', \\
Z'V_s^{-1}Z &= U - UD_s(I_q + D_sU)^{-1}U, \\
Z'V_s^{-1}e_s &= T' - R'\beta_s - UD_s(I_q + D_sU)^{-1}(T' - R'\beta_s).
\end{aligned} \tag{4.4}$$

The restricted log-likelihood function, given by Equation (2.5), can similarly be rewritten in terms of product forms as follows:

$$l_R(\theta_s) = -\frac{1}{2}\Bigg\{(n-p)\log(\sigma_s^2) + \log|I + D_sU| + \log|P - RD_s(I + UD_s)^{-1}R'| +$$

$$\sigma_s^{-2}\Big(S - 2Q'\beta_s + \beta_s'P\beta_s + (T' - R'\beta_s)'D_s(I + UD_s)^{-1}(T - R'\beta_s)\Big)\Bigg\}.$$

To choose initial values for the optimization procedure, BLMM adopts the methods proposed in Section 3.3.2 of Chapter 3. To be specific, BLMM employs the OLS estimators given by Equation (2.2) as starting estimates for β and σ^2, and Equation (3.20) as a starting estimate of $\text{vec}(D_k)$. As with the previous expressions, Equations (2.2) and (3.20) may be evaluated using only the product forms and do not require use of the original matrices X, Y and Z.

Successful convergence of the FSFS algorithm occurs when the difference in log-likelihood observed between successive iterations becomes less than a predefined tolerance (10^{-6} by default) whilst convergence failure is deemed to have occured if the maximum number of iterations (10^4 by default) is exceeded. The predefined tolerance and maximum number of iterations may be changed by the user via the "tol" and "maxnit" options, respectively. We note here that the default value of 10^4 for maxnit is likely over-cautious as the simulations of Chapter 3 found that, for a range of well-specified designs, the FSFS algorithm typically converged within $5-30$ iterations (c.f. Section 3.4.2.1).

Pseudocode for the ReML FSFS algorithm is provided by Algorithm 7. In certain instances, further improvements in terms of computational performance can be obtained by utilising structural features of the analysis design which simplify the above expressions. In particular, BLMM has been optimized to give faster performance for models that contain (i) one random effect grouped by one random factor and (ii) multiple random effects grouped by one random factor. Further detail and discussion of the improvements employed for such models can be found in Appendix B.3.

In the fMRI analysis setting, careful attention must be given to computational efficiency during parameter estimation, as parameters must be estimated for every voxel in the analysis mask. As noted in Section 4.1.1, computation that is performed independently for one voxel at a time can result in prohibitively slow computation speeds. It follows that, in order to execute LMM parameter estimation for fMRI data in a practical time frame, computation must be streamlined to allow for parameter estimation to be performed concurrently across multiple voxels at once.

Algorithm 7: ReML Full Simplified Fisher Scoring Pseudocode

```
Assign β, σ² and {D_k}_{k∈{1,...r}} to initial estimates using Equations (2.2) and (3.20).
while current l_R(θ) and previous l_R(θ) differ by more than a predefined tolerance do
    Update β and σ² using Equation (3.14).
    for k ∈ {1, ...r} do
        Update vec(D_k) using Equation (4.3).
        Project D_k to the space of non-negative definite matrices using an
          eigendecomposition.
    end
    Recompute l_R(θ) using Equation (2.5).
    Assign α = α/2 if l_R(θ) has decreased in value.
end
```

On a single node, parameter estimation may be parallelised across voxels by using broadcasted computation, which exploits the repetitive nature of simplistic operations to streamline calculation. A considerable advantage in using the FSFS algorithm is that it relies upon only conceptually simplistic operations (such as matrix multiplication, matrix inversion and the eigendecomposition), for which a wealth of broadcasted support already exists in modern programming languages such as MATLAB and Python. By utilizing this support, the FSFS algorithm may be executed for multiple voxels concurrently in order to achieve quick and efficient computational performance (c.f. Section 4.4.1 for an assessment of BLMM computation time).

If multiple nodes are available, parameter estimation may also be parallelised further by partitioning the analysis mask into "batches" of voxels and having each node perform parameter estimation for an individual batch. This approach is also provided by BLMM and is referred to as "voxel-wise batching". The ability to distribute computation in this manner is advantageous in situations where the analysis design is large and the product forms cannot be read into memory for many voxels at once. However, this additional layer of parallelisation may not be necessary for smaller designs and is therefore offered optionally in the BLMM package (as shown in Fig. 4.1).

4.3.1.4 Inference and Output

The final stage of the BLMM pipeline is to perform inference on the fixed effects parameters and output the analysis results in NIfTI format. To support null-hypothesis testing for the fixed effects parameter vector, β, BLMM adopts an approach that is similar to that taken by the popular univariate LMM packages lmerTest, MIXED and PROC-MIXED (c.f. Section 4.1.2). In this approach, the ReML estimates of the parameter vector, $(\hat{\beta}, \hat{\sigma}^2, \hat{D})$, are used to construct T or F test statistics. To obtain corresponding p-values, a WSDF-based approach is then employed to model the sampling distribution of the T- and F-statistics.

Assuming the user has provided a contrast vector, L, to specify a null hypothesis $H_0 : L\beta = 0$, BLMM will compute the corresponding T-statistic or F-statistic for each voxel as follows:

$$T = \frac{L\hat{\beta}}{\sqrt{\hat{\sigma}^2 L(X'\hat{V}^{-1}X)^{-1}L'}}, \qquad F = \frac{\hat{\beta}'L'[L(X'\hat{V}^{-1}X)^{-1}L']^{-1}L\hat{\beta}}{\hat{\sigma}^2 \text{rank}(L)},$$

where $\hat{V} = I_n + Z\hat{D}Z'$. BLMM assumes that the distributions of T and F are reasonably approximated with a student's t- or F-distribution. As the distributional assumptions for T and F are approximate, and not exact, the degrees of freedom are unknown and must be estimated (see Section 3.3.5 and Appendix A.9). The estimation is performed using a WSDF-based approach based on the Welch-Satterthwaite equation;

$$v(\hat{\eta}) = \frac{2(S^2(\hat{\eta}))^2}{\text{Var}(S^2(\hat{\eta}))},$$

where $\hat{\eta}$ represents an estimate of the variance parameters $\eta = (\sigma^2, D_1, \ldots D_r)$ and $S^2(\hat{\eta}) = \hat{\sigma}^2 L(X'\hat{V}^{-1}X)^{-1}L'$. The numerator of the above expression may be evaluated directly. However, an approximation, obtained using a second order Taylor expansion, must be employed to estimate the unknown variance term in the denominator (see Section 3.3.5 for further detail).

While other tools (lmerTest, MIXED, PROC-MIXED) require numerical opti-

misation to obtain this denominator, we make use of the closed-form expressions presented in Section 3.3.5 which may be used to evaluate the terms of the Taylor expansion directly. These expressions not only provide a computationally efficient alternative to the use of numerical optimisation, but also can be expressed purely in terms of the product forms. As a consequence of this, using a similar approach to that of Section 4.3.1.3, BLMM is able to perform degrees of freedom estimation by utilising only the product forms and parameter estimates at each voxel, and by employing operations that can be broadcasted or further accelerated through batch-wise parallelisation.

Once an analysis has been executed using BLMM, parameter estimate images for β, σ^2 and D, contrast images of the form $L\hat{\beta}$ and statistic images with corresponding $p-$value images (which must then be corrected for multiple comparisons) are output. Correction for multiple comparisons is essential, though the non-linear estimation process precludes the use of standard random field theory and, while permutation or wild bootstrap methods are available for mixed models, such methods add substantial computational burden. Hence, either control of the Family-Wise Error rate (FWE) with the Bonferroni correction (Dunn (1961)) or the False Discovery Rate (FDR) with the Benjamini-Hochberg method (Benjamini and Hochberg (1995)) should be used to account for multiple comparisons.

Following the conclusion of a BLMM analysis, BLMM also provides Likelihood Ratio Testing (LRT) for comparison of LMM analyses which contain a single-random factor and differ only by the inclusion of random effects (c.f. Section 2.2.4). This process involves first computing the LRT statistic given by Equation (2.8). It is then assumed that the LRT statistic follows an even mixture distribution of χ^2 distributions (see Verbeke and Molenberghs (2001) for further detail). Under this distributional assumption, BLMM may then be used to generate uncorrected p-value significance images for the LRT statistic. Examples of this method in practice are provided by Section 4.4.2.

4.3.2 Simulation Methods

In order to quantitatively assess and demonstrate the computational accuracy and efficiency of BLMM, extensive simulations were conducted. Simulated data was generated for nine simulation settings: three sample sizes ($n = 200, 500$ and 1000, respectively), each generated under three experimental designs. The experimental designs considered for simulation in this chapter reflect the two settings for which parameter estimation in BLMM has been explicitly optimized (c.f. Section 4.3.1.3 and Appendix B.3), as well as the most general model specification that BLMM caters to. These were the settings in which the experimental design contains; (i) one random factor which groups one random effect, (ii) one random factor which groups multiple random effects and (iii) multiple random factors, respectively. These models correspond to common use cases designs employing, for example, (i) a subject-level random intercept in a repeated measures setting, (ii) a subject-level random intercept and slope in a repeated measures setting, and (iii) a subject-level intercept and site-level intercept for repeated measures taken from multiple subjects across multiple sites. For each simulation setting, 1000 individual simulation instances were performed, and all reported results are given as averages taken across the 1000 instances.

In each simulation setting, the fixed effects parameter vector, β, and the fixed effects variance, σ^2, were fixed across simulation instances and given by $\beta = [4, 3, 2, 1, 0]'$ and $\sigma^2 = 1$, respectively. The fixed effects design matrix, X, contained an intercept and four regressors, each of which varied across simulation instances and consisted of values generated according to a uniform $[-0.5, 0.5]$ distribution. Each experimental design enforced a different structure on the random effects design and covariance matrices, Z and D. The first experimental design (Design 1) included a single factor which grouped one random effect into 100 levels (i.e. $r = 1$, $q_1 = 1$ and $l_1 = 100$). The second experimental design (Design 2) included a single factor which grouped two random effects into 50 levels ($r = 1$, $q_1 = 2$, $l_1 = 50$). The third experimental design (Design 3) included two crossed factors, the first of which grouped two random effects into 20 levels and the second of which grouped one random effect into 10

levels (i.e. $r = 2, q_1 = 2, q_2 = 1, l_1 = 20$ and $l_2 = 10$). In all simulation instances, the first random effect appearing in Z was an intercept. Any additional random effects regressors varied across simulation instances and were generated according to a uniform $[-0.5, 0.5]$ distribution. For each factor, observations were assigned to levels uniformly at random so that the the probability of an observation belonging to any specific level was the same for all levels. The diagonal and off-diagonal elements of the random effects covariance matrix for each factor were held fixed across simulation instances and given as 1 and 0.5, respectively.

The spatially varying random terms, ϵ and b were generated as images of Gaussian noise, with the appropriate covariance between images induced for b. The response images, Y, were then calculated using X, Z, β, b and ϵ according to Equation (2.3). An isotropic Gaussian filter with a Full Width Half Maximum (FWHM) of 5 was then applied to the response images, Y, to induce spatial correlation across voxels. Following this, for each response image, random perturbations were applied to the FSL standard 2mm MNI brain mask in order to generate a corresponding "random mask" image, thus simulating missingness near the edge of the brain (c.f. Supplementary Material B Section S2). The "random" masks were applied to the response images to finally obtain a masked smoothed random response, roughly resembling null fMRI data. A visual overview of the data generation process is provided by Figure 4.2. All NIfTI volumes generated for simulation were of dimension $(100 \times 100 \times 100)$ voxels. It must be stressed that this process was not designed to simulate realistic fMRI data rigorously, but rather data that had approximately the same shape, size, smoothness, and degree of missingness as might be observed in real fMRI data.

In order to assess the accuracy and performance of the parameter estimation in each simulation instance, the R function lmer was also used to obtain parameter estimates for each voxel in the analysis mask. To measure computation speed, we define the 'Serial Computation Time' (SCT) for BLMM parameter estimation as the time in seconds that would have been spent executing the FSFS algorithm if the computation performed by each node had been run back-to-back in serial. The serial computation time for lmer parameter estimation is similarly defined as the total amount of time

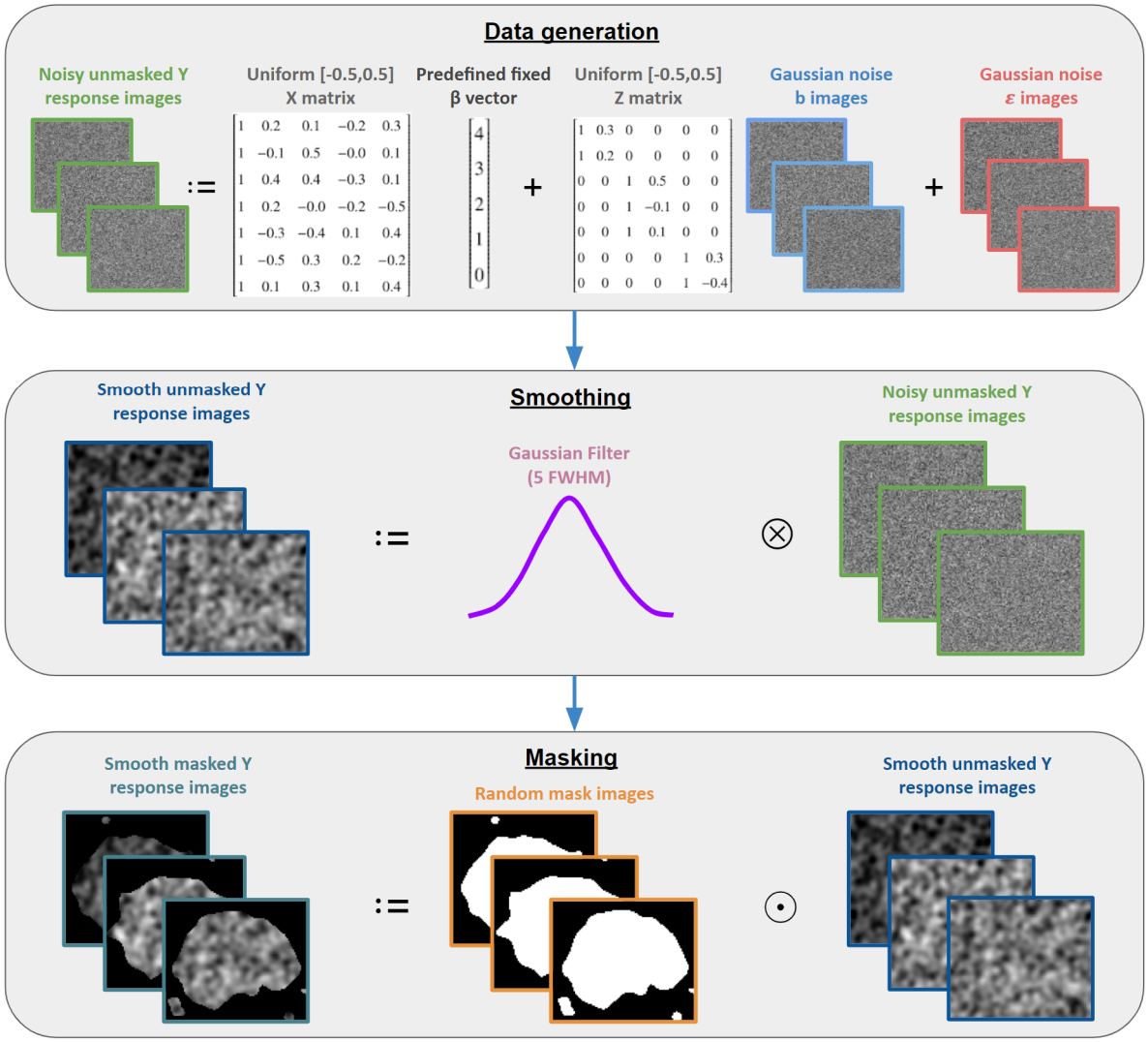

Figure 4.2: A visual representation of the pipeline employed to generate the simulated data of Section 4.3.2. The first box depicts the model which was used for data generation, notably highlighting that ϵ and b varied across space. The second box details the smoothing process, with the $\otimes$ symbol representing convolution in this instance. The third box details the masking stage, with $\odot$ representing the Hadamard (element-wise) product.

that would have been spent executing the function 'optimizeLmer' had all computation been performed in serial. The performance of lmer and BLMM were contrasted in terms of SCT averaged across simulation instances, whilst the parameter estimates produced by lmer and BLMM were compared in terms of image-wide mean absolute difference to assess the correctness of the FSFS algorithm employed by BLMM. It is noted here that the computation of product forms is not included in our evaluation of SCT for BLMM. However, we do not believe this has impacted the results of our analysis as preliminary testing demonstrated that the time spent performing product form computation was many orders of magnitude (approximately 10^5) smaller than the time spent performing ReML estimation using the FSFS algorithm.

The primary purpose in choosing lmer as a baseline for comparison was to demonstrate the substantial computational benefits of using the BLMM pipeline for mass-univariate analysis in the place of naive computation via 'for loops' and lmer. We stress that it is not the author's intention for the results of this analysis to be interpreted as a reflection on the ability of the lme4 package, which is not designed for use in the mass-univariate setting, but rather the efficiency of the FSFS algorithm when combined with vectorised computation. To ensure the missingness capabilities of BLMM were exhaustively tested, each simulated analysis employed a lenient missingness threshold of 50%.

In sum, the simulations we have described assess BLMM's (i) correctness in terms of parameter estimation, (ii) ability to handle missing data, and (iii) computation speed for parameter estimation. All reported results were obtained using an SGE cluster with Intel(R) Xeon(R) Gold 6126 2.60GHz processors each with 16GB RAM.

4.3.3 Real Data Methods

As a demonstration of the large-scale capabilities of BLMM, here we provide an example involving a much larger model than those considered during the simulations of Section 4.3.2. In this example, we utilise data from the UK Biobank, in which repeated measures were recorded for 2461 subjects, each of whom completed a "faces vs

shapes" task twice across separate visits. In FSL, a first-level analysis was conducted independently for each visit for each subject. In each first-level analysis, the task design was regressed onto Blood Oxygenation Level Dependent (BOLD) response. From each first-level design, a Contrast Parameter Estimate (COPE) map was generated that represented the average difference in the BOLD response between subjects being shown images of faces and images of abstract shapes.

At the group level, BLM and BLMM were used to perform parameter estimation for three models, each designed to estimate the average group level 'faces>shapes' response. In each model, the response images, Y, were the COPE maps that were output by FSL during the first-level analysis, registered to MNI space. The fixed-effects design matrix, X, in each model included an intercept, the cross-sectional effect of age (average age per subject), longitudinal time (inter-visit time in years, centred by subject), sex, a cross-sectional age-sex interaction effect and the Townsend deprivation index (a measure of socio-economic status).

The first of the three group-level analyses (Model 1) was a linear regression model estimated using BLM. As it was a standard linear regression, this model included no random-effects design matrix, Z. The second group-level model (Model 2) was analysed using BLMM and included a subject-level random intercept in the random-effects design matrix. The third of the group-level models (Model 3) was run using BLMM and included both a subject-level random intercept and a subject-level random slope for longitudinal time in the random-effects design matrix. For each model, approximate T-statistics were computed, using the methods of Section 4.3.1.4, for the model intercept (the group-level average response to the 'faces>shapes' task), the cross-sectional age and longitudinal time. Once the parameters of each model had been estimated, to assess goodness of fit, model comparison was performed using the LRT method detailed in Sections 2.2.4 and 4.3.1.4.

It must be noted that, as each model contained two observations per subject and model 2 contained two random effects per subject, model 2 contained the same number of observations as random effects. As a result of this, model 2 may be expected to be unidentifiable for many voxels, as parameter estimation for any voxel

with missing data will attempt to model at least two random effects using less than two visits. This choice of model is deliberate as model 2 is an extreme use-case that both stress-tests the BLMM code and serves as a clear example of a model that is expected to be rejected by the LRT procedure. It must be stressed that, by including model 2 in this example, it is not our intention to endorse estimation of models that have been ill-specified in this manner. Model 2 is provided purely for the purposes of demonstration and can only be estimated in BLMM by turning off a 'safeMode' setting during input specification.

The primary purpose of the analyses described above is to demonstrate BLMM's usage in practice and highlight BLMM's efficiency and scalability via a worked example. In order to assess computational efficiency, model 2 was also estimated voxel-wise using lmer and, as in Section 4.3.2, serial computation times were recorded for parameter estimation for both lmer and BLMM. However, the same comparison could not be performed for model 3, as the default settings in lmer will not allow for models with equal numbers of random effects and observations to be estimated. In Section 4.4.2, results are reported for both Likelihood Ratio Tests (model 1 vs model 2 and model 2 vs model 3), with p-value significance maps for the approximate t-tests then being provided for the selected model. All reported significance regions were obtained using a 5% Bonferonni corrected threshold. All analysis results were obtained using an SGE cluster with Intel(R) Xeon(R) Gold 6126 2.60GHz processors, each with 16GB RAM.

4.4 Results

4.4.1 Simulation Results

Across the nine simulation settings outlined in Section 4.4.1 (three designs across three sample sizes), all parameter estimates and maximised restricted likelihood criteria produced by BLMM were near identical to those produced by lmer. In particular, extremely strong agreement was observed between the parameter estimates produced

by BLMM and lmer for both voxels with missing observations and voxels with all observations present, illustrating BLMM's capacity for handling missing data. For each experimental design, the largest mean absolute difference for parameter estimation was observed in the estimation of the random effects covariance parameters (the unique non-zero elements of D) in the $n = 200$ setting. The observed mean absolute difference in the random effects covariance parameter estimates produced by BLMM and lmer for this setting were 6.86×10^{-9}, 4.39×10^{-5} and 6.02×10^{-3} for designs 1, 2 and 3 respectively. Assuming double precision and a machine level tolerance of $2^{-26} \approx 1.49 \times 10^{-8}$, it can be seen that this means the estimates produced by BLMM and lmer were indistinguishable at the machine precision level for design 1 and very similar for designs 2 and 3. Similar results may also be observed across simulations for the maximised ReML criterions produced by BLMM and lmer (see Supplementary Material B Sections S3-S6).

The observed SCTs for each simulated setting are presented in Fig. 4.3. As can be seen from Fig. 4.3, BLMM significantly outperformed computation via 'for loops' and lmer and, notably, appeared to maintain an approximately constant computation time as the number of observations, n, increased. In contrast, computation time for lmer appeared to increase as the number of observations increased. These results match expectation as, as noted in Section 4.3.1.2, the BLMM pipeline begins by computing the product forms and discarding any model matrices which had dimensions scaling with n. This means that during parameter estimation (the most intensive stage of the BLMM pipeline, which dominates the computation time) it is expected that computation should not scale as a function of n. As computation time does scale as a function of q, however, it should be noted that, as shown in Fig. 4.3, when q and n are close in magnitude the benefits of the product form approach to computation are less pronounced.

The results of Fig. 4.3 demonstrate BLMM's strong computational efficiency for analysing the designs which were simulated. However, it should be noted that the difference in SCT between lmer and BLMM may be smaller than those observed in these simulations for analyses involving designs in which the second dimension

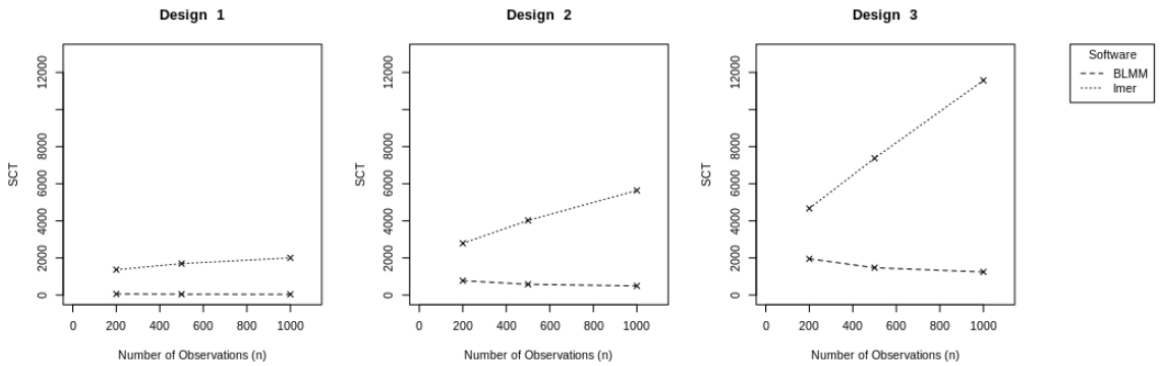

Figure 4.3: Observed serial computation times for each experimental design, displayed as a function of the number of observations, n. Displayed are the SCT in seconds for BLMM (dashed) and lmer (dotted).

of the random effects design matrix, q, is very large. The reason for this is that, as q increases, the storage requirements associated with each voxel increase and, as a result, parameter estimation can be performed for fewer voxels concurrently via vectorisation. In the simulations presented in this section, q was set to 100 in designs 1 and 2 and set to 50 in design 3. We note that it would be of benefit to conduct further detailed analysis of BLMM's performance compared to that of lmer for models with higher values of q. However, at present, such comparative simulations are practically infeasible due to the inordinate computation time they would require (the reported simulations required several months to run, generated over five million simulated brain images and executed approximately 10^7 univariate LMM analyses in lmer). Whilst such extensive simulations may not be possible, we highlight here that the analyses presented in Section 4.4.2 provide further evidence of BLMM's strong performance for two models in which q is much larger than considered in the simulations presented in this section ($q = 2461$ and $q = 4922$ in models 1 and 2, respectively).

4.4.2 Real Data Results

The results of the UK Biobank real data example of Section 4.3.3 are given by Fig. 4.4. The first and second rows of Fig. 4.4 display the regions of significance obtained from LRTs for model comparison between models 1 and 2, and models 2 and 3, respec-

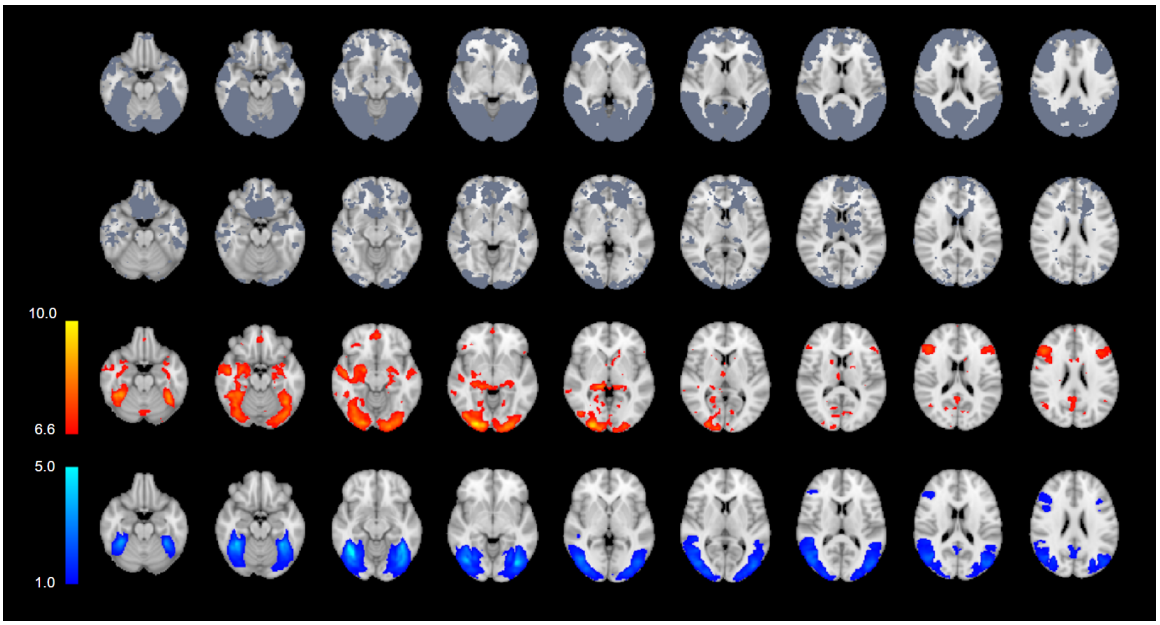

Figure 4.4: First row: LRT results for comparison of model 1 and model 2; for this row, regions highlighted demonstrate where evidence was found that inclusion of a random subject intercept significantly affected the results of the analysis. Second row: LRT results for comparison of model 2 and model 3; regions highlighted indicate where the inclusion of a random slope for longitudinal time significantly impacted the analysis. Third row: Effect of "Faces > Shapes" p-values, given on the $-\log_{10}(p)$ scale, are shown; for this row, voxels shown are those Bonferroni-significant for the "Faces > Shapes" contrast. Fourth row: Variance estimates for the subject-level random intercept, thresholded at 1. Voxels shown are those which displayed notable variability at the subject-level for the random intercept included in the model.

tively. The results of the first LRT, comparing model 2 (the random intercept model) and model 1 (the linear regression), highlight most of occipital, parietal and frontal lobes. This observation suggests that, in these regions, evidence was observed that the inclusion of subject-level random intercepts substantially influenced the outcome of the analysis. This conclusion is reasonable and reflects regions where there is non-negligible BOLD response to this task.

The second LRT, comparing model 3 (the random intercept and slope model) to model 2, only identifies two different regions: orbitofrontal areas, which are often subject to signal loss, and white matter areas in the centre of brain, which have the poorest SNR with this multicoil acquisition. The sporadic and noisy appearance of the regions identified here could indicate that these regions were influenced by

idiosyncratic temporal changes due to variation in head placement or simply by large random temporal changes. In either case, it can be seen that the inclusion of a random slope in the analysis had little impact on anatomical regions of practical relevance to the 'faces vs shapes' task. The conclusion of the LRTs is, therefore, that model 2 should be chosen as the suitable model upon which inference can be performed using approximate T-tests.

Of the three contrasts that were estimated for model 2 using the approximate T-test, only the first contrast (the main effect of the "Faces vs Shapes" task) reported significant regions following a 5% Bonferroni-corrected threshold. The thresholded p-value significance map for this contrast, reported on the $-\log_{10}(p)$ scale, is displayed on the third row of Fig. 4.4. In this instance, the occipital lobe, known for its role in processing perceptual information (c.f. Rehman and Khalili (2019)), has been identified as significant. This result matches expectation, given the visual nature of the faces vs shapes task.

Also provided in the fourth row of Fig. 4.4 is an image representing the estimated subject-level variability for the random intercept. This image serves as a useful diagnostic, as it highlights regions at which inclusion of the random intercept notably contributed to the analysis results. In this example, it can be seen that the regions at which the subject-level variance estimate exceeded 1 exhibit similarity to those regions which were deemed significant during the approximate T-test. This illustrates the vital role that the inclusion of a random intercept played in obtaining the results of this analysis and reinforces the conclusion of the LRT testing procedure.

In total, computation time in BLMM for this analysis took approximately 55 minutes for model 2 and 4 hours for model 3, using 500 nodes. The extreme computation time observed for model 3 can be attributed to the identifiability problems noted in Section 4.3.3, with parameter estimation for approximately 150 voxels exceeding the default maximum iteration limit. By way of comparison, for model 2, the SCT observed for BLMM was approximately 77.6% of that observed for lmer. As noted previously, this difference is less pronounced than those observed in the simulations

of Section 4.4.1, as the second dimension of the random effects design matrix is extremely large ($q = 2461$) and close in magnitude to the value of n ($n = 4922$). The linear regression model, model 1, was run using BLM and took approximately 5 minutes using 60 computational nodes. The full analysis results for all three models are publicly available and may be accessed on NeuroVault (see Section 4.6).

4.5 Discussion and Conclusion

In this chapter, we have detailed and presented BLMM, a freely available software package for performing LMM parameter estimation and inference on large-sample fMRI datasets. LMM computation for fMRI is an extremely computationally intensive task and, as a result, the work presented in this chapter is both informed and limited by the currently available support in terms of software and technology. For this reason, BLMM is a continually evolving project, and it is expected that much of the methodology currently implemented in BLMM may gradually change over time.

A large driving force motivating the approaches taken throughout this chapter is the current lack of available support for broadcasted sparse matrix operations. While the available support for sparse matrix methodology in programming languages such as MatLab and Python has substantially improved in recent years, currently, there is little support for performing sparse matrix operations concurrently many times. Unfortunately, this substantially limits the utility of sparsity-based approaches to LMM estimation in the context of fMRI analysis. This is due to the substantial overheads which would be accrued if LMM estimation were performed independently for each voxel in an image. An undesirable ramification of this is that the BLMM code, which has been explicitly optimised to account for the patterns of sparsity present in single-factor models (see Appendix B.3), may not provide comparable performance for multi-factor models in which q is very large. However, as programming languages evolve and new support becomes available, this is expected to change, and we predict that future development of BLMM is highly likely to incorporate sparse matrix methodology into its approach to analysing multi-factor models. For this reason, we

suggest that the incorporation of sparse matrix methodology into BLMM may form a substantial basis for future development.

Another potential avenue for future development focuses on how BLMM currently handles file input and output on SGE clusters. In recent years, in Python especially, there has been a strong movement towards standardising and streamlining cluster-based computation. A project of particular note is the Dask Python package, which acts as a standardised specification to encode parallel algorithms and file I/O (Rocklin (2015)). As noted in Section 4.1.1, Dask is heavily employed by the AFNI package 3dLME and provides substantial benefits both in terms of computation time and ease of distribution. Although many of the broadcasted operations employed by BLMM are yet to be supported by Dask, we suggest that incorporating the approach of AFNI's 3dLME by integrating Dask with the existing BLMM code-base may provide further speed improvements and could form the foundation of future development within BLMM.

There is also much room for further theoretical development of BLMM. In Section 3.3.4 of Chapter 3, we presented a general approach for performing constrained optimisation to enforce structure in the random effects covariance matrix. This concept may be of utility in a wide range of commonplace applications, as structured covariance matrices are a feature of many popular statistical models. Such applications include, for example, modelling genetic components in twin studies and auto-correlation in time-series analyses. Whilst the approaches outlined in Chapter 3 are sufficiently general to perform such analyses in the univariate setting, it is not immediately apparent whether they may be feasibly executed on a mass-univariate scale. For this reason, we suggest that the potential for enforcing covariance structure in BLMM merits further investigation.

4.6 Data and Code Availability

In this chapter, we have used data from The UK Biobank (Allen et al. (2012)). The BLMM toolbox, as well as the code used for the simulations and timing comparisons

described in Sections 4.3.2 and 4.3.3, are available at:

https://github.com/TomMaullin/BLMM

The BLM toolbox, which is also referenced several times throughout this work, may be found at:

https://github.com/TomMaullin/BLM

The images generated by BLMM for the three models discussed in Sections 4.3.3 and 4.4.2 may be found at:

Model 1: https://identifiers.org/neurovault.collection:9881
Model 2: https://identifiers.org/neurovault.collection:10448
Model 3: https://identifiers.org/neurovault.collection:10450

The results of the LRT's for model comparison discussed in Sections 4.3.3 and 4.4.2 may also be found at:

https://identifiers.org/neurovault.collection:10451

Chapter 5

Spatial Confidence Regions for Combinations of Excursion Sets in Image Analysis

In this chapter, we develop a theory of confidence regions for intersections and unions of excursion sets. This work was developed under the supervision of Professor Armin Schwartzman and Professor Thomas E. Nichols. A manuscript corresponding to the work in this chapter is currently in submission for publication. A full proof of the theorems stated in this chapter is provided by Appendix C. Extensive simulation results are documented in the 'Supplementary Material C' document provided alongside this thesis.

5.1 Introduction

The collection and analysis of imaging data, modelled as a random field sampled on a spatial domain, is central to a broad range of scientific disciplines such as neuroimaging, climatology, and cosmology. Often, spatial data drawn from n observations of a random field are combined to obtain an estimate, $\hat{\mu}_n$, of some spatially varying target function μ. The target function μ is defined on a closed spatial domain, $S \subset \mathbb{R}^N$,

and maps spatial locations to some variable of interest in $\mathbb{R}$ (e.g. seismic activity, infrared heat, changes in blood oxygenation in the brain, etc.). In such applications, it is typically desirable to ask "At which locations does the target function exceed a certain value, c?". For example, "Where has a significant change in temperature occurred?"; "Where do significant heat readings indicate the presence of celestial bodies?". Such questions are addressed by the study of excursion sets, i.e. sets of the form $\{s \in S : \mu(s) \geq c\}$.

A wealth of literature has previously focused on documenting the geometric and topological properties of excursion sets of random functions (c.f. Adler (1981), Torres (1994), Cao (1999), Azas and Wschebor (2009), Adler and Taylor (2009)). These properties include, for example, the Euler characteristic (a measure of topological structure), the Hausdorff dimension (indicating how fractal the set may be), Lipschitz Killing curvatures (describing high-dimensional volumes and areas) and Betti Numbers (describing how many stationary points appear above the threshold, c) (Worsley (1996), Adler (1977), Adler et al. (2017), Pranav et al. (2019)). Much of this work, though, is limited to homogeneous stationary processes, i.e. those with zero mean and a spatial covariance structure which is dependent upon only distance. Only the most recent literature, including the work of this chapter, concerns processes with a non-zero mean function and a potentially heterogeneous covariance function.

In addition, at present there is little published concerning how excursion sets may be compared to one another. In many applications, it is typical for a researcher to collect data from two or more study conditions and contrast the results to draw some meaningful inference (e.g. temperature changes in winter vs summer, heat measurements gathered using different imaging modalities, brain activity in healthy subjects across a range of tasks, etc.). Such settings are akin to having M spatially aligned, potentially correlated, estimates, $\hat{\mu}_n^1, \hat{\mu}_n^2, ...\hat{\mu}_n^M : S \to \mathbb{R}$, of M target functions, $\mu^1, \mu^2, ...\mu^M$.

Often, pertinent questions that arise in such settings may be formally expressed using logical statements which involve conjunctions (logical 'and's), negations (logical 'not's) and disjunctions (logical 'or's). For example, a climatologist may ask where

significant changes in temperature were observed during either winter *or* summer (i.e. "Where does μ^1 *or* μ^2 exceed c?'). Alternatively, a neuroscientist may ask "At which locations in the brain did subjects exhibit signs of cognitive behaviour during one task *and not* another?" (e.g. "At which locations does μ^1, *and not* μ^2, exceed c?").

One approach to answering such questions is to employ null-hypothesis testing. A hypothesis test of a disjunction of nulls to assess evidence for a conjunction of alternative hypotheses can be conducted with an 'intersection-union' test. An intersection-union test rejects at level α when all of the individual nulls are rejected at level α. This approach has proven particularly useful for tests of bioequivalence (Berger and Hsu (1996), Berger (1997)), but does not consider the spatial aspect that is our focus here. Here, we provide novel theory which allows the quantification of *spatial uncertainty* present in the locale of intersections and unions of random excursion sets.

It is common practice for researchers to compare images corresponding to different study conditions via rudimentary visual inspection. (In fMRI, just a few recent examples of qualitative assessment of 'overlap' and 'differences' are Zhang et al. (2021), Dijkstra et al. (2021), Ferreira et al. (2021)). Such a subjective practice may lead to biased or misleading results. In recent years, much literature has been published on the theory of confidence regions, which focuses on quantifying spatial uncertainty by providing, with a fixed confidence, sub- and super-sets bounding the excursion set of a single target function (Sommerfeld et al. (2018), Bowring et al. (2019), Bowring et al. (2021)) (see Section 2.3.5 for further detail). However, the theory of confidence regions does not currently offer a means for investigating logical conjunctions and disjunctions of exceedance statements (i.e. intersections and unions of excursion sets). The work of this chapter addresses this issue by first proposing a theory of spatial confidence regions for 'conjunction inference' (i.e. inference for logical statements involving conjunctions) in image analysis. Following this, the proposed method is extended to allow the investigation of statements containing disjunctions and negations. This work actively contributes towards the literature on confidence regions by developing a new theory allowing for the generation of confidence regions for 'conjunction

inference' (i.e. inference for logical statements involving conjunctions) in image analysis. The proposed method is further extended to allow the investigation of statements containing disjunctions and negations.

It must be noted that alternative approaches have been developed for generating confidence regions for latent Gaussian models (c.f. Bolin and Lindgren (2015), Mejia et al. (2020)). Such methods are similar in nature to those proposed by Sommerfeld et al. (2018) but instead adopt a Bayesian perspective and utilize computationally-intensive numerical integration techniques to achieve the desired confidence level. We shall not discuss such approaches in detail in this work but note here that, in a similar manner to those proposed by Sommerfeld et al. (2018), the methods of Bolin and Lindgren (2015) may not be employed for the conjunction and disjunction inference which is the primary focus of this work.

Formally, the work of this chapter primarily focuses on the intersection of M excursion sets (i.e. the region over which the conjunction statement "$\mu^1(s) \geq c$ *and* $\mu^2(s) \geq c$ *and ... and* $\mu^M(s) \geq c$" holds). We denote $\mathcal{M} = \{1, ..., M\}$, and for each $i \in \mathcal{M}$ (i.e. for each study condition) define the i^{th} true excursion set and i^{th} estimated excursion set as:

$$\mathcal{A}_c^i = \{s \in S : \mu^i(s) \geq c\} \quad \text{and} \quad \hat{\mathcal{A}}_c^i = \{s \in S : \hat{\mu}_n^i(s) \geq c\},$$

respectively. Next, we define $\mathcal{F}_c$ to be the intersection of the sets $\{\mathcal{A}_c^i\}_{i \in \mathcal{M}}$, and $\hat{\mathcal{F}}_c$ to be the intersection of the sets $\{\hat{\mathcal{A}}_c^i\}_{i \in \mathcal{M}}$. In other words, the spatial region we are interested in, and our estimate of that region, are given by:

$$\mathcal{F}_c = \left\{ s \in S : \min_{i \in \mathcal{M}} \mu^i(s) \geq c \right\} \quad \text{and} \quad \hat{\mathcal{F}}_c = \left\{ s \in S : \min_{i \in \mathcal{M}} \hat{\mu}_n^i(s) \geq c \right\},$$

respectively. An illustration of $\mathcal{F}_c$ in a setting in which $M = 2$ is provided by Fig. 5.1.

We note here that the above construction can be adjusted to allow c to vary across study conditions (i.e. "$\mu^1(s) \geq c_1$ *and ... and* $\mu^M(s) \geq c_M$" for $c_1, ..., c_M \in \mathbb{R}$). Such

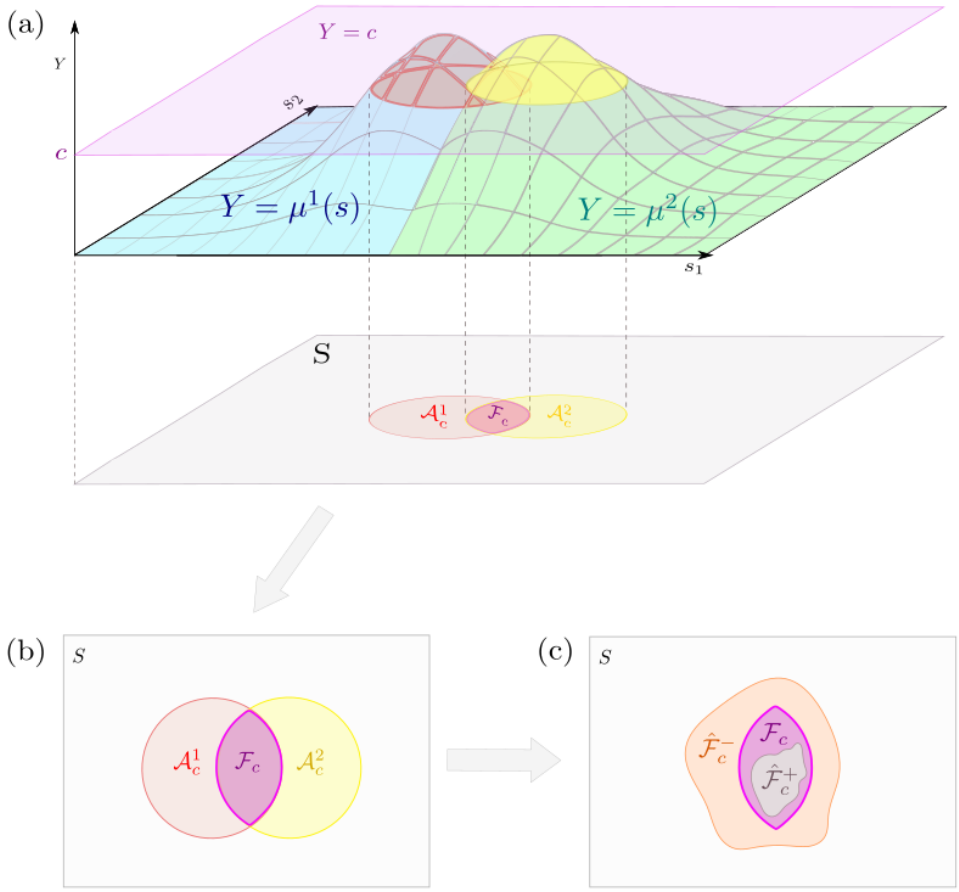

Figure 5.1: Illustration of the set $\mathcal{F}_c$ for a setting in which $M = 2$. Shown in (a) are two spatially-varying target functions, $\mu^1(s)$ (blue) and $\mu^2(s)$ (green), overlaid and thresholded at the level c. In (b), the regions at which $\mu^1(s) \geq c$ and $\mu^2(s) \geq c$ are displayed as red and yellow circles, respectively, with the intersection set, $\mathcal{F}_c$, illustrated in purple. In (c), potential confidence regions, $\hat{\mathcal{F}}_c^+$ (grey) and $\hat{\mathcal{F}}_c^-$ (orange), for $\mathcal{F}_c$ are illustrated.

an adjustment would involve simply translating each field upwards or downwards (e.g. substituting $\{\mu^i\}_{i\in\mathcal{M}}$ for $\{\mu^i - c + c_i\}_{i\in\mathcal{M}}$). For ease in the proceeding text, we shall assume c is equal for all study conditions.

Our goal is to obtain confidence regions $\hat{\mathcal{F}}_c^+$ and $\hat{\mathcal{F}}_c^-$ such that the below probability holds asymptotically;

$$\mathbb{P}\left[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-\right] = 1 - \alpha, \tag{5.1}$$

for a predefined tolerance level α ($\alpha = 0.05$, for example). To generate such regions, we build on the theory of Sommerfeld et al. (2018) (c.f. Section 2.3.5), to show that under an appropriate definition of $\hat{\mathcal{F}}_c^+$ and $\hat{\mathcal{F}}_c^-$, the above probability may be approximated using quantiles of a well-defined random variable. Using a wild $t-$bootstrapping procedure, we will then demonstrate that the relationship between this random variable and the above probability can be used to generate $\hat{\mathcal{F}}_c^+$ and $\hat{\mathcal{F}}_c^-$ for any desired value of α. Unlike much of the previous literature on random excursion sets, in this work no model is assumed for the processes' means or spatial covariance structure, and we do not make any assumption of between-'study condition' independence (i.e. $\{\hat{\mu}_n^i\}_{i\in\mathcal{M}}$ may be correlated with one another).

We emphasize here that, to our knowledge, no prior work exists allowing confidence regions to be generated for excursion sets derived from conjunctions or disjunctions of exceedance statements. A notable point of differentiation between this work and the previous literature is that the excursion sets considered here may possess sharp, discontinuous corners. Whilst previous work has centered on excursion sets that possess continuous curves for boundaries, here our problem concerns excursion set boundaries that are piece-wise continuous, i.e. composed of numerous continuous segments (see Section 5.2.2 for further detail). The combinatoric nature of the boundary compositions considered in this work adds an additional layer of complexity to the results and proofs provided, and has not previously been considered in the confidence regions literature.

In the following sections, we first describe the notation and assumptions upon which our theory relies. Following this, we provide the central theoretical result of this work, which relates Equation (5.1) to an exceedance statement for a well-defined random variable. Next, via the use of the wild $t-$bootstrap, we detail how this result may be employed to obtain confidence regions for conjunction inference in the setting of linear regression modelling. Finally, we validate our results with simulations and a real data example based on fMRI data taken from the Human Connectome Project (Van Essen et al. (2013)). Proofs of the theory presented in this chapter, alongside

further illustration, are provided in Appendix C while extensive simulation results may be found in Supplementary Material C.

5.2 Confidence Regions for Excursion Set Combinations

5.2.1 Notation

In the following sections, we shall need notation to describe the numerous possible low dimensional sub-manifolds which can arise from intersecting the boundaries of M excursion sets. To do so, we first denote the power set of a finite set, A, as $\mathcal{P}(A)$, and define $\mathcal{P}^+(A)$ as $\mathcal{P}^+(A) = \mathcal{P}(A) \setminus \{\emptyset\}$. For each $i \in \mathcal{M}$, we define $\partial\mathcal{A}_c^i$ as the level set $\{s \in S : \mu^i(s) = c\}$. Similarly, we define $\partial\mathcal{F}_c$ as the level set $\{s \in S : \min_{i\in\mathcal{M}} \mu^i(s) = c\}$. For a spatial set, $B \subseteq S$, its complement in S is denoted $B^c = S \setminus B$, its topological closure is denoted $\overline{B}$, its interior is denoted B° and its boundary is denoted ∂B. In addition, for closed B, the notation B^1 and B^{-1} shall represent B and $\overline{B^c}$, respectively. We note here that the notation $\partial\mathcal{A}_c^i$ and ∂B may conflict with one another. This conflict is resolved, however, by the assumptions of Section 5.2.2 which ensure that the boundaries of $\{\mathcal{A}_c^i\}_{i\in\mathcal{M}}$ and $\mathcal{F}_c$ are equal to the level sets $\{\partial\mathcal{A}_c^i\}_{i\in\mathcal{M}}$ and $\partial\mathcal{F}_c$ (at least locally, in the vicinity of $\partial\mathcal{F}_c$, which is sufficient for our purposes).

We can now partition the level set $\partial\mathcal{F}_c$ into sub-manifolds using the set of possible intersections of the M excursion set boundaries. For all $\alpha \in \mathcal{P}^+(\mathcal{M})$, we define the boundary segment $\partial^\alpha\mathcal{F}_c$ as follows;

$$\partial^\alpha\mathcal{F}_c = \partial\mathcal{F}_c \cap \left(\bigcap_{i\in\alpha} \partial\mathcal{A}_c^i\right) \cap \left(\bigcap_{j\in\mathcal{M}\setminus\alpha} (\partial\mathcal{A}_c^j)^c\right).$$

We note it is possible that, for some $\alpha \in \mathcal{P}^+(\mathcal{M})$, the boundary segment $\partial^\alpha\mathcal{F}_c$ is empty (in fact, this is often the case in practice). This notation is crucial to the

statement of Theorem 5.1 and is illustrated by Fig. 5.2.

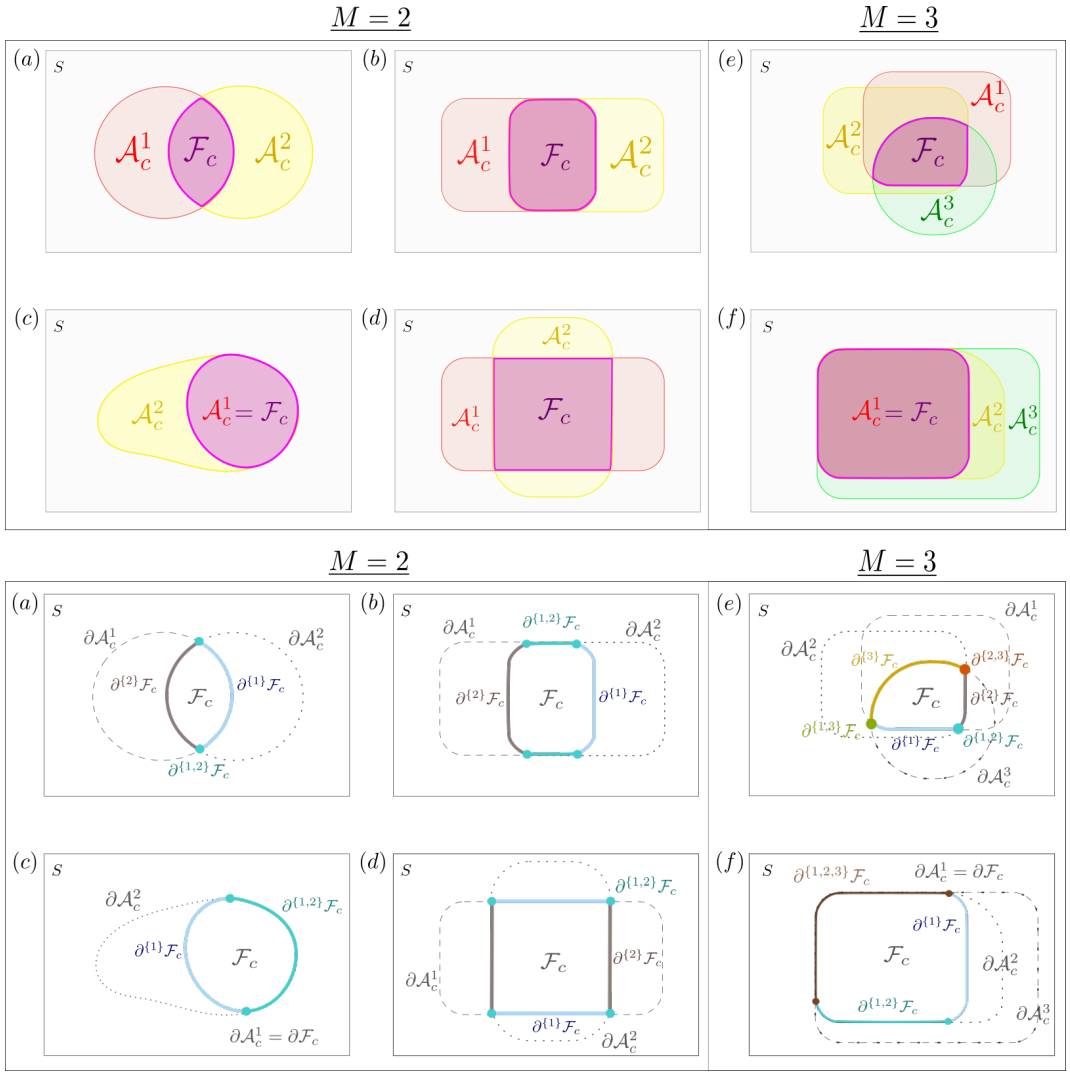

Figure 5.2: Annotated images of 'overlapping' excursion sets (top) with corresponding illustrations of boundary partitions (bottom) for settings in which the number of study conditions, M, is equal to 2 (left) or 3 (right).

5.2.2 Assumptions

In this section, we outline and discuss the assumptions upon which our theory relies. Additional discussion of why each assumption is required, alongside illustration, is given in Appendix C.1.

Assumption 5.1. For each $i \in \mathcal{M}$, there exists a bounded function $\sigma^i : S \to \mathbb{R}^+$ and positive sequence $\tau_n \to 0$, such that the below joint Central Limit Theorem (CLT) holds:

$$\left\{ \frac{\hat{\mu}_n^i(s) - \mu^i(s)}{\tau_n \sigma^i(s)} \right\}_{s \in S, i \in \mathcal{M}} \xrightarrow{d} \{G^i(s)\}_{s \in S, i \in \mathcal{M}}$$

where $\{G^i(s)\}_{s \in S, i \in \mathcal{M}}$ is a well-defined multivariate random field with continuous sample paths in S and $\xrightarrow{d}$ represents convergence in distribution.

Remark. This assumption introduces the notation $\sigma^i(s)$ and τ_n. In typical applications, $\sigma^i(s)$ represents standard deviation across observations (for study condition i, at spatial location s) and τ_n is a decreasing function of sample size. For conciseness in the following text, we now define $\{g^i\}_{i \in \mathcal{M}}$ and $\{\hat{g}_n^i\}_{i \in \mathcal{M}}$ as;

$$g^i(s) = \frac{\mu^i(s) - c}{\sigma^i(s)} \quad \text{and} \quad \hat{g}_n^i(s) = \frac{\hat{\mu}_n^i(s) - c}{\sigma^i(s)}.$$

Assumption 5.2. We assume that:

(a) The functions $\{g^i\}_{i \in \mathcal{M}}$ are continuous on S.

(b) For n large enough, the functions $\{\hat{g}_n^i\}_{i \in \mathcal{M}}$ are continuous on S.

Remark. In practice, assumptions of this form are already common in the imaging literature, as in many applications $\{\mu^i\}_{i \in \mathcal{M}}$, $\{\hat{\mu}_n^i\}_{i \in \mathcal{M}}$ and $\{\sigma^i\}_{i \in \mathcal{M}}$ are assumed to be continuous across space (c.f. Brett et al. (2004), Adler and Taylor (2009)). For further remarks, see Appendix C.1.2.

Assumption 5.3. We assume that:

(a) Every open ball around a point on the boundary $\partial \mathcal{F}_c$ has a non-empty intersection with $(\mathcal{F}_c)^\circ$. In addition, for all $\alpha \in \mathcal{P}^+(\mathcal{M})$ every open ball around a point in $\partial^\alpha \mathcal{F}_c$ has a non-empty intersection with the set:

$$\mathcal{J}_c^\alpha := \left(\bigcap_{i \in \alpha} (\mathcal{A}_c^i)^c \right) \cap \left(\bigcap_{j \in \mathcal{M} \setminus \alpha} (\mathcal{A}_c^j)^\circ \right),$$

where, if $\mathcal{M} \setminus \alpha$ is empty, the last term is defined to be equal to S.

(b) There is an open neighbourhood of $\partial\mathcal{F}_c$ over which the functions $\{g^i\}_{i\in\mathcal{M}}$ and, for n large enough, $\{\hat{g}_n^i\}_{i\in\mathcal{M}}$ are C^1 with finite, non-zero, gradients.

(c) For every point $s \in \partial\mathcal{F}_c$ and $\alpha \in \mathcal{P}^+(\mathcal{M})$, if $s \in \partial\big(\bigcap_{i\in\alpha} \mathcal{A}_c^i\big)$ then the set $\partial\big(\bigcap_{i\in\alpha} \mathcal{A}_c^i\big)$ partitions every sufficiently small open ball around s into exactly two components, each of which is path connected.

Remark. The above statements ensure that $\mathcal{F}_c$ is non-empty and has a well defined boundary which is equal to the level set $\partial\mathcal{F}_c = \{s \in S : \min_{i\in\mathcal{M}} g^i(s) = 0\}$. All three statements ensure that no changes in topology occur at the level c by requiring that $\partial\mathcal{F}_c$ does not contain plateaus (spatial regions of non-zero measure over which $\min_{i\in\mathcal{M}} g^i(s) = 0$ exactly), local minima or maxima. In addition, each statement handles pathological cases which the other statements do not. For brevity, we do not list such cases here but instead provide further discussion in Appendix C.1.3.

Each of the examples presented in Fig. 5.2 satisfy Assumptions 5.2 and 5.3 and highlight several features worthy of note. Firstly, we do not assume $\partial\mathcal{F}_c$ is a C^1 curve, but rather piece-wise C^1 (c.f. examples (d) and (e)). Secondly, the assumptions allow for the possibility that the excursion sets could be nested within one another (c.f. examples (c) and (f)). Thirdly, the assumptions permit the excursion sets to share common boundaries (see examples (b), (c) and (f)). The last two observations are of practical relevance as, in many applications, it may be expected that the target functions for different study conditions exhibit a strong similarity. For instance, in the neuroimaging example, it may be expected that some anatomical regions of the brain display evidence of activation for all of the study conditions. We note here that there is no requirement that $\mathcal{F}_c$ be (globally) path-connected; $\mathcal{F}_c$ may consist of several disconnected components without violating the assumptions of this section.

5.2.3 Theory

In this section, we present the central result of the work of this chapter, Theorem 5.1, alongside two corollaries, Corollary 5.1 and Corollary 5.2. To do so, we now define the nested sets $\hat{\mathcal{F}}_c^-$ and $\hat{\mathcal{F}}_c^+$ by thresholding the statistic $\tau_n^{-1} \min_{i \in \mathcal{M}} \hat{g}_n^i(s)$:

$$\hat{\mathcal{F}}_c^- := \hat{\mathcal{F}}_c^-(a) = \{s \in S : \tau_n^{-1} \min_{i \in \mathcal{M}} \hat{g}_n^i(s) \geq -a\},$$

$$\hat{\mathcal{F}}_c^+ := \hat{\mathcal{F}}_c^+(a) = \{s \in S : \tau_n^{-1} \min_{i \in \mathcal{M}} \hat{g}_n^i(s) \geq +a\},$$

using some constant $a \in \mathbb{R}^+$. To find an appropriate value of a, such that Equation (5.1) holds for a desired tolerance level (e.g. $\alpha = 0.05$), Theorem 5.1 is required.

Theorem 5.1. *Under the assumptions of Section 5.2.2, the below holds:*

$$\lim_{n \to \infty} \mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-] = \mathbb{P}[H \leq a], \tag{5.2}$$

where the variable H is defined as follows;

$$H = \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \partial^\alpha \mathcal{F}_c} \left| \min_{i \in \alpha} (G^i(s)) \right| \right).$$

Proof of Theorem 5.1 is provided in Appendix C.2. Conceptually, the event $\{H > a\}$ may be thought of as the situation in which, at some point, s, inside some boundary segment, $\partial^\alpha \mathcal{F}_c$, a large value was observed for the statistic $|\min_{i \in \alpha}(G^i(s))|$. Therefore, Theorem 5.1 tells us the probability that the statement $\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-$ is violated is directly determined by such points. If we can find a value of a such that $\mathbb{P}[H \leq a] = 1 - \alpha$ then asymptotically, the inclusion probability, $\mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-]$, will also equal $1 - \alpha$. In other words, if the $(1-\alpha)^{th}$ quantile of the distribution of H is known, then confidence regions $\hat{\mathcal{F}}_c^-$ and $\hat{\mathcal{F}}_c^+$ may be constructed which satisfy Equation (5.1). We shall take up the task of evaluating the quantiles of H in Section 5.3.

Theorem 5.1 is key to generating confidence regions for conjunction inference (intersections of excursion sets). However, Theorem 5.1 may also be extended to

allow investigation of other logical statements involving negations or disjunctions. Below we provide two corollaries, each of which is illustrated via a worked example. Corollary 5.1 extends Theorem 5.1 to account for logical negations, whilst Corollary 5.2 provides an equivalent statement for inference performed upon logical disjunctions (unions of excursion sets).

Example 5.2.1. Suppose, given two target functions, μ^1 and μ^2, we were interested in the region defined by the logical conjunction '$\mu^1 \geq c$ *and* $\mu^2 \leq c$' (i.e. "Where did a variable of interest exceed a threshold under one study condition *and not* under another?"). In the notation of Section 5.2.1, this region is readily seen to be given by $(\mathcal{A}_c^1)^1 \cap (\mathcal{A}_c^2)^{-1}$.

To apply the result of Theorem 5.1, we first define $\tilde{\mu}^1 = \mu^1$ and $\tilde{\mu}^2 = 2c - \mu^2$ and note that the logical statement '$\mu^1 \geq c$ *and* $\mu^2 \leq c$' is identical to the statement '$\tilde{\mu}^1 \geq c$ *and* $\tilde{\mu}^2 \geq c$'. By using the latter statement, Theorem 5.1 may now be applied to obtain the desired confidence regions (assuming that an appropriate mechanism exists for evaluating the quantiles of H). However, during this process, the limiting variables $\{G^i\}_{i\in\{1,2\}}$, which correspond to the target functions $\{\mu^i\}_{i\in\{1,2\}}$, must be replaced by variables corresponding to the functions $\{\tilde{\mu}^i\}_{i\in\{1,2\}}$, $\{\tilde{G}^i\}_{i\in\{1,2\}}$. By noting the definitions of $\{\tilde{\mu}^i\}_{i\in\{1,2\}}$ and $\{\tilde{G}^i\}_{i\in\{1,2\}}$, it can be seen that $\tilde{G}^1 = G^1$ and $\tilde{G}^2 = -G^2$.

In summary, a procedure for generating confidence regions corresponding to the logical conjunction '$\mu^1 \geq c$ *and* $\mu^2 \leq c$' would be identical to that previously described, except that G^2 would be substituted for $-G^2$ and $\partial\mathcal{F}_c$ would be substituted for $\partial((\mathcal{A}_c^1)^1 \cap (\mathcal{A}_c^2)^{-1})$. In general, this example may be extended to allow for inference to be performed upon arbitrary conjunctions of statements and negated statements. To achieve this, the definitions of $\mathcal{F}_c$, $\hat{\mathcal{F}}_c^-$ and $\hat{\mathcal{F}}_c^+$ must be extended as follows.

Corollary 5.1. *Let* $\{\delta_i\}_{i\in\mathcal{M}}$ *be an arbitrary sequence of integers in* $\{-1, 1\}$ *and define* $\mathcal{F}_c$ *as* $\mathcal{F}_c = \bigcap_{i\in\mathcal{M}} (\mathcal{A}_c^i)^{\delta_i}$. *Similarly, extend the definition of the sets* $\hat{\mathcal{F}}_c^+$ *and* $\hat{\mathcal{F}}_c^-$ *to:*

$$\hat{\mathcal{F}}_c^+ = \bigcap_{i\in\mathcal{M}} \{s \in S : \tau_n^{-1}\hat{g}_n^i(s) \geq +a\}^{\delta_i} \quad \textit{and} \quad \hat{\mathcal{F}}_c^- = \bigcap_{i\in\mathcal{M}} \{s \in S : \tau_n^{-1}\hat{g}_n^i(s) \geq -a\}^{\delta_i},$$

respectively and, for each $\alpha \in \mathcal{P}^+(\mathcal{M})$*, define* $\partial^\alpha \mathcal{F}_c$ *as before. If* $\mathcal{F}_c$ *satisfies the assumptions of Section 5.2.2, then the below statement holds asymptotically:*

$$\lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-] = \mathbb{P}\left[\max_{\alpha\in\mathcal{P}^+(\mathcal{M})} \left(\sup_{s\in\partial^\alpha \mathcal{F}_c} \left| \min_{i\in\alpha}(\delta_i G^i(s)) \right| \right) \leq a \right].$$

Example 5.2.2. Suppose, given two target functions, μ^1 and μ^2, interest lay in the logical disjunction of '$\mu^1 \geq c$ *or* $\mu^2 \geq c$' (i.e. "Where did *at least one* of the variables of interest exceed the threshold?"). The region of space over which this statement holds is given by $\mathcal{A}_c^1 \cup \mathcal{A}_c^2$, and shall be denoted $\mathcal{G}_c$.

To generate confidence regions for $\mathcal{G}_c$, we first assume that $\mathcal{G}_c$ satisfies Assumptions 5.1-5.3. By Assumptions 5.2 and 5.3 and De Morgan's law, it can be seen that $\mathcal{G}_c = ((\mathcal{A}_c^1)^{-1} \cap (\mathcal{A}_c^2)^{-1})^{-1}$. It must be noted that Assumptions 5.2 and 5.3 are necessary for this statement to hold due to the fact that $(\mathcal{A}_c^i)^{-1} = \overline{(\mathcal{A}_c^i)^c}$ and not $(\mathcal{A}_c^i)^c$. By employing Corollary 5.1, for a fixed value of α, confidence regions, $\hat{\mathcal{F}}_c^-$ and $\hat{\mathcal{F}}_c^+$, may be obtained which satisfy the following;

$$\lim_{n\to\infty} \mathbb{P}[(\hat{\mathcal{F}}_c^-)^c \subseteq ((\mathcal{A}_c^1)^{-1} \cap (\mathcal{A}_c^2)^{-1})^c \subseteq (\hat{\mathcal{F}}_c^+)^c] = 1 - \alpha.$$

By taking the closure of each of the sets in the above probability, and again via careful consideration of Assumptions 5.2 and 5.3, the below is obtained;

$$\lim_{n\to\infty} \mathbb{P}[(\hat{\mathcal{F}}_c^-)^{-1} \subseteq \mathcal{G}_c \subseteq (\hat{\mathcal{F}}_c^+)^{-1}] = 1 - \alpha.$$

In other words, the sets $(\hat{\mathcal{F}}_c^-)^{-1}$ and $(\hat{\mathcal{F}}_c^+)^{-1}$ serve as confidence regions for disjunction inference on μ^1 and μ^2. Note that, as in the previous example, great attention must be paid to the signs which appear in front of the variables $\{G^i\}_{i\in\mathcal{M}}$ in the definition of H. As this example began by generating confidence regions for $(\mathcal{A}_c^1)^{-1} \cap (\mathcal{A}_c^2)^{-1}$, it can be seen that the variables G^1 and G^2 must be appended with negative signs (i.e. when applying the result of Corollary 5.1, both δ_1 and δ_2 were set to -1). This example may be expanded upon to obtain similar results for larger values of M, as

shown by Corollary 5.2.

Corollary 5.2. *Let $\{\delta_i\}_{i\in\mathcal{M}}$ be an arbitrary sequence of integers in $\{-1,1\}$ and define $\mathcal{G}_c$ as $\mathcal{G}_c = \bigcup_{i\in\mathcal{M}}(\mathcal{A}_c^i)^{\delta_i}$. Define the sets $\hat{\mathcal{G}}_c^+$ and $\hat{\mathcal{G}}_c^-$ as:*

$$\hat{\mathcal{G}}_c^+ = \left(\bigcap_{i\in\mathcal{M}} \{s \in S : \tau_n^{-1}\hat{g}_n^i(s) \geq -a\}^{\delta_i} \right)^{-1} \text{and} \quad \hat{\mathcal{G}}_c^- = \left(\bigcap_{i\in\mathcal{M}} \{s \in S : \tau_n^{-1}\hat{g}_n^i(s) \geq +a\}^{\delta_i} \right)^{-1},$$

respectively and, for each $\alpha \in \mathcal{P}^+(\mathcal{M})$, define $\partial^\alpha\mathcal{G}_c$ analogously to $\partial^\alpha\mathcal{F}_c$ in the previous sections. If $\mathcal{G}_c$ satisfies the assumptions of Section 5.2.2, then the below statement holds asymptotically:

$$\lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{G}}_c^+ \subseteq \mathcal{G}_c \subseteq \hat{\mathcal{G}}_c^-] = \mathbb{P}\left[\max_{\alpha\in\mathcal{P}^+(\mathcal{M})} \left(\sup_{s\in\partial^\alpha\mathcal{G}_c} \left| \min_{i\in\alpha}(-\delta_i G^i(s)) \right| \right) \leq a \right].$$

5.3 Spatial Conjunction Inference for the Linear Model

In this section, we build upon the work of Sommerfeld et al. (2018) and Bowring et al. (2019), in which methods for generating confidence regions were presented for a single target function, $\mu(s)$, derived from a linear regression model. In Section 5.3.1 we formally describe the spatially-varying Linear Model (LM) for M study conditions and, in Section 5.3.2, we explain how the wild t-bootstrap may be applied to estimate quantiles of the variable H. Further implementation details, for when data are sampled from a discrete lattice rather than across a continuous space, alongside pseudocode, are provided in Supplementary Material C Section S2. This section focuses solely on conjunction inference. However, the methods described here may equally be applied to perform inference on other logical statements via the use of the corollaries and examples presented in Section 5.2.3.

5.3.1 Model Specification

For each $i \in \mathcal{M}$ (i.e. for each study condition), we assume that the data follow a spatially-varying Linear Model (LM), possessing an arbitrary covariance structure, defined at location $s \in S$ as:

$$Y^i(s) = X^i\beta^i(s) + \epsilon^i(s), \quad \epsilon^i(s) \sim N(0, \Sigma^i(s)). \tag{5.3}$$

The known quantities in the model are; the $(n \times 1)$ vector of responses, $Y^i(s)$, and the $(n \times p)$ design matrix, X^i. The unknown model parameters are; the $(p \times 1)$ vector of regression coefficients, $\beta^i(s)$, and the $(n \times n)$ random error covariance matrix, $\Sigma^i(s)$. For each spatial location, $s \in S$, the Generalized Least Squares (GLS) estimator of the parameter vector $\beta^i(s)$, $\hat{\beta}^i(s)$, is given by:

$$\hat{\beta}^i(s) = (X^{i'}\Sigma^i(s)^{-1}X^i)^{-1}X^{i'}\Sigma^i(s)^{-1}Y^i(s).$$

In this context, under the i^{th} study condition, at each spatial location s, our interest lies in assessing whether linear relationships hold between the elements of the parameter vector $\beta^i(s)$. Such relationships may be expressed using a contrast vector L^i of dimension $(p \times 1)$. In the notation of the previous sections, the target and estimator functions for the spatially-varying LM are given as $\mu^i(s) = L^{i'}\beta^i(s)$ and $\hat{\mu}^i_n(s) = L^{i'}\hat{\beta}^i(s)$ respectively, and $\tau_n^{-1}\sigma^i(s)$ is the contrast standard error, given as:

$$\tau_n^{-1}\sigma^i(s) = \left[L^{i'}(X^{i'}\Sigma^i(s)^{-1}X^i)^{-1}L^i\right]^{\frac{1}{2}}.$$

Whilst in the above notation we have presented $\{\hat{\beta}^i(s)\}_{i\in\mathcal{M}}$ as being estimated separately using M different models, we note that nothing in our theory prevents the same model being used across all M study conditions. For example, it is possible for the model matrices, $\{Y^i(s)\}_{i\in\mathcal{M}}$ and $\{X^i\}_{i\in\mathcal{M}}$, and parameter vectors, $\{\beta^i(s)\}_{i\in\mathcal{M}}$, to be equal for all $i \in \mathcal{M}$, with the only distinction between study conditions being represented by the contrast vector L^i. This observation is noteworthy as, in many

applications, it is common for researchers to compile all study conditions into a single LM to obtain a heightened statistical power.

To apply Theorem 5.1 to the above LM, we now assume that Assumptions 5.1-5.3 hold. Such assumptions are commonly satisfied by standard conditions placed on the continuity and boundedness of the increments and moments of $\{\beta^i(s)\}_{s\in S}$ and $\{\epsilon^i(s)\}_{s\in S}$. A sufficient list of such conditions is provided in Appendix C.3 (further detail may be found in Sommerfeld et al. (2018)). We note here that the covariance matrix, $\Sigma^i(s)$, is usually unknown in practice, meaning the function $\tau_n^{-1}\sigma^i(s)$ cannot be calculated. As demonstrated in Sommerfeld et al. (2018), however, $\Sigma^i(s)$ can be replaced by any consistent estimator $\hat{\Sigma}^i(s)$ and the assumptions given in Section 5.2.2 are still satisfied. Such estimation is common in the statistics literature and is typically achieved by assuming that some fixed correlation structure applies to $\Sigma^i(s)$ (e.g. diagonal independence, auto-regressive, etc.) so that the number of independent variance parameters which must be estimated is small (c.f. Section 3.3.4 and Appendix A.7).

5.3.2 The Wild $t-$bootstrap

In practice, the distribution of H is unknown and must be estimated. In this section, for the model specification described in Section 5.3.1, we employ a wild $t-$bootstrap resampling scheme to obtain the quantile, a, of H which satisfies $\mathbb{P}[H \leq a] = 1 - \alpha$.

To do so, we first define the $(n \times 1)$-dimensional decorrelated residual vector, R^i, for the LM of the i^{th} study condition ($i \in \mathcal{M}$) as:

$$[R_1^i(s), ..., R_n^i(s)]' = R^i(s) = [\hat{\Sigma}^i(s)]^{-\frac{1}{2}}(Y^i(s) - X^i\hat{\beta}^i(s)).$$

The wild $t-$bootstrap proceeds by, in each bootstrap instance, generating n i.i.d. Rademacher random variables, $\{r_1, ..., r_n\}$, (random variables which take the values -1 and $+1$ with equal probability) independently of the data. For the i^{th} study condition, at spatial location s, a new bootstrap sample is then obtained by multiplying

the elements of the decorrelated residual vector by the Rademacher variables (i.e. the new sample is given by $\{r_1 R_1^i(s), ..., r_n R_n^i(s)\}$). Once the boostrap sample has been generated, the standard deviation of the bootstrap sample, $\hat{\sigma}^{*,i}(s)$, is calculated. A bootstrap instance of the variable G^i, $\tilde{G}^i$, is now generated as follows;

$$\tilde{G}^i(s) = n^{-\frac{1}{2}} \sum_{l=1}^{n} \frac{r_l R_l^i(s)}{\hat{\sigma}^{*,i}(s)}.$$

A bootstrap instance of the variable H, $\tilde{H}$, may then be computed using the bootstrap variables $\{\tilde{G}^i(s)\}_{s\in S, i\in\mathcal{M}}$ as follows:

$$\tilde{H} = \max_{\alpha\in\mathcal{P}^+(\mathcal{M})} \left(\sup_{s\in\partial^\alpha \hat{\mathcal{F}}_c} \left| \min_{i\in\alpha}(\tilde{G}^i(s)) \right| \right). \tag{5.4}$$

Quantiles of the distribution of H may now be estimated empirically from the observed sampling distribution of $\tilde{H}$. Discussion of how the above supremum, taken across continuous space, is evaluated in practice is deferred to Supplementary Material C Section S2.1. However, we do briefly note here that, as the location of the true boundary segment $\partial^\alpha \mathcal{F}_c$ is in practice unknown, the boundary segment $\partial^\alpha \mathcal{F}_c$ has been replaced by an estimate $\partial^\alpha \hat{\mathcal{F}}_c$. Substitutions of this form have previously been discussed at great length and demonstrated to provide reasonable empirical coverage for similar 'confidence region'-based methods in the works of Sommerfeld et al. (2018), Bowring et al. (2019) and Bowring et al. (2021). In keeping with these publications, all simulations discussed in this work have been replicated using both the true and estimated boundary segments (e.g. $\{\partial^\alpha \mathcal{F}_c\}_{\alpha\in\mathcal{M}}$ and $\{\partial^\alpha \hat{\mathcal{F}}_c\}_{\alpha\in\mathcal{M}}$) with a full comparison being provided in Supplementary Material C Sections S5-S6. It should be noted that, as a result of this substitution, our proposed method may not be applied when $\hat{\mathcal{F}}_c$ is empty.

Several strong references exist that argue and verify the correctness of using the wild $t-$bootstrap for estimating quantiles of maxima distributions in the above manner. Of particular note are the works of Chang et al. (2017) and Bowring et al. (2019), in which the wild $t-$bootstrap is proposed for the estimation of Gaussian

maxima distributions and the generation of confidence regions, respectively. Discussion of the wild $t-$bootstrap in an imaging context may also be found in Telschow and Schwartzman (2021). Other significant references which serve as precursors to this work include Chernozhukov et al. (2013) and Sommerfeld et al. (2018), in which the Gaussian multiplier bootstrap was investigated for estimating Gaussian maxima distributions and generating confidence regions, respectively. More general introductions to such bootstrapping procedures may be found in Wu (1986), Efron and Tibshirani (1994) and Hesterberg (2014).

In the form that it has been presented here, the wild $t-$bootstrap requires that $\{G^i\}_{i\in\mathcal{M}}$ be symmetric. Further, as noted above, the wild $t-$bootstrap has been extensively verified for estimating maxima distributions only in settings in which $\{G^i\}_{i\in\mathcal{M}}$ are Gaussian. Here we stress that, whilst the theory presented in Section 5.2 makes no assumptions on the distribution of $\{G^i\}_{i\in\mathcal{M}}$, the bootstrap theory presented throughout Section 5.3 requires $\{G^i\}_{i\in\mathcal{M}}$ to be Gaussian. In order to utilise the results of Section 5.2 when the random variables $\{G^i\}_{i\in\mathcal{M}}$ are not Gaussian, alternative methods must be employed for estimating the quantiles of H.

5.4 Simulations

5.4.1 Simulation Settings

To assess the correctness and performance of the method, a range of simulations were conducted using synthetic data. Each simulation was designed to investigate how the empirical coverage (i.e. the proportion of simulation instances in which the inclusion $\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-$ held) was affected by various secondary factors of practical interest. In this section, we describe three simulations designed to investigate how the method's performance was influenced by; (1) the degree of overlap between excursion sets, (2) the presence of correlation between the noise fields for different study conditions, and (3) the magnitude of the spatial rate of change of the signal.

In all simulations, for $i \in \mathcal{M}$, the model used to generate the synthetic data took the form;

$$Y^i(s) = \mu^i(s) + \epsilon^i(s), \quad \epsilon^i(s) \sim N(0, \sigma^i(s) I_n), \tag{5.5}$$

where n was allowed to vary from 40 to 500 (in increments of 20). To induce spatial correlation between observations, $\epsilon^i(s)$ was smoothed using an isotropic Gaussian filter with a Full-Width Half Maximum (FWHM) of 3 pixels. The aim, across all simulations, was to assess the performance of the method when employed for conjunction inference on the true signal $\{\mu^i\}_{i \in \mathcal{M}}$ (i.e. when asked the question "Where do *all* of the $\{\mu^i\}_{i \in \mathcal{M}}$ exceed a predefined threshold?").

The mechanism used to generate the true signal, $\{\mu^i\}_{i \in \mathcal{M}}$, varied across simulations. In Simulations 1 and 2, $\{\mu^i\}_{i \in \mathcal{M}}$ were generated using two binary images of circles and squares, respectively, positioned in a 'Venn diagram' arrangement. Each binary image was scaled by a predefined amount and then smoothed using an isotropic Gaussian filter with a Full-Width Half Maximum (FWHM) of 5 pixels. In Simulation 3, μ^1 and μ^2 were simulated as a horizontal and vertical linear ramp, respectively, with predefined gradients.

Each simulation was performed twice, once using noisier 'low-Signal-to-Noise Ratio (SNR)' synthetic data and again using less noisy 'high-SNR' synthetic data. To generate the high-SNR synthetic data for Simulations 1 and 2, all simulated $\{\mu^i\}_{i \in \mathcal{M}}$ were scaled by a factor of 3 prior to smoothing. This data generation process is identical to that employed in Bowring et al. (2019), in which it is noted that synthetic data generated using these parameters strongly resembles that of an fMRI analysis when voxels are of dimension 2mm^3. In Simulation 3, for the high-SNR data, the linear ramps were initially simulated with a gradient of 8 per 50 pixels. Across all simulations, the low-SNR data was generated by reducing the signal magnitude of the high-SNR data by a factor of 4. In all simulations, thresholds of $c = 1/2$ and $c = 2$ were employed for the low-SNR data and high-SNR data, respectively.

As Simulation 1 aimed to investigate the method's performance as the overlap between excursion sets increased, in this simulation, the distance between the centre

of the circles was varied, ranging from 0 pixels (full overlap) to 50 pixels (barely overlapping) in increments of 2 pixels. Similarly, as Simulation 2 aimed to assess the effect of correlation between ϵ^1 and ϵ^2, in this simulation, this correlation was varied between -1 and 1 in increments of 0.1 (in all other simulations, the noise fields were generated independently). As Simulation 3 served to assess how sensitive the empirical coverage was to the rate of change of μ^1 and μ^2 over space, in this simulation, the initial ramp gradients were multiplied by a factor of k, where k was varied from 0.25 to 1.75 in 0.05 increments.

All simulation results were obtained as averages taken across 2500 simulation instances and compared to the nominal coverage using 95% binomial confidence intervals. In all simulation instances, the image dimensions were (100×100) pixels, confidence regions were computed for a tolerance level of $\alpha = 0.05$ using 5000 bootstrap realizations, $\sigma^i(s)$ was computed using the ordinary least squares estimator and τ_n was computed as $\tau_n = n^{-0.5}$. In each simulation instance, the assessment of whether the inclusion statement, $\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-$, held was verified using the interpolation-based methods of Bowring et al. (2019). In this approach, the inclusion statement was deemed to hold if, and only if, the sets of pixels which were identified as $\hat{\mathcal{F}}_c^-$, $\mathcal{F}_c$ and $\hat{\mathcal{F}}_c^+$ were appropriately nested and, in addition, interpolation along the boundary $\partial\mathcal{F}_c$ confirmed that no violations of the inclusion statement had occurred.

A further six simulations which investigated the influence of other secondary factors of interest were also conducted. The secondary factors considered by these simulations include; the presence of spatial structure in the noise, the effect of $\partial^{\{1\}}\mathcal{F}_c$ and $\partial^{\{2\}}\mathcal{F}_c$ having differing lengths, the effect of ϵ^1 having a much larger variance than ϵ^2 and the impact of varying the value of M. A full discussion of these simulations is deferred to Supplementary Material C Section S4. In addition, in keeping with the previous literature, tolerance levels of $\alpha = 0.1$ and $\alpha = 0.2$, as well as the substitution of $\partial\mathcal{F}_c$ for $\partial\hat{\mathcal{F}}_c$ in Equation (5.4), were also considered for simulation (c.f. Sommerfeld et al. (2018), Bowring et al. (2019), Bowring et al. (2021)). Again, the results of these simulations may be found in the Supplementary Material C document in Sections S5 and S6.

5.4.2 Simulation Results

For a majority of the simulations conducted, the empirical coverage estimates were tightly distributed around the expected nominal coverage. In particular, for the high-SNR data, across all simulations, the 95% binomial confidence intervals consistently captured the nominal coverage at the predicted rate. However, when the simulations employed the low-SNR data, it can be seen that the confidence regions produced by the method were conservative when the data were generated under certain unfavourable conditions.

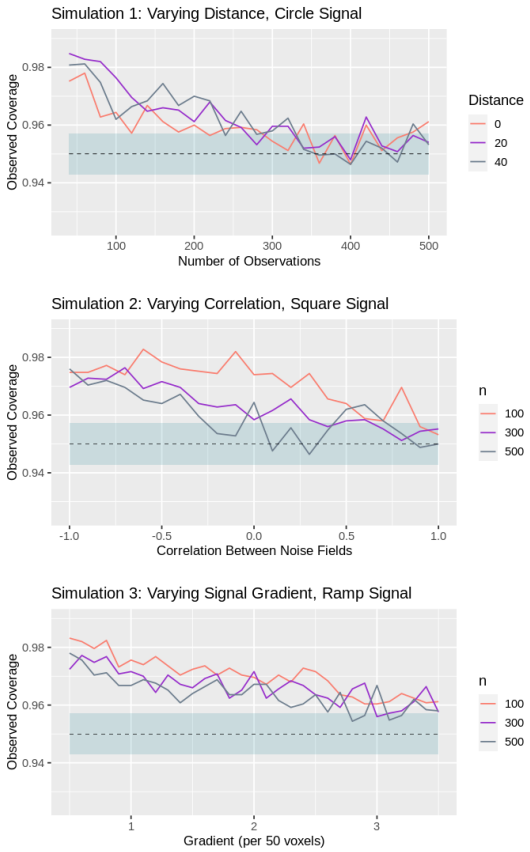

Figure 5.3: Simulation results generated using the low-SNR synthetic data. Nominal coverage is shown as a dashed line, with a corresponding binomial confidence interval also shaded. All data points are averages taken across 2500 simulation instances.

For example, in Fig. 5.3, it can be seen that some over-coverage was observed for Simulation 1 when the data was noisy, and the number of subjects was low. This observation is corroborated by the results of the remaining low-SNR synthetic

data simulations (c.f. Supplementary Material C Section S5). However, in all cases, the degree of agreement between the empirical and nominal coverage improved as the sample size increased. Furthermore, no such over-coverage was observed for the equivalent high-SNR simulations. This observation matches expectation, as Theorem 5.1 holds only asymptotically, and is supported by the previous literature on confidence regions, in which similar results were observed for the single-'study condition' setting (c.f. Sommerfeld et al. (2018), Bowring et al. (2019)).

The presence of strong negative correlation between the study conditions and the rate at which the signal varied across space both also appeared to influence the method's performance (c.f. Fig 5.3, Simulations 2 and 3). In both instances, it is likely that the observed over-coverage is due to difficulties in estimating the true location of the boundary $\partial\mathcal{F}_c$, as each of these factors make it harder to identify the locations at which small changes in $\min_{i\in\mathcal{M}}\mu^i$ occur. When high-SNR data was used in the place of low-SNR data, both of these factors had a substantially reduced impact on the empirical coverage.

Despite the above observations, the simulation results overwhelmingly supported the claim that the proposed method is robust to a range of secondary factors. Included in the list of such factors are; variation in the shape and size of $\mathcal{F}_c$, spatial correlation in the noise, variation in the lengths of individual boundary segments, between-'study condition' correlation, realistic levels of variation in the magnitude of the noise, large numbers of study conditions and situations in which boundary segments are shared by many excursion sets. For further detail, see Supplementary Material C Sections S5 and S6.

In terms of time efficiency, the wild $t-$bootstrap provided extremely fast computational performance. For instance, using an Intel(R) Xeon(R) Gold 6126 2.60GHz processor with 16GB RAM, and averaging over the 2500 simulation instances conducted for $n = 500$ high-SNR observations, in Simulation 1, the time taken to generate confidence regions for a circle separation of 20 pixels was 5.98 seconds. For a comprehensive summary of computation times, see Supplementary Material C Section S7.

5.5 Real Data Application

To demonstrate the method in practice, we apply it to fMRI data drawn from the Human Connectome Project (HCP) dataset (Van Essen et al. (2013)). In this example, task fMRI data were collected as part of a block design from 80 healthy, unrelated, young adults as part of a working memory task that had four distinct components, each using a different stimuli type: pictures of places, tools, faces and body parts. In Supplementary Material C Section S3, we provide a brief overview of the imaging acquisition protocol, task paradigm, preprocessing stages and first-level analysis employed for generating this dataset (for exhaustive detail, see Barch et al. (2013) and Glasser et al. (2013)). Following first-level analysis, the data consisted of 4 images for each of the 80 subjects, measuring the %BOLD response to each of the 4 stimuli types. In this example, our interest lies in identifying, at the group-level, which regions of the brain are associated with working memory regardless of stimuli type (i.e. "Which regions are active in response to *all* four working memory stimuli types?").

As the work in this chapter has predominantly focused on two-dimensional excursion sets, a single slice of the brain was chosen for analysis ($z = 46$mm, covering portions of the frontal gyrus involved in working memory). For each stimuli type, an $n = 80$ group-level linear model was constructed. Using the proposed method, confidence regions were then generated at the 5% confidence level to assess where the group-level percentage BOLD change exceeded 1% for *all* four stimuli types. To compare the proposed method to standard fMRI inference procedures, a group-level contrast was also generated using the Big Linear Model toolbox (c.f. Appendix B.1) for each of the four stimuli types. In addition, single-'study condition' confidence regions were also generated using the methods of Sommerfeld et al. (2018) for each of the four stimuli types.

The results are shown in Fig. 5.4. The conjunction inference identified several localized regions, including the Superior Frontal Gyrus, which is well known for its involvement in working memory (c.f. Boisgueheneuc et al. (2006), Vogel et al. (2016),

Alagapan et al. (2019)), and the Angular Gyri, which are well known to be involved in a range of tasks including memory retrieval, attention and spatial cognition (c.f. Seghier (2013), Bréchet et al. (2018)). Whilst the confidence regions for conjunction

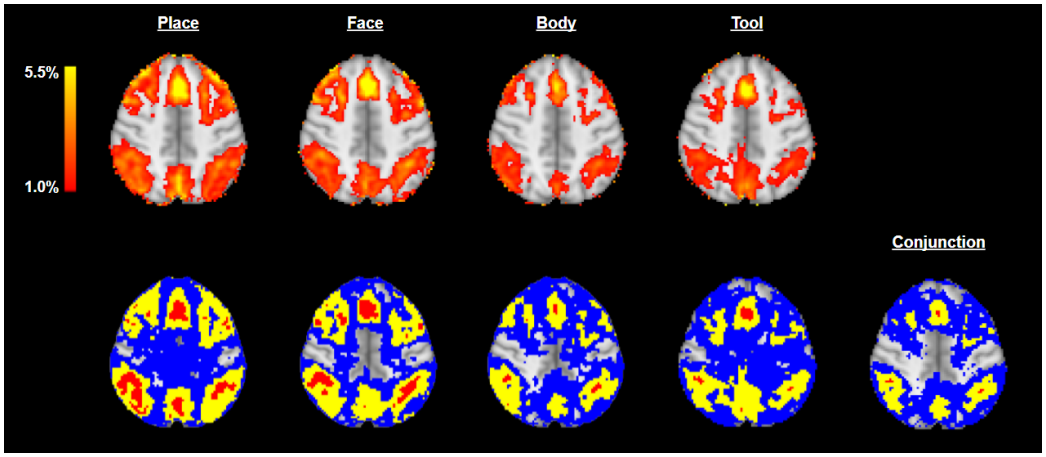

Figure 5.4: %BOLD images (top) and 95% confidence regions (bottom left) for the HCP working memory task for each of the four visual stimuli types, alongside conjunction confidence regions assessing the overlap of the four thresholded %BOLD images (bottom right). Displayed is axial slice $z = 46$mm. The thresholds employed were $c = 1\%$ BOLD change, for all images, and $\alpha = 0.05$, for the confidence regions. Upper and lower confidence regions are displayed in red and blue, respectively. Across the bottom row, the yellow sets are the point estimates $\hat{\mathcal{A}}_c^1, ...\hat{\mathcal{A}}_c^4$ and $\hat{\mathcal{F}}_c$ respectively. The red conjunction set has localized regions in the Superior Frontal Gyrus (top) and the Angular Gyri (left/right), for which we can assert with 95% confidence that, for all four study conditions, there was (at least) a 1.0% change in BOLD response.

inference illustrated in Fig. 5.4 exhibit a clear resemblance to the four contrast images, we note that the red set, $\hat{\mathcal{F}}_c^+$, is very small in comparison to the blue set, $\hat{\mathcal{F}}_c^-$, which is large and diffuse. This observation conveys important information about the spatial variability of the regions identified. In particular, it can be seen that, in the regions immediately surrounding the Superior Frontal Gyrus and Angular Gyri, there is a much higher resemblance between the blue ($\hat{\mathcal{F}}_c^-$) and yellow ($\hat{\mathcal{F}}_c$) sets than is seen across the rest of the brain. This resemblance provides an insight into the degree of spatial variation present surrounding these regions. In this case, the strength and localization of this resemblance suggest that the spatial variability of $\hat{\mathcal{F}}_c$ is less severe in and around the aforementioned anatomical regions than it is across the rest of the brain. It may be concluded that the estimated yellow 'blobs' corresponding to the

Superior Frontal Gyrus and the Angular Gyri have been more reliably localized than the other yellow 'blobs' appearing in the image.

In general, the single-'study condition' confidence regions can be seen to be much larger than those observed for the conjunction inference. This observation matches expectation as the overlap of the four excursion sets is smaller than each set individually. In addition, in this example, the conjunction method employed the entire experimental design for analysis (as opposed to the single-'study condition' method which employed only the data concerning the stimuli of interest) and is therefore expected to have a higher statistical power and, thus, tighter confidence bounds.

5.6 Discussion and Conclusion

In this chapter, we have produced a method for generating confidence regions for intersections and unions of multiple excursion sets. The confidence regions generated by the method generalise the notion of confidence intervals to arbitrary spatial dimensions and possess the usual frequentist interpretation that, were the 0.95 procedure repeated on numerous samples, the proportion of confidence regions, $\hat{\mathcal{F}}_c^+$ and $\hat{\mathcal{F}}_c^-$, that correctly enclose $\mathcal{F}_c$ would tend to 0.95. Such confidence regions serve as an indicator of the reliability of $\hat{\mathcal{F}}_c$ as an estimate of $\mathcal{F}_c$. Confidence regions which closely resemble the set $\hat{\mathcal{F}}_c$ may be interpreted as signifying that $\hat{\mathcal{F}}_c$ is a reliable estimate of $\mathcal{F}_c$, whilst little resemblance may suggest that there is a high degree of spatial variability present in the data and that the estimated shape, size and locale of $\hat{\mathcal{F}}_c$ are not particularly reliable. Such statements are of particular value in imaging applications where the aim of a statistical analysis is typically to assess how reliably some form of activity may be localized to a particular spatial region.

We stress here that, whilst $\hat{\mathcal{F}}_c$ is equal to the intersection of $\{\hat{\mathcal{A}}_c^i\}_{i\in\mathcal{M}}$, the confidence regions $\hat{\mathcal{F}}_c^\pm$ are not the intersections of the single-'study condition' confidence regions $\{\hat{\mathcal{A}}_c^{\pm,i}\}_{i\in\mathcal{M}}$ which would be obtained using the methods of Sommerfeld et al. (2018). This can be seen by noting that $\{\hat{\mathcal{A}}_c^{\pm,i}\}_{i\in\mathcal{M}}$ are defined using separate thresholds from different bootstraps and can be heavily influenced by the behavior

of $\{\partial\mathcal{A}_c^i\}_{i\in\mathcal{M}}$ in spatial regions far from $\mathcal{F}_c$. In fact, treating the naive intersection of the $(1-\alpha)$ confidence regions $\{\hat{\mathcal{A}}_c^{\pm,i}\}_{i\in\mathcal{M}}$ as confidence regions for $\mathcal{F}_c$ is not a valid method in general and can result in undesirable asymptotic coverage lying anywhere inside the range $[1-M\alpha, 1]$. This claim is further discussed in Appendix C.4. To our knowledge, the only valid approach for generating confidence regions for conjunction inference is that outlined in this chapter.

One potential limitation of the specific methods of Section 5.3 is that they allow for confidence regions to be generated only when the target functions, $\{\mu^i\}_{i\in\mathcal{M}}$, are linear combinations of regression parameter estimates. As noted by Bowring et al. (2021), it is often preferable for an analysis to consider standardized effect sizes, such as Cohen's d (Cohen (2013)) or Hedges' G (Hedges (1981)), instead of regression parameter estimates, as standardized effect sizes allow for information about statistical power to be incorporated into the analysis. To this end, Bowring et al. (2021) outlined an approach for generating confidence regions (in the single-'study condition' setting) for images of Cohen's d effect sizes. Here, we suggest that a potential avenue for future work is to combine the methods of Bowring et al. (2021) and those proposed in this chapter to allow for conjunction, or disjunction, inference to be performed on standardized effect sizes rather than regression parameter estimates.

Another limitation noted here is that, whilst our proposed method theoretically works for data of arbitrary dimensions, the simulations of Section 5.4 verify the performance of our method only for two-dimensional data. Whilst it is typical for many practical applications to focus on two-dimensional data (e.g. climate maps, surface images), higher-dimensional datasets are abundant in a wealth of disciplines (e.g. fMRI scans, astrological maps). For this reason, we intend to investigate further the performance of the method in higher dimensions in future work.

5.7 Data and Code Availability

Data used in this chapter were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil;

1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The code used to obtain the results of Sections 5.4 and 5.5 is freely available and may be found at:

https://github.com/TomMaullin/ConfSets

Chapter 6

Concluding Remarks

For a long time, a common idiom in the statistical literature has been *"the plural of anecdote is not data"*. Broadly speaking, this expression served to indicate that the sample sizes in statistics were insufficient, and that bigger datasets were required. Through an inadvertent turn of phrase, the opinion piece by Frégnac (2017) cited at the start of Chapter 1 succinctly described the motivations of this thesis in five words: *"big data is not knowledge"*. The field of fMRI now possesses big data, but requires new methods to analyse it.

To contribute towards the growing literature on large-sample fMRI, this thesis has pursued two avenues of methods development. The first of these centered on providing computationally efficient methods for LMM analysis, and the second focused on developing spatial confidence regions for between-'study condition' comparison. Whilst these avenues are disparate in content, they are united in motivation; both aim to improve the inference methodology available for large-sample fMRI analysis.

In Chapter 3, we provided a Fisher Scoring approach to performing estimation of and inference upon LMMs containing multiple, potentially crossed, random factors. This work was developed, in large part, to facilitate the vectorisation of LMM parameter estimation and inference across voxels in large-sample fMRI analysis. However, much of the work in this chapter is also applicable in a broader statistical context. In particular, we have provided novel expressions for the LMM score vector and Fisher

Information matrix. We have demonstrated these expressions possess utility for several topical issues in statistics, including the modelling of covariance structure and estimation of degrees of freedom. The methods presented in this chapter face an issue common to much of the LMM statistics literature; computation scales with the second dimension of the random effects covariance matrix. However, we highlight that there is a strong potential for the approaches described in this work to be combined with current sparse matrix methodology to obtain improved computational performance.

In Chapter 4, we developed the BLMM framework; a tool for large-sample fMRI LMM analysis, designed for usage on SGE HPCs. To do so, we employed a spatially-varying design to model the missingness caused by mask variability and utilised the statistical methods developed in Chapter 3 to streamline computation across voxels. Computational performance was further improved via careful consideration of the matrices stored and employed for calculation and by accounting for the patterns of sparsity present in single-factor designs. BLMM offers hypothesis testing for both fixed effects parameters and random effects parameters, with fixed effects inference utilising the WS-based methodology of Chapter 3. One notable absence in the current implementation of the BLMM toolbox is the constrained covariance methodology that was outlined in Chapter 3. The ability to model customized covariance structure is of strong utility to the fMRI literature, and for this reason, we suggest the implementation of this methodology could form a strong basis for future work.

In Chapter 5, we developed a method for generating confidence regions for intersections, and unions, of multiple excursion sets. The primary motivation for this work was to assess the spatial variability in the overlap of regions of brain activation obtained under different study conditions in fMRI. However, in this chapter, we have demonstrated that the proposed method also has wide applicability in a range of other imaging disciplines such as climatology and cosmology. The strong utility of the method arises from it's ability to assess how reliably some form of activity has been localized to a particular spatial region. As such localization is of primary importance to fMRI, we predict that, as sample sizes grow, such spatial confidence methods will play an increasingly central role in interpreting the results of fMRI analysis. We note

that practical details of the Wild $t-$bootstrap approach employed in Chapter 5 limit the application of the proposed method to only situations in which the signal of interest is derived from a spatially varying LM. However, we highlight that the theoretical results given in this chapter are much more generally applicable. For this reason we suggest that extending the proposed method to a broader range of scenarios, such as performing inference on standardized effect sizes rather than regression parameter estimates, could form the basis of future work.

Appendix A

Linear Mixed Models

A.1 Score Vectors

In this appendix, we provide full derivations for the derivatives (3.5) and (3.9). For derivatives (3.3) and (3.4), we note that the derivations for the multi-factor LMM are identical to those given for the single-factor setting in Demidenko (2013). As a result, we refer the reader to this source for proofs. To obtain the derivatives (3.5) and (3.9), two lemmas, Lemma A.1 and Lemma A.2, are required. Proofs for Lemma A.1 and Lemma A.2, alongside discussion of the definition of derivative employed throughout Chapter 3 and Appendix A, are provided in Appendix A.2.

Lemma A.1. *Let g be a column vector, A be a square matrix and $\{B_s\}$ be a set of arbitrary matrices of equal size. Let K be an unstructured matrix which none of g, A or any of the $\{B_s\}$ depend on. Further, assume $A, \{B_s\}, g$ and K can be multiplied as required. The below matrix derivative expression now holds;*

$$\frac{\partial}{\partial K}\left[g'(A+\sum_t B_t K B_t')^{-1}g\right] = \\ -\sum_s B_s'\big(A'+\sum_t B_t K B_t'\big)^{-1} g g' \big(A'+\sum_t B_t K B_t'\big)^{-1} B_s. \tag{A.1}$$

Lemma A.2. *Let $A, \{B_s\}$ and K be defined as in Lemma A.1. Then the following is true:*

$$\frac{\partial}{\partial K} \log |A + \sum_t B_t K B_t'| = \sum_s B_s' (A + \sum_t B_t K' B_t')^{-1} B_s. \tag{A.2}$$

We first derive the partial derivative matrix $\frac{\partial l}{\partial D_k}$ (i.e. the matrix derivative with respect to D_k which does not account for the equal elements of D_k induced by symmetry).

Theorem A.1. *The partial derivative matrix $\frac{\partial l}{\partial D_k}$ is given by the following.*

$$\frac{\partial l(\theta)}{\partial D_k} = \frac{1}{2} \sum_{j=1}^{l_k} Z_{(k,j)}' V^{-1} \left(\frac{ee'}{\sigma^2} - V \right) V^{-1} Z_{(k,j)}. \tag{A.3}$$

Proof. To derive Equation (A.3), we use the expression for the log-likelihood of the LMM, given by Equation (2.4). As the first term inside the brackets of Equation (2.4) does not depend on D_k, we need only consider the second and third term for differentiation. We now note that, by the construction of the random effects design matrix, Z, and the block diagonal structure of D, it can be seen that:

$$V = I + ZDZ' = I + \sum_{k=1}^{r} \sum_{j=1}^{l_r} Z_{(k,j)} D_k Z_{(k,j)}'. \tag{A.4}$$

By substituting Equation (A.4) into the second term of Equation (2.4), and taking the partial derivative matrix with respect to D_k using Lemma A.1, the below can be obtained:

$$\frac{\partial}{\partial D_k} \left[\sigma^{-2} e' V^{-1} e \right] = \sigma^{-2} \sum_{j=1}^{l_k} Z_{(k,j)}' V^{-1} e e' V^{-1} Z_{(k,j)}.$$

Similarly, by substituting Equation (A.4) into the third term of Equation (2.4), and taking the partial derivative matrix with respect to D_k using Lemma A.2, the below can be obtained:

$$\frac{\partial}{\partial D_k} \big[\log |V| \big] = \sum_{j=1}^{l_k} Z_{(k,j)}' V^{-1} Z_{(k,j)}.$$

By combining the previous two derivative expressions, Equation (A.3) is obtained. □

Through applying the vectorisation operator to Equation (A.3), the following corollary is obtained. This is the result stated by Equation (3.9) in Section 3.3.1.2.

Corollary A.1. *The partial derivative vector of the log-likelihood with respect to* $vec(D_k)$ *is given as:*

$$\frac{\partial l(\theta)}{\partial vec(D_k)} = \frac{1}{2} vec\bigg(\sum_{j=1}^{l_k} Z'_{(k,j)} V^{-1} \bigg(\frac{ee'}{\sigma^2} - V \bigg) V^{-1} Z_{(k,j)} \bigg).$$

Using Theorem 5.12 of Turkington (2013), which states that, in our notation:

$$\frac{dl(\theta)}{d\text{vec}(D_k)} = \mathcal{D}_{q_k} \mathcal{D}'_{q_k} \frac{\partial l(\theta)}{\partial \text{vec}(D_k)},$$

the following corollary is now obtained.

Corollary A.2. *The total derivative vector of the log-likelihood with respect to* $vec(D_k)$ *is given as:*

$$\frac{dl(\theta)}{dvec(D_k)} = \frac{1}{2} \mathcal{D}_{q_k} \mathcal{D}'_{q_k} vec\bigg(\sum_{j=1}^{l_k} Z'_{(k,j)} V^{-1} \bigg(\frac{ee'}{\sigma^2} - V \bigg) V^{-1} Z_{(k,j)} \bigg).$$

Finally, by noting that the vectorisation and half-vectorisation operators satisfy $\text{vec}(D_k) = \mathcal{D}^+_{q_k} \text{vech}(D_k)$, the following corollary is obtained. This is the result stated by Equation (3.5) in Section 3.3.1.1.

Corollary A.3. *The total derivative vector of the log-likelihood with respect to* $vech(D_k)$ *is given as:*

$$\frac{dl(\theta)}{dvech(D_k)} = \frac{1}{2} \mathcal{D}'_{q_k} vec\bigg(\sum_{j=1}^{l_k} Z'_{(k,j)} V^{-1} \bigg(\frac{ee'}{\sigma^2} - V \bigg) V^{-1} Z_{(k,j)} \bigg).$$

A.2 Matrices and Differentiation

In this section, additional information is provided to support the proofs of Appendix A.1. Section A.2.1 provides an explicit definition for, and discussion of, the notion

of derivative employed throughout Chapter 3 and Appendix A. Section A.2.2 proves the lemmas employed in the arguments of Appendix A.1.

A.2.1 Derivative Definition

Many different notions of matrix derivative currently exist in the statistical literature. In this section, we explicitly state the definition of derivative which is employed throughout Chapter 3 and Appendix A. Throughout these sections, for arbitrary matrices A and B of dimensions $(p \times q)$ and $(m \times n)$ respectively, the following definition of matrix (partial) derivative has been implicitly employed:

$$\frac{\partial \text{vec}(A)}{\partial \text{vec}(B)} = \begin{bmatrix} \frac{\partial A_{[1,1]}}{\partial B_{[1,1]}} & \cdots & \frac{\partial A_{[p,1]}}{\partial B_{[1,1]}} & \cdots & \frac{\partial A_{[p,q]}}{\partial B_{[1,1]}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial A_{[1,1]}}{\partial B_{[m,1]}} & \cdots & \frac{\partial A_{[p,1]}}{\partial B_{[m,1]}} & \cdots & \frac{\partial A_{[p,q]}}{\partial B_{[m,1]}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial A_{[1,1]}}{\partial B_{[m,n]}} & \cdots & \frac{\partial A_{[p,1]}}{\partial B_{[m,n]}} & \cdots & \frac{\partial A_{[p,q]}}{\partial B_{[m,n]}} \end{bmatrix},$$

with the equivalent definition holding for the total derivative. In addition, for a scalar value, a, and matrix, B, defined as above, we employ the below matrix (partial) derivative definition:

$$\frac{\partial a}{\partial B} = \begin{bmatrix} \frac{\partial a}{\partial B_{[1,1]}} & \cdots & \frac{\partial a}{\partial B_{[1,n]}} \\ \vdots & \ddots & \vdots \\ \frac{\partial a}{\partial B_{[m,1]}} & \cdots & \frac{\partial a}{\partial B_{[m,n]}} \end{bmatrix},$$

with, again, the equivalent definition holding for the total derivative. The above notation can be useful since by definition, the two derivatives defined above satisfy the below relation:

$$\text{vec}\left(\frac{\partial a}{\partial B}\right) = \frac{\partial a}{\partial \text{vec}(B)}.$$

The matrix derivative definitions we have used are commonly employed in the mathematical and statistical literature, but are not a universal standard. For example, sources such as Neudecker and Wansbeek (1983) and Magnus and Neudecker

(1999) define $\partial \text{vec}(A)/\partial \text{vec}(B)$ as the transpose of the definition supplied above. This discrepancy between definitions is noteworthy as some of the referenced results employed in Appendix A.1 are the transpose of those found in the original source. For a full account and in-depth discussion of the different definitions of matrix derivatives used in the literature, we recommend the work of Turkington (2013).

A.2.2 Derivative Lemmas

In this section, Lemma A.1 and Lemma A.2, which were employed in Appendix A.1 are proven rigorously. We begin by proving the result of Lemma A.1.

Proof. To begin, denote $M = A + \sum_s B_s K B_s'$. For clarity and convenience, in this proof, we shall briefly depart from the usual notation of $H_{[i,j]}$ for the $(i,j)^{th}$ element of an arbitrary matrix H and instead employ Ricci calculus notation, in which the $(i,j)^{th}$ element of H is given by H^i_j and all summations are made implicit. The left hand side of Equation (A.1) is now given, using Ricci calculus notation, as;

$$\frac{\partial}{\partial K^m_n}\left[(g')_i (M^{-1})^i_j g^j\right] = (g')_i g^j \frac{\partial}{\partial K^m_n}\left[M^{-1}\right]^i_j. \tag{A.5}$$

We employ a result from Giles (2008) which states, in Ricci calculus notation;

$$\frac{\partial [F^{-1}]^i_j}{\partial x} = -(F^{-1})^i_q \frac{\partial F^q_p}{\partial x}(F^{-1})^p_j.$$

Applying this result to Equation (A.5), it can be seen that the right hand side of Equation (A.5) is equal to:

$$-(g')_i g^j (M^{-1})^i_q \frac{\partial M^q_p}{\partial K^m_n}(M^{-1})^p_j.$$

Evaluating the derivative in the above expression and rearranging the terms gives the

below:

$$\begin{aligned}&- (B'_s)^n_p (M^{-1})^p_j g^j (g')_i (M^{-1})^i_q B^q_{sm} \\ &= -[B'_s M^{-1} g g' M^{-1} B_s]^n_m.\end{aligned}$$

By applying a transposition to the matrices in the above expression, in order to interchange the indices m and n, the result follows. □

We now shall prove Lemma A.2.

Proof. As in the proof of Lemma A.1, the notation of $M = A + \sum_t B_t K' B'_t$ will be adopted and the proof will be given in Ricci calculus notation. Expanding the derivative with respect to K^m_n using the partial derivatives of the elements of M gives the following:

$$\frac{\partial}{\partial K^m_n} \log |M| = \frac{\partial \log |M|}{\partial M^i_j} \frac{\partial M^i_j}{\partial K^m_n}. \tag{A.6}$$

We now make use of the following well-known result, which can be found in Giles (2008).

$$\frac{\partial \log |X|}{\partial X^i_j} = (X'^{-1})^i_j. \tag{A.7}$$

Inserting Equation (A.7) into Equation (A.6) and noting that A^i_j does not depend on K^m_n, yields:

$$\frac{\partial}{\partial K^m_n} \log |M| = (M'^{-1})^i_j \frac{\partial (B_s K (B_s)')^i_j}{\partial K^m_n}.$$

Evaluating the derivative on the right hand side of the above now returns:

$$\frac{\partial}{\partial K^m_n} \log |M| = (M'^{-1})^i_j (B_s)^i_m (B'_s)^n_j = (B'_s)^m_i (M'^{-1})^i_j (B_s)^j_n.$$

Combining the summation terms in the above expression yields the following.

$$\frac{\partial}{\partial K^m_n} \log |M| = (B'_s M'^{-1} B_s)^m_n.$$

Finally, converting the above into matrix notation now gives Equation (A.2) as desired. □

A.3 Fisher Information Matrix

In this appendix, we provide derivations of the components of the Fisher Information matrix of θ^f which relate to D_k, for some factor f_k. To derive these results, we will follow a similar argument to that of Demidenko (2013) in the single-factor setting. The derivation of the elements $\mathcal{I}^f_\beta$, $\mathcal{I}^f_{\beta,\sigma^2}$ and $\mathcal{I}^f_{\sigma^2}$ for the multi-factor LMM is identical to that used for the single factor LMM given in Demidenko (2013) and, therefore, will not be repeated here. Theorems A.2 and A.3 provide derivation of Equation (3.10) and Theorem A.4 provides derivation of Equation (3.11). Following this, Corollaries A.4-A.6 detail the derivation of Equations (3.7) and (3.8).

Theorem A.2. *For any arbitrary integer k between 1 and r, the covariance of the partial derivatives of $l(\theta^f)$ with respect to β and vec(D_k) is given by:*

$$\mathcal{I}^f_{\beta,vec(D_k)} = cov\left(\frac{\partial l(\theta^f)}{\partial \beta}, \frac{\partial l(\theta^f)}{\partial vec(D_k)}\right) = \mathbf{0}_{p,q_k^2}.$$

Proof. First, let u and $T_{(k,j)}$ denote the following quantities:

$$u = \sigma^{-1}V^{-\frac{1}{2}}e, \quad T_{(k,j)} = Z'_{(k,j)}V^{-\frac{1}{2}}. \tag{A.8}$$

As $e \sim N(0, \sigma^2 V)$, it follows that $u \sim N(0, I_n)$. Let c be an arbitrary column vector of length q_k. Noting that the derivative of the log-likelihood function has mean zero, the below can be seen to be true:

$$\text{cov}\left(\frac{\partial l(\theta|y)}{\partial \beta}, c'\frac{\partial l(\theta|y)}{\partial \text{vec}(D_k)}c\right) = \mathbb{E}\left[\frac{\partial l(\theta|y)}{\partial \beta}c'\frac{\partial l(\theta|y)}{\partial \text{vec}(D_k)}c\right].$$

By rewriting in terms of u and $T_{(k,j)}$, and noting that $\mathbb{E}[u] = 0$, the right hand side

of the above can be seen to simplify to:

$$\mathbb{E}\left[\sigma^{-1}XV^{-\frac{1}{2}}uc'\left(\frac{1}{2}\sum_{j=1}^{l_k}(T_{(k,j)}u)(T_{(k,j)}u)'\right)c\right]$$
$$=\frac{1}{2}\sigma^{-1}XV^{-\frac{1}{2}}\sum_{j=1}^{l_k}\mathbb{E}\left[uc'T_{(k,j)}uu'\right]T'_{(k,j)}c=\mathbf{0}_{p,q_k^2}.$$

That the above is equal to a matrix of zeros, follows directly from the third moment of the Normal distribution being 0. The result of the theorem now follows. □

Theorem A.3. *For any arbitrary integer k between 1 and r, the covariance of the partial derivatives of $l(\theta^f)$ with respect to σ^2 and $vec(D_k)$ is given by:*

$$\mathcal{I}^f_{\sigma^2,vec(D_k)}=cov\left(\frac{\partial l(\theta^f)}{\partial\sigma^2},\frac{\partial l(\theta^f)}{\partial vec(D_k)}\right)=\frac{1}{2\sigma^2}vec'\left(\sum_{j=1}^{l_k}Z'_{(k,j)}V^{-1}Z_{(k,j)}\right).$$

Proof. To begin, we rewrite the covariance in Theorem A.3 in terms of Equation (A.8) and remove constant terms to obtain:

$$\frac{1}{4\sigma^2}\text{cov}\left(u'u,\text{vec}\left(\sum_{j=1}^{l_k}(T_{(k,j)}u)(T_{(k,j)}u)'\right)\right).$$

Noting that the Kronecker product satisfies the property $\text{vec}(aa')=a\otimes a$, for all column vectors a, we obtain that the above is equal to:

$$\frac{1}{4\sigma^2}\text{cov}\left(u'u,\sum_{j=1}^{l_k}\left[(T_{(k,j)}u)\otimes(T_{(k,j)}u)\right]\right).$$

We now note that the Kronecker product satisfies $\text{vec}(ABC)=(C'\otimes A)\text{vec}(B)$ for arbitrary matrices A, B and C of appropriate dimensions. Utilizing this, applying the mixed product property of the Kronecker product and then moving constant values

outside of the covariance function now gives:

$$\frac{1}{4\sigma^2}\text{vec}'(I_n)\text{cov}(u\otimes u)\sum_{j=1}^{l_k}\left[(T_{(k,j)}\otimes T_{(k,j)})\right]'.$$

Noting that $u \sim N(0, I_n)$ we now employ a result from Magnus and Neudecker (1986) which states that $\text{cov}(u \otimes u) = 2N_n$. Substituting this result into the previous expression gives the below:

$$\frac{1}{2\sigma^2}\text{vec}'(I_n)N_n\sum_{j=1}^{l_k}\left[(T_{(k,j)}\otimes T_{(k,j)})\right]'.$$

By a result given in Magnus and Neudecker (1986), the matrix N_n satisfies $(A \otimes A)N_k = N_n(A \otimes A)$ for all A, n and k such that the resulting matrix multiplications are well defined. Applying this result to the above expression, and again using the relationship $\text{vec}(ABC) = (C' \otimes A)\text{vec}(B)$, the above can be seen to reduce to:

$$\frac{1}{2\sigma^2}\text{vec}'\left(\sum_{j=1}^{l_k}T_{(k,j)}T'_{(k,j)}\right)N_{q_k}.$$

By the definition of $T_{(k,j)}$, and noting that the matrix N_{q_k} satisfies the relationship $\text{vec}'(A)N_{q_k} = \text{vec}'(A)$ for all appropriately sized symmetric matrices A, the result now follows. □

Theorem A.4. *For any arbitrary integers k_1 and k_2 between 1 and r, the covariance of the partial derivatives of $l(\theta^f)$ with respect to $vec(D_{k_1})$ and $vec(D_{k_2})$ is given by:*

$$\mathcal{I}^f_{vec(D_{k_1}),vec(D_{k_2})} = cov\left(\frac{\partial l(\theta^f)}{\partial vec(D_{k_1})}, \frac{\partial l(\theta^f)}{\partial vec(D_{k_2})}\right) =$$
$$\frac{1}{2}N_{q_{k_1}}\sum_{j=1}^{l_{k_2}}\sum_{i=1}^{l_{k_1}}(Z'_{(k_1,i)}V^{-1}Z_{(k_2,j)}\otimes Z'_{(k_1,i)}V^{-1}Z_{(k_2,j)}).$$

Proof. We begin by substituting the result of Corollary A.1 into the covariance expression appearing in Theorem A.4. By converting to the notation introduced in

Equation (A.8) and removing constant terms, the below is obtained:

$$\frac{1}{4}\text{cov}\bigg(\sum_{j=1}^{l_{k_1}}\text{vec}\Big[(T_{(k_1,j)}u)(T_{(k_1,j)}u)'\Big],\sum_{j=1}^{l_{k_2}}\text{vec}\Big[(T_{(k_2,j)}u)(T_{(k_2,j)}u)'\Big]\bigg).$$

Through similar arguments to those used in the proof of the previous theorem, Theorem A.3, the above can be seen to be equal to:

$$\frac{1}{2}N_{q_{k_1}}\sum_{j=1}^{l_{k_2}}\sum_{i=1}^{l_{k_1}}\Big[(T_{(k_1,i)}T'_{(k_2,j)})\otimes(T_{(k_1,i)}T'_{(k_2,j)})\Big].$$

From the definition of $T_{(k,j)}$, it can be seen that the above is equal to the result of Theorem 4. □

We now turn attention to the derivation of Equations (3.7) and (3.8). As in the proofs of Corollaries A.2 and A.3 of Appendix A.1, we begin by noting that:

$$\frac{dl(\theta)}{d\text{vech}(D_k)}=\mathcal{D}^+_{q_k}\frac{dl(\theta)}{d\text{vec}(D_k)}=\mathcal{D}^+_{q_k}\mathcal{D}_{q_k}\mathcal{D}'_{q_k}\frac{\partial l(\theta)}{\partial\text{vec}(D_k)}=\mathcal{D}'_{q_k}\frac{\partial l(\theta)}{\partial\text{vec}(D_k)},$$

where the first equality follows from the definition of the duplication matrix and the second equality follows from Theorem 5.12 of Turkington (2013). Applying the above identity to Theorems A.2, A.3 and A.4 and moving the matrix $\mathcal{D}'_{q_k}$ outside the covariance function in each, leads to the following three corollaries which, when taken in combination, provide Equations (3.7) and (3.8).

Corollary A.4. *For any arbitrary integer k between 1 and r, the covariance of the total derivatives of $l(\theta^h)$ with respect to β and vech(D_k) is given by:*

$$\mathcal{I}^h_{\beta,vech(D_k)}=\mathbf{0}_{p,q_k(q_k+1)/2}.$$

Corollary A.5. *For any arbitrary integer k between 1 and r, the covariance of the*

total derivatives of $l(\theta^h)$ *with respect to* σ^2 *and* $vech(D_k)$ *is given by:*

$$\mathcal{I}^h_{\sigma^2, vech(D_k)} = \frac{1}{2\sigma^2} vec' \left(\sum_{j=1}^{l_k} Z'_{(k,j)} V^{-1} Z_{(k,j)} \right) \mathcal{D}_{q_k}.$$

Corollary A.6. *For any arbitrary integers* k_1 *and* k_2 *between* 1 *and* r*, the covariance of the total derivatives of* $l(\theta^h)$ *with respect to* $vech(D_{k_1})$ *and* $vech(D_{k_2})$ *is given by:*

$$\mathcal{I}^h_{vech(D_{k_1}), vech(D_{k_2})} = \frac{1}{2} \mathcal{D}'_{q_{k_1}} \sum_{j=1}^{l_{k_2}} \sum_{i=1}^{l_{k_1}} (Z'_{(k_1,i)} V^{-1} Z_{(k_2,j)} \otimes Z'_{(k_1,i)} V^{-1} Z_{(k_2,j)}) \mathcal{D}_{q_{k_2}}.$$

We note that the results of Corollaries A.4, A.5 and A.6 do not contain the matrix N_{q_k}, which appears in the corresponding theorems (Theorems A.2, A.3 and A.4). This is due to another result of Magnus and Neudecker (1986), which states that $\mathcal{D}'_k N_k = \mathcal{D}'_k$ for any integer k. This concludes the derivations of Fisher Information matrix expressions given in Sections 3.3.1.1-3.3.1.4.

A.4 Full Fisher Scoring

In this section, we prove that Equations (3.12) and (3.16) are valid Fisher Scoring rules of the form given by Equation (3.2). To achieve this, we first require the following lemma.

Lemma A.3. *Let* n, p *and* k *be arbitrary positive integers with* $p < n$*. Let* $\tilde{N}$ *be of dimension* $(n \times n)$ *with* $rank(\tilde{N}) < n$*,* $\tilde{K}$ *be of dimension* $(n \times p)$ *with* $rank(\tilde{K}) = p$*,* C *be of dimension* $(n \times n)$ *with* $rank(C) = n$ *and* δ *be of size* $(n \times k)$*. If the following properties are satisfied:*

1. $\tilde{K}\tilde{K}^+ = \tilde{N}$*, where* $\tilde{K}^+ = (\tilde{K}'\tilde{K})^{-1}\tilde{K}'$ *is the generalized inverse of* $\tilde{K}$*.*
2. $\tilde{N}$ *is symmetric and idempotent.*
3. $\tilde{N}C = C\tilde{N} = \tilde{N}C\tilde{N}$*.*

4. $\tilde{N}\delta = \delta$.

then it follows that:

$$(\tilde{N}C)^{+}\delta = C^{-1}\delta.$$

Proof. In proving the above lemma, we make use of the following result given by Barnett (1990);

- If A is a matrix of dimension $(m \times l)$ with $\text{rank}(A) = l < m$, and B is a matrix of dimension $(l \times m)$ with $\text{rank}(B) = l$, then;

$$(AB)^{+} = B'(BB')^{-1}(A'A)^{-1}A'. \tag{A.9}$$

By property 1, we have that:

$$(\tilde{N}C)^{+} = (\tilde{K}\tilde{K}^{+}C)^{+}. \tag{A.10}$$

By applying Equation (A.9) to the above, with $A = \tilde{K}$ and $B = \tilde{K}^{+}C$, and simplifying the result, the following is obtained:

$$(\tilde{K}\tilde{K}^{+}C)^{+} = C'\tilde{K}^{+'}(\tilde{K}^{+}CC'\tilde{K}^{+'})^{-1}\tilde{K}^{+}. \tag{A.11}$$

We now consider the inversion in the center of the above expression, i.e. the inversion of $(\tilde{K}^{+}CC'\tilde{K}^{+'})$. This inversion is given by;

$$(\tilde{K}^{+}CC'\tilde{K}^{+'})^{-1} = \tilde{K}'C^{-1'}C^{-1}\tilde{K}.$$

We will demonstrate this by showing that the product of the matrix inside the inversion on the left hand side and the matrix on the right hand side is the identity matrix:

$$(\tilde{K}^{+}CC'\tilde{K}^{+'})(\tilde{K}'C^{-1'}C^{-1}\tilde{K}) = \tilde{K}^{+}CC'\tilde{K}^{+'}\tilde{K}'C^{-1'}C^{-1}\tilde{K}.$$

Using properties 1, 2 and 3 we see that the above is equal to:

$$\tilde{K}^{+}\tilde{N}CC'C^{-1'}C^{-1}\tilde{K}.$$

Simplifying and applying property 1 now yields that the above is equal the below:

$$\tilde{K}^{+}\tilde{N}\tilde{K} = \tilde{K}^{+}\tilde{K} = I,$$

where I is the identity matrix as required. Returning to Equation (A.11) and applying properties 1 and 2, we now have that:

$$(\tilde{N}C)^{+} = C'\tilde{N}C^{-1'}C^{-1}\tilde{N}.$$

By applying properties 2 and 3, the above can now be reduced to:

$$(\tilde{N}C)^{+} = \tilde{N}C^{-1}\tilde{N}. \tag{A.12}$$

We now note that, by property 3, $\tilde{N}C = C\tilde{N}$. By multiplying the left and right side of this identity by C^{-1} it can be seen that:

$$C^{-1}\tilde{N} = \tilde{N}C^{-1}.$$

Multiplying the right hand side of the above by $\tilde{N}$ and applying property 2 we obtain:

$$C^{-1}\tilde{N} = \tilde{N}C^{-1}\tilde{N}.$$

Substituting the above into Equation (A.12) gives:

$$(\tilde{N}C)^{+} = C^{-1}\tilde{N}.$$

The result of Lemma A.3 now follows by multiplying the right-hand side by δ and

applying property 4. □

Theorem A.5. *The below two update rules are equivalent, where $F(\theta_s^f)$ is the matrix defined in Section 3.3.1.2 and $\mathcal{I}(\theta_s^f)$ is the Fisher Information matrix of θ_s^f.*

$$\theta_{s+1}^f = \theta_s^f + \alpha_s F(\theta_s^f)^{-1} \frac{\partial l(\theta_s^f)}{\partial \theta},$$

$$\theta_{s+1}^f = \theta_s^f + \alpha_s \mathcal{I}(\theta_s^f)^{+} \frac{\partial l(\theta_s^f)}{\partial \theta}.$$

Proof. To prove Theorem A.5, it suffices to prove the below equality:

$$\mathcal{I}(\theta_s^f)^{+} \frac{\partial l(\theta_s^f)}{\partial \theta} = F(\theta_s^f)^{-1} \frac{\partial l(\theta_s^f)}{\partial \theta}.$$

To prove the above, we employ Lemma A.3. To achieve this, we proceed by defining:

- $\tilde{N}$ to be the matrix:
$$\begin{bmatrix} I_{p+1} & 0 & 0 & \dots & 0 \\ 0 & N_{q_1} & 0 & \dots & 0 \\ 0 & 0 & N_{q_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & N_{q_r} \end{bmatrix},$$
- $\tilde{K}$ to be the matrix:
$$\begin{bmatrix} I_{p+1} & 0 & 0 & \dots & 0 \\ 0 & K_{q_1,q_1} & 0 & \dots & 0 \\ 0 & 0 & K_{q_2,q_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & K_{q_r,q_r} \end{bmatrix},$$
- δ to be the column vector:
$$\frac{\partial l(\theta^f)}{\partial \theta^f},$$
- C to be the matrix $F(\theta^f)$,

and claim that $F(\theta^f)\tilde{N} = \mathcal{I}(\theta^f)$. To prove this equality, we first note the following two properties of the matrix N_k, given by Magnus and Neudecker (1986).

1. $N_{k_1}(A \otimes A) = N_{k_1}(A \otimes A)N_{k_2} = (A \otimes A)N_{k_2}$ for any arbitrary matrix A and scalars k_1 and k_2, such that the matrix multiplications are well defined.

2. $N_k \text{vec}(A) = \text{vec}(A)$ for any arbitrary symmetric matrix A of appropriate dimension.

The first of the above properties, in combination with Equations (3.11) and (3.13), can be seen to imply that, for any k_1 and k_2 between 1 and r:

$$\mathcal{I}^f_{\text{vec}(D_{k_1}),\text{vec}(D_{k_2})} = F_{\text{vec}(D_{k_1}),\text{vec}(D_{k_2})} N_{q_{k_2}}.$$

In a similar manner, the second of the above properties, in combination with Equation (3.10), can be seen to imply that, for any k between 1 and r:

$$\mathcal{I}^f_{\sigma^2,\text{vec}(D_k)} = F_{\sigma^2,\text{vec}(D_k)} N_{q_k}.$$

Expanding the multiplication $F(\theta^f_s)\tilde{N}$ block-wise demonstrates that the two equations above are sufficient to prove the equality $F(\theta^f_s)\tilde{N} = \mathcal{I}(\theta^f_s)$. The matrices $\tilde{N}, \tilde{K}$ and C can be seen to satisfy properties 1-3 of Lemma A.3 by noting standard results concerning the matrices $\mathcal{D}_k$, $K_{m,n}$ and N_n, provided in Magnus and Neudecker (1986). Property 4 of Lemma A.3 can be seen to hold for the matrix $\tilde{N}$ and column vector δ by noting the symmetry of $\frac{\partial l(\theta^f_s)}{\partial D_k}$ (see Theorem A.1). By combining the previous arguments and applying Lemma A.3, the below is obtained:

$$\mathcal{I}(\theta^f)^+\delta = (F(\theta^f)\tilde{N})^+\delta = F(\theta^f)^{-1}\delta.$$

The result of Theorem A.5 now follows.

□

Theorem A.6. *For any fixed integer k between 1 and r, the below two update rules are equivalent, where $F_{vec(D_{k,s})}$ is as defined in Section 3.3.1.2 and $\mathcal{I}^f_{vec(D_{k,s})}$ is the Fisher Information matrix of $vec(D_k)$.*

$$vec(D_{k,s+1}) = vec(D_{k,s}) + \alpha_s F^{-1}_{vec(D_{k,s})} \frac{\partial l(\theta^f_s)}{\partial vec(D_{k,s})},$$

$$vec(D_{k,s+1}) = vec(D_{k,s}) + \alpha_s \big(\mathcal{I}^f_{vec(D_{k,s})}\big)^{+} \frac{\partial l(\theta^f_s)}{\partial vec(D_{k,s})}.$$

Proof. The proof of the above proceeds by defining $\tilde{N} = N_{q_k}$, $\tilde{K} = K_{q_k,q_k}$, $C = F_{\text{vec}(D_k)}$ and δ to be the partial derivative appearing in the statement of the theorem. Following this, the proof follows the same structure as that of Theorem A.5. □

A.5 Restricted Maximum Likelihood Estimation

In this appendix, we describe how the methods from Section 3.3.1 may be adapted to use an alternative likelihood criteria: the criteria employed by Restricted Maximum Likelihood (ReML) estimation. A well-documented issue for ML estimation is that the variance estimates produced using ML are biased. ReML addresses this issue by maximizing the log-likelihood function of the residual vector, e, instead of the response vector, Y. Neglecting constant terms, the Restricted Maximum log-likelihood function, l_R, satisfies:

$$l_R(\theta^h) = l(\theta^h) - \frac{1}{2}\bigg(- p\log(\sigma^2) + \log|X'V^{-1}X|\bigg),$$

where $l(\theta^h)$ is given in Equation (2.4). To derive the ReML-based FS algorithm, akin to that described in Section 3.3.1.1, the following adjustments to the score vectors

given by Equations (3.4) and (3.5) must be used.

$$\frac{dl_R(\theta^h)}{d\beta} = \frac{dl(\theta^h)}{d\beta}, \qquad \frac{dl_R(\theta^h)}{d\sigma^2} = \frac{dl(\theta^h)}{d\sigma^2} + \frac{1}{2}p\sigma^{-2},$$
$$\frac{dl_R(\theta^h)}{d\text{vech}(D_k)} = \frac{dl(\theta^h)}{d\text{vech}(D_k)} + \frac{1}{2}\mathcal{D}'_{q_k}\text{vec}\Bigg(\sum_{j=1}^{l_k} Z'_{(k,j)}V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}Z_{(k,j)}\Bigg).$$

The latter result can be derived through similar means to that of Theorem A.1 and Corollaries A.1-A.3. Derivation of the ReML score vectors of β and σ^2 can be found in Demidenko (2013) where proofs may be found for the single-factor LMM. As these proofs do not depend on the number of factors in the model, they can be seen to also apply in the multi-factor LMM setting without further adjustment.

Due to the asymptotic equivalence of the ML and ReML estimates, the parameter estimates produced by ML and ReML have the same asymptotic covariance matrix. Consequently, the Fisher Information matrix for the parameter estimates produced by ReML, which is the inverse of the asymptotic covariance matrix, is identical to that of the parameter estimates produced by ML. As a result, the ReML-based FS algorithm utilizes the Fisher Information matrix specified by Equations (3.6)-(3.8) and requires no further derivation. To summarize, the ReML-based FS algorithm for the multi-factor LMM is almost identical to that given in Algorithm 1. The only adaptations required occur on line 3 of Algorithm 1 where the score vectors must be substituted for their ReML counterparts, provided above. Analogous adjustments can be made for the algorithms presented in Sections 3.3.1.2-3.3.1.5. Notably, in Chapter 4, we employ the ReML adaptation of the FSFS algorithm for parameter estimation of large-sample fMRI LMM analyses.

A.6 Ensuring Non-negative Definiteness

In order to ensure numerical optimization produces a valid covariance matrix when performing LMM parameter estimation, optimization must be constrained to ensure that for each factor, f_k, D_k is non-negative definite. This is a well-documented obsta-

cle for numerical approaches to LMM parameter estimation for which many solutions have been suggested. In this section, we describe how non-negative definiteness was ensured for the covariance matrix estimates produced by the Fisher Scoring algorithms described in Chapter 3.

A common approach, utilized in Bates et al. (2015) for example, is to reparameterize optimization in terms of the Cholesky factor of the covariance matrix, Λ_k, as oppose to the covariance matrix itself, D_k. This approach is adopted by the CSFS method, described in Section 3.3.1.5. However, the methods described in Sections 3.3.1.1-3.3.1.4 require an alternative approach as optimization does not account for the constraint that D_k must be non-negative definite. For the FS, FFS, SFS and FSFS algorithms, we use an approach described by Demidenko (2013), for the single factor LMM. Denoting the Eigendecomposition of $\{D_k\}_{k\in\{1,\dots,r\}}$ as $D_k = U_k\Sigma_k U_k'$, we project D_k onto the space of all $(q_k \times q_k)$ non-negative definite matrices by, following each update of the covariance matrix, D, replacing $\{D_k\}_{k\in\{1,..r\}}$ with $\{D_{+,k}\}_{k\in\{1,..r\}}$,

$$D_{+,k} = U_k\Sigma_{+,k}U_k',$$

where $\Sigma_{+,k} = \max(\Sigma_k, 0)$ with 'max' representing the element-wise maximum. For each factor, f_k, this ensures D_k, and as a result D, is non-negative definite.

We note here it could be argued that using the Eigendecomposition in this manner defeats our objective to employ only simplistic vectorizable operations for parameter estimation. In response to this potential criticism, however, we highlight that vectorized Eigendecomposition operations are widely available in practice and can be found in many programming languages such as, for example, MatLab and Python.

A.7 Constraint Matrices

In this section, we provide further detail on the approach to constrained optimization which was outlined in Section 3.3.4. We begin by providing examples of how the

notation described in Section 3.3.4 may be used in practice. For a given k, we introduce the notation $\{d_i\}$ and $\{\tilde{d}_i\}$ to represent the set of unique parameters required to specify the constrained and unconstrained representations of D_k, respectively. For example, if D_k is (3×3) in dimension and we wish to induce Toeplitz structure onto D_k, then D_k can be considered to take the following "constrained" and "unconstrained" representations:

$$D_k = \begin{bmatrix} d_1 & d_4 & d_7 \\ d_2 & d_5 & d_8 \\ d_3 & d_6 & d_9 \end{bmatrix} = \begin{bmatrix} \tilde{d}_1 & \tilde{d}_2 & \tilde{d}_3 \\ \tilde{d}_2 & \tilde{d}_1 & \tilde{d}_2 \\ \tilde{d}_3 & \tilde{d}_2 & \tilde{d}_1 \end{bmatrix}.$$

The notation vecu(A) is used to denote the column vector of unique variables in the matrix A, given in order of first occurrence as listed in column major order. In the above example, this implies vecu$(D_k) = [\tilde{d}_1, \tilde{d}_2, \tilde{d}_3]'$ whilst vec$(D_k) = [d_1, \ldots, d_9]'$. Using this notation, the constraint matrix $\mathcal{C}_k$ may be defined element-wise. To see this, note that the chain rule for the total derivative is given by:

$$\frac{dl(\theta^{con})}{d\tilde{d}_j} = \sum_i \frac{\partial d_i}{\partial \tilde{d}_j} \frac{\partial l(\theta^{con})}{\partial d_i}.$$

Comparing this with the derivative expression provided in Equation (3.25), it can be seen that the elements of $\mathcal{C}_k$ are given by:

$$\mathcal{C}_{k[j,i]} = \frac{\partial d_i}{\partial \tilde{d}_j}.$$

In general, derivation of the constraint matrix for the k^{th} factor is relatively straightforward for most practical applications. Table A.1 supports this statement by providing examples of the constraint matrix for the k^{th} factor operating under the assumption that D_k exhibits several commonly employed covariance structures. In addition, we note that two instances in which a constraint matrix approach was employed implicitly can already be found in Chapter 3. In Section 3.3.1.1, we took $\mathcal{C}_k = \mathcal{D}_{q_k}$ to enforce symmetry in D_k and, in Section 3.3.1.5, we took $\mathcal{C}_k = (d\text{vech}(D_k)/d\text{vech}(\Lambda_k))\mathcal{D}_{q_k}$ in order to derive the Fisher Scoring update in

Covariance Structure	**Constraint matrix, $\mathcal{C}_k$**
Diagonal $\begin{bmatrix} \tilde{d}_1 & 0 & 0 \\ 0 & \tilde{d}_1 & 0 \\ 0 & 0 & \tilde{d}_1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$
Variance components $\begin{bmatrix} \tilde{d}_1 & 0 & 0 \\ 0 & \tilde{d}_2 & 0 \\ 0 & 0 & \tilde{d}_3 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
Toeplitz $\begin{bmatrix} \tilde{d}_1 & \tilde{d}_2 & \tilde{d}_3 \\ \tilde{d}_2 & \tilde{d}_1 & \tilde{d}_2 \\ \tilde{d}_3 & \tilde{d}_2 & \tilde{d}_1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$
Compound Symmetry $\begin{bmatrix} 1 & \tilde{d}_1 & \tilde{d}_1 \\ \tilde{d}_1 & 1 & \tilde{d}_1 \\ \tilde{d}_1 & \tilde{d}_1 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$
AR(1) $\begin{bmatrix} 1 & \tilde{d}_1 & \tilde{d}_1^2 \\ \tilde{d}_1 & 1 & \tilde{d}_1 \\ \tilde{d}_1^2 & \tilde{d}_1 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 2\tilde{d}_1 & 1 & 0 & 1 & 2\tilde{d}_1 & 1 & 0 \end{bmatrix}$

Table A.1: The constraint matrix for the k^{th} factor, $\mathcal{C}_k$, in the setting in which the k^{th} factor groups 3 random effects, under various covariance structures.

terms of the Cholesky factor, Λ_k.

It is also worth noting that this approach to constrained optimization allows for different covariance structures to be enforced on different factors in the LMM. For instance, a researcher may enforce the first factor in an LMM to have an AR(1) structure while allowing the second factor to exhibit a completely different structure such as compound symmetry. As noted in Section 3.3.4, this approach can also be extended further to allow all elements of the vectorized covariance matrices, $\text{vec}(D_1), \ldots \text{vec}(D_r)$, to be expressed as functions of one set of parameters, ρ_D, common to all $\{\text{vec}(D_k)\}_{k\in\{1,\ldots,r\}}$. To further detail this, we define $v(D)$ and extend the definition θ^{con} as:

$$v(D) = \begin{bmatrix} \text{vec}(D_1) \\ \vdots \\ \text{vec}(D_r) \end{bmatrix}, \qquad \theta^{con} = \begin{bmatrix} \beta \\ \sigma^2 \\ \rho_D \end{bmatrix}.$$

Through the same process used to obtain the equations presented in Section 3.3.4, the below can be obtained;

$$\frac{dl(\theta^{con})}{d\rho_D} = \mathcal{C}\frac{\partial l(\theta^{con})}{\partial v(D)}, \qquad \mathcal{I}^{con}_{\rho_D} = \mathcal{C}\mathcal{I}^{f}_{\text{v}(D)}\mathcal{C}', \tag{A.13}$$

where $\mathcal{C}$ is the constraint matrix defined to be the Jacobian matrix of partial derivatives of $v(D)$ taken with respect to ρ_D. The score vector and Fisher Information matrix associated to $v(D)$ are readily seen to be given by block-wise combinations of the matrix expressions (3.9) and (3.11).

A.8 Cholesky Fisher Scoring

In this appendix, we prove the identity stated by the below theorem, Theorem A.7, which was utilized in Section 3.3.1.5.

Theorem A.7. *Let D_k be a square symmetric positive-definite matrix with Cholesky decomposition given by $D_k = \Lambda_k\Lambda_k'$. The below expression gives the derivative of*

vech(D_k) *with respect to vech*(Λ_k)*.*

$$\frac{\partial vech(D_k)}{\partial vech(\Lambda_k)} = \mathcal{L}_{q_k}(\Lambda_k' \otimes I_{q_k})(I_{q_k^2} + K_{q_k})\mathcal{D}_{q_k}.$$

Proof. By the chain rule for vector-valued functions, as stated by Turkington (2013), the derivative in Theorem A.7 can be expanded in the following manner:

$$\frac{\partial \text{vech}(D_k)}{\partial \text{vech}(\Lambda_k)} = \frac{\partial \text{vec}(\Lambda_k)}{\partial \text{vech}(\Lambda_k)} \frac{\partial \text{vec}(D_k)}{\partial \text{vec}(\Lambda_k)} \frac{\partial \text{vech}(D_k)}{\partial \text{vec}(D_k)}.$$

The first and third derivatives in the above product are given by Theorems 5.9 and 5.10 of Turkington (2013), respectively, which state:

$$\frac{\partial \text{vec}(\Lambda_k)}{\partial \text{vech}(\Lambda_k)} = \mathcal{L}_{q_k}, \qquad \frac{\partial \text{vech}(D_k)}{\partial \text{vec}(D_k)} = \mathcal{D}_{q_k}.$$

The second derivative in the product is given by a result of Magnus and Neudecker (1999), which states that:

$$\frac{\partial \text{vec}(\Lambda_k \Lambda_k')}{\partial \text{vec}(\Lambda_k)} = (\Lambda_k' \otimes I_{q_k})(I_{q_k^2} + K_{q_k}).$$

Combining the above derivatives yields the desired result. □

A.9 Satterthwaite Degrees of Freedom Estimation

In this section, we provide an in-depth accounting of Satterthwaite estimation for degrees of freedom in the multi-factor LMM setting. In Section A.9.1, we detail how the expression described by Equation (3.26) is derived for estimating the degrees of freedom of the approximate T-statistic. Following this, in Section A.9.2, we describe how this approach is extended to degrees of freedom estimation for the approximate F-statistic. Finally, in Section A.9.3, we prove the derivative result stated in Section 3.3.5 and extend it to the setting of the constrained covariance matrices discussed

in Section 3.3.4. The first two appendices follow the proofs outlined in the work of Kuznetsova et al. (2017). The reader is referred to this work for a more in-depth discussion of Satterthwaite degrees of freedom estimation for the LMM.

A.9.1 Satterthwaite Estimation for the Approximate T-statistic

To derive the estimator given by Equation (3.26), we first begin by noting that the approximate T-statistic used for LMM null hypothesis testing is given by:

$$T = \frac{L\hat{\beta}}{\sqrt{L\widehat{\text{Var}}(\hat{\beta})L'}},$$

where $\hat{\beta}$ is the estimator of β obtained from numerical optimization and $\widehat{\text{Var}}(\hat{\beta})$ is an estimate of the variance of the β estimator, given by:

$$\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2(X'\hat{V}^{-1}X)^{-1}.$$

We highlight that the above expression differs from the T-statistic typically employed in the setting of LM hypothesis testing (c.f. Section 2.2.4). Under the standard assumptions for the LM, it is possible to derive an estimate for the variance of the estimator $\hat{\beta}$, given by $\hat{\sigma}^2(X'X)^{-1}$, which ensures that T is student's t-distributed. However, the above estimator, $\widehat{\text{Var}}(\hat{\beta})$, differs from this estimate as it includes an estimate of V which has an unknown distribution. As a direct result, no equivalent theory is available to obtain the distribution of $\widehat{\text{Var}}(\hat{\beta})$ in the setting of the LMM. A consequence of this is that, whilst T follows a student's t_{n-p}-distribution in the setting of linear regression, the distribution of T in the LMM setting is unknown. The aim of this section is to approximate the degrees of freedom of T assuming that it follows a students t-distribution.

To meet this aim, we now denote $\text{Var}(\hat{\beta})$ as the unknown true variance of the $\hat{\beta}$ estimator. By multiplying both the numerator and denominator of the T-statistic by $(L\text{Var}(\hat{\beta})L)^{1/2}$, the below expression for T is obtained:

$$T = \frac{Z}{\sqrt{\frac{L\widehat{\text{Var}}(\hat{\beta})L'}{L\text{Var}(\hat{\beta})L'}}},$$

where Z is a random variable with standard normal distribution, given by $Z = L\hat{\beta}/\sqrt{L\text{Var}(\hat{\beta})L'}$. We now consider the definition of the students t-distribution, which states that a random variable is t-distributed with v degrees of freedom if and only if it takes the following form:

$$T = \frac{Z}{\sqrt{\frac{X}{v}}},$$

where $Z \sim N(0,1)$ and $X \sim \chi^2_v$. When the degrees of freedom v are unknown, as both X and T are distributed with v degrees of freedom, it suffices to estimate the degrees of freedom of the chi-square variable X instead of the t-distributed variable T. By equating the approximate T-statistic with the defining form of a t-distributed random variable, under the assumption that the T-statistic is truly t-distributed, the below expression for X can be obtained:

$$X = \frac{vL\widehat{\text{Var}}(\hat{\beta})L'}{L\text{Var}(\hat{\beta})L'} \sim \chi^2_v.$$

We now define the notation $S^2(\hat{\eta})$ to represent the estimated variance of $L\hat{\beta}$ as a function of the estimated variance parameters given by $\hat{\eta} = (\hat{\sigma}^2, \hat{D}_1, \ldots, \hat{D}_r)$. Similarly, we define Σ^2 to be the true unknown variance of $L\hat{\beta}$. Noting that $\Sigma^2 = L\text{Var}(\hat{\beta})L'$ and by rearranging the above, the below can be obtained:

$$S^2(\hat{\eta}) = L\widehat{\text{Var}}(\hat{\beta})L' \sim \frac{\Sigma^2}{v}\chi^2_v.$$

By considering the variance of the above expression and noting that, as $X \sim \chi^2_v$, $\text{Var}(X) = 2v$, the below expression is obtained.

$$\text{Var}(S^2(\hat{\eta})) = \frac{(\Sigma^2)^2}{v^2}\text{Var}(X) = \frac{2(\Sigma^2)^2}{v}.$$

To obtain an estimator of v in terms of $S^2(\hat{\eta})$, the above is rearranged and the approximation $\Sigma^2 \approx S^2(\hat{\eta})$ is used. This yields the below approximation.

$$\hat{v}(\hat{\eta}) = \frac{2(S^2(\hat{\eta}))^2}{\text{Var}(S^2(\hat{\eta}))}.$$

Under the assumption that $T \sim t_v$, $\hat{v}(\hat{\eta})$ is an estimator for the degrees of freedom of X and, therefore, for the degrees of freedom of T. This concludes the derivation of Equation (3.26).

A.9.2 Satterthwaite Estimation for the Approximate F-statistic

In this section, we detail how the Satterthwaite method described in Section A.9.1 may be extended in order to estimate the degrees of freedom of the approximate F-statistic. The approximate F-statistic for the LMM takes the below form:

$$F = \frac{\hat{\beta}' L' [L\widehat{\text{Var}}(\hat{\beta})L']^{-1} L\hat{\beta}}{r},$$

where, throughout this section only, the notation F will represent the approximate F-statistic and r will be used to denote $r = \text{rank}(L)$. The aim of this section is to approximate F with an $F_{r,v}$ distribution where v, the denominator degrees of freedom, is estimated using a Satterthwaite method of approximation. To simplify the notation we will first consider Q, which is the numerator of F;

$$Q = rF = \hat{\beta}' L' [L\text{Var}(\hat{\beta})L']^{-1} L\hat{\beta}.$$

We begin by considering the eigendecomposition of $L\text{Var}(\hat{\beta})L'$, which we will denote:

$$L\text{Var}(\hat{\beta})L' = U\Lambda U',$$

where U is an orthonormal matrix of eigenvectors and Λ is a diagonal matrix of eigenvalues. Using standard properties of the eigendecomposition the below is obtained:

$$Q = (U'L\hat{\beta})'\Lambda^{-1}(U'L\hat{\beta}) = \sum_{i=1}^{r} \frac{(U'L\hat{\beta})_i^2}{\lambda_i},$$

where $(U'L\hat{\beta})_i$ is the i^{th} element of the vector $U'L\hat{\beta}$ and λ_i is the i^{th} diagonal element (eigenvalue) of Λ, $\Lambda_{[i,i]}$. For purposes which will become clear, we define $\tilde{L}$ as $U'L$. By noting that $\lambda_i = (U'L\text{Var}(\hat{\beta})L'U)_i$, the above can be seen to be equal to:

$$Q = \sum_{i=1}^{r} \left(\frac{\tilde{L}_i\hat{\beta}}{\sqrt{\tilde{L}_i\text{Var}(\hat{\beta})\tilde{L}_i'}} \right)^2,$$

where $\tilde{L}_i$ represents the i^{th} row of $\tilde{L}$. Noting now the definition of $\tilde{L}$, it can be seen that the above is a sum of independent squared T-statistics of the form discussed in Section A.9.1. By denoting the unknown degrees of freedom of each of the T-statistics in the above sum by v_i, it follows that:

$$\mathbb{E}(Q) = \sum_{i=1}^{r} \mathbb{E}\left[\left(\frac{\tilde{L}_i\hat{\beta}}{\sqrt{\tilde{L}_i\text{Var}(\hat{\beta})\tilde{L}_i'}} \right)^2 \right] = \sum_{i=1}^{r} \frac{v_i}{v_i - 2},$$

where the second equality follows from the fact that a squared T-statistic with degrees of freedom v_i is an F_{1,v_i}-statistic, which in turn has expectation $\frac{v_i}{v_i-2}$. Noting that, in practice, the true variance of $\hat{\beta}$ is unknown, the methods of A.9.1 must be employed to estimate each of the v_i in the above sum.

To use the above expression for $\mathbb{E}(Q)$ to estimate the denominator degrees of freedom of F, three cases are considered; the cases $r > 2, r = 2$ and $r = 1$. Each of these will now be considered in turn.

Case 1 ($\mathbf{r > 2}$): By recalling $Q = rF$, and by the standard formula for the expectation of an F distribution with numerator degrees of freedom v greater than 2, the below is obtained.

$$\mathbb{E}(Q) = \mathbb{E}(rF) = \frac{rv}{v-2}.$$

Setting the above equal to the previous expression for $\mathbb{E}(Q)$ and rearranging, the

below expression for v is obtained:

$$v = \frac{2\sum_{i=1}^{r} \frac{v_i}{v_i - 2}}{\left(\sum_{i=1}^{r} \frac{v_i}{v_i - 2}\right) - r}.$$

By using the Satterthwaite degrees of freedom approximation method described in Section A.9.1 to estimate $\{v_i\}_{i\in\{1,\ldots,r\}}$, it follows that the above formula can be used to estimate the degrees of freedom of the approximate F-statistic when $r > 2$.

Case 2 ($\mathbf{r = 2}$): When $r = 2$, the expectation of Q is infinite and, resultantly, the equality used in the argument of case 1 no longer holds. A common convention for estimation of v in the case $r = 2$ is based on the observation that for $r > 2$ the expectation of Q satisfies the following relationship with v:

$$v = \frac{2\mathbb{E}(Q)}{\mathbb{E}(Q) - r}.$$

By taking the limit from above of both sides of this expression, and noting that $\mathbb{E}(Q) \to \infty$ as $r \downarrow 2$, the following estimate for v is obtained.

$$v = \lim_{r\downarrow 2} \frac{2\mathbb{E}(Q)}{\mathbb{E}(Q) - r} = 2.$$

Case 3 ($\mathbf{r = 1}$): For the case $r = 1$, it is well known that if $F \sim F_{1,v_1}$ for some integer v_1, then F is the square of a T-statistic with the same degrees of freedom, v_1. Resultantly, in this case, the unknown denominator degrees of freedom of the F-statistic, v_1, can be estimated using the methods of Section A.9.1.

A.9.3 Derivative of $S^2(\hat{\eta}^h)$ with respect to vech$(\hat{D}_k)$

In this appendix, proof of the derivative result which was given in Section 3.3.5 is provided. Following this, an extension of this result is provided for models containing

constrained covariance matrices of the form described in Section 3.3.4.

Theorem A.8. *Define $\hat{\eta}^h$ as in Section 3.3.5 and define the function $S^2(\hat{\eta}^h)$ by:*

$$S^2(\hat{\eta}^h) = \sigma^2 L(X'\hat{V}^{-1}X)^{-1}L'.$$

The derivative of $S^2(\hat{\eta}^h)$ with respect to the half vectorisation of $\hat{D}_k$ is given by:

$$\frac{dS^2(\hat{\eta}^h)}{dvech(\hat{D}_k)} = \hat{\sigma}^2 \mathcal{D}'_{q_k}\Bigg(\sum_{j=1}^{l_k} \hat{B}_{(k,j)} \otimes \hat{B}_{(k,j)}\Bigg),$$

where $\hat{B}_{(k,j)}$ is given by $\hat{B}_{(k,j)} = Z'_{(k,j)}\hat{V}^{-1}X(X'\hat{V}^{-1}X)^{-1}L'$.

Proof. By the chain rule for vector valued functions, as stated by Turkington (2013), and by noting that the function $S^2(\hat{\eta}^h)$ outputs a (1×1) scalar value, it can be seen that:

$$\frac{\partial S^2(\hat{\eta}^h)}{\partial \text{vec}(\hat{D}_k)} = \hat{\sigma}^2 \frac{\partial\big(L(X'\hat{V}^{-1}X)^{-1}L'\big)}{\partial \text{vec}(\hat{D}_k)} =$$
$$\hat{\sigma}^2 \frac{\partial \text{vec}(\hat{V})}{\partial \text{vec}(\hat{D}_k)} \frac{\partial \text{vec}(\hat{V}^{-1})}{\partial \text{vec}(\hat{V})} \frac{\partial \text{vec}(X'\hat{V}^{-1}X)}{\partial \text{vec}(\hat{V}^{-1})} \frac{\partial\big(L(X'\hat{V}^{-1}X)^{-1}L'\big)}{\partial \text{vec}(X'\hat{V}^{-1}X)}.$$

Using the expansion given in Equation (A.4) the first derivative in the product is evaluated to the below.

$$\frac{\partial \text{vec}(\hat{V})}{\partial \text{vec}(\hat{D}_k)} = \sum_{j=1}^{l_k} Z'_{(k,j)} \otimes Z'_{(k,j)}.$$

For the second derivative in the product, we apply result (4.17) from Turkington (2013), which states:

$$\frac{\partial \text{vec}(\hat{V}^{-1})}{\partial \text{vec}(\hat{V})} = -\hat{V}^{-1} \otimes \hat{V}^{-1}.$$

For the third term of the product, a result stated in Chapter 5.7 of Turkington (2013) gives the following:

$$\frac{\partial \text{vec}(X'\hat{V}^{-1}X)}{\partial \text{vec}(\hat{V}^{-1})} = X \otimes X.$$

By a variant of (4.17) from Turkington (2013), given on the line following the statement of (4.17), the below is obtained:

$$\frac{\partial\big(L(X'\hat{V}^{-1}X)^{-1}L'\big)}{\partial \text{vec}(X'\hat{V}^{-1}X)} = -((X'\hat{V}^{-1}X)^{-1}L') \otimes ((X'\hat{V}^{-1}X)^{-1}L').$$

Following this, application of the mixed product property of the Kronecker product and multiplication by the transposed duplication matrix yields the result of Theorem A.8. □

Theorem A.9. *Define $\rho_{\hat{D}}$ and $\mathcal{C}$ as in Section 3.3.4 and $\hat{B}_{(k,j)}$ and $\hat{\eta}^h$ as in Theorem A.8. The derivative of $S^2(\hat{\eta}^h)$ with respect to $\rho_{\hat{D}}$ is given by:*

$$\frac{dS^2(\hat{\eta}^h)}{d\rho_{\hat{D}}} = \sigma^2 \mathcal{C}\hat{\mathcal{B}},$$

where $\hat{\mathcal{B}}$ is defined by:

$$\hat{\mathcal{B}} = \left[\left(\sum_{j=1}^{l_1} \hat{B}'_{(1,j)} \otimes \hat{B}'_{(1,j)}\right), \ldots, \left(\sum_{j=1}^{l_r} \hat{B}'_{(r,j)} \otimes \hat{B}'_{(r,j)}\right)\right]'.$$

Proof. From the proof of Theorem A.8, it can be seen that the partial derivative of $S^2(\hat{\eta}^h)$ with respect to $\text{vec}(\hat{D}_k)$ is given by:

$$\frac{\partial S^2(\hat{\eta}^h)}{\partial \text{vec}(\hat{D}_k)} = \hat{\sigma}^2 \sum_{j=1}^{l_k} \hat{B}_{(k,j)} \otimes \hat{B}_{(k,j)}.$$

By defining $v(\hat{D})$ as in Section 3.3.4, it can be seen from the above that the partial derivative of $S^2(\hat{\eta}^h)$ with respect to $v(\hat{D})$ is $\hat{\sigma}^2\hat{\mathcal{B}}$. By the chain rule for vector valued functions, as stated by Turkington (2013), the below can now be obtained;

$$\frac{dS^2(\hat{\eta}^h)}{d\rho_{\hat{D}}} = \frac{\partial v(\hat{D})}{\partial \rho_{\hat{D}}} \frac{\partial S^2(\hat{\eta}^h)}{\partial v(\hat{D})} = \mathcal{C}\frac{\partial S^2(\hat{\eta}^h)}{\partial v(\hat{D})},$$

where the second equality follows from the definition of $\mathcal{C}$. Substituting the partial

derivative of $S^2(\hat{\eta}^h)$ with respect to $v(\hat{D})$ into the above completes the proof. □

A.10 The ACE Model

In this appendix, further detail concerning the ACE model of Section 3.5.1.2 is provided. First, Appendix A.10.1 details how the random effects vector, b, and covariance matrix, D, may be constructed for the ACE model. Following this, Appendix A.10.2 details how the kinship matrices, which model the additive genetic and common environmental variance components of the ACE model, are defined. Next, Appendix A.10.3 provides an overview of the constrained optimization procedure adopted for the ACE model, alongside an expression for the constraint matrix employed for optimization. Finally, Appendix A.10.4 provides detail on how the parameter estimation method that was employed to obtain the results of Section 3.5.2.2 was efficiently implemented.

A.10.1 Specification of Random Effects

In this appendix, we provide detail on how the random effects vector, b, and covariance matrix, D, are defined for the ACE model. To do so, we first describe the covariance of the random terms $\gamma_{k,j,i}$ from Section 3.5.1.2. Following this, we use the $\gamma_{k,j,i}$ terms to construct the random effects vector, b. To simplify notation, we assume that for each family structure type, subjects within family units which exhibit such a structure are ordered consistently. For example, if the family structure describes "families containing one twin-pair and one half-sibling", we assume that the members of every family who exhibit such a structure are given in the same order: (twin, twin, half-sibling).

As noted in Section 3.5.1.2, $\gamma_{k,j,i}$ models the within-"family unit" covariance. As such, $\text{cov}(\gamma_{k,j_1,i}, \gamma_{k,j_2,i}) = 0$, for any two distinct family units, $j_1 \neq j_2$. Within any individual family unit (e.g. family unit j of structure type k), the random effects

$\{\gamma_{k,j,i}\}_{i\in\{1,\ldots,q_k\}}$ are defined to have the below covariance matrix:

$$\text{cov}\left(\left[\gamma_{k,j,1}, \quad \ldots, \quad \gamma_{k,j,q_k}\right]'\right) = \sigma_a^2\mathbf{K}_k^a + \sigma_c^2\mathbf{K}_k^c,$$

where $\mathbf{K}_k^a$ and $\mathbf{K}_k^c$ are the known, predefined kinship (additive genetic) and common environmental effects matrices, respectively (See Appendix A.10.2 for more details). The random effects vector b, is constructed by vertical concatenation of the random $\gamma_{k,j,i}$ terms, i.e. $b = [\gamma_{1,1,1}, \gamma_{1,1,2}, ..., \gamma_{r,l_r,q_r}]'$. To derive an expression for the covariance matrix of b, we note from Equation (2.3) that $\sigma_e^2 D$ is equal to $\text{cov}(b)$. Equating this with the previous expression, it may now be seen that D is block diagonal, with its k^{th} unique diagonal block given by $D_k = \sigma_e^{-2}(\sigma_a^2\mathbf{K}_k^a + \sigma_c^2\mathbf{K}_k^c)$.

A.10.2 Kinship and Environmental Matrices

In twin studies, twin pairs are typically described as either Monozygotic (MZ) or Dizygotic (DZ). MZ twin pairs are identical twins, whilst DZ twins are non-identical. The additive genetic kinship matrix, $\mathbf{K}_k^a$, describes the additive genetic similarity between members of a family unit. For a family of q_k members, $\mathbf{K}_k^a$, is a $(q_k \times q_k)$ dimensional constant matrix with its $(i,j)^{th}$ element given by:

$$(\mathbf{K}_k^a)_{[i,j]} = \begin{cases} 1 & \text{if subjects } i \text{ and } j \text{ are MZ twins or } i = j. \\ \frac{1}{2} & \text{if subjects } i \text{ and } j \text{ are DZ twins or full siblings.} \\ \frac{1}{4} & \text{if subjects } i \text{ and } j \text{ are half siblings.} \\ 0 & \text{if subjects } i \text{ and } j \text{ are unrelated.} \end{cases}$$

The common environmental matrix, $\mathbf{K}_k^c$, describes the similarity between members of a family unit which is attributed to individuals' being reared in a shared environment. $\mathbf{K}_k^c$ is a $(q_k \times q_k)$ dimensional constant matrix with its $(i,j)^{th}$ element

given by:

$$(\mathbf{K}_k^a)_{[i,j]} = \begin{cases} 1 & \text{if subjects } i \text{ and } j \text{ were reared together.} \\ 0 & \text{if subjects } i \text{ and } j \text{ were not reared together.} \end{cases}$$

For more information on the construction of kinship and common environmental matrices for twin studies, the reader is referred to Lawlor et al. (2009).

A.10.3 Constrained Optimization for the ACE Model

In this appendix, a brief overview of the constrained optimization procedure which was adopted for parameter estimation of the ACE model is provided. A derivation of the constraint matrix, $\mathcal{C}$ is then provided by Theorem A.10.

To perform parameter estimation for the ACE model, an approach based on the ReML criterion described in Appendix A.5 and the constrained covariance structure methods outlined in Section 3.3.4 was adopted. The resulting optimization procedure was performed in terms of the parameter vector $\theta^{ACE} = (\beta, \tau_a, \tau_c, \sigma_e^2)$, where $\tau_a = \sigma_e^{-1}\sigma_a$ and $\tau_c = \sigma_e^{-1}\sigma_c$. β and σ_e^2 were updated according to the GLS update rules provided by Equation (3.14), whilst updates for the parameter vector $[\tau_a, \tau_c]'$ were performed via a Fisher Scoring update rule. The Fisher Scoring update rule employed was of the form (3.1) with θ substituted for $[\tau_a, \tau_c]'$. To obtain the gradient vector and information matrix required to perform this update step, a constraint based approach (c.f. Section 3.3.4) was employed. An expression for the required constraint matrix, $\mathcal{C}$, alongside derivation, is given by Theorem A.10 below.

Theorem A.10. *For the ACE model, the constraint matrix (Jacobian) which maps a partial derivative vector taken with respect to $v(D)$ to a total derivative vector taken with respect to $\tau = [\tau_a, \tau_c]'$ is given by:*

$$\mathcal{C} = \left(\mathbb{1}_{(1,r)} \otimes \begin{bmatrix} 2\tau_a & 0 \\ 0 & 2\tau_c \end{bmatrix} \right) \left(\bigoplus_{k=1}^{r} \begin{bmatrix} vec(\mathbf{K}_k^a)' \\ vec(\mathbf{K}_k^c)' \end{bmatrix} \right),$$

where $\oplus$ represents the direct sum.

Proof. We first define $\tilde{\tau}_{a,1}, \ldots \tilde{\tau}_{a,r}$ and $\tilde{\tau}_{c,1}, \ldots \tilde{\tau}_{c,r}$ as variables which satisfy the below expressions, for all $k \in \{1, \ldots, r\}$.

$$\tau_a = \tilde{\tau}_{a,k}, \qquad \tau_c = \tilde{\tau}_{c,k}, \qquad \text{vec}(D_k) = \tilde{\tau}_{a,k}^2 \text{vec}(\mathbf{K}_k^a) + \tilde{\tau}_{c,k}^2 \text{vec}(\mathbf{K}_k^c).$$

We now define the vector $\tilde{\tau} = [\tilde{\tau}_{a,1}, \tilde{\tau}_{c,1}, \ldots, \tilde{\tau}_{a,r}, \tilde{\tau}_{c,r}]'$. By the chain rule and definition of $\mathcal{C}$, it can be seen that:

$$\mathcal{C} = \frac{dv(D)}{d\tau} = \frac{\partial \tilde{\tau}}{\partial \tau} \frac{\partial v(D)}{\partial \tilde{\tau}}. \tag{A.14}$$

The first partial derivative in the product is given by:

$$\frac{\partial \tilde{\tau}}{\partial \tau} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & \ldots 0 \\ 0 & 1 & 0 & 1 & 0 & \ldots 1 \end{bmatrix} = \mathbb{1}_{(1,r)} \otimes I_2.$$

To evaluate the second derivative in the product, we first consider the partial derivative of $\text{vec}(D_k)$ with respect to $[\tilde{\tau}_{a,k}, \tilde{\tau}_{c,k}]'$ for arbitrary $k \in \{1, \ldots, r\}$. From the definitions of $\tilde{\tau}_{a,k}$ and $\tilde{\tau}_{c,k}$, it can be seen that:

$$\frac{\partial \text{vec}(D_k)}{\partial [\tilde{\tau}_{a,k}, \tilde{\tau}_{c,k}]'} = \begin{bmatrix} 2\tilde{\tau}_{a,k} \text{vec}(\mathbf{K}_k^a)' \\ 2\tilde{\tau}_{c,k} \text{vec}(\mathbf{K}_k^c)' \end{bmatrix}.$$

By similar reasoning it can be seen, for arbitrary $k_1, k_2 \in \{1, \ldots, r\}$, such that $k_1 \neq k_2$, that the below is true:

$$\frac{\partial \text{vec}(D_{k_1})}{\partial [\tilde{\tau}_{a,k_2}, \tilde{\tau}_{c,k_2}]'} = \mathbf{0}_{(2, q_{k_1}^2)},$$

where $\mathbf{0}_{(2,q_{k_1}^2)}$ is the $(2 \times q_{k_1}^2)$ dimensional matrix of zero elements. By combining the above expressions and noting the definitions of $v(D)$ and $\tilde{\tau}$, it can now be seen that

the derivative of $v(D)$ with respect to $\tilde{\tau}$ is given by the below.

$$\begin{bmatrix} \begin{bmatrix} 2\tilde{\tau}_{a,1}\text{vec}(\mathbf{K}_1^a)' \\ 2\tilde{\tau}_{c,1}\text{vec}(\mathbf{K}_1^c)' \end{bmatrix} & \mathbf{0}_{(2,q_2^2)} & \cdots & \mathbf{0}_{(2,q_r^2)} \\ \mathbf{0}_{(2,q_1^2)} & \begin{bmatrix} 2\tilde{\tau}_{a,2}\text{vec}(\mathbf{K}_2^a)' \\ 2\tilde{\tau}_{c,2}\text{vec}(\mathbf{K}_2^c)' \end{bmatrix} & \cdots & \mathbf{0}_{(2,q_r^2)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{(2,q_1^2)} & \mathbf{0}_{(2,q_2^2)} & \cdots & \begin{bmatrix} 2\tilde{\tau}_{a,r}\text{vec}(\mathbf{K}_r^a)' \\ 2\tilde{\tau}_{c,r}\text{vec}(\mathbf{K}_r^c)' \end{bmatrix} \end{bmatrix}$$

$$= \bigoplus_{k=1}^{r} \begin{bmatrix} 2\tilde{\tau}_{a,k}\text{vec}(\mathbf{K}_k^a)' \\ 2\tilde{\tau}_{c,k}\text{vec}(\mathbf{K}_k^c)' \end{bmatrix}.$$

By substituting the above partial derivative results into Equation (A.14), substituting $\tau_a = \tilde{\tau}_{a,k}$ and $\tau_c = \tilde{\tau}_{c,k}$ and rearranging, the result stated in Theorem A.10 can now be obtained. □

A.10.4 Computation and the ACE Model

A key aspect in which the ACE model differs from other LMM's considered in this thesis is that, in the ACE model, the second dimension of the random effects design matrix, q, is equal to the number of observations, n. Resultantly, the random effects design matrix, Z, which is given by the $(n \times n)$ identity matrix, has notably large dimensions. Due to the large size of Z, the methods of Section 3.3.3, which assume that the second dimension of Z is much smaller than n (i.e. $q << n$), cannot be applied to the ACE model in order to improve computation speed. In this section, we briefly outline how the FSFS algorithm of Section 3.3.1.4 may be implemented efficiently, in conjunction with the ReML adjustments described in Appendix A.5 and the constrained optimization methods of Section 3.3.4 to perform parameter estimation for the ACE model.

To begin, we consider the GLS updates, given by Equation (3.14), which are employed for updating the parameters β and σ^2. By defining the matrix, $\bar{D}_k$ by $\bar{D}_k = I_{q_k} + D_k$, and noting that the matrix Z is the $(n \times n)$ identity matrix, the below

matrix equality can be seen to hold for the ACE model:

$$X'V^{-1}X = \sum_{k=1}^{r}\sum_{j=1}^{l_k} X'_{(k,j)}\bar{D}_k^{-1}X_{(k,j)},$$

where $X_{(k,j)}$ represents the design matrix for the j^{th} family unit which possesses family structure of type k. Via well-known properties of the Kronecker product, the above equality can be seen to be equivalent to the following:

$$\text{vec}(X'V^{-1}X) = \sum_{k=1}^{r}\left(\sum_{j=1}^{l_k} X'_{(k,j)} \otimes X'_{(k,j)}\right)\text{vec}(\bar{D}_k^{-1}).$$

The above equality is noteworthy as the matrix expression which appears inside the large brackets on the right-hand side is fixed and does not depend on β, σ^2 or D. Further, the matrix expression inside the brackets is typically small in size and has dimensions $(p^2 \times q_k^2)$. As a result, this expression can be calculated and stored prior to the iterative procedure of the FSFS algorithm. This means that the above equality offers a quick and convenient method for repeated evaluation of the matrix, $X'V^{-1}X$, which is employed for many calculations throughout the FSFS algorithm. Similar logic can be applied to the evaluation of the vector, $X'V^{-1}Y$, and the scalar value, $Y'V^{-1}Y$. Using this approach, it can now be seen that the GLS estimators, (3.14), can be calculated using only $\{D_k\}_{k\in\{1,\ldots,r\}}$ and pre-calculated matrices which are small in dimension. Since the matrices $\{D_k\}_{k\in\{1,\ldots,r\}}$ are also typically small in dimension, this calculation is extremely quick and computationally efficient.

Next, we consider the matrix $F_{\text{vec}(D_k)}$ and the partial derivative vector of $l_R(\theta^f)$, taken with respect to $\text{vec}(D_k)$. Expressions for these quantities can be obtained using Equations (3.13) and (3.9), respectively, in combination with the adjustments described in Appendix A.5. Again noting that the random effects design matrix, Z, is equal to the identity matrix, the ReML equivalent of Equation (3.9) simplifies

substantially to the below expression:

$$\frac{\partial l_R(\theta^f)}{\partial \text{vec}(D_k)} = \frac{1}{2}\text{vec}\bigg(\bar{D}_k^{-1}\bigg(\frac{E_k E_k'}{\sigma^2} - l_k\bar{D}_k + \sum_{j=1}^{l_k} X_{(k,j)}(X'V^{-1}X)^{-1}X_{(k,j)}'\bigg)\bar{D}_k^{-1}\bigg),$$

where $E_k = \text{vec}_{q_k}(e_{(k)}')'$, vec_m is defined as the generalized vectorization operator described in Section 3.3.3 and $e_{(k)}$ is defined as the residual vector corresponding to subjects who belong to a family unit with family structure type k. Using Equation (3.13), the sub-matrix of F corresponding to $\text{vec}(D_k)$ can also be seen to be given by the following simplified expression.

$$F_{\text{vec}(D_k)} = \frac{1}{2}l_k(\bar{D}_k^{-1} \otimes \tilde{D}_k^{-1}).$$

The two expressions above can be evaluated computationally in an extremely efficient manner primarily for two reasons. The first of these is that the matrices involved in computation are typically small; D_k, E_k and $X_{(k,j)}$ have dimensions $(q_k \times q_k)$, $(q_k \times l_k q_k)$ and $(q_k \times p)$, respectively. The second reason is that, excluding the inversions of $\bar{D}_k$ and $X'V^{-1}X$, the only operations required to evaluate the above expressions are computationally quick to perform. These operations are: matrix multiplication, the reshape operation, matrix addition and the Kronecker product.

Lastly, we consider the evaluation of the score vector and Fisher Information matrix associated to $\tau = [\tau_a, \tau_c]' = \sigma_e^{-1}[\sigma_a, \sigma_c]'$, which shall be denoted as $\frac{dl(\theta^{ACE})}{d\tau}$ and $\mathcal{I}_\tau^{ACE}$, respectively. We define $\mathbf{K}_k = [\text{vec}(\mathbf{K}_k^a), \text{vec}(\mathbf{K}_k^c)]'$. By employing Equation (A.13) of Section A.7, and using the constraint matrix derived in Appendix A.10.3, the score vector and Fisher Information matrix associated to τ can be reduced to the following:

$$\mathcal{I}_\tau^{ACE} = (\tau\tau') \odot \bigg(\sum_{k=1}^{r} \mathbf{K}_k F_{\text{vec}(D_k)} \mathbf{K}_k'\bigg),$$
$$\frac{dl_R(\theta^{ACE})}{d\tau} = \tau \odot \bigg(\sum_{k=1}^{r} \mathbf{K}_k \frac{\partial l_R(\theta^f)}{\partial \text{vec}(D_k)}\bigg),$$

where $\odot$ represents the Hadamard (entry-wise) product. The results of Section 3.5.2.2 were produced by using the GLS updates for the parameters β and σ_e^2 and by em-

ploying the two expressions above to perform Fisher Scoring update steps for the parameter vector τ.

It should be noted that, in the ACE model, misspecification of variance components can result in a low-rank Fisher Information matrix. This can be seen by noting the positioning of the Hadamard product in the above Fisher Information matrix expression and by considering the case in which one of the elements of τ equals 0. If not accounted for appropriately, this issue may, in practice, result in failure of the algorithm to converge. A simple, practical solution to this issue is to execute the algorithm multiple times using different initial starting points; once with a starting point where τ_a is set to zero, once with τ_c set to zero and once with neither τ_a nor τ_c set to zero. The results of Section 3.5.2.2 were obtained using this approach.

Appendix B

The BLMM Toolbox

B.1 BLM: Parallelised Computing for Big Linear Models

In Section 4.3.1.2, a "product form"-based method for performing parameter estimation and inference for the Linear Model was referenced. This method served as the initial motivation for BLMM's approach to parallelised parameter estimation and inference for the Linear Mixed Model and is implemented as a sister-project to BLMM under the acronym BLM (Big Linear Models). In this appendix, we give a brief outline of the computational stages of BLM, highlighting the similarity between the approaches of BLM and BLMM.

Much like BLMM, BLM is designed to account for missingness observed as a result of mask-variability. As such, BLM assumes a model setup that is analogous to that described in Section 4.2.1 given by;

$$Y_v = M_v X\beta_v + \epsilon_v, \qquad \epsilon_v \sim N(0, \sigma_v^2 M_v)$$

where M_v is the $(n \times n)$-dimensional 'missingness' matrix described in Section 4.2.1 and the missing values in Y_v are encoded as zero. Mirroring the approach of Section

4.3.1.2, to reduce memory consumption, computation in BLM begins by evaluating the below product forms for each voxel in the analysis mask;

$$P_v = X_v' X_v, \;\; Q_v = X_v' Y_v, \;\; S_v = Y_v' Y_v,$$

where $X_v = M_v X$. Following computation of P_v, Q_v and S_v, the matrices X_v and Y_v are redundant and discarded from memory. By employing vectorized computation in a manner similar to that described in Section 4.3.1.3, BLM then evaluates the ordinary least squares estimators using the product forms as follows:

$$\hat{\beta}_v = P_v^{-1} Q_v, \qquad \hat{\sigma}_v^2 = S_v - 2Q_v' \hat{\beta}_v + \hat{\beta}_v' P_v \hat{\beta}_v$$

Finally, in a similar manner to that described in Section 4.3.1.4, BLM supports null-hypothesis testing via T- and F-statistics. Neglecting the voxel subscript v, T- and F-statistics are calculated using the product forms as follows:

$$T = \frac{L\hat{\beta}}{\sqrt{\hat{\sigma}^2 L P^{-1} L'}}, \qquad F = \frac{\hat{\beta}' L' (L P^{-1} L')^{-1} L \hat{\beta}}{\hat{\sigma}^2 \text{rank}(L)},$$

where L is a fixed and known contrast vector. For further information on the BLM toolbox, see:

https://github.com/TomMaullin/BLM

B.2 Computational Efficiency for Product Form Evaluation

In Section 4.3.1.2, pseudocode demonstrates how BLMM evaluates the product forms in a computationally efficient manner. For clarity and conciseness, however, several practical details were neglected from Algorithm 6. Therefore, in this section, to

complement the discussion of Section 4.3.1.2, we provide a brief overview of several practical implementation details which were neglected from Algorithm 6.

Firstly, we highlight that in Algorithm 6, several operations were conceptualized as involving the missingness matrix $M_v^{(b)}$ which has dimension $(\frac{n}{B} \times \frac{n}{B})$. While the use of the matrix $M_v^{(b)}$ in this situation is theoretically justified, construction of such a matrix is computationally infeasible due to the large amount of memory required in order to store $M_v^{(b)}$ for all voxels. Here we note that, in practice, construction of the matrix $M_v^{(b)}$ is unnecessary. Instead, multiplications such as $M_v^{(b)} X_v^{(b)}$ can be evaluated incrementally by replacing rows of $X_v^{(b)}$ with zeros in an appropriate fashion.

Secondly, as a result of between-voxel design commonality (c.f. Section 4.2.1), we highlight that typically P_v, R_v and U_v will take the same values for many voxels. As such, P_v, R_v, and U_v do not need to be computed and stored separately for every voxel. Although not noted in Algorithm 6, BLMM utilises this between-voxel commonality during the product form stage of computation to reduce storage costs by storing only the unique instances of P_v, R_v and U_v. As it is often the case that there are large contiguous regions of the analysis mask over which there is little missingness, storing P_v, R_v and U_v in this manner is expected to be much more memory efficient than storing them for each voxel individually.

Thirdly, in Algorithm 6, it appears there are many "for loops" which are being serially executed across voxels. Here we highlight that the operations inside these loops utilise only conceptually simplistic operations such as matrix multiplication and addition. Consequently, in practice, these loops can be realised in a quick, efficient parallelised manner by using vectorised computation (c.f. Section 4.3.1.3). It must also be noted, however, that in cases when R_v and U_v are particularly large, on each node, BLMM may not be able to evaluate R_v and U_v for all voxels concurrently due to the high RAM usage that would be required. In such cases, BLMM will split $\{X_v^{(b)}\}$ and $\{Z_v^{(b)}\}$ voxel-wise into further batches and, on each node, consider each batch in serial.

Finally, it is also noted that for models which contain only one random factor, U_v becomes block diagonal. In such situations, substantial gains in computation time and

memory consumption may be made via careful consideration of the structure of U_v. The benefits of this approach are most extreme when the model design contains only one random factor and one random effect. In this case, U_v is diagonal, and BLMM needs only to evaluate and store the diagonal elements of U_v. Such considerations are discussed in greater detail in the following section, Section B.3.

B.3 Computational Efficiency for Single-Factor Designs

This section outlines two common scenarios in which structural features of the random effects covariance and design matrices, D and Z, allow for further improvements in terms of computational efficiency in the BLMM pipeline. To illustrate how BLMM exploits the structure of D and Z to achieve improved computational performance in each of these scenarios, we shall use Equation (4.4) as an informative example throughout this section. For the remaining expressions listed in Sections 4.3.1.1-4.3.1.4, we note that similar adjustments are possible but, for brevity, will not be listed here.

The first scenario in which BLMM accounts for the structure of Z and D computationally is the setting in which the experimental design contains a single random factor which groups multiple random effects (i.e. $r = 1$ and $q_1 > 1$). In such instances, U (which is equal to $Z'Z$) and D are both block diagonal, with D consisting of only one block, D_1, which is of dimension $(q_1 \times q_1)$, repeated l_1 times along it's diagonal (i.e. $D = I_{l_1} \otimes D_1$). By noting the structure of D and U, and neglecting the subscript s for iteration number, Equation (4.4) may be rewritten in this setting as follows:

$$
\begin{aligned}
X'V^{-1}X &= P - \sum_{j=1}^{l_1} R_{(j)} D_1 (I + D_1 U_{(j)})^{-1} R'_{(j)} \\
Z'V^{-1}Z &= \bigoplus_{j=1}^{l_1} \left(U_{(j)} - U_{(j)} D_1 (I_{q_1} + D_1 U_{(j)})^{-1} U_{(j)} \right) \\
Z'V^{-1}e &= T' - R'\beta - \bigoplus_{j=1}^{l_1} \left(U_{(j)} D_1 (I_{q_1} + D_1 U_{(j)})^{-1} \right) (T' - R'\beta)
\end{aligned}
\tag{B.1}
$$

where $U_{(j)}$ and $R_{(j)}$ are shorthand for $Z'_{(1,j)}Z_{(1,j)}$ and $X'Z_{(1,j)}$ respectively, and $\oplus$ is the direct sum. Despite appearing notationally convoluted, the above expressions are much more convenient than Equation (4.4) for computational purposes. This is due to the notably smaller dimensions of the matrices involved in the right hand side of Equation (B.1) in comparison to those employed in Equation (4.4).

For example, consider the dimensions of the matrices which are inverted in Equation (B.1), $\{(I_{q_1} + D_1 U_{(j)})\}_{j \in \{1,\ldots,l_1\}}$, and the matrix which is inverted in Equation (4.4), $(I_q + DU)$. The former matrices have dimension $(q_1 \times q_1)$ whilst the latter has dimension $(q \times q)$. To illustrate the significance of this observation, consider a setting in which 2 random effects are modelled across 100 levels (subjects or time-points, for example). In such a scenario, $q = 200$ whilst $q_1 = 2$ and, by using the above reformulation of Equation (4.4), the inversion of a matrix of dimension (200×200) can be substituted for the much faster operation of inverting 100 matrices of dimension (2×2). When applied in the mass-univariate setting, such improvements in efficiency can reduce computation time from hours to seconds.

The second scenario to which computation in the BLMM pipeline has been tailored is the setting in which the experimental design contains a single random factor that groups a single random effect (i.e. $r = 1$ and $q_1 = 1$). In this instance, the matrices D and U are diagonal, with D containing a single scalar value, d, repeated l_1 times along the diagonal. Consequently, by denoting u as the vector composed of the diagonal

elements of U, Equation (4.4) can be rewritten in this setting as follows:

$$\begin{aligned} X'V^{-1}X &= P - d(RR' \odot \text{diag}(1 \oslash (1 + du)) \\ Z'V^{-1}Z &= \text{diag}(u \odot (1 - du) \oslash (1 + du)) \\ Z'V^{-1}e &= (1 - du) \oslash (1 + du) \odot (T' - R'\beta) \end{aligned} \tag{B.2}$$

where $\odot$ and $\oslash$ represent Hadamard (element-wise) multiplication and division, respectively, and diag is the unique function which maps an arbitrary $(m \times 1)$ vector, x, to an $(m \times m)$ diagonal matrix with the elements of x listed along its diagonal. As in the previous scenario, in which the use of Equation (B.1) greatly reduced the time complexity of computation, the use of Equation (B.2) can also provide strong improvements in terms of computational speed. A notable example of this can be seen by contrasting the time complexity of the evaluation of $Z'V^{-1}Z$ obtained using the right-hand side of Equation (B.2) against that obtained using Equation (4.4). The time complexity of the former is $O(q)$, whereas the time complexity of the latter is greater than $O(q^2)$. As a consequence, in this scenario, the matrix $Z'V^{-1}Z$, which is utilised many times throughout the BLMM pipeline, may be evaluated in a much faster manner via the use of Equation (B.2) than by use of Equation (4.4).

In each of the above scenarios, it must further be noted that, by utilising the internal structure of Z, memory consumption can also be significantly reduced. For example, when storing the product form $U = Z'Z$, only the non-zero entries of U must be recorded. This reduces the storage cost associated to U from a complexity of $O(q^2)$ to; $O(q_1 q)$ in the scenario in which U is block-diagonal, and $O(q)$ in the scenario in which U is diagonal. We note here that sparse matrix methods could theoretically also be employed to obtain further improvements in computational time and storage. However, due to the present lack of vectorised support available for sparse matrix operations, such improvements are currently infeasible in practice for mass-univariate analyses. For this reason, BLMM currently does not utilise sparse matrix methodology.

Appendix C

Spatial Confidence Regions Theory

C.1 Discussion and Illustration of Assumptions

C.1.1 Assumption 5.1: The Central Limit Theorem

Below is a brief example demonstrating the usage of the notation introduced by Assumption 5.1. We note that, in general, the theory presented in Section 5.2.3 requires no assumption of Gaussianity for the variables $\{G^i(s)\}_{i\in\mathcal{M},s\in S}$.

Example C.1.1. Suppose that, for some spatial location s and study condition $i \in \mathcal{M}$, $\hat{\mu}^i_n(s)$ had been calculated as the mean of n independent and identically distributed (i.i.d.) Gaussian observations $\{X^i_1(s), ..., X^i_n(s)\}$. In this instance, Assumption 5.1 would be satisfied by the classical CLT with $\mu^i(s) = \mathbb{E}[X^i_1(s)], \sigma^i(s) = \sqrt{\text{Var}(X^i_1(s))}$ and $\tau_n = n^{-\frac{1}{2}}$. In this example, $G^i(s)$ would be a $N(0, 1)$ random variable.

C.1.2 Assumption 5.2: Continuity

Assumption 5.2 ensures that throughout the proofs of Section C.2, the limits taken across space are well defined. However, this assumption alone does not provide any

guarantee that the sets $\{\mathcal{A}_c^i\}_{i\in\mathcal{M}}$ will resemble the well-defined 'blobs' which may be expected.

In fact, under Assumption 5.2, without further assumptions, it is still possible for the excursion sets $\{\mathcal{A}_c^i\}_{i\in\mathcal{M}}$ to appear extremely erratic, possessing jagged edges and displaying fractal-like behaviour. For instance, a well-known example of a random field that satisfies Assumption 5.2 but which is not well-behaved in the sense we desire is the highly fractal Wiener process (also known as Brownian motion, c.f. Adler (1981)). To restrict our attention to situations that are more common to imaging studies, the additional conditions provided by Assumption 5.3 are required.

C.1.3 Assumption 5.3

Each of the three statements provided by Assumption 5.3 serves a separate purpose, handling different use cases that the others do not. Statement (a) handles cases in which the excursion set boundaries $\{\partial\mathcal{A}_c^i\}_{i\in\mathcal{M}}$ coincide with ∂S, Statement (b) ensures that the notion of gradient is well defined in a neighbourhood of $\partial\mathcal{F}_c$ and Statement (c) handles unusual, potentially unintuitive, fractal-like behaviour which is not ruled out by the first two statements. Each of these claims is discussed further in the following subsections.

C.1.3.1 The Ball Assumption

Statement (a), which we shall refer to as 'the ball assumption', ensures that $\partial\mathcal{F}_c$ does not coincide exactly with ∂S by explicitly requiring well-defined regions to exist on either side of $\partial\mathcal{F}_c$. To ensure this criterion is met, the ball assumption defines the set $\mathcal{J}_c^\alpha$ which, broadly speaking, may be thought of as the set which we would 'naturally expect' to lie on the outside of $\partial^\alpha\mathcal{F}_c$. To understand the definition of $\mathcal{J}_c^\alpha$, consider a brief example in which $M = 2$. In this setting, if $s \in \partial^{\{1\}}\mathcal{F}_c$, then $g^1(s) = 0$ and $g^2(s) > 0$. In a neighbourhood of s, we wish to assume that $g^1 > 0$ on one 'side' of $\partial\mathcal{F}_c$, and $g^1 < 0$ on the other. As $g^2(s) > 0$, it follows that $g^2 > 0$ everywhere in a small enough neighbourhood of s and, therefore, such a neighbourhood intersects

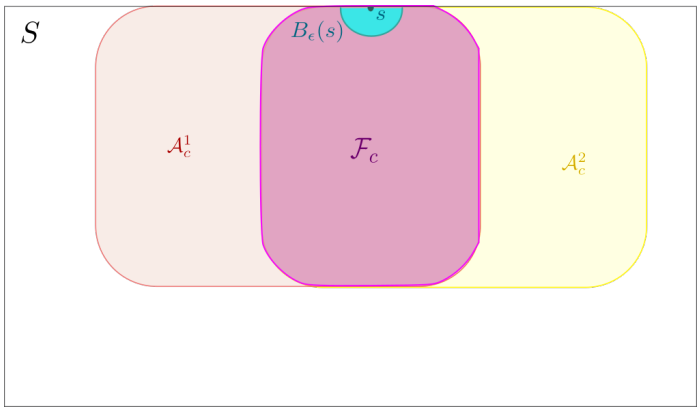

Figure C.1: An example in which Assumption 5.3 Statements (b) and (c) are satisfied but the ball assumption is not. In this example, $\partial\mathcal{F}_c$ coincides perfectly with ∂S, and as such, at the point $s \in \partial^{\{1,2\}}\mathcal{F}_c$, it is possible to draw a ball, $B_\epsilon(s)$, which does not intersect $\mathcal{J}_c^{\{1,2\}}$.

$(\mathcal{F}_c)^\circ$ (where $g^1 > 0$ and $g^2 > 0$), and $\mathcal{J}_c^{\{1\}}$ (where $g^1 < 0$ and $g^2 > 0$). These are the sets stated by the ball assumption.

The ball assumption is necessary to the proofs of Section C.2 as we shall often need to define sequences of points, which lie either inside or outside $\mathcal{F}_c$, that tend to $\partial\mathcal{F}_c$. We note that the ball assumption is naturally satisfied in many practical applications which involve set intersections (or unions) (c.f. Fig. 5.2 of Chapter 5) and rarely impacts the applicability of our results. Similar assumptions may also be found in precursors to this work (e.g. Assumption 2.1 in Chapter 2, c.f. Sommerfeld et al. (2018)). There are some use cases that the ball assumption removes from consideration which the other statements provided by Assumption 5.3 do not. These are the cases in which $\partial\mathcal{F}_c$ perfectly aligns with ∂S, (see Fig. C.1). We emphasize that this assumption does not forbid $\partial\mathcal{F}_c$ from touching ∂S in general, but instead restricts the degree to which the two sets may coincide.

C.1.3.2 The Gradient Assumption

Statement (b), which we shall refer to as 'the gradient assumption', ensures that, in a neighbourhood of $\partial\mathcal{F}_c$, the functions $\{g^i\}_{i\in\mathcal{M}}$ are C^1. By applying implicit function theorem to the functions $\{g^i\}_{i\in\mathcal{M}}$, it can be seen that this statement ensures that

$\{\partial\mathcal{A}_c^i\}_{i\in\mathcal{M}}$ are well-defined C^1 curves. As it is also required that, along the boundary $\partial\mathcal{F}_c$, the gradients are non-zero and finite, this assumption also ensures that the functions $\{g^i\}_{i\in\mathcal{M}}$ neither flatten out nor become directly vertical, in a neighbourhood of $\partial\mathcal{F}_c$. This observation is particularly relevant in the proofs of Section C.2 when we consider spatial regions, which are defined using the functions $\{g^i\}_{i\in\mathcal{M}}$, that 'shrink' to the boundary segments $\{\partial^\alpha\mathcal{F}_c\}_{\alpha\in\mathcal{P}^+(\mathcal{M})}$. Without bounding the rate of change of the functions $\{g^i\}_{i\in\mathcal{M}}$ in this manner, the steps in the proof which involve such 'shrinking' spatial regions would not be possible.

C.1.3.3 The Path-Connectedness Assumption

Statement (c), which we shall refer to as 'the path-connectedness assumption', ensures that near the boundary of $\mathcal{F}_c$, the boundary sets $\{\partial(\cap_{i\in\alpha}\mathcal{A}_c^i)\}_{\alpha\in\mathcal{P}^+(\mathcal{M})}$ are not self-intersecting and do not exhibit fractal-like behaviour. Most notably, as $\partial\mathcal{F}_c$ is a set of the form $\partial(\cap_{i\in\alpha}\mathcal{A}_c^i)$ (with $\alpha=\mathcal{M}$), it follows that this assumption prevents $\partial\mathcal{F}_c$ from crossing itself or exhibiting fractal-like properties.

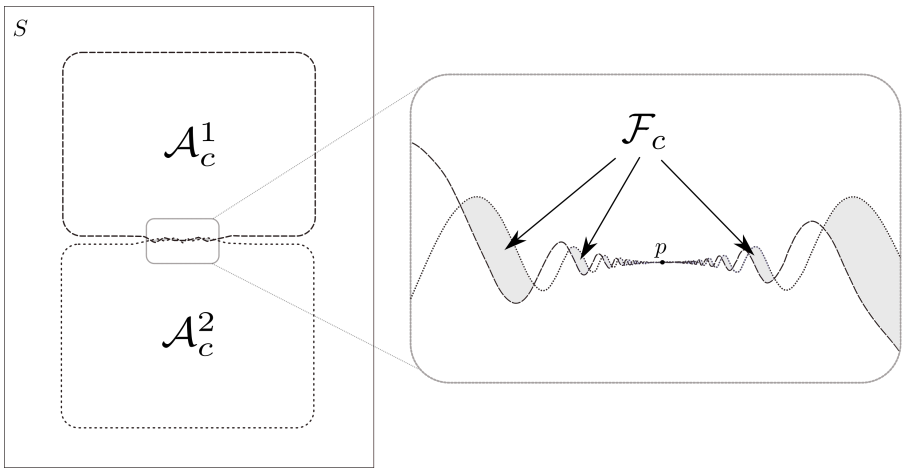

Figure C.2: An example in which $N=2$ and Assumption 5.3 Statements (a) and (b) are satisfied but the path-connectedness assumption is not. In the neighbourhood of p shown on the right, $\partial\mathcal{A}_c^1$ and $\partial\mathcal{A}_c^2$ resemble the functions $y=x^2\sin(\frac{\pi}{2}-\frac{\pi}{x})$ and $y=x^2\sin(\frac{\pi}{x})$ respectively (Note that both of these functions are C^1 and therefore consistent with Assumption 5.3 Statement (b)). However, the path-connectedness assumption does not hold at the point p and, as a result, $(\mathcal{F}_c)^\circ$ possesses the undesirable property of being a fractal.

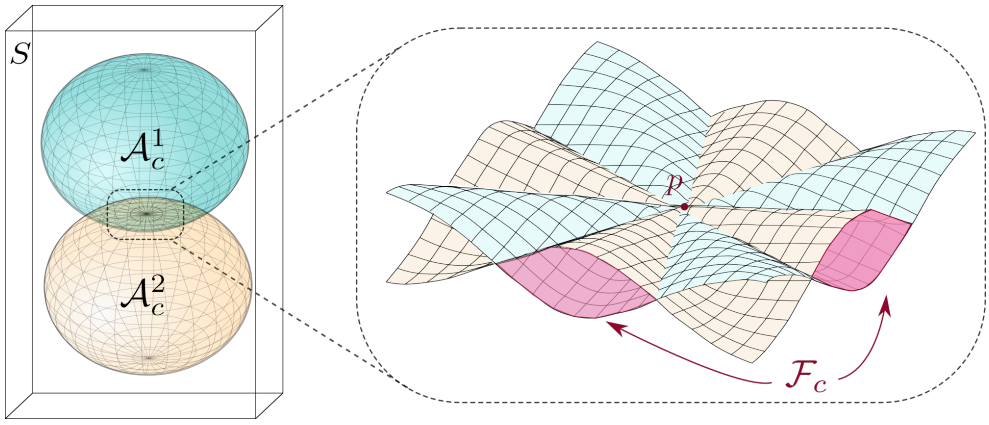

Figure C.3: An example in which $N = 3$ and Assumption 5.3 Statements (a) and (b) are satisfied but the path-connectedness assumption is not. Here, the path-connectedness assumption does not hold at the point p, but (allowing artistic license) it may be assumed that all derivatives are continuous at p and meet the requirements of Assumption 5.3 Statement (b). In this instance, $(\mathcal{F}_c)^\circ$ possesses the undesirable property of consisting of multiple disconnected components (the conic regions highlighted in red) in all neighbourhoods of the point p. The theory presented in Chapter 5 is not equipped to handle topologies of this form.

Assumption 5.3 Statement (c) is necessary as it requires 'path-connectedness'. This requirement ensures that, for arbitrary $p \in \partial\mathcal{F}_c$, all boundaries of the form $\partial(\cap_{i\in\alpha}\mathcal{A}_c^i)$ which contain p (including $\partial\mathcal{F}_c$) partition a neighbourhood of p into exactly two components (e.g. one component inside $(\cap_{i\in\alpha}\mathcal{A}_c^i)^\circ$ and one component inside $(\cap_{i\in\alpha}\mathcal{A}_c^i)^c$). Scenarios which satisfy Assumption 5.3 Statements (a) and (b) but violate the path-connectedness assumption are often difficult to visualise and rarely of practical interest. However, neglecting the path-connectedness assumption from consideration can have significant consequences. Fig. C.2 provides an example in which both Assumption 5.3 Statements (a) and (b) are satisfied, but $\mathcal{F}_c$ is a fractal. Similarly, Fig. C.3 provides an example, this time in 3 dimensions, in which the set $(\mathcal{F}_c)^\circ$ is not path-connected everywhere but instead consists of four conic regions whose apex meet at a single point $p \in \partial\mathcal{F}_c$.

C.2 Proof

In this section, we provide a full proof of Theorem 5.1. To do so, we shall prove each of the below statements in turn:

$$\liminf_{n\to\infty} \mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-] \geq \mathbb{P}\left[\max_{\alpha\in\mathcal{P}^+(\mathcal{M})} \left(\sup_{s\in\partial^\alpha \mathcal{F}_c} \left|\min_{i\in\alpha}(G^i(s))\right|\right) \leq a\right], \tag{C.1}$$

$$\limsup_{n\to\infty} \mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-] \leq \mathbb{P}\left[\max_{\alpha\in\mathcal{P}^+(\mathcal{M})} \left(\sup_{s\in\partial^\alpha \mathcal{F}_c} \left|\min_{i\in\alpha}(G^i(s))\right|\right) \leq a\right]. \tag{C.2}$$

It is clear that, when taken together, Equations (C.1) and (C.2) imply the result of Theorem 5.1. In Sections C.2.1 and C.2.2 we shall outline the proofs of Equations (C.1) and (C.2), respectively. Both Section C.2.1 and Section C.2.2 describe the 'broad strokes' of the proof, whilst the 'heavy lifting' is relegated to the theorems listed in Sections C.2.4-C.2.11, with supporting lemmas provided in Section C.2.3.

Before proceeding, it must be noted that the excursion sets defined in Section 5.1 may be reformulated in terms of the functions $\{g^i\}_{i\in\mathcal{M}}$ and $\{\hat{g}_n^i\}_{i\in\mathcal{M}}$ (c.f. Section 5.2.2) as:

$$\mathcal{A}_c^i = \{s \in S : g^i(s) \geq 0\} \quad \text{and} \quad \hat{\mathcal{A}}_c^i = \{s \in S : \hat{g}_n^i(s) \geq 0\}.$$

In a similar manner, the sets $\mathcal{F}_c$ and $\hat{\mathcal{F}}_c$ may be reformulated as:

$$\mathcal{F}_c = \{s \in S : \min_{i\in\mathcal{M}} g^i(s) \geq 0\} \quad \text{and} \quad \hat{\mathcal{F}}_c = \{s \in S : \min_{i\in\mathcal{M}} \hat{g}_n^i(s) \geq 0\},$$

respectively. By noting the above, the notations μ^i, $\hat{\mu}_n^i$ and σ^i may dispensed with and the following proof may be given solely in terms of the functions $\{g^i\}_{i\in\mathcal{M}}$ and $\{\hat{g}_n^i\}_{i\in\mathcal{M}}$.

C.2.1 Part I: Proof of Equation (C.1)

In this section, we shall prove Equation (C.1). To do so, we begin by defining $\{\eta_n\}_{n\in\mathbb{N}}$ and $\{K_n\}_{n\in\mathbb{N}}$ as arbitrary sequences of positive integers which satisfy $\eta_n \to 0, \tau_n^{-1}\eta_n \to \infty, K_n \to \infty, K_n^M \eta_n \to 0$ and $K_n \tau_n^{-1}\eta_n \to \infty$ as $n \to \infty$. Lemma C.1 demonstrates that it is always possible to find such sequences. Next, we define the inflated boundary for $\mathcal{F}_c$ as:

$$\mathcal{F}_c^n = \{s \in S : |\min_{i\in\mathcal{M}}(g^i(s))| \leq \eta_n\}.$$

And similarly, for $k \in \mathcal{M}$, we define the β-inflated boundary for $\mathcal{A}_c^k$ as:

$$\mathcal{A}_c^{\beta,k} = \{s \in S : |g^k(s)| \leq \beta\},$$

for arbitrary $\beta \in \mathbb{R}^+$. Intuitively, $\mathcal{F}_c^n$ may be expected to form an inflated 'band' around $\partial\mathcal{F}_c$ which shrinks to $\partial\mathcal{F}_c$ as $n \to \infty$. Similarly, the η_n-inflated boundary for $\partial\mathcal{A}_c^k$, $\mathcal{A}_c^{\eta_n,k}$, forms an inflated 'band' around $\partial\mathcal{A}_c^k$ and shrinks to $\partial\mathcal{A}_c^k$ as $n \to \infty$. Using the above definitions, our aim is now to build similar 'inflated' boundaries for the sets $\{\partial^\alpha\mathcal{F}_c\}_{\alpha\in\mathcal{P}^+(\mathcal{M})}$.

For each $\alpha \in \mathcal{P}^+(\mathcal{M})$, we shall meet this aim by defining two 'components' of the inflated boundary for $\partial^\alpha\mathcal{F}_c$; an 'inner' component (i.e. the component which lies inside the set $\mathcal{F}_c$) and an 'outer' component (e.g. the component which lies outside the set $\mathcal{F}_c$). For all $\alpha \in \mathcal{P}^+(\mathcal{M})$, the inner component, $\mathcal{I}_c^{n,\alpha}$, is defined as;

$$\mathcal{I}_c^{n,\alpha} = \mathcal{F}_c \cap \left(\bigcap_{i\in\alpha} \mathcal{A}_c^{(K_n)^{|\alpha|-1}\eta_n,i}\right) \Bigg\backslash \left(\bigcup_{\substack{\alpha'\in\mathcal{P}^+(\mathcal{M}) \\ \text{Ord}(\alpha')>\text{Ord}(\alpha)}} \mathcal{I}_c^{n,\alpha'}\right), \tag{C.3}$$

where 'Ord' represents the index of a set in $\mathcal{P}^+(\mathcal{M})$ when $\mathcal{P}^+(\mathcal{M})$ is listed in lexicographic order. For example, if $M = 3$ then Ord maps the sets $(\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, ...)$, to the integers $(1, 2, 3, 4, 5, ...)$.

The outer component, which will be denoted $\mathcal{O}_c^{n,\alpha}$, is defined for $\alpha \in \mathcal{P}^+(\mathcal{M})$ as:

$$\mathcal{O}_c^{n,\alpha} = \mathcal{F}_c^n \cap \left(\bigcap_{i \in \alpha} (\mathcal{A}_c^i)^c \right) \cap \left(\bigcap_{j \in \mathcal{M} \setminus \alpha} \mathcal{A}_c^j \right), \tag{C.4}$$

where, if $\mathcal{M} \setminus \alpha$ is empty, the last term is defined to be equal to S. To aid understanding, in Section C.2.1.1, we provide further discussion of the definitions of $\mathcal{I}_c^{n,\alpha}$ and $\mathcal{O}_c^{n,\alpha}$. For now, it suffices to state that $\mathcal{I}_c^{n,\alpha}$ and $\mathcal{O}_c^{n,\alpha}$ are sets, located on the inside and outside of $\mathcal{F}_c$ respectively, each of which increasingly resembles $\partial^\alpha \mathcal{F}_c$ as $n \to \infty$. Illustrations of both sets are provided in Section C.2.1.1 (c.f. Figs. C.4 and C.5).

Next, we define the sets $\mathcal{O}_c^{n,\emptyset}$ and $\mathcal{I}_c^{n,\emptyset}$ as the regions outside and inside $\mathcal{F}_c$, which are not covered by the sets $\{\mathcal{O}_c^{n,\alpha}\}_{\alpha \in \mathcal{P}^+(\mathcal{M})}$ and $\{\mathcal{I}_c^{n,\alpha}\}_{\alpha \in \mathcal{P}^+(\mathcal{M})}$, respectively. I.e:

$$\mathcal{O}_c^{n,\emptyset} = (\mathcal{F}_c)^c \Bigg\backslash \left(\bigcup_{\alpha \in \mathcal{P}^+(\mathcal{M})} \mathcal{O}_c^{n,\alpha} \right), \qquad \mathcal{I}_c^{n,\emptyset} = \mathcal{F}_c \Bigg\backslash \left(\bigcup_{\alpha \in \mathcal{P}^+(\mathcal{M})} \mathcal{I}_c^{n,\alpha} \right).$$

It now follows that the sets $\{\mathcal{O}_c^{n,\alpha}\}_{\alpha \in \mathcal{P}(\mathcal{M})}$ and $\{\mathcal{I}_c^{n,\alpha}\}_{\alpha \in \mathcal{P}(\mathcal{M})}$ cover $(\mathcal{F}_c)^c$ and $\mathcal{F}_c$, respectively. We now define the 'inflated' boundary for $\partial^\alpha \mathcal{F}_c$ as $\mathcal{F}_c^{n,\alpha} = \mathcal{O}_c^{n,\alpha} \cup \mathcal{I}_c^{n,\alpha}$.

Finally, we define the following events, for $\alpha \in \mathcal{P}^+(\mathcal{M})$:

$$\mathcal{E}_{\mathcal{I}}^{\alpha,n} = \left\{ \sup_{s \in \mathcal{I}_c^{n,\alpha}} \left| \tau_n^{-1} \min \left(\min_{i \in \alpha} (\hat{g}_n^i(s) - g^i(s)), \min_{j \in \mathcal{M} \setminus \alpha} (\hat{g}_n^j(s)) \right) \right| \leq a \right\},$$

$$\mathcal{E}_{\mathcal{O}}^{\alpha,n} = \left\{ \sup_{s \in \mathcal{O}_c^{n,\alpha}} \left| \tau_n^{-1} \min_{i \in \alpha} (\hat{g}_n^i(s) - g^i(s)) \right| \leq a \right\},$$

and, for the empty set, $\emptyset$;

$$\mathcal{E}_{\mathcal{I}}^{\emptyset,n} = \left\{ \sup_{s \in \mathcal{I}_c^{n,\emptyset}} \max_{i \in \mathcal{M}} \left(\tau_n^{-1} |\hat{g}_n^i(s) - g^i(s)| \right) \leq a + \tau_n^{-1} \eta_n \right\},$$

$$\mathcal{E}_{\mathcal{O}}^{\emptyset,n} = \left\{ \sup_{s \in \mathcal{O}_c^{n,\emptyset}} \max_{i \in \mathcal{M}} \left(\tau_n^{-1} |\hat{g}_n^i(s) - g^i(s)| \right) \leq a + \tau_n^{-1} \eta_n \right\}.$$

In the following text, if $\mathcal{M} \setminus \alpha$ is empty, it will be assumed that any minimum function

taken over $\mathcal{M} \setminus \alpha$ is defined as $+\infty$ (this is the equivalent of 'removing' such terms from the proof entirely when $\alpha = \mathcal{M}$).

We now are in a position to begin the proof. The first step of the proof is to show that if the events $\mathcal{E}_{\mathcal{I}}^{\alpha,n}$ and $\mathcal{E}_{\mathcal{O}}^{\alpha,n}$ occur for all $\alpha \in \mathcal{P}(\mathcal{M})$, then the inclusion $\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-$ holds. This implication is proven by Theorem C.1 in Section C.2.4. From this result and by using De Morgan's law, it can now be seen that:

$$\mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-] \geq \mathbb{P}\left[\bigcap_{\alpha \in \mathcal{P}^+(\mathcal{M})} \{\mathcal{E}_{\mathcal{I}}^{\alpha,n} \wedge \mathcal{E}_{\mathcal{O}}^{\alpha,n}\}\right] + \mathbb{P}[\mathcal{E}_{\mathcal{I}}^{\emptyset,n} \wedge \mathcal{E}_{\mathcal{O}}^{\emptyset,n}] - 1.$$

Next, we apply $\liminf_{n\to\infty}$ to both sides of the above expression. By the definitions of $\mathcal{E}_{\mathcal{I}}^{\emptyset,n}$ and $\mathcal{E}_{\mathcal{O}}^{\emptyset,n}$ it can be seen that the second term on the right hand side becomes:

$$\mathbb{P}[\mathcal{E}_{\mathcal{I}}^{\emptyset,n} \wedge \mathcal{E}_{\mathcal{O}}^{\emptyset,n}] = \mathbb{P}\left[\sup_{s \in \mathcal{F}_c^{n,\emptyset}} \max_{i \in \mathcal{M}} \left(\tau_n^{-1} |\hat{g}_n^i(s) - g^i(s)|\right) \leq a + \tau_n^{-1}\eta_n\right] \xrightarrow{n\to\infty} 1.$$

The above convergence follows from the fact that $\tau_n^{-1}\eta_n \to \infty$ as $n \to \infty$, whilst the expression involving the suprema and maxima converges in distribution to a well-defined random variable by Assumption 5.1. It can now be seen that:

$$\liminf_{n\to\infty} \mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-] \geq \liminf_{n\to\infty} \mathbb{P}\left[\bigcap_{\alpha \in \mathcal{P}^+(\mathcal{M})} \{\mathcal{E}_{\mathcal{I}}^{\alpha,n} \wedge \mathcal{E}_{\mathcal{O}}^{\alpha,n}\}\right]. \tag{C.5}$$

We now define the following shorthand:

$$A_\alpha^{n,\mathcal{I}} = \sup_{s \in \mathcal{I}_c^{n,\alpha}} \left|\tau_n^{-1} \min\left(\min_{i \in \alpha}(\hat{g}_n^i(s) - g^i(s)), \min_{j \in \mathcal{M}\setminus\alpha}(\hat{g}_n^j(s))\right)\right|,$$

$$B_\alpha^{n,\mathcal{I}} = \sup_{s \in \mathcal{I}_c^{n,\alpha}} \left|\tau_n^{-1} \min_{i \in \alpha}(\hat{g}_n^i(s) - g^i(s))\right|, \quad B_\alpha^{n,\mathcal{O}} = A_\alpha^{n,\mathcal{O}} = \sup_{s \in \mathcal{O}_c^{n,\alpha}} \left|\tau_n^{-1} \min_{i \in \alpha}(\hat{g}_n^i(s) - g^i(s))\right|.$$

And define the following four quantities:

$$A_\alpha^n = \max(A_\alpha^{n,\mathcal{I}}, A_\alpha^{n,\mathcal{O}}), \qquad B_\alpha^n = \max(B_\alpha^{n,\mathcal{I}}, B_\alpha^{n,\mathcal{O}}),$$

$$C_\alpha^n = \sup_{s \in \partial^\alpha \mathcal{F}_c} \left| \tau_n^{-1} \min_{i \in \alpha} (\hat{g}_n^i(s) - g^i(s)) \right|, \qquad D_\alpha = \sup_{s \in \partial^\alpha \mathcal{F}_c} |\min_{i \in \alpha}(G^i(s))|.$$

Using theorems given in Sections C.2.5 and C.2.6, alongside Assumption 5.1, we now note that the below relations hold:

$$\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (A_\alpha^n) - \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (B_\alpha^n) \xrightarrow{p} 0, \qquad \text{(By the result of Theorem C.2)}$$

$$\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (B_\alpha^n) - \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (C_\alpha^n) \xrightarrow{p} 0, \qquad \text{(By the result of Theorem C.3)}$$

$$\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (C_\alpha^n) - \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (D_\alpha) \xrightarrow{d} 0, \qquad \text{(By Assumption 5.1)}$$

where $\xrightarrow{p}$ and $\xrightarrow{d}$ represent convergence in probability and distribution, respectively. By combining the above results, it can now be seen that:

$$\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (A_\alpha^n) \xrightarrow{d} \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (D_\alpha).$$

It now follows that:

$$\liminf_{n \to \infty} \mathbb{P}\left[\bigcap_{\alpha \in \mathcal{P}^+(\mathcal{M})} \{\mathcal{E}_{\mathcal{I}}^{\alpha,n} \wedge \mathcal{E}_{\mathcal{O}}^{\alpha,n}\} \right] = \liminf_{n \to \infty} \mathbb{P}\left[\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (A_\alpha^n) \leq a \right]$$

$$= \mathbb{P}\left[\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} (D_\alpha) \leq a \right] = \mathbb{P}\left[\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \partial^\alpha \mathcal{F}_c} |\min_{i \in \alpha}(G^i(s))| \right) \leq a \right],$$

where the first equality follows from the definition of A_α^n, the second follows from the previous argument, and the third follows from the definition of D_α. By combining the above with Equation (C.5), it can now be seen that Statement (C.1) holds, as desired.

C.2.1.1 Intuition for $\mathcal{I}_c^{n,\alpha}$ and $\mathcal{O}_c^{n,\alpha}$

Here, we provide further discussion of the definitions of the sets $\{\mathcal{I}_c^{n,\alpha}\}_{\alpha \in \mathcal{P}^+(\mathcal{M})}$ and $\{\mathcal{O}_c^{n,\alpha}\}_{\alpha \in \mathcal{P}^+(\mathcal{M})}$. Illustrations of $\{\mathcal{I}_c^{n,\alpha}\}_{\alpha \in \mathcal{P}^+(\mathcal{M})}$ and $\{\mathcal{O}_c^{n,\alpha}\}_{\alpha \in \mathcal{P}^+(\mathcal{M})}$ are provided by Figs. C.4 and C.5, respectively, with a corresponding illustration of $\{\partial^\alpha \mathcal{F}_c\}_{\alpha \in \mathcal{P}^+(\mathcal{M})}$

given by Fig. C.6.

To begin, we fix $\alpha \in \mathcal{P}^+(\mathcal{M})$ and describe why each of the terms appearing in the definition of $\mathcal{I}_c^{n,\alpha}$, Equation (C.3), is necessary. The first term in Equation (C.3) is $\mathcal{F}_c$, whose inclusion guarantees that $\mathcal{I}_c^{n,\alpha}$ lies inside $\mathcal{F}_c$ (thus living up to the name "inner"). The second term is $\cap_{i\in\alpha}\mathcal{A}_c^{(K_n)^{|\alpha|-1}\eta_n,i}$. To understand the role of this term note that, by construction, $(K_n)^{|\alpha|-1}\eta_n \to 0$ as $n \to \infty$. From the definition of the β-inflated boundary for $\mathcal{A}_c^i$, it follows that, for all $i \in \mathcal{M}$, $\mathcal{A}_c^{(K_n)^{|\alpha|-1}\eta_n,i}$ must 'shrink' towards $\partial\mathcal{A}_c^i$ as $n \to \infty$. By consequence, the set $\mathcal{F}_c \cap (\cap_{i\in\alpha}\mathcal{A}_c^{(K_n)^{|\alpha|-1}\eta_n,i})$ is expected to 'shrink'[1] to $\mathcal{F}_c \cap (\cap_{i\in\alpha}\partial\mathcal{A}_c^i) = \cup_{\alpha'\supseteq\alpha}\partial^{\alpha'}\mathcal{F}_c$ as $n \to \infty$.

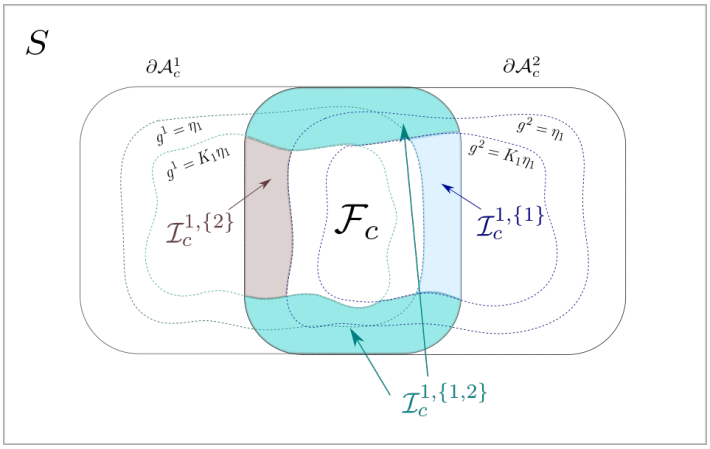

Figure C.4: Illustration of the inner inflated sets in a setting in which $n = 1$, $M = 2$ and the excursion sets $\mathcal{A}_c^1$ and $\mathcal{A}_c^2$ both resemble rectangles with rounded corners. Shown are the inner inflated sets for $\alpha = \{1\}$ (light blue), $\alpha = \{2\}$ (light brown) and $\alpha = \{1, 2\}$ (green). Also depicted are the contours $g^1 = \eta_1$, $g^2 = \eta_1$, $g^1 = K_1\eta_1$ and $g^2 = K_1\eta_1$.

In the following sections, we require $\mathcal{I}_c^{n,\alpha}$ to shrink to a set resembling $\partial^\alpha\mathcal{F}_c = \cup_{\alpha'\supseteq\alpha}\partial^{\alpha'}\mathcal{F}_c \setminus \cup_{\alpha'\supset\alpha}\partial^{\alpha'}\mathcal{F}_c$. To remove $\cup_{\alpha'\supset\alpha}\partial^{\alpha'}\mathcal{F}_c$ from consideration, the union which appears following the set minus in Equation (C.3) is required. This term ensures that $\mathcal{I}_c^{n,\alpha}$ does not intersect the set[2] $\cup_{\alpha'\supset\alpha}\mathcal{I}_c^{n,\alpha'}$. However, if $\cup_{\alpha'\supset\alpha}\mathcal{I}_c^{n,\alpha'}$ were to shrink to $\cup_{\alpha'\supset\alpha}\partial^{\alpha'}\mathcal{F}_c$ this would still not necessarily guarantee that $\mathcal{I}_c^{n,\alpha} \setminus \cup_{\alpha'\supset\alpha}\mathcal{I}_c^{n,\alpha'}$ would

[1]We note here that we have not defined the word 'shrink'. In the following sections, 'shrink' shall mean 'converge in Hausdorff distance'. For ease of reading, in this section we shall continue to use 'shrink'.

[2]In fact, the inclusion of the union guarantees something slightly stronger; that $\mathcal{I}_c^{n,\alpha}$ does not intersect $\cup_{\mathrm{Ord}(\alpha')>\mathrm{Ord}(\alpha)}\mathcal{I}_c^{n,\alpha'}$. This slight deviation is to ensure the sets $\{\mathcal{I}_c^{n,\alpha}\}_{\alpha\in\mathcal{P}^+(\mathcal{M})}$ are disjoint.

shrink to a set which does not contain $\cup_{\alpha' \supset \alpha} \partial^{\alpha'} \mathcal{F}_c$. For this reason, it is an additional requirement that $\cup_{\alpha' \supset \alpha} \mathcal{I}_c^{n,\alpha'}$ must shrink to $\cup_{\alpha' \supset \alpha} \partial^{\alpha'} \mathcal{F}_c$ at a much 'slower rate' than the rate at which $\mathcal{I}_c^{n,\alpha}$ shrinks to $\partial^{\alpha} \mathcal{F}_c$. This requirement is handled by the constant $(K_n)^{|\alpha|-1}$ which appears earlier in the definition of the inflated boundaries. As, for all $\alpha' \supset \alpha$, $|\alpha'| \geq |\alpha| + 1$, it follows that the rate at which $\mathcal{I}_c^{n,\alpha}$ shrinks is, in some sense, at least a factor of K_n faster than the rate at which $\cup_{\alpha' \supset \alpha} \mathcal{I}_c^{n,\alpha'}$ shrinks. As $K_n \to \infty$ as $n \to \infty$, it may be expected that this ensures $\mathcal{I}_c^{n,\alpha}$ cannot shrink to a region containing $\cup_{\alpha' \supset \alpha} \partial^{\alpha'} \mathcal{F}_c$. This intuition is argued formally in Section C.2.6.

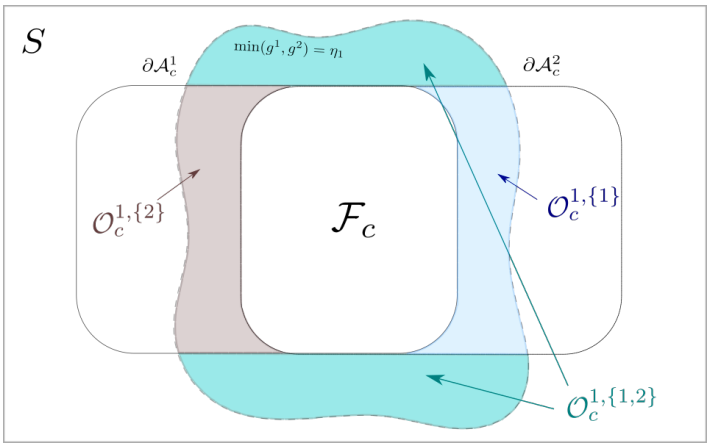

Figure C.5: Illustration of the outer inflated sets in a setting in which $n = 1$, $M = 2$ and the excursion sets $\mathcal{A}_c^1$ and $\mathcal{A}_c^2$ both resemble rectangles with rounded corners. Shown are the outer inflated sets for $\alpha = \{1\}$ (light blue), $\alpha = \{2\}$ (light brown) and $\alpha = \{1, 2\}$ (green). Also depicted is the contour $\min(g^1, g^2) = \eta_1$, which corresponds to the outer boundary of the set $\mathcal{F}_c^1$.

Visually, this concept may be understood by considering Fig. C.4 and taking $\alpha = \{1\}$ and $\alpha' = \{1, 2\}$. In this case, for $n = 1$, $\mathcal{I}_c^{n,\alpha}$ is represented by the light blue region, whilst $\mathcal{I}_c^{n,\alpha'}$ is represented by the green region. In this example, $\mathcal{I}_c^{n,\alpha}$ is much 'thinner' than $\mathcal{I}_c^{n,\alpha'}$. In general, as $K_n \to \infty$ when $n \to \infty$, it is expected that the contour $g^1 = \eta_n$ will 'expand' outwards towards $\partial \mathcal{A}_c^1$ at a much faster rate than the rate at which the contour $g^2 = K_n \eta_n$ will 'expand' outwards towards $\partial \mathcal{A}_c^2$. By consequence, in the limit, $\mathcal{I}_c^{n,\alpha}$ will shrink 'faster' than $\mathcal{I}_c^{n,\alpha'}$ (the light blue region will continue to be much 'thinner' than the green region). As a result, it is not possible for $\mathcal{I}_c^{n,\alpha}$ to 'shrink' to a region which includes the set $\partial^{\alpha'} \mathcal{F}_c$ (the light blue region

cannot 'get close' to arbitrary points belonging to the the top or the bottom of $\partial\mathcal{F}_c$).

It is worth briefly recapping the roles of the sequences $\{\tau_n\}_{n\in\mathbb{N}}, \{\eta_n\}_{n\in\mathbb{N}}$ and $\{K_n\}_{n\in\mathbb{N}}$. The sequence $\{\tau_n\}_{n\in\mathbb{N}}$ is typically a measurement of sample size or statistical power and is dictated by application. In contrast, $\{\eta_n\}_{n\in\mathbb{N}}$ and $\{K_n\}_{n\in\mathbb{N}}$ are synthetic sequences constructed for the purposes of proof. As $n \to \infty$, $\eta_n \to 0$ in order to ensure that the inner inflated sets, $\{\mathcal{I}_c^{n,\alpha}\}_{\alpha\in\mathcal{P}^+(\mathcal{M})}$, shrink to resemble the partitioned boundary sets, $\{\partial^\alpha\mathcal{F}_c\}_{\alpha\in\mathcal{P}^+(\mathcal{M})}$. Conversely, $K_n \to \infty$ as $n \to \infty$ and is used to distinguish between the rates of convergence of different inner inflated sets.

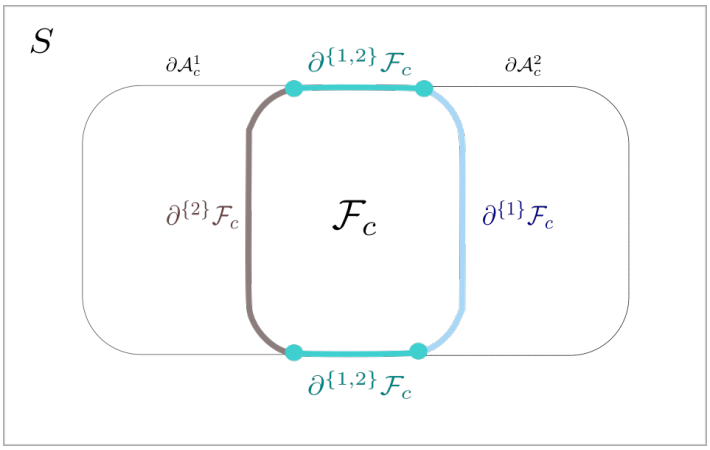

Figure C.6: Illustration of the boundary segments in a setting in which the excursion sets $\mathcal{A}_c^1$ and $\mathcal{A}_c^2$ both resemble rectangles with rounded corners. Shown are the boundary segments for $\alpha = \{1\}$ (light blue), $\alpha = \{2\}$ (light brown) and $\alpha = \{1, 2\}$ (green).

We now consider the definition of the 'outer' inflated boundaries $\{\mathcal{O}_c^{n,\alpha}\}_{\alpha\in\mathcal{P}^+(\mathcal{M})}$, given by Equation (C.4). The first term on the right hand side of Equation (C.4) is the set $\mathcal{F}_c^n$ whose inclusion guarantees that the set $\mathcal{O}_c^{n,\alpha}$ must shrink to some subset of $\partial\mathcal{F}_c$ as $n \to \infty$. The second and third terms strongly resemble the definition of $\mathcal{J}_c^\alpha$ given in Section 5.2.2 of Chapter 5. The intuition behind the construction of $\mathcal{J}_c^\alpha$ has already been discussed in Section C.1.3.1 and shall not recounted again here. Here, it suffices to note that, for fixed $\alpha \in \mathcal{P}^+(\mathcal{M})$, the second and third terms describe the region outside $\mathcal{F}_c$ (thus living up to the name 'outer') which we may 'naturally expect' to border the boundary segment $\partial^\alpha\mathcal{F}_c$.

In summary, for fixed $\alpha \in \mathcal{P}^+(\mathcal{M})$, intuition suggests that the sets $\mathcal{O}_c^{n,\alpha}$ and $\mathcal{I}_c^{n,\alpha}$ border the boundary segment $\partial^\alpha \mathcal{F}_c$ and shrinks towards it as $n \to \infty$. This intuition can still fail, however, as in certain circumstances the sets $\mathcal{I}_c^{n,\alpha}$ and $\mathcal{O}_c^{n,\alpha}$ can shrink to sets which contain $\partial^\alpha \mathcal{F}_c$ and have equal measure to $\partial^\alpha \mathcal{F}_c$, but do not equal $\partial^\alpha \mathcal{F}_c$ (c.f. Fig. C.7 in Section C.2.6). Such circumstances are explicitly handled by the theorems presented in Sections C.2.8 and C.2.9.

C.2.2 Part II: Proof of Equation (C.2)

In this section, we shall prove Equation (C.2). To do so, we first define, for every $\alpha \in \mathcal{P}^+(\mathcal{M})$, the 'shrunken' boundary for $\partial^\alpha \mathcal{F}_c$ as:

$$\mathcal{F}_c^{-n,\alpha} = \{s \in S : \forall i \in \alpha, g^i(s) = 0 \text{ and } \forall j \in \mathcal{M} \setminus \alpha, g^j(s) > \eta_n\}.$$

$\mathcal{F}_c^{-n,\alpha}$ is a subset of $\partial^\alpha \mathcal{F}_c$ (thus living up to the name 'shrunken') and is expected to increasingly resemble $\partial^\alpha \mathcal{F}_c$ as $n \to \infty$. Next, we define the shrunken boundary for $\partial \mathcal{F}_c$, as $\mathcal{F}_c^{-n} = \cup_{\alpha \in \mathcal{P}^+(\mathcal{M})} \mathcal{F}_c^{-n,\alpha}$.

Now, suppose the below statement were true:

$$\exists \delta > 0, s \in \mathcal{F}_c^{-n} \text{ such that } \tau_n^{-1} \min_{i \in \mathcal{M}}(\hat{g}_n^i(s)) \geq a + \delta. \tag{C.6}$$

As, for n large enough, $\tau_n^{-1} \min_{i \in \mathcal{M}}(\hat{g}_n^i)$ is continuous, it follows that, for n large enough, there must be an $s \in \mathcal{F}_c^{-n}$ such that $\tau_n^{-1} \min_{i \in \mathcal{M}}(\hat{g}_n^i) > a$ in some open neighbourhood of s. As such a neighbourhood must intersect $\mathcal{J}_c^\alpha$ by the ball assumption, the below now follows:

$$\exists \tilde{s} \in \mathcal{J}_c^\alpha \subseteq (\mathcal{F}_c)^c \text{ such that } \tau_n^{-1} \min_{i \in \mathcal{M}}(\hat{g}_n^i(\tilde{s})) > a.$$

And therefore, by the definitions of $\mathcal{F}_c$ and $\hat{\mathcal{F}}_c^+$, it follows that there exists an $\tilde{s} \in (\mathcal{F}_c)^c$ which also belongs to $\hat{\mathcal{F}}_c^+$. Therefore, $\hat{\mathcal{F}}_c^+ \not\subseteq \mathcal{F}_c$.

Through similar logic, it can be seen that if the below statement is true:

$$\exists \delta > 0, s \in \mathcal{F}_c^{-n} \text{ such that } \tau_n^{-1} \min_{i \in \mathcal{M}}(\hat{g}_n^i(s)) \leq -a - \delta, \tag{C.7}$$

then $\mathcal{F}_c \not\subseteq \hat{\mathcal{F}}_c^-$. By combining Statements (C.6) and (C.7) and the above arguments it can now be seen that if the below statement holds:

$$\exists \delta > 0, s \in \mathcal{F}_c^{-n} \text{ such that } \tau_n^{-1} |\min_{i \in \mathcal{M}}(\hat{g}_n^i(s))| \geq a + \delta,$$

then it cannot be the case that the inclusion statement, $\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-$, holds. Therefore:

$$1 - \mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-] \geq \mathbb{P}\left[\exists \delta > 0 \text{ such that } \sup_{s \in \mathcal{F}_c^{-n}} |\tau_n^{-1} \min_{i \in \mathcal{M}}(\hat{g}_n^i(s))| \geq a + \delta\right]$$

$$= \mathbb{P}\left[\sup_{s \in \mathcal{F}_c^{-n}} |\tau_n^{-1} \min_{i \in \mathcal{M}}(\hat{g}_n^i(s))| > a\right].$$

By rearranging the above, the following may be obtained:

$$\mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-] \leq \mathbb{P}\left[\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \mathcal{F}_c^{-n,\alpha}} |\tau_n^{-1} \min_{i \in \mathcal{M}}(\hat{g}_n^i(s))|\right) \leq a\right].$$

The aim is now to show that, as $n \to \infty$, the limit supremum of the probability appearing on the right hand side of the above tends to the right hand side of Equation (C.2). To achieve this, we define the following four quantities, for $\alpha \in \mathcal{P}^+(\mathcal{M})$:

$$W_\alpha^n = \sup_{s \in \mathcal{F}_c^{-n,\alpha}} \left|\tau_n^{-1} \min_{i \in \mathcal{M}}(\hat{g}_n^i(s))\right|, \qquad X_\alpha^n = \sup_{s \in \mathcal{F}_c^{-n,\alpha}} \left|\tau_n^{-1} \min_{i \in \alpha}(\hat{g}_n^i(s) - g^i(s))\right|,$$

$$Y_\alpha^n = \sup_{s \in \partial^\alpha \mathcal{F}_c} \left|\tau_n^{-1} \min_{i \in \alpha}(\hat{g}_n^i(s) - g^i(s))\right|, \qquad Z_\alpha = \sup_{s \in \partial^\alpha \mathcal{F}_c} |\min_{i \in \alpha}(G^i(s))|.$$

Using theorems given in Sections C.2.10 and C.2.11, alongside Assumption 5.1, we

now note that the below relations hold:

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(W_\alpha^n) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(X_\alpha^n) \xrightarrow{p} 0, \qquad \text{(By the result of Theorem C.7)}$$

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(X_\alpha^n) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(Y_\alpha^n) \xrightarrow{p} 0, \qquad \text{(By the result of Theorem C.8)}$$

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(Y_\alpha^n) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(Z_\alpha) \xrightarrow{d} 0, \qquad \text{(By Assumption 5.1)}$$

where $\xrightarrow{p}$ and $\xrightarrow{d}$ represent convergence in probability and distribution, respectively. By combining the above results, it can now be seen that:

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(W_\alpha^n) \xrightarrow{d} \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(Z_\alpha).$$

Combining the preceding arguments, it now follows that:

$$\limsup_{n\to\infty} \mathbb{P}[\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-] \leq \limsup_{n\to\infty} \mathbb{P}\left[\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(W_\alpha^n) \leq a\right]$$

$$= \mathbb{P}\left[\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(Z_\alpha) \leq a\right] = \mathbb{P}\left[\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}\left(\sup_{s\in\partial^\alpha\mathcal{F}_c}\left|\min_{i\in\alpha}(G^i(s))\right|\right) \leq a\right],$$

where the inequality follows from the definition of W_α^n, the first equality follows from the argument above, and the second equality follows from the definition of Z_α. It now follows that Equation (C.2) holds. □

C.2.3 Supporting Lemmas

Lemma C.1. *Given an arbitrary positive sequence $\{\tau_n\}_{n\in\mathbb{N}}$ such that $\tau_n \to 0$ as $n \to \infty$, it is possible to define two positive sequences $\{\eta_n\}_{n\in\mathbb{N}}$ and $\{K_n\}_{n\in\mathbb{N}}$ such that, as $n \to \infty$:*

$$\eta_n \to 0, \quad \tau_n^{-1}\eta_n \to \infty, \quad K_n \to \infty, \quad K_n^M\eta_n \to 0 \quad \textit{and} \quad K_n\tau_n^{-1}\eta_n \to \infty.$$

Proof. The above limits are satisfied by choosing $\eta_n = \tau_n^{\frac{1}{2}}$ and $K_n = \eta_n^{-\frac{1}{2M}}$. □

Lemma C.2. *For arbitrary a, b, c and d, it is true that:*

$$\min(a-c, b-d) \leq \min(a,b) - \min(c,d) \leq \max(a-c, b-d).$$

Lemma C.3. *If $\{A_n\}_{n\in\mathbb{N}}, \{\tilde{A}_n\}_{n\in\mathbb{N}}$ and $\{B_n\}_{n\in\mathbb{N}}$ are sequences of random variables and $A_n - \tilde{A}_n \xrightarrow{p} 0$ then the below convergence in probability holds:*

$$\max(A_n, B_n) - \max(\tilde{A}_n, B_n) \xrightarrow{p} 0.$$

Lemma C.4. *For functions a and b acting on a closed set X, it is true that:*

$$\left| \sup_{x\in X} |a(x)| - \sup_{x\in X} |b(x)| \right| \leq \sup_{x\in X} \left| a(x) - b(x) \right|.$$

Lemma C.5. *If $\{A_n\}_{n\in\mathbb{N}}$ and $\{B_n\}_{n\in\mathbb{N}}$ are sequences of random variables which jointly tend in probability to random variables A and B, respectively, then the below convergence in probability holds:*

$$\max(A_n, B_n) \xrightarrow{p} \max(A, B).$$

Lemma C.6. *For every $\alpha \in \mathcal{P}^+(\mathcal{M})$, it is true that;*

$$\forall s \in \mathcal{I}_c^{n,\alpha}, \min_{j\in\mathcal{M}\setminus\alpha} (g^j(s)) \geq K_n^{|\alpha|}\eta_n.$$

Proof. We can discount the case $\alpha = \mathcal{M}$ as the minimum taken over the empty set is equal to $+\infty$ (see Section C.2.1). For $\alpha \neq \mathcal{M}$, suppose that the negation of the above were true and fix $s \in \mathcal{I}_c^{n,\alpha}$ and $j \in \mathcal{M} \setminus \alpha$ such that $g^j(s) < K_n^{|\alpha|}\eta_n$. By the definition of $\mathcal{I}_c^{n,\alpha}$, it is also true that for all $i \in \alpha$, $g^i(s) < K_n^{|\alpha|}\eta_n$. Define $\tilde{\alpha} = \alpha \cup \{j\}$. As $|\tilde{\alpha}| = |\alpha| + 1$, it now follows that $g^k(s) < K_n^{|\tilde{\alpha}|-1}\eta_n$ for all $k \in \tilde{\alpha}$. By the definition of the $\beta-$inflated boundary for $\mathcal{A}_c^k$, it now follows that $s \in \mathcal{A}_c^{K_n^{(|\tilde{\alpha}|-1)}\eta_n,k}$ for all $k \in \tilde{\alpha}$. However, by the definition of $\mathcal{I}_c^{n,\tilde{\alpha}}$, this implies that either $s \in \mathcal{I}_c^{n,\tilde{\alpha}}$ or $s \in \mathcal{I}_c^{n,\tilde{\alpha}'}$ for some $\tilde{\alpha}'$ such that $\mathrm{Ord}(\tilde{\alpha}') > \mathrm{Ord}(\tilde{\alpha})$. In either case, $s \in \mathcal{I}_c^{n,\alpha'}$ for some α' such that $\mathrm{Ord}(\alpha') > \mathrm{Ord}(\alpha)$. From the definition of $\mathcal{I}_c^{n,\alpha}$, it follows that $s \notin \mathcal{I}_c^{n,\alpha}$. This is a contradiction, as $s \in \mathcal{I}_c^{n,\alpha}$ by construction. The result of the lemma now follows. □

Lemma C.7. *Define the modulus of continuity, ω, of a function, f, and a distance, δ, by:*

$$\omega(f, \delta) = \sup_{s_1, s_2 \in S : |s_1 - s_2| < \delta} |f(s_1) - f(s_2)|.$$

If $\{f_n\}_{n \in \mathbb{N}}$ is a sequence of continuous random variables on S which converges in distribution to a function f which has continuous sample paths, we have that:

$$\forall \epsilon > 0, \lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}[\omega(f_n, \delta) \geq \epsilon] = 0.$$

Proof. See Theorem 3.3.1 of Khoshnevisan (2002). □

C.2.4 Theorem C.1: The Events Implication Theorem

Theorem C.1. *Defining the events* $\{\mathcal{E}_{\mathcal{I}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ *and* $\{\mathcal{E}_{\mathcal{O}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ *as in Section C.2.1, the following implication holds:*

$$\bigcap_{\alpha\in\mathcal{P}(\mathcal{M})}\{\mathcal{E}_{\mathcal{I}}^{\alpha,n}\wedge\mathcal{E}_{\mathcal{O}}^{\alpha,n}\}\implies\{\hat{\mathcal{F}}_c^{+}\subseteq\mathcal{F}_c\subseteq\hat{\mathcal{F}}_c^{-}\}.$$

Proof. We shall first show that if the events $\{\mathcal{E}_{\mathcal{I}}^{\alpha,n}\wedge\mathcal{E}_{\mathcal{O}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ occur then $\{\mathcal{F}_c\subseteq\hat{\mathcal{F}}_c^{-}\}$ holds. In other words, we will show that, under the assumption that the events occur, if $s\in\mathcal{F}_c$, then $s\in\hat{\mathcal{F}}_c^{-}$. As $\mathcal{F}_c=\cup_{\alpha\in\mathcal{P}(\mathcal{M})}\mathcal{I}_c^{n,\alpha}$, we split the statement $s\in\mathcal{F}_c$ into two cases; $s\in\mathcal{I}_c^{n,\emptyset}$ (Case 1), and $s\in\mathcal{I}_c^{n,\alpha}$ for some $\alpha\in\mathcal{P}^{+}(\mathcal{M})$ (Case 2).

Case 1:

Let $s\in\mathcal{I}_c^{n,\emptyset}$. As $\mathcal{I}_c^{n,\emptyset}\subseteq\mathcal{F}_c$, it follows that $s\in\mathcal{F}_c$. Now, suppose that for some $i\in\mathcal{M}$, $|g^i(s)|\leq\eta_n$. This would mean that $s\in\mathcal{A}_c^{\eta_n,i}$ by the definition of $\mathcal{A}_c^{\eta_n,i}$. By considering the definition of $\mathcal{I}_c^{n,\{i\}}$, it can be seen that this implies that either $s\in\mathcal{I}_c^{n,\{i\}}$ or $s\in\mathcal{I}_c^{n,\alpha'}$ for some $\alpha'\in\mathcal{P}^{+}(\mathcal{M})$ such that $\mathrm{Ord}(\alpha')>\mathrm{Ord}(\{i\})$. However, this is a contradiction, as $s\in\mathcal{I}_c^{n,\emptyset}$ which is disjoint from the sets $\{\mathcal{I}_c^{n,\alpha}\}_{\alpha\in\mathcal{P}^{+}(\mathcal{M})}$. Therefore, it follows that for all $i\in\mathcal{M}$, $|g^i(s)|>\eta_n$.

As $s\in\mathcal{F}_c$, $g^i(s)\geq 0$ for all $i\in\mathcal{M}$. Therefore, for all $i\in\mathcal{M}$, $g^i(s)>\eta_n$. Consequently, for all $i\in\mathcal{M}$, $\eta_n-g^i(s)<0$ and:

$$\tau_n^{-1}\min_{i\in\mathcal{M}}(\hat{g}_n^i(s))>\tau_n^{-1}\min_{i\in\mathcal{M}}\left(\hat{g}_n^i(s)+\eta_n-g^i(s)\right)$$

$$=\tau_n^{-1}\min_{i\in\mathcal{M}}\left(\hat{g}_n^i(s)-g^i(s)\right)+\tau_n^{-1}\eta_n\geq-\tau_n^{-1}\max_{i\in\mathcal{M}}\left|\hat{g}_n^i(s)-g^i(s)\right|+\tau_n^{-1}\eta_n$$

$$\geq-a-\tau_n^{-1}\eta_n+\tau_n^{-1}\eta_n=-a,$$

where the second inequality follows from basic properties of 'min' and 'max' and the third from the definition of $\mathcal{E}_{\mathcal{I}}^{\emptyset,n}$. By the definition of $\hat{\mathcal{F}}_c^-$, it follows that $s \in \hat{\mathcal{F}}_c^-$ as required. Therefore, if the events $\{\mathcal{E}_{\mathcal{I}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ and $\{\mathcal{E}_{\mathcal{O}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ occur then $\mathcal{I}_c^{n,\emptyset} \subseteq \hat{\mathcal{F}}_c^-$. In other words;

$$\bigcap_{\alpha\in\mathcal{P}(\mathcal{M})} \{\mathcal{E}_{\mathcal{I}}^{\alpha,n} \wedge \mathcal{E}_{\mathcal{O}}^{\alpha,n}\} \implies \{\mathcal{I}_c^{n,\emptyset} \subseteq \hat{\mathcal{F}}_c^-\}.$$

Case 2:

Let $s \in \mathcal{I}_c^{n,\alpha}$ for some $\alpha \in \mathcal{P}^+(\mathcal{M})$. Again, as $s \in \mathcal{F}_c$ it follows that $g^i(s) \geq 0$ for all $i \in \alpha$. It therefore follows that;

$$\tau_n^{-1} \min_{i\in\mathcal{M}}(\hat{g}_n^i(s)) = \tau_n^{-1} \min\left(\min_{i\in\alpha}(\hat{g}_n^i(s)), \min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s)) \right)$$

$$\geq \tau_n^{-1} \min\left(\min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s)), \min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s)) \right) \geq -a,$$

where the last inequality follows from the definition of $\mathcal{E}_{\mathcal{I}}^{\alpha,n}$. Using similar logic to that employed in case 1, it now follows that if $\{\mathcal{E}_{\mathcal{I}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ and $\{\mathcal{E}_{\mathcal{O}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ occur then, for any $\alpha \in \mathcal{P}^+(\mathcal{M})$, it is true that $\mathcal{I}_c^{n,\alpha} \subseteq \hat{\mathcal{F}}_c^-$. And therefore, for all $\alpha \in \mathcal{P}^+(\mathcal{M})$;

$$\bigcap_{\alpha\in\mathcal{P}(\mathcal{M})} \{\mathcal{E}_{\mathcal{I}}^{\alpha,n} \wedge \mathcal{E}_{\mathcal{O}}^{\alpha,n}\} \implies \{\mathcal{I}_c^{n,\alpha} \subseteq \hat{\mathcal{F}}_c^-\}.$$

By combining the results of both cases, it can now be seen that:

$$\bigcap_{\alpha\in\mathcal{P}(\mathcal{M})} \{\mathcal{E}_{\mathcal{I}}^{\alpha,n} \wedge \mathcal{E}_{\mathcal{O}}^{\alpha,n}\} \implies \{\mathcal{F}_c \subseteq \hat{\mathcal{F}}_c^-\}. \tag{C.8}$$

We shall now show that if $\{\mathcal{E}_{\mathcal{I}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ and $\{\mathcal{E}_{\mathcal{O}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ occur then it follows that $\hat{\mathcal{F}}_c^+ \subseteq \mathcal{F}_c$. To do this, we shall show that, if the events hold and $s \in (\mathcal{F}_c)^c$ then $s \in (\hat{\mathcal{F}}_c^+)^c$. We begin by noting that $(\mathcal{F}_c)^c = \cup_{\alpha\in\mathcal{P}(\mathcal{M})}\mathcal{O}_c^{n,\alpha}$ and again split the statement $s \in (\mathcal{F}_c)^c$ into two cases; $s \in \mathcal{O}_c^{n,\emptyset}$ (Case 1), and $s \in \mathcal{O}_c^{n,\alpha}$ for some $\alpha \in \mathcal{P}^+(\mathcal{M})$ (Case 2).

Case 1:

Let $s \in \mathcal{O}_c^{n,\emptyset}$. As $\mathcal{O}_c^{n,\emptyset} \subseteq (\mathcal{F}_c)^c$, it follows that $\min_{i\in\mathcal{M}} g^i(s) < 0$. Furthermore, by the definition of $\mathcal{F}_c^n$ and the fact that $\mathcal{O}_c^{n,\emptyset}$ and $\mathcal{F}_c^n$ are disjoint, it follows that $|\min_{i\in\mathcal{M}} g^i(s)| > \eta_n$. Combining these facts we have that $\min_{i\in\mathcal{M}} g^i(s) < -\eta_n$ and therefore $-\eta_n - \min_{i\in\mathcal{M}} g^i(s) > 0$. It now follows that:

$$\tau_n^{-1} \min_{i\in\mathcal{M}} \hat{g}_n^i(s) < \tau_n^{-1}\left(\min_{i\in\mathcal{M}} \hat{g}_n^i(s) - \min_{i\in\mathcal{M}} g^i(s)\right) - \tau_n^{-1}\eta_n$$

$$\leq \tau_n^{-1} \max_{i\in\mathcal{M}} \left(\hat{g}_n^i(s) - g^i(s)\right) - \tau_n^{-1}\eta_n \leq a + \tau_n^{-1}\eta_n - \tau_n^{-1}\eta_n = a,$$

where the second inequality follows from Lemma C.2 and the third follows from the definition of $\mathcal{E}_{\mathcal{O}}^{\emptyset,n}$. From the definition of $\hat{\mathcal{F}}_c^+$, it follows that $s \in (\hat{\mathcal{F}}_c^+)^c$ as required. Therefore, if $\{\mathcal{E}_{\mathcal{I}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ and $\{\mathcal{E}_{\mathcal{O}}^{\alpha,n}\}_{\alpha\in\mathcal{P}(\mathcal{M})}$ occur then $\mathcal{O}_c^{n,\emptyset} \subseteq (\hat{\mathcal{F}}_c^+)^c$. In other words, the below holds:

$$\bigcap_{\alpha\in\mathcal{P}(\mathcal{M})} \{\mathcal{E}_{\mathcal{I}}^{\alpha,n} \wedge \mathcal{E}_{\mathcal{O}}^{\alpha,n}\} \implies \{\mathcal{O}_c^{n,\emptyset} \subseteq (\hat{\mathcal{F}}_c^+)^c\}.$$

Case 2:

Let $s \in \mathcal{O}_c^{n,\alpha}$ for some $\alpha \in \mathcal{P}^+(\mathcal{M})$. By the definition of $\mathcal{O}_c^{n,\alpha}$ it follows that $g^i(s) < 0$ for all $i \in \alpha$. Therefore, it follows that:

$$\tau_n^{-1} \min_{i\in\mathcal{M}} \hat{g}_n^i(s) \leq \tau_n^{-1} \min_{i\in\alpha} \hat{g}_n^i(s) < \tau_n^{-1} \min_{i\in\alpha} \left(\hat{g}_n^i(s) - g^i(s)\right) \leq a,$$

where the first inequality follows as $\alpha \subseteq \mathcal{M}$ and the third follows from the definition of $\mathcal{E}_{\mathcal{O}}^{\alpha,n}$ for $\alpha \in \mathcal{P}^+(\mathcal{M})$. From the definition of $\hat{\mathcal{F}}_c^+$, it now follows that $s \in (\hat{\mathcal{F}}_c^+)^c$. By using similar logic to that employed in case 1, it can now be seen that:

$$\bigcap_{\alpha\in\mathcal{P}(\mathcal{M})} \{\mathcal{E}_{\mathcal{I}}^{\alpha,n} \wedge \mathcal{E}_{\mathcal{O}}^{\alpha,n}\} \implies \{\mathcal{O}_c^{n,\alpha} \subseteq (\hat{\mathcal{F}}_c^+)^c\}.$$

By combining this with the result of case 1, it can be seen that:

$$\bigcap_{\alpha\in\mathcal{P}(\mathcal{M})}\{\mathcal{E}_{\mathcal{I}}^{\alpha,n}\wedge\mathcal{E}_{\mathcal{O}}^{\alpha,n}\}\implies\{(\mathcal{F}_c)^c\subseteq(\hat{\mathcal{F}}_c^+)^c\}\equiv\{\hat{\mathcal{F}}_c^+\subseteq\mathcal{F}_c\},\tag{C.9}$$

as required. The result of Theorem C.1 now follows directly from Statements (C.8) and (C.9). □

C.2.5 Theorem C.2: Convergence of $\max(A_\alpha^n)-\max(B_\alpha^n)$

Theorem C.2. *Defining A_α^n and B_α^n as in Section C.2.1, it is true that:*

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(A_\alpha^n)-\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(B_\alpha^n)\xrightarrow{p}0.$$

Proof. By Lemma C.5, it suffices to show that, for arbitrary $\alpha\in\mathcal{P}^+(\mathcal{M})$, $A_\alpha^n-B_\alpha^n\xrightarrow{p}0$. Further, by Lemma C.3, it suffices to show that $A_\alpha^{n,\mathcal{I}}-B_\alpha^{n,\mathcal{I}}\xrightarrow{p}0$. From the definitions of $A_\alpha^{n,\mathcal{I}}$ and $B_\alpha^{n,\mathcal{I}}$ it follows that if $\alpha=\mathcal{M}$ then $A_\alpha^{n,\mathcal{I}}=B_\alpha^{n,\mathcal{I}}$ for all $n\in\mathbb{N}$ and the convergence trivially holds. To prove the convergence for $\alpha\neq\mathcal{M}$, we employ Lemma C.4 to see that:

$$|A_\alpha^{n,\mathcal{I}}-B_\alpha^{n,\mathcal{I}}|\leq\tau_n^{-1}\sup_{s\in\mathcal{I}_c^{n,\alpha}}\left|\min\left(\min_{i\in\alpha}(\hat{g}_n^i(s)-g^i(s)),\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s))\right)-\min_{i\in\alpha}(\hat{g}_n^i(s)-g^i(s))\right|$$

$$=\tau_n^{-1}\sup_{s\in\mathcal{I}_c^{n,\alpha}}\left|\min\left(0,\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s))-\min_{i\in\alpha}(\hat{g}_n^i(s)-g^i(s))\right)\right|.$$

Therefore, for arbitrary $\epsilon>0$, it follows that:

$$\mathbb{P}\left[|A_\alpha^{n,\mathcal{I}}-B_\alpha^{n,\mathcal{I}}|\geq\epsilon\right]\leq\mathbb{P}\left[\tau_n^{-1}\sup_{s\in\mathcal{I}_c^{n,\alpha}}\left|\min\left(0,\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s))-\min_{i\in\alpha}(\hat{g}_n^i(s)-g^i(s))\right)\right|\geq\epsilon\right]$$

$$= \mathbb{P}\Bigg[\exists s \in \mathcal{I}_c^{n,\alpha} \text{ such that } \tau_n^{-1}\Bigg(\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\Bigg) \leq -\epsilon\Bigg]. \quad \text{(C.10)}$$

Now, suppose that an $s \in \mathcal{I}_c^{n,\alpha}$ exists which satisfies the inequality appearing inside the above probability. This would trivially imply that;

$$\tau_n^{-1}\Bigg(\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s)) - \min_{j\in\mathcal{M}\setminus\alpha}(g^j(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\Bigg) \leq -\epsilon - \tau_n^{-1}\min_{j\in\mathcal{M}\setminus\alpha}(g^j(s)).$$

However, by Lemma C.2, it can be seen that the above implies:

$$\tau_n^{-1}\Bigg(\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s) - g^j(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\Bigg) \leq -\epsilon - \tau_n^{-1}\min_{j\in\mathcal{M}\setminus\alpha}(g^j(s)).$$

We now note that by Lemma C.6, it follows that for all $s \in \mathcal{I}_c^{n,\alpha}$ and $j \in \mathcal{M} \setminus \alpha$ it is true that $g^j(s) \geq K_n^{|\alpha|}\eta_n$. Therefore, the above now implies that;

$$\tau_n^{-1}\Bigg(\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s) - g^j(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\Bigg) \leq -\epsilon - \tau_n^{-1}K_n^{|\alpha|}\eta_n.$$

Therefore, Equation (C.10) must be less than or equal to the following:

$$\mathbb{P}\Bigg[\exists s \in \mathcal{I}_c^{n,\alpha} \text{ such that } \tau_n^{-1}\Bigg(\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s) - g^j(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\Bigg) \leq -\epsilon - \tau_n^{-1}K_n^{|\alpha|}\eta_n\Bigg]$$

$$= \mathbb{P}\Bigg[\tau_n^{-1}\inf_{s\in\mathcal{I}_c^{n,\alpha}}\Bigg(\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s) - g^j(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\Bigg) \leq -\epsilon - \tau_n^{-1}K_n^{|\alpha|}\eta_n\Bigg].$$

The left hand side of the expression inside the above probability tends in distribution to a well defined random variable by Assumption 5.1. However, the right hand side converges to $-\infty$ by the definition of K_n. Therefore, the above probability converges to 0. It now follows that, for arbitrary $\epsilon > 0$, $\mathbb{P}[|A_\alpha^{n,\mathcal{I}} - B_\alpha^{n,\mathcal{I}}| \geq \epsilon] \to 0$ as $n \to \infty$. Therefore, $A_\alpha^{n,\mathcal{I}} - B_\alpha^{n,\mathcal{I}} \xrightarrow{p} 0$ as required. □

C.2.6 Theorem C.3: Convergence of $\max(B_\alpha^n) - \max(C_\alpha^n)$

Theorem C.3. *Defining B_α^n and C_α^n as in Section C.2.1, it is true that:*

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(B_\alpha^n) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(C_\alpha^n) \xrightarrow{p} 0.$$

Proof. By Lemma C.5, it suffices to show the following two statements:

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(B_\alpha^{n,\mathcal{I}}) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(C_\alpha^n) \xrightarrow{p} 0, \tag{C.11}$$

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(B_\alpha^{n,\mathcal{O}}) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(C_\alpha^n) \xrightarrow{p} 0. \tag{C.12}$$

To prove each statement we shall employ arguments based on convergence in Hausdorff distance. However, these arguments are complicated by the fact that the sets $\mathcal{I}_c^{n,\alpha}$ and $\mathcal{O}_c^{n,\alpha}$ may not converge in Hausdorff distance to $\partial^\alpha\mathcal{F}_c$ (or, in the case of $\mathcal{I}_c^{n,\alpha}$, converge at all). To understand why, define $\mathcal{D}_\alpha$ as;

$$\mathcal{D}_\alpha = \left(\bigcup_{\substack{\alpha_1\in\mathcal{P}^+(\mathcal{M})\\ \alpha_1\subset\alpha}} \overline{\partial^{\alpha_1}\mathcal{F}_c}\right) \cap \left(\bigcup_{\substack{\alpha_2\in\mathcal{P}^+(\mathcal{M})\\ \alpha_2\supset\alpha}} \overline{\partial^{\alpha_2}\mathcal{F}_c}\right).$$

$\mathcal{D}_\alpha$ can be perceived as a set of "difficult" points, at which a section of $\partial^\alpha\mathcal{F}_c$ has shrunk to a point and essentially 'disappeared' (i.e. is no longer considered part of $\partial^\alpha\mathcal{F}_c$, see Fig. C.7).

Proof of Statement (C.11):

Due to the resemblance between $\mathcal{D}_\alpha$ and $\partial^\alpha\mathcal{F}_c$, $\mathcal{I}_c^{n,\alpha}$ may converge in Hausdorff distance to a set containing a point in $\mathcal{D}_\alpha$ or to a set disjoint from $\mathcal{D}_\alpha$. It is also possible that $\mathcal{I}_c^{n,\alpha}$ converges to neither option but instead oscillates infinitely between the two, thus not converging at all. To handle this issue define $\tilde{\epsilon}_\alpha(s) = d(s, \partial^\alpha\mathcal{F}_c)$

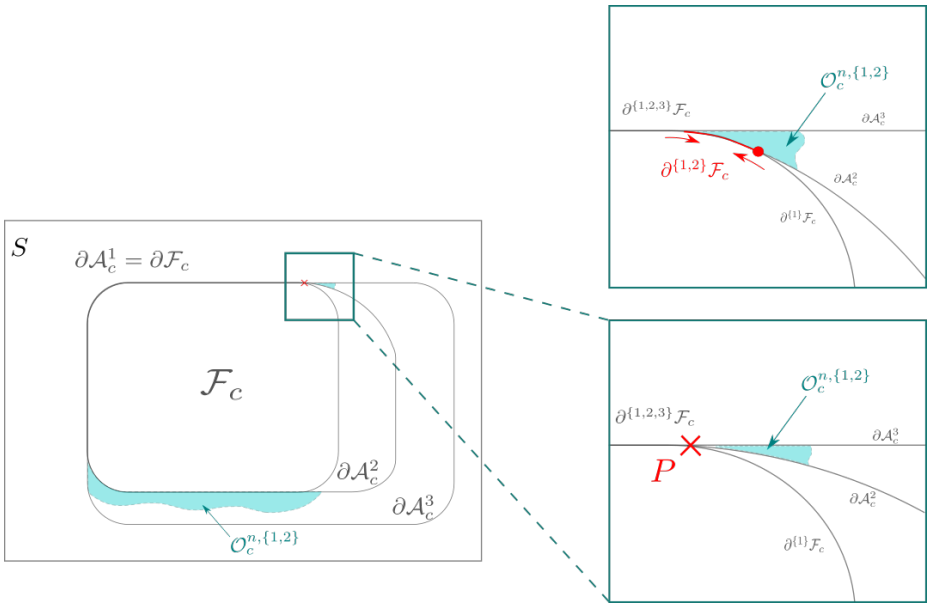

Figure C.7: An example in which $\mathcal{O}_c^{n,\{1,2\}}$ may converge to $\partial^{\{1,2\}}\mathcal{F}_c \cup P$ where $P \in \mathcal{D}_\alpha$. In this instance, it is possible for the set $\mathcal{O}_c^{n,\{1,2\}}$ to border the point P (see bottom right) despite P being located far from the boundary $\partial^{\{1,2\}}\mathcal{F}_c$. This situation has arisen as P strongly resembles a segment of $\partial^{\{1,2\}}\mathcal{F}_c$ which has become vanishingly small (see top right).

(where $d(x, A)$ is the minimum distance from the point x to the set A) and $T_\delta(\mathcal{D}_\alpha)$ as:

$$T_\delta(\mathcal{D}_\alpha) = \bigcup_{s \in \mathcal{D}_\alpha} B_{\min(\delta, \tilde{\epsilon}_\alpha(s))}(s),$$

where $B_\epsilon(s)$ is the open ϵ-ball about s. To illustrate the role of $T_\delta(\mathcal{D}_\alpha)$, consider an example in 3 dimensions. In this example, each boundary segment $\partial^\alpha\mathcal{F}_c$ may appear as a plane and $\mathcal{D}_\alpha$ consists of regions at which two planes intersect. Suppose that $\mathcal{D}_\alpha$ is a line segment. In this case, $T_\delta(\mathcal{D}_\alpha)$ is an open tube of width δ around $\mathcal{D}_\alpha$ which tapers to a point whenever one of the tubes endpoints gets close to $\partial^\alpha\mathcal{F}_c$. Fix $\delta > 0$ and define:

$$L_\alpha^n = \sup_{s \in \mathcal{I}_c^{n,\alpha} \setminus T_\delta(\mathcal{D}_\alpha)} \left| \tau_n^{-1} \min_{i \in \alpha} (\hat{g}_n^i(s) - g^i(s)) \right|, \qquad U_\alpha^n = \sup_{s \in \mathcal{I}_c^{n,\alpha} \cup \mathcal{D}_\alpha} \left| \tau_n^{-1} \min_{i \in \alpha} (\hat{g}_n^i(s) - g^i(s)) \right|.$$

L_α^n and U_α^n are lower and upper bounds for $B_\alpha^{n,\mathcal{I}}$ (i.e. $L_\alpha^n \leq B_\alpha^{n,\mathcal{I}} \leq U_\alpha^n$). Therefore, in order to show Statement (C.11), it suffices to show that $\max_{\alpha \in \mathcal{P}^+(\mathcal{M})}(L_\alpha^n) -$

$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(C^n_\alpha) \xrightarrow{p} 0$ and $\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(U^n_\alpha) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(C^n_\alpha) \xrightarrow{p} 0$. To prove convergence for the lower bound we note that, by the definitions of L^n_α and C^n_α:

$$|L^n_\alpha - C^n_\alpha| = \left| \sup_{s\in\mathcal{I}^{n,\alpha}_c\setminus T_\delta(\mathcal{D}_\alpha)} \left|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}^i_n(s) - g^i(s))\right| - \sup_{s\in\partial^\alpha\mathcal{F}_c} \left|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}^i_n(s) - g^i(s))\right|\right|$$

$$\leq \omega\left(|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}^i_n(s) - g^i(s))|, \delta^{\alpha,L}_n\right),$$

where $\delta^{\alpha,L}_n$ is the Hausdorff distance between $\mathcal{I}^{n,\alpha}_c \setminus T_\delta(\mathcal{D}_\alpha)$ and $\partial^\alpha\mathcal{F}_c$, and ω is the modulus of continuity (c.f. Lemma C.7). By Lemma C.7, it now follows that, for any $\epsilon > 0$;

$$\lim_{n\to\infty}\mathbb{P}[|L^n_\alpha - C^n_\alpha| \geq \epsilon] \leq \limsup_{n\to\infty}\mathbb{P}\left[\omega\left(|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}^i_n(s) - g^i(s))|, \delta^{\alpha,L}_n\right) \geq \epsilon\right] = 0,$$

if $\delta^{\alpha,L}_n \to 0$ as $n \to \infty$. Therefore, if $\delta^{\alpha,L}_n \to 0$ as $n \to \infty$, the convergence of the lower bound holds. We now must show that $\delta^{\alpha,L}_n \to 0$ as $n \to \infty$. As S is compact, by standard properties of the Hausdorff distance, it suffices to show that:

$$\limsup_{n\to\infty}\left(\mathcal{I}^{n,\alpha}_c \setminus T_\delta(\mathcal{D}_\alpha)\right) = \liminf_{n\to\infty}\left(\mathcal{I}^{n,\alpha}_c \setminus T_\delta(\mathcal{D}_\alpha)\right) = \overline{\partial^\alpha\mathcal{F}_c}.$$

To do so, first fix $t^* \in \partial^\alpha\mathcal{F}_c$. Suppose that $t^* \in \mathcal{I}^{n,\alpha'}_c$ for some $\alpha' \in \mathcal{P}^+(\mathcal{M})$ such that $\text{Ord}(\alpha') > \text{Ord}(\alpha)$. By the definition of Ord, it follows that $\alpha' \cap (\mathcal{M}\setminus\alpha)$ is non-empty. Fix $j \in \alpha' \cap (\mathcal{M}\setminus\alpha)$. As $K_n^{|\alpha'|-1}\eta_n \to 0$ as $n \to \infty$ and $g^j(t^*) > 0$ (by the definition of $\partial^\alpha\mathcal{F}_c$), it follows that, for n large enough, $g^j(t^*) > K_n^{|\alpha'|-1}\eta_n$. Therefore, for n large enough, $t^* \notin \mathcal{A}_c^{K_n^{|\alpha'|-1}\eta_n,j}$. However, this contradicts the definition of $\mathcal{I}^{n,\alpha'}_c$. It therefore follows, that for n large enough, $t^* \notin \mathcal{I}^{n,\alpha'}_c$ for any $\alpha' \in \mathcal{P}^+(\mathcal{M})$ such that $\text{Ord}(\alpha') > \text{Ord}(\alpha)$.

By the definition of $\partial^\alpha\mathcal{F}_c$, $t^* \in \mathcal{F}_c$ and $t^* \in \mathcal{A}_c^{(K_n)^{|\alpha|-1}\eta_n,i}$ for all $i \in \alpha$. By the

definition of $T_\delta(\mathcal{D}_\alpha)$, $t^* \notin T_\delta(\mathcal{D}_\alpha)$. Combining these observations, for n large enough:

$$t^* \in \mathcal{F}_c \cap \left(\bigcap_{i \in \alpha} \mathcal{A}_c^{(K_n)^{|\alpha|-1}\eta_n, i} \right) \setminus \left(T_\delta(\mathcal{D}_\alpha) \cup \bigcup_{\substack{\alpha' \in \mathcal{P}^+(\mathcal{M}) \\ \mathrm{Ord}(\alpha') > \mathrm{Ord}(\alpha)}} \mathcal{I}_c^{n,\alpha'} \right) = \mathcal{I}_c^{n,\alpha} \setminus T_\delta(\mathcal{D}_\alpha).$$

Therefore, for every $t^* \in \partial^\alpha \mathcal{F}_c$, if n is large enough $t^* \in \mathcal{I}_c^{n,\alpha} \setminus T_\delta(\mathcal{D}_\alpha)$. It now follows that;

$$\overline{\partial^\alpha \mathcal{F}_c} \subseteq \liminf_{n \to \infty} \left(\mathcal{I}_c^{n,\alpha} \setminus T_\delta(\mathcal{D}_\alpha) \right).$$

Now, fix $s^* \in \limsup_{n\to\infty} \left(\mathcal{I}_c^{n,\alpha} \setminus T_\delta(\mathcal{D}_\alpha) \right)$. It follows that there exists a sequence of points $\{s_{n_m}\}_{m \in \mathbb{N}}$ such that $s_{n_m} \in \mathcal{I}_c^{n_m,\alpha} \setminus T_\delta(\mathcal{D}_\alpha)$ and $s_{n_m} \to s^*$ as $m \to \infty$. As $s_{n_m} \in \mathcal{I}_c^{n_m,\alpha}$, it follows that for each $m \in \mathbb{N}$, s_{n_m} belongs in either $(\mathcal{F}_c)^\circ$ or $\partial^{\tilde{\alpha}} \mathcal{F}_c$ for some $\tilde{\alpha} \subseteq \alpha$ (c.f. Lemma C.6). For ease, we shall now assume that the points $\{s_{n_m}\}_{m \in \mathbb{N}}$ lie only in $(\mathcal{F}_c)^\circ$ and not in $\partial^{\tilde{\alpha}} \mathcal{F}_c$ for any $\tilde{\alpha} \subseteq \alpha$ and return to the latter possibility later in the proof. As $s_{n_m} \in \mathcal{I}_c^{n_m,\alpha}$, it follows that for all $i \in \alpha$, $g^i(s_{n_m}) < (K_{n_m})^{|\alpha|-1}\eta_{n_m}$. Therefore, for all $i \in \alpha$;

$$0 \leq g^i(s^*) = \lim_{m \to \infty} g^i(s_{n_m}) \leq \lim_{m \to \infty} (K_{n_m})^{|\alpha|-1}\eta_{n_m} = 0.$$

And therefore, $g^i(s^*) = 0$ for all $i \in \alpha$. This implies that $s^* \in \partial^{\alpha_1} \mathcal{F}_c$ for some $\alpha_1 \supseteq \alpha$. Now suppose that $s^* \notin \overline{\partial^\alpha \mathcal{F}_c}$. This would imply that $s^* \in \cup_{\alpha_1 \supset \alpha} \partial^{\alpha_1} \mathcal{F}_c \setminus \overline{\partial^\alpha \mathcal{F}_c}$.

Suppose further that $s^* \in \mathcal{D}_\alpha$. As for all $m \in \mathbb{N}$, $s_{n_m} \in \mathcal{I}_c^{n_m,\alpha} \setminus T_\delta(\mathcal{D}_\alpha)$, it follows that $s_{n_m} \notin T_\delta(\mathcal{D}_\alpha)$. By the definition of $T_\delta(\mathcal{D}_\alpha)$, this means that, for all $m \in \mathbb{N}$, $s_{n_m} \notin B_\epsilon(s^*)$ for all $\epsilon < \min(\delta, \tilde{\epsilon}_\alpha(s^*))$. This is a contradiction as $s_{n_m} \to s^*$. Consequently $s^* \notin \mathcal{D}_\alpha$. Thus;

$$s^* \in \bigcup_{\alpha_1 \supset \alpha} \partial^{\alpha_1} \mathcal{F}_c \setminus \left(\overline{\partial^\alpha \mathcal{F}_c} \cup \mathcal{D}_\alpha \right) = \bigcup_{\alpha_1 \supset \alpha} \partial^{\alpha_1} \mathcal{F}_c \setminus \bigcup_{\alpha_2 \subseteq \alpha} \overline{\partial^{\alpha_2} \mathcal{F}_c},$$

where the equality follows from the definition of $\mathcal{D}_\alpha$. Now, by the definition of $\mathcal{I}_c^{n,\alpha}$ and Lemma C.6, it follows that for any $i \in \alpha$ and $j \in \mathcal{M} \setminus \alpha$, $0 < g^i(s_{n_m}) \leq (K_{n_m})^{|\alpha|-1}\eta_{n_m}$ and $g^j(s_{n_m}) \geq (K_{n_m})^{|\alpha|}\eta_{n_m}$. It follows that, for any $i \in \alpha$ and $j \in \mathcal{M} \setminus \alpha$, $K_{m_n} \leq$

$g^j(s_{n_m})/g^i(s_{n_m})$. And, therefore;

$$a_{n_m} := \min_{\substack{i\in\alpha \\ j\in\mathcal{M}\setminus\alpha}} \left(\frac{g^j(s_{n_m})}{g^i(s_{n_m})}\right) \geq K_{n_m}.$$

As $K_{n_m} \to \infty$ as $m \to \infty$, it follows that $\{a_n\}_{n\in\mathbb{N}}$ is unbounded. However, this contradicts Theorem C.4 which states that $\{a_n\}_{n\in\mathbb{N}}$ is bounded. Therefore, it cannot be the case that $s^* \in \limsup_{n\to\infty}\left(\mathcal{I}_c^{n,\alpha}\setminus T_\delta(\mathcal{D}_\alpha)\right)$ unless $s^* \in \overline{\partial^\alpha \mathcal{F}_c}$. It therefore follows that;

$$\limsup_{n\to\infty}\left(\mathcal{I}_c^{n,\alpha} \setminus T_\delta(\mathcal{D}_\alpha)\right) \subseteq \overline{\partial^\alpha \mathcal{F}_c}.$$

We now return to our earlier assumption, in which we discounted sequences $\{s_{n_m}\}_{m\in\mathbb{N}}$ which lay inside $\partial^{\tilde{\alpha}}\mathcal{F}_c$ for some $\tilde{\alpha} \subseteq \alpha$ from consideration. Trivially, it is true that if infinitely many $\{s_{n_m}\}_{m\in\mathbb{N}}$ belonged to $\partial^\alpha \mathcal{F}_c$ the above result would still hold. To handle the cases in which infinitely many $\{s_{n_m}\}_{m\in\mathbb{N}}$ belong to $\partial^{\tilde{\alpha}}\mathcal{F}_c$ for some $\tilde{\alpha} \subset \alpha$, the above argument may be employed with α substituted for $\bar{\alpha} = \alpha \setminus \tilde{\alpha}$ and $\mathcal{M}$ substituted for $\bar{\mathcal{M}} = \mathcal{M} \setminus \tilde{\alpha}$ throughout.

We now have that the limit superior and inferior agree and therefore $\delta_n^{\alpha,L} \to 0$ as $n \to \infty$ as desired. The convergence for the lower bound follows.

To adapt the proof to show convergence for the upper bound, we now define $\hat{C}_\alpha^n$ as follows;

$$\hat{C}_\alpha^n = \sup_{s\in\partial^\alpha\mathcal{F}_c\cup\mathcal{D}_\alpha}\left|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\right|.$$

By employing an identical argument to the above, but neglecting the steps of the proof which involved $T_\delta(\mathcal{D}_\alpha)$, and noting that trivially $\overline{\mathcal{D}_\alpha} \subseteq \liminf_{n\to\infty}(\mathcal{I}_c^{n,\alpha} \cup \mathcal{D}_\alpha)$, it can be seen that;

$$\liminf_{n\to\infty}\left(\mathcal{I}_c^{n,\alpha} \cup \mathcal{D}_\alpha\right) = \limsup_{n\to\infty}\left(\mathcal{I}_c^{n,\alpha} \cup \mathcal{D}_\alpha\right) = \overline{\partial^\alpha \mathcal{F}_c} \cup \overline{\mathcal{D}}_\alpha.$$

Therefore, $\mathcal{I}_c^{n,\alpha} \cup \mathcal{D}_\alpha$ converges to $\partial^\alpha \mathcal{F}_c \cup \mathcal{D}_\alpha$ in Hausdorff distance. By Lemma C.7, it can be seen that the above implies that $U_\alpha^n - \hat{C}_\alpha^n \xrightarrow{p} 0$ for every $\alpha \in \mathcal{P}^+(\mathcal{M})$.

Finally, noting the result of Theorem C.5, which states that $\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(C^n_\alpha) = \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(\hat{C}^n_\alpha)$, it can be seen that $\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(U^n_\alpha) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(C^n_\alpha) \xrightarrow{p} 0$.

As we have now shown convergence for both the lower and upper bounds, we can conclude that $\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(B^{n,\mathcal{I}}_\alpha) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(C^n_\alpha) \xrightarrow{p} 0$ as desired.

Proof of Statement (C.12):

To begin, we claim that for any $\alpha \in \mathcal{P}^+(\mathcal{M})$ there must exist a set $\tilde{\mathcal{D}}_\alpha \subseteq \mathcal{D}_\alpha$ such that $\delta^{\alpha,\mathcal{O}}_n := d_H(\mathcal{O}^{n,\alpha}_c, \partial^\alpha\mathcal{F}_c \cup \tilde{\mathcal{D}}_\alpha) \to 0$ as $n \to \infty$. Proof of this claim is provided by Theorem C.6. Fix α and $\tilde{\mathcal{D}}_\alpha$, and define $\tilde{C}^n_\alpha$ as:

$$\tilde{C}^n_\alpha = \sup_{s\in\partial^\alpha\mathcal{F}_c\cup\tilde{\mathcal{D}}_\alpha} \left|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}^i_n(s) - g^i(s))\right|.$$

By the definitions of $B^{n,\mathcal{O}}_\alpha$ and $\tilde{C}^n_\alpha$ we have that:

$$|B^{n,\mathcal{O}}_\alpha - \tilde{C}^n_\alpha| = \left|\sup_{s\in\mathcal{O}^{n,\alpha}_c}\left|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}^i_n(s) - g^i(s))\right| - \sup_{s\in\partial^\alpha\mathcal{F}_c\cup\tilde{\mathcal{D}}_\alpha}\left|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}^i_n(s) - g^i(s))\right|\right|$$

$$\leq \omega\left(|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}^i_n(s) - g^i(s))|, \delta^{\alpha,\mathcal{O}}_n\right).$$

As $\delta^{\alpha,\mathcal{O}}_n \to 0$ as $n \to \infty$, we can apply Lemma C.7 to see that:

$$\forall\epsilon > 0, \mathbb{P}[|B^{n,\mathcal{O}}_\alpha - \tilde{C}^n_\alpha| \geq \epsilon] \leq \mathbb{P}\left[\omega\left(|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}^i_n(s) - g^i(s))|, \delta^{\alpha,\mathcal{O}}_n\right) \geq \epsilon\right] \xrightarrow{n\to\infty} 0,$$

and therefore, for every $\alpha \in \mathcal{P}^+(\mathcal{M})$, $B^{n,\mathcal{O}}_\alpha - \tilde{C}^n_\alpha \xrightarrow{p} 0$. Again, by employing Lemma C.5 and the result of Theorem C.5, this can now be seen to imply that Statement (C.12) holds. □

C.2.7 Theorem C.4: Boundedness of $\{a_n\}_{n\in\mathbb{N}}$

Theorem C.4. *If $\alpha \in \mathcal{P}^+(\mathcal{M})$ and $\{s_n\}_{n\in\mathbb{N}} \in (\mathcal{F}_c)^\circ$ is a convergent sequence with limit s^* where;*

$$s^* \in \bigcup_{\alpha_1 \supset \alpha} \partial^{\alpha_1}\mathcal{F}_c \setminus \bigcup_{\alpha_2 \subseteq \alpha} \overline{\partial^{\alpha_2}\mathcal{F}_c},$$

then the sequence a_n, defined in Section C.2.6, is bounded above and below.

Whilst its final form no longer bears a strong resemblance, we would like to acknowledge that the following proof drew heavy inspiration from the seminal proofs of Lawlor (2020).

Proof. As $s^* \in \partial^{\alpha_1}\mathcal{F}_c$ for some $\alpha_1 \supset \alpha$, we can choose $\tilde{i} \in \alpha$ such that $g^{\tilde{i}}(s^*) = 0$. As $s^* \notin \overline{\partial^{\alpha_2}\mathcal{F}_c}$ for any $\alpha_2 \subseteq \alpha$, we can define an open neighbourhood of s^*, $\mathcal{N}_0$, which does not intersect $\overline{\partial^{\alpha_2}\mathcal{F}_c}$ for any $\alpha_2 \subseteq \alpha$. Further, by Assumptions 5.2 and 5.3, we can define an open neighbourhood $\mathcal{N}_1 \subseteq \mathcal{N}_0$ such that the functions $\{g^i\}_{i\in\mathcal{M}}$ are continuously differentiable inside $\mathcal{N}_1$.

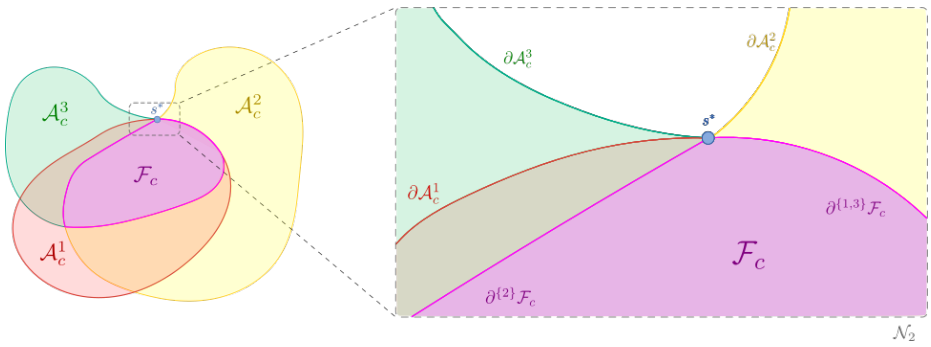

Figure C.8: An example in which $M = 3$ and $s^* \in \partial\mathcal{F}_c$ lies at the intersection of $\partial\mathcal{A}_c^1$, $\partial\mathcal{A}_c^2$ and $\partial\mathcal{A}_c^3$, and therefore belongs to $\partial^{\{1,2,3\}}\mathcal{F}_c$. It may also be seen above that $s^* \in \overline{\partial^{\{1,3\}}\mathcal{F}_c}$ and $s^* \in \overline{\partial^{\{2\}}\mathcal{F}_c}$. By taking $\alpha = \{1\}$, it follows that s^* lies inside the set specified by Theorem C.4. $\mathcal{N}_2$ is represented as a gray dashed rectangle.

By Assumption 5.3 Statement (b), $\nabla g^i(s^*) \neq 0$ for all $i \in \mathcal{M}$. Therefore, the directional derivative of $g^{\tilde{i}}$ in the direction of $\nabla g^{\tilde{i}}(s^*)$, $g^{\tilde{i}}_{\nabla g^{\tilde{i}}(s^*)}(s^*)$, must be strictly

greater than zero. As $g^{\tilde{i}}_{\nabla g^{\tilde{i}}(s)}(s) = \nabla g^{\tilde{i}}(s) \cdot \nabla g^{\tilde{i}}(s)$ is a continuous function of s inside $\mathcal{N}_1$ we can further restrict our attention to a convex open neighbourhood of s^*, $\mathcal{N}_2 \subseteq \mathcal{N}_1$, over which $g^{\tilde{i}}_{\nabla g^{\tilde{i}}} > 0$. An example, which we shall refer to throughout the proof, is provided by Fig. C.8.

Now, define the projection of s_n to $\partial\mathcal{A}^i_c$ as;

$$\mathfrak{p}_i(s_n) = \arg\inf_{s \in \partial\mathcal{A}^i_c} d(s, s_n).$$

Assume for ease, that $\mathfrak{p}_i(s_n)$ is uniquely defined. Define $L(s_n)$ to be the closed line segment which runs from $\mathfrak{p}_{\tilde{i}}(s_n)$ to s_n and $\vec{L}(s_n)$ to be the corresponding vector running from $\mathfrak{p}_{\tilde{i}}(s_n)$ to s_n. Define the projection of s_n to $\partial\mathcal{F}_c$ as:

$$\mathfrak{p}_{\partial\mathcal{F}_c}(s_n) = \arg\inf_{s \in L(s_n) \cap \partial\mathcal{F}_c} d(s, s_n).$$

The existence of such a projection is guaranteed by fact that $s_n \in (\mathcal{F}_c)^\circ$, $\mathfrak{p}_{\tilde{i}}(s_n) \notin (\mathcal{F}_c)^\circ$, and $\{g^j\}_{j \in \mathcal{M}}$ are continuous along $L(s_n)$. From the definition of $\mathfrak{p}_{\tilde{i}}(s_n)$ and noting that $s^* \in \partial\mathcal{A}^{\tilde{i}}_c$, it can be seen that $d(s_n, \mathfrak{p}_{\tilde{i}}(s_n)) \to 0$ as $n \to \infty$ and therefore $\mathfrak{p}_{\tilde{i}}(s_n) \to s^*$. Similarly, as $\mathfrak{p}_{\partial\mathcal{F}_c}(s_n) \in L(s_n)$, it follows that $\mathfrak{p}_{\partial\mathcal{F}_c}(s_n) \to s^*$. Using the example given in Fig. C.8, these sequences are illustrated by Fig. C.9.

Assume that n is large enough such that $\mathfrak{p}_{\tilde{i}}(s_n) \in \mathcal{N}_2$. This means that $\mathfrak{p}_{\tilde{i}}(s_n)$ lies on the (locally) continuous curve $\partial\mathcal{A}^{\tilde{i}}_c$. As $\mathcal{N}_2$ is open and $g^{\tilde{i}}$ is continuously differentiable over $\mathcal{N}_2$, it follows that in an open neighbourhood of $\mathfrak{p}_{\tilde{i}}(s_n)$, $g^{\tilde{i}}$ is continuously differentiable. By construction, $\mathfrak{p}_{\tilde{i}}(s_n)$ is the point s which locally minimizes the distance function $d(s_n, s)$ subject to the constraint $g^{\tilde{i}}(s) = 0$. Therefore, $\mathfrak{p}_{\tilde{i}}(s_n)$ is a local minima of the Lagrangian function;

$$\mathcal{L}(s) = d(s_n, s) + \lambda_n g^{\tilde{i}}(s).$$

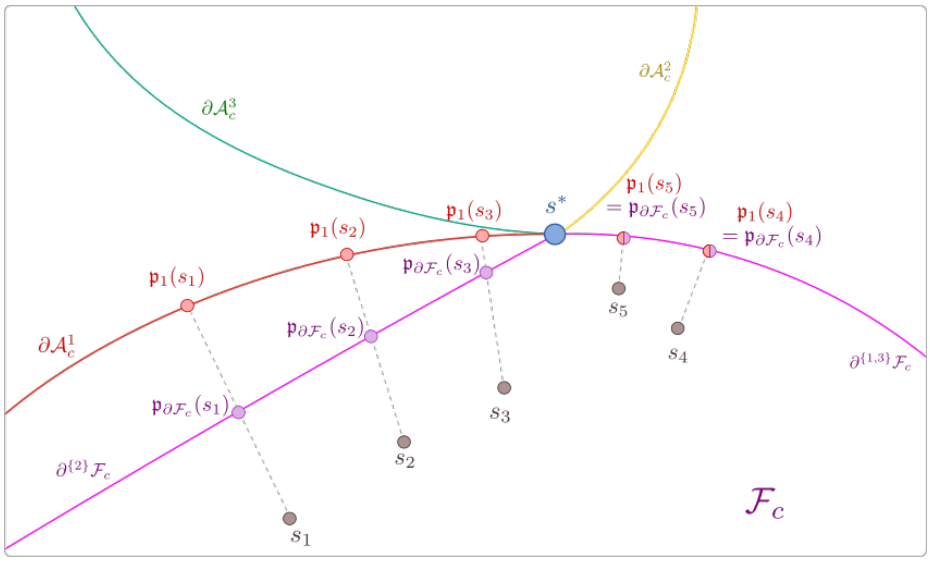

Figure C.9: An illustration of the sequences $\{s_n\}_{n\in\mathbb{N}}$ (shown in gray), $\{\mathfrak{p}_{\tilde{i}}(s_n)\}_{n\in\mathbb{N}}$ (red) and $\{\mathfrak{p}_{\partial\mathcal{F}_c}(s_n)\}_{n\in\mathbb{N}}$ (purple) for an example in which $\alpha = \{1\}$ and $\tilde{i} = 1$. Also displayed are dashed gray lines representing the line segments $\{L(s_n)\}_{n\in\mathbb{N}}$ and the point s^* in dark blue.

By setting $\nabla\mathcal{L}(s)$ equal to zero, we can see that $\mathfrak{p}_{\tilde{i}}(s_n)$ satisfies:

$$2(s_n - \mathfrak{p}_{\tilde{i}}(s_n)) = \lambda_n \nabla g^{\tilde{i}}(\mathfrak{p}_{\tilde{i}}(s_n)).$$

We note that λ_n must be non-negative as if it were negative, it would be possible to find a point $p \in L(s_n)$ which is closer to s_n than $\mathfrak{p}_{\tilde{i}}(s_n)$ such that $g^{\tilde{i}}(p) = 0$. The existence of such a point p would contradict the definition of $\mathfrak{p}_{\tilde{i}}(s_n)$. If λ_n were equal to zero, it would follow that $s_n = \mathfrak{p}_{\tilde{i}}(s_n)$. However, this cannot happen as, by construction, $s_n \in (\mathcal{F}_c)^\circ$ and $\mathfrak{p}_{\tilde{i}}(s_n) \notin (\mathcal{F}_c)^\circ$. Therefore, $\lambda_n > 0$.

We shall now show that for n large enough, $g^{\tilde{i}}$ is strictly increasing along $\vec{L}(s_n)$ from $\mathfrak{p}_{\tilde{i}}(s_n)$ to s_n. To show this, we proceed by proof by contradiction and assume the negation:

$$\forall N \in \mathbb{N}, \exists n \geq N \text{ such that } \exists c_n \in L(s_n) \text{ such that } g^{\tilde{i}}_{\vec{L}(s_n)}(c_n) \leq 0.$$

Let $\{c_{n_m}\}_{m\in\mathbb{N}}$ be a sequence satisfying the above. Trivially, as $c_{n_m} \in L(s_{n_m})$, it follows

that $c_{n_m} \to s^*$ as $m \to \infty$. By noting $\vec{L}(s_n)$ is a positive multiple of $\nabla g^{\tilde{i}}(\mathfrak{p}_{\tilde{i}}(s_n))$, it follows that;

$$g^{\tilde{i}}_{\nabla g^{\tilde{i}}(s^*)}(s^*) = \lim_{m\to\infty} g^{\tilde{i}}_{\nabla g^{\tilde{i}}(\mathfrak{p}_{\tilde{i}}(s_{n_m}))}(c_{n_m}) \leq 0.$$

However, this is a contradiction as;

$$g^{\tilde{i}}_{\nabla g^{\tilde{i}}(s^*)}(s^*) = \nabla g^{\tilde{i}}(s^*) \cdot \nabla g^{\tilde{i}}(s^*) > 0.$$

Therefore, for n large enough, $g^{\tilde{i}}$ is strictly increasing along $\vec{L}(s_n)$ from $\mathfrak{p}_{\tilde{i}}(s_n)$ to s_n and;

$$0 = g^{\tilde{i}}(\mathfrak{p}_{\tilde{i}}(s_n)) \leq g^{\tilde{i}}(\mathfrak{p}_{\partial\mathcal{F}_c}(s_n)) < g^{\tilde{i}}(s_n),$$

which implies;

$$0 < g^{\tilde{i}}(s_n) - g^{\tilde{i}}(\mathfrak{p}_{\partial\mathcal{F}_c}(s_n)) \leq g^{\tilde{i}}(s_n) - g^{\tilde{i}}(\mathfrak{p}_{\tilde{i}}(s_n)).$$

As $\mathcal{N}_2$ does not intersect $\overline{\partial^{\alpha_2}\mathcal{F}_c}$ for any $\alpha_2 \subseteq \alpha$, it follows that for n large enough:

$$\mathfrak{p}_{\partial\mathcal{F}_c}(s_n) \in \partial\mathcal{F}_c \setminus \bigcup_{\alpha_2\subseteq\alpha} \partial^{\alpha_2}\mathcal{F}_c = \bigcup_{\alpha_1\not\subseteq\alpha} \partial^{\alpha_1}\mathcal{F}_c.$$

For fixed n large enough, this means that $\mathfrak{p}_{\partial\mathcal{F}_c}(s_n) \in \partial^{\alpha_1}\mathcal{F}_c$ for some $\alpha_1 \not\subseteq \alpha$. As $\alpha_1 \not\subseteq \alpha$, it follows that $\alpha_1 \setminus \alpha$ is non-empty. Let $\tilde{j}_n \in \alpha_1 \setminus \alpha$. From the definition of $\partial^{\alpha_1}\mathcal{F}_c$, it now follows that $g^{\tilde{j}_n}(\mathfrak{p}_{\partial\mathcal{F}_c}(s_n)) = 0$.

By applying the generalized mean value theorem in the direction of $\vec{L}(s_n)$ between $\mathfrak{p}_{\partial\mathcal{F}_c}(s_n)$ and s_n, it can be seen that there must exist a point $t_n \in L(s_n)$ between $\mathfrak{p}_{\partial\mathcal{F}_c}(s_n)$ and s_n which satisfies;

$$\frac{g^{\tilde{j}_n}_{\vec{L}(s_n)}(t_n)}{g^{\tilde{i}}_{\vec{L}(s_n)}(t_n)} = \frac{g^{\tilde{j}_n}(s_n) - g^{\tilde{j}_n}(\mathfrak{p}_{\partial\mathcal{F}_c}(s_n))}{g^{\tilde{i}}(s_n) - g^{\tilde{i}}(\mathfrak{p}_{\partial\mathcal{F}_c}(s_n))} = \frac{g^{\tilde{j}_n}(s_n)}{g^{\tilde{i}}(s_n) - g^{\tilde{i}}(\mathfrak{p}_{\partial\mathcal{F}_c}(s_n))}$$

$$\geq \frac{g^{\tilde{j}_n}(s_n)}{g^{\tilde{i}}(s_n) - g^{\tilde{i}}(\mathfrak{p}_{\tilde{i}}(s_n))} = \frac{g^{\tilde{j}_n}(s_n)}{g^{\tilde{i}}(s_n)},$$

where the second and third equality follow by noting that $g^{\tilde{j}_n}(\mathfrak{p}_{\partial\mathcal{F}_c}(s_n)) = 0$ and

$g^{\tilde{i}}(\mathfrak{p}_{\tilde{i}}(s_n)) = 0$, whilst the inequality follows by the previous argument. The sequence $\{t_n\}_{n\in\mathbb{N}}$ is illustrated by Fig. C.10, alongside the sequences $\{s_n\}_{n\in\mathbb{N}}$ and $\{\mathfrak{p}_{\partial\mathcal{F}_c}(s_n)\}_{n\in\mathbb{N}}$.

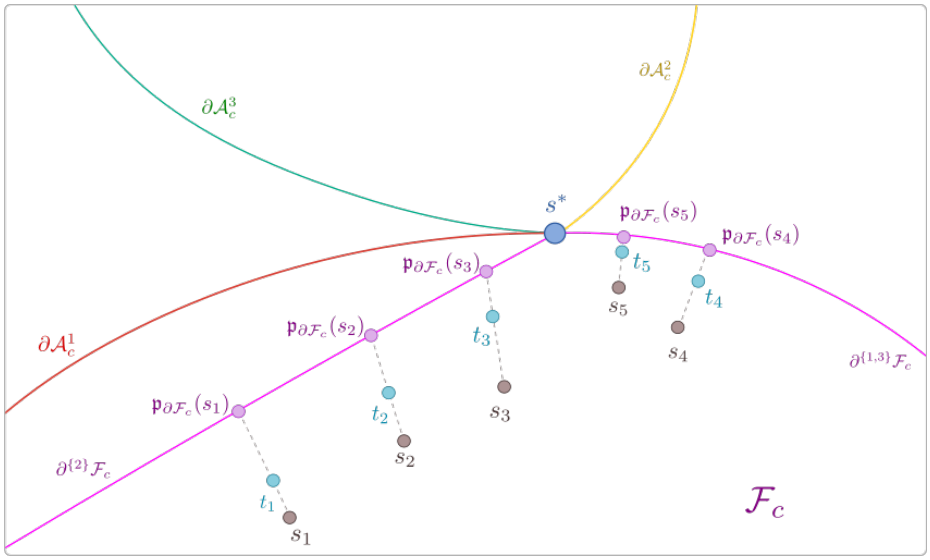

Figure C.10: An illustration of the sequences $\{s_n\}_{n\in\mathbb{N}}$ (shown in gray), $\{t_n\}_{n\in\mathbb{N}}$ (light blue) and $\{\mathfrak{p}_{\partial\mathcal{F}_c}(s_n)\}_{n\in\mathbb{N}}$ (purple) for an example in which $\alpha = \{1\}$ and $\tilde{i} = 1$. Also displayed are dashed gray lines representing segments of $\{L(s_n)\}_{n\in\mathbb{N}}$ and the point s^* in dark blue. In this example, a possible choice for the sequence $\{\tilde{j}_n\}_{n\in\mathbb{N}}$ is given by $\tilde{j}_1 = \tilde{j}_2 = \tilde{j}_3 = 2$ and $\tilde{j}_4 = \tilde{j}_5 = 3$.

We have shown that for n sufficiently large, there must exist an $\tilde{i} \in \alpha$ and $\tilde{j}_n \in \mathcal{M} \setminus \alpha$ such that;

$$\frac{g^{\tilde{j}_n}_{\vec{L}(s_n)}(t_n)}{g^{\tilde{i}}_{\vec{L}(s_n)}(t_n)} \geq \frac{g^{\tilde{j}_n}(s_n)}{g^{\tilde{i}}(s_n)}.$$

It follows that for n sufficiently large, there exists an $\tilde{i} \in \alpha$ and $\tilde{j}_n \in \alpha_1 \setminus \alpha \subseteq \mathcal{M} \setminus \alpha$ such that;

$$\frac{g^{\tilde{j}_n}_{\vec{L}(s_n)}(t_n)}{g^{\tilde{i}}_{\vec{L}(s_n)}(t_n)} \geq \min_{\substack{i\in\alpha \\ j\in\mathcal{M}\setminus\alpha}} \left(\frac{g^j(s_n)}{g^i(s_n)}\right) = a_n > 0,$$

where the second inequality follows from the fact that $s_n \in (\mathcal{F}_c)^\circ$. As $\tilde{j}_n$ is dependent on n, we note that:

$$\max_{j\in\mathcal{M}\setminus\alpha} \left(\frac{g^j_{\vec{L}(s_n)}(t_n)}{g^{\tilde{i}}_{\vec{L}(s_n)}(t_n)}\right) \geq a_n > 0.$$

Now we note that, by construction $t_n \to s^*$ as $n \to \infty$ (c.f. Fig. C.10). Therefore, the below convergence holds:

$$\max_{j \in \mathcal{M} \setminus \alpha} \left(\frac{g^j_{\vec{L}(s_n)}(t_n)}{g^{\tilde{i}}_{\vec{L}(s_n)}(t_n)} \right) \to \max_{j \in \mathcal{M} \setminus \alpha} \left(\frac{\nabla g^{\tilde{i}}(s^*) \cdot \nabla g^j(s^*)}{||\nabla g^{\tilde{i}}(s^*)||^2} \right) \text{ as } n \to \infty.$$

As $\nabla g^{\tilde{i}}(s^*)$ is non-zero by assumption the above limit is finite and it follows that we have bounded $\{a_n\}_{n \in \mathbb{N}}$ above and below with convergent sequences.

□

C.2.8 Theorem C.5: Equality of Maxima Involving $\mathcal{D}_\alpha$

Theorem C.5. *Let $\tilde{\mathcal{D}}_\alpha$ be an arbitrary set satisfying $\tilde{\mathcal{D}}_\alpha \subseteq \mathcal{D}_\alpha$. For n large enough, the following equality is true:*

$$\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \partial^\alpha \mathcal{F}_c} \left| \min_{i \in \alpha}(G^i_n(s)) \right| \right) = \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \partial^\alpha \mathcal{F}_c \cup \tilde{\mathcal{D}}_\alpha} \left| \min_{i \in \alpha}(G^i_n(s)) \right| \right),$$

where $G^i_n(s)$ is shorthand for $\hat{g}^i_n(s) - g^i(s)$.

Proof. To begin, note that the right hand side of the statement in Theorem C.5 is equal to:

$$\max \left[\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \tilde{\mathcal{D}}_\alpha} \left| \min_{i \in \alpha}(G^i_n(s)) \right| \right), \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \partial^\alpha \mathcal{F}_c} \left| \min_{i \in \alpha}(G^i_n(s)) \right| \right) \right].$$

It therefore follows that, in order to prove the result of the theorem, we need only show that:

$$\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \tilde{\mathcal{D}}_\alpha} \left| \min_{i \in \alpha}(G^i_n(s)) \right| \right) \leq \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \partial^\alpha \mathcal{F}_c} \left| \min_{i \in \alpha}(G^i_n(s)) \right| \right). \tag{C.13}$$

To show the above inequality is true, we begin by noting that, as $\tilde{\mathcal{D}}_\alpha \subseteq \mathcal{D}_\alpha$:

$$\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \tilde{\mathcal{D}}_\alpha} \left| \min_{i \in \alpha} (G_n^i(s)) \right| \right) \leq \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \mathcal{D}_\alpha} \left| \min_{i \in \alpha} (G_n^i(s)) \right| \right).$$

Now, let $\alpha^* \in \mathcal{P}^+(\mathcal{M})$ and $s^* \in \mathcal{D}_{\alpha^*}$ be the values which maximize the expression on the right-hand side of the above. In other words, let α^* and s^* be the values of α and s which satisfy $s^* \in \mathcal{D}_{\alpha^*}$ and:

$$\max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \mathcal{D}_\alpha} \left| \min_{i \in \alpha} (G_n^i(s)) \right| \right) = \left| \min_{i \in \alpha^*} (G_n^i(s^*)) \right|.$$

Now, as $s^* \in \mathcal{D}_{\alpha^*}$, we know that there must exist an α_1 and α_2 such that $\alpha_1 \subset \alpha^* \subset \alpha_2$, $s^* \in \overline{\partial^{\alpha_1} \mathcal{F}_c}$ and $s^* \in \overline{\partial^{\alpha_2} \mathcal{F}_c}$. Fix such α_1 and α_2 and note that, from trivial properties of the minimum function and the fact that $\alpha_1 \subset \alpha^* \subset \alpha_2$, it must be true that either:

$$\left| \min_{i \in \alpha^*} (G_n^i(s^*)) \right| \leq \left| \min_{i \in \alpha_1} (G_n^i(s^*)) \right| \text{ or } \left| \min_{i \in \alpha^*} (G_n^i(s^*)) \right| \leq \left| \min_{i \in \alpha_2} (G_n^i(s^*)) \right|.$$

In other words;

$$\left| \min_{i \in \alpha^*} (G_n^i(s^*)) \right| \leq \max \left(\left| \min_{i \in \alpha_1} (G_n^i(s^*)) \right|, \left| \min_{i \in \alpha_2} (G_n^i(s^*)) \right| \right) = \max_{\alpha \in \{\alpha_1, \alpha_2\}} \left(\left| \min_{i \in \alpha} (G_n^i(s^*)) \right| \right).$$

As, by the construction of $\mathcal{D}_{\alpha^*}$, we know that $s^* \in \overline{\partial^{\alpha_1} \mathcal{F}_c}$ and $s^* \in \overline{\partial^{\alpha_2} \mathcal{F}_c}$, it therefore follows that the above is less than or equal to:

$$\max_{\alpha \in \{\alpha_1, \alpha_2\}} \left(\sup_{s \in \overline{\partial^{\alpha} \mathcal{F}_c}} \left| \min_{i \in \alpha} (G_n^i(s)) \right| \right).$$

As all functions $\{G_n^i\}_{i \in \mathcal{M}}$ are continuous for n large enough it follows that suprema taken over $\partial^\alpha \mathcal{F}_c$ attain their limit points, and therefore the above is equal to;

$$\max_{\alpha \in \{\alpha_1, \alpha_2\}} \left(\sup_{s \in \partial^{\alpha} \mathcal{F}_c} \left| \min_{i \in \alpha} (G_n^i(s)) \right| \right) \leq \max_{\alpha \in \mathcal{P}^+(\mathcal{M})} \left(\sup_{s \in \partial^{\alpha} \mathcal{F}_c} \left| \min_{i \in \alpha} (G_n^i(s)) \right| \right).$$

This is the expression we set out to prove. We have now shown that:

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}\left(\sup_{s\in\tilde{\mathcal{D}}_\alpha}\left|\min_{i\in\alpha}(G_n^i(s))\right|\right)\leq\left|\min_{i\in\alpha^*}(G_n^i(s^*))\right|\leq\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}\left(\sup_{s\in\partial^\alpha\mathcal{F}_c}\left|\min_{i\in\alpha}(G_n^i(s))\right|\right),$$

as required.

□

C.2.9 Theorem C.6: Convergence of $\mathcal{O}_c^{n,\alpha}$

Theorem C.6. *For all $\alpha\in\mathcal{P}^+(\mathcal{M})$, there exists a set $\tilde{\mathcal{D}}_\alpha\subseteq\mathcal{D}_\alpha$ such that $d_H(\mathcal{O}_c^{n,\alpha},\partial^\alpha\mathcal{F}_c\cup\tilde{\mathcal{D}}_\alpha)\to 0$ as $n\to\infty$.*

Proof. Suppose $\alpha=\mathcal{M}$. As $\mathcal{D}_\mathcal{M}=\emptyset$, the only way for the theorem to be satisfied is if $d_H(\mathcal{O}_c^{n,\mathcal{M}},\partial^\mathcal{M}\mathcal{F}_c)\to 0$ as $n\to\infty$. As $\{\mathcal{O}_c^{n,\mathcal{M}}\}_{n\in\mathbb{N}}$ are strictly decreasing it is true that;

$$d_H\left(\mathcal{O}_c^{n,\mathcal{M}},\bigcap_{m\geq 1}\overline{\mathcal{O}_c^{m,\mathcal{M}}}\right)\to 0\text{ as }n\to\infty.$$

We shall now show that $\partial^\mathcal{M}\mathcal{F}_c=\cap_{m\geq 1}\overline{\mathcal{O}_c^{m,\mathcal{M}}}$. Let $s^*\in\cap_{m\geq 1}\overline{\mathcal{O}_c^{m,\mathcal{M}}}$. It follows that there must exist a sequence $\{s_m\}_{m\in\mathbb{N}}$ such that $s_m\in\overline{\mathcal{O}_c^{m,\mathcal{M}}}$ and $s_m\to s^*$ as $m\to\infty$. As $\overline{\mathcal{O}_c^{m,\mathcal{M}}}\subseteq\mathcal{F}_c^m$, it follows that $s_m\in\mathcal{F}_c^m$ and therefore;

$$|\min_{i\in\mathcal{M}}g^i(s^*)|=\lim_{m\to\infty}|\min_{i\in\mathcal{M}}g^i(s_m)|\leq\lim_{m\to\infty}\eta_m=0.$$

Therefore, $\min_{i\in\mathcal{M}}g^i(s^*)=0$ and, as a result, for all $i\in\mathcal{M}$, $g^i(s^*)\geq 0$. As $\overline{\mathcal{O}_c^{m,\mathcal{M}}}\subseteq\cap_{i\in\mathcal{M}}\overline{(\mathcal{A}_c^i)^c}$, it also follows that $s_m\in\cap_{i\in\mathcal{M}}\overline{(\mathcal{A}_c^i)^c}$. Therefore, $g^i(s_m)\leq 0$ for all $i\in\mathcal{M}$, and;

$$g^i(s^*)=\lim_{m\to\infty}g^i(s_m)\leq 0.$$

Therefore $g^i(s^*) = 0$ for all $i \in \mathcal{M}$ and, thus, $s^* \in \partial^{\mathcal{M}}\mathcal{F}_c$. As a result, $\cap_{m\geq 1}\overline{\mathcal{O}_c^{m,\mathcal{M}}} \subseteq \partial^{\mathcal{M}}\mathcal{F}_c$.

Let $t^* \in \partial^{\mathcal{M}}\mathcal{F}_c$ and fix $m \in \mathbb{N}$. As $\min_{i\in\mathcal{M}} g^i(t^*) = 0$, we can define a neighbourhood of t^* over which $|\min_{i\in\mathcal{M}} g^i| < \eta_m$. Such a neighbourhood must be a subset of $\mathcal{F}_c^m$ by definition. By the ball assumption, any neighbourhood of t^* must intersect $\mathcal{J}_c^{\mathcal{M}} = \cap_{i\in\mathcal{M}}(\mathcal{A}_c^i)^c$. Therefore, for arbitrary m, every neighbourhood of t^* intersects $\cap_{i\in\mathcal{M}}(\mathcal{A}_c^i)^c \cap \mathcal{F}_c^m = \mathcal{O}_c^{m,\mathcal{M}}$, and thus $t^* \in \cap_{m\geq 1}\overline{\mathcal{O}_c^{m,\mathcal{M}}}$. Therefore, $\partial^{\mathcal{M}}\mathcal{F}_c \subseteq \cap_{m\geq 1}\overline{\mathcal{O}_c^{m,\mathcal{M}}}$. Combining this with the previous, it follows that $\partial^{\mathcal{M}}\mathcal{F}_c = \cap_{m\geq 1}\overline{\mathcal{O}_c^{m,\mathcal{M}}}$, as desired.

Suppose $\alpha \neq \mathcal{M}$. Let $s^* \in \partial^\alpha\mathcal{F}_c$ and fix $n \in \mathbb{N}$. By the definition of $\partial^\alpha\mathcal{F}_c$, $g^i(s^*) = 0$ for all $i \in \alpha$ and $g^j(s^*) > 0$ for all $j \in \mathcal{M} \setminus \alpha$. Therefore, for ϵ small enough, the open ϵ-ball about s^*, $B_\epsilon(s^*)$, satisfies the below:

$$\forall s \in B_\epsilon(s^*), g^j(s) > 0 \text{ and } -\eta_n < g^i(s) < \eta_n,$$

for all $i \in \alpha$ and $j \in \mathcal{M} \setminus \alpha$. By the ball assumption, $B_\epsilon(s^*) \cap \mathcal{J}_c^\alpha$ is non-empty. Therefore, by the definition of $\mathcal{J}_c^\alpha$, for ϵ small enough, there must exist an $s \in B_\epsilon(s^*) \cap \mathcal{J}_c^\alpha$ such that $-\eta_n < g^i(s) < 0 < g^j(s)$ for all $i \in \alpha$ and $j \in \mathcal{M} \setminus \alpha$. For such s it follows that $|\min_{i\in\mathcal{M}} g^i(s)| < \eta_n$ and therefore $s \in \mathcal{J}_c^\alpha \cap \mathcal{F}_c^n \subseteq \mathcal{O}_c^{n,\alpha}$. Thus, $s \in \mathcal{O}_c^{n,\alpha}$. As $\epsilon > 0$ was arbitrary, it follows that $s^* \in \overline{\mathcal{O}_c^{n,\alpha}}$. As $s^* \in \partial^\alpha\mathcal{F}_c$ was arbitrary, it follows that $\partial^\alpha\mathcal{F}_c \subseteq \overline{\mathcal{O}_c^{n,\alpha}}$. As n was arbitrary, $\partial^\alpha\mathcal{F}_c \subseteq \cap_{n\geq 1}\overline{\mathcal{O}_c^{n,\alpha}}$.

Define $\tilde{\mathcal{D}}_\alpha$ as $\tilde{\mathcal{D}}_\alpha = \bigcap_{n\geq 1}\overline{\mathcal{O}_c^{n,\alpha}} \setminus \overline{\partial^\alpha\mathcal{F}_c}$. It now follows that $\overline{\partial^\alpha\mathcal{F}_c} \cup \tilde{\mathcal{D}}_\alpha = \bigcap_{n\geq 1}\overline{\mathcal{O}_c^{n,\alpha}}$. As $\{\overline{\mathcal{O}_c^{n,\alpha}}\}_{n\in\mathbb{N}}$ is sequence of closed decreasing sets, it now follows that $d_H(\mathcal{O}_c^{n,\alpha}, \partial^\alpha\mathcal{F}_c \cup \tilde{\mathcal{D}}_\alpha) \to 0$ as $n \to \infty$. We must now show that $\tilde{\mathcal{D}}_\alpha \subseteq \mathcal{D}_\alpha$. By the definition of $\mathcal{D}_\alpha$ (c.f. Section C.2.6), this means we must show that, for arbitrary $s^* \in \tilde{\mathcal{D}}_\alpha$, there exists α_1 and α_2 such that $\alpha_1 \subset \alpha \subset \alpha_2$ and $s^* \in \overline{\partial^{\alpha_1}\mathcal{F}_c} \cap \overline{\partial^{\alpha_2}\mathcal{F}_c}$. In other words, we will prove the following two statements:

$$\exists \alpha_2 \supset \alpha \text{ such that } s^* \in \overline{\partial^{\alpha_2}\mathcal{F}_c}, \tag{C.14}$$

$$\exists \alpha_1 \subset \alpha \text{ such that } s^* \in \overline{\partial^{\alpha_1} \mathcal{F}_c}. \tag{C.15}$$

It can be seen that, as $\tilde{\mathcal{D}}_\alpha \subseteq \cap_{n\geq 1} \overline{\mathcal{O}_c^{n,\alpha}}$, it follows that $\tilde{\mathcal{D}}_\alpha \subseteq \partial \mathcal{F}_c$ and therefore $s^* \in \partial \mathcal{F}_c$. To show Statement (C.14), we note that, by the ball assumption, any open neighbourhood of s^* has a non-empty intersection with $(\mathcal{F}_c)^\circ$. Further, by the construction of $\tilde{\mathcal{D}}_\alpha$, it follows that any open neighbourhood of s^* has a non-empty intersection with $\mathcal{O}_c^{n,\alpha}$ for arbitrary n. Fix n. By the definitions of $(\mathcal{F}_c)^\circ$ and $\mathcal{O}_c^{n,\alpha}$, it now follows that for any $i \in \alpha$;

$$\forall \epsilon > 0, \quad B_\epsilon(s^*) \cap \{s \in S : g^i(s) > 0\} \neq \emptyset \quad \text{and} \quad B_\epsilon(s^*) \cap \{s \in S : g^i(s) < 0\} \neq \emptyset.$$

As $B_\epsilon(s^*)$ is path connected, it follows that for all $i \in \alpha$ and $\epsilon > 0$, there exists an $s \in B_\epsilon(s^*)$ such that $g^i(s) = 0$. As ϵ is arbitrary, it now follows that $g^i(s^*) = 0$ for all $i \in \alpha$. Therefore, by the definition of $\partial^\alpha \mathcal{F}_c$, $s^* \in \partial^{\alpha_2} \mathcal{F}_c$ for some $\alpha_2 \supseteq \alpha$. As by construction $\tilde{\mathcal{D}}_\alpha \cap \overline{\partial^\alpha \mathcal{F}_c} = \emptyset$, it follows that $s^* \in \overline{\partial^{\alpha_2} \mathcal{F}_c}$ for some $\alpha_2 \supset \alpha$. This concludes the proof of Statement (C.14).

To prove Statement (C.15), we recall from the previous argument $s^* \in \partial^{\alpha_2} \mathcal{F}_c$ for some $\alpha_2 \supset \alpha$. From the definition of $\partial^{\alpha_2} \mathcal{F}_c$, it is true that for all k in $\mathcal{M} \setminus \alpha_2$, $g^k(s^*) > 0$. Therefore, we can choose $\epsilon > 0$ small enough such that;

$$\forall s \in B_\epsilon(s^*), \forall k \in \mathcal{M} \setminus \alpha_2, g^k(s) > 0. \tag{C.16}$$

Similarly, from the definition of $\partial^{\alpha_2} \mathcal{F}_c$, it is true that for all $i \in \alpha$, $g^i(s^*) = 0$. Therefore, for arbitrary n and ϵ small enough;

$$\forall s \in B_\epsilon(s^*), \forall i \in \alpha, |g^i(s)| < \eta_n. \tag{C.17}$$

Fix $n \in \mathbb{N}$ and let ϵ satisfy the above. As $s^* \in \partial^{\alpha_2} \mathcal{F}_c$, $g^j(s^*) = 0$ for all $j \in \alpha_2 \setminus \alpha$. Therefore, $s^* \in S_0 := \{s \in B_\epsilon(s^*) | \min_{j \in \alpha_2 \setminus \alpha}(g^j(s)) = 0\}$. By the path-connectedness

assumption, for ϵ small enough, S_0 must partition $B_\epsilon(s^*)$ into two path-connected components; S_1 and S_2.

As S_1 and S_2 are path-connected, $\min_{j\in\alpha_2\setminus\alpha} g^j$ cannot change sign within S_1 or within S_2. By the ball assumption, $B_\epsilon(s^*)$ must intersect $(\mathcal{F}_c)^\circ$ (where $\min_{j\in\alpha_2\setminus\alpha} g^j > 0$) and $\mathcal{J}_c^{\alpha_2}$ (where $\min_{j\in\alpha_2\setminus\alpha} g^j < 0$). It therefore follows that in one of S_1 or S_2, $\min_{j\in\alpha_2\setminus\alpha} g^j < 0$, whilst in the other set $\min_{j\in\alpha_2\setminus\alpha} g^j > 0$. Without loss of generality, assume that:

$$S_1 = \left\{ s \in B_\epsilon(s^*) \middle| \min_{j\in\alpha_2\setminus\alpha} (g^j(s)) > 0 \right\} \quad \text{and} \quad S_2 = \left\{ s \in B_\epsilon(s^*) \middle| \min_{j\in\alpha_2\setminus\alpha} (g^j(s)) < 0 \right\}.$$

By construction $s^* \in \overline{\mathcal{O}_c^{m,\alpha}}$ for arbitrary m and, therefore, $B_\epsilon(s^*) \cap \mathcal{O}_c^{n,\alpha}$ is non-empty. Suppose $B_\epsilon(s^*) \cap \mathcal{O}_c^{n,\alpha} \subseteq S_0$ and let $t^* \in B_\epsilon(s^*) \cap \mathcal{O}_c^{n,\alpha}$. By the definition of $\mathcal{O}_c^{n,\alpha}$, $g^i(t^*) < 0$ for all $i \in \alpha$. Therefore, we can choose $\delta > 0$ such that $B_\delta(t^*) \subset B_\epsilon(s^*)$ and for all $t \in B_\delta(t^*)$, $g^i(t) < 0$ for all $i \in \alpha$. This construction is illustrated by Fig. C.11.

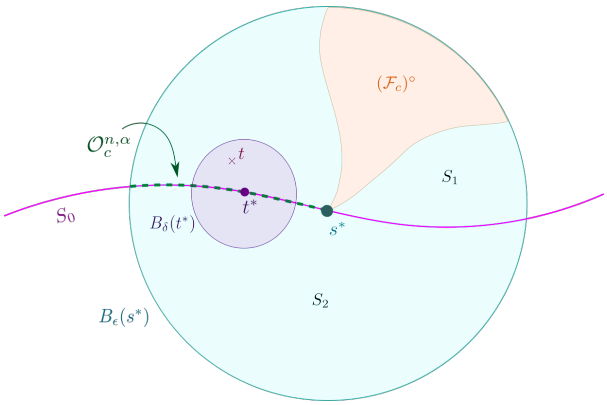

Figure C.11: Illustration of the situation in which $B_\epsilon(s^*) \cap \mathcal{O}_c^{n,\alpha} \subseteq S_0$. Displayed are the sets S_0 (light purple), $B_\epsilon(s^*)$ (light blue), $\mathcal{O}_c^{n,\alpha}$ (dashed green), $(\mathcal{F}_c)^\circ$ (orange) and $B_\delta(t^*)$ (dark purple). s^* and t^* are labelled with circles and t is depicted as a red cross. Also labelled in black are the regions S_1 and S_2.

As $B_\delta(t^*) \cap S_1$ is non-empty (this can be seen by noting that S_0 is composed of segments of C^1 curves, and that S_1 is non-empty), we can fix $t \in B_\delta(t^*) \cap S_1$. By the construction of $B_\delta(t^*)$, $g^i(t) < 0$ for all $i \in \alpha$. By the definition of S_1, $g^j(t) > 0$ for all $j \in \alpha_2 \setminus \alpha$. By the definition of $B_\epsilon(s^*)$, $g^k(t) > 0$ for all $k \in \mathcal{M} \setminus \alpha_2$ and

$-\eta_n < g^i(t)$ for all $i \in \alpha$ (c.f. Statements (C.16) and (C.17)). By combining these facts and noting the definition of $\mathcal{O}_c^{n,\alpha}$, it can now be seen that $t \in \mathcal{O}_c^{n,\alpha}$.

Therefore, $t \in B_\epsilon(s^*) \cap \mathcal{O}_c^{n,\alpha} \subseteq S_0$ and $t \in S_1$. This is a contradiction as S_0 and S_1 are disjoint. Therefore, it cannot be true that $B_\epsilon(s^*) \cap \mathcal{O}_c^{n,\alpha} \subseteq S_0$. By construction, $\mathcal{O}_c^{n,\alpha} \cap S_2$ is empty so it must be that $\mathcal{O}_c^{n,\alpha} \cap S_1$ is non-empty. By the ball assumption, $B_\epsilon(s^*) \cap (\mathcal{F}_c)^\circ$ is non-empty and therefore, by the definitions of S_0, S_1 and S_2, $(\mathcal{F}_c)^\circ \cap S_1$ is non-empty.

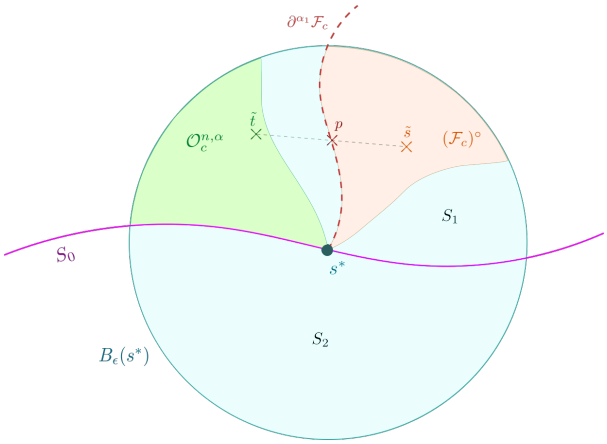

Figure C.12: Illustration of the situation in which $B_\epsilon(s^*) \cap \mathcal{O}_c^{n,\alpha} \not\subseteq S_0$. Displayed are the sets S_0 (light purple), $B_\epsilon(s^*)$ (light blue), $\mathcal{O}_c^{n,\alpha}$ (green), $(\mathcal{F}_c)^\circ$ (orange) and $\partial^{\alpha_1}\mathcal{F}_c$ (red dashed). Labelled with crosses are the points $\tilde{s}$, p and $\tilde{t}$, with the path from $\tilde{s}$ to $\tilde{t}$ displayed as a grey dotted line. Also labelled in black are the regions S_1 and S_2.

Let $\tilde{s} \in S_1 \cap (\mathcal{F}_c)^\circ$ and $\tilde{t} \in S_1 \cap \mathcal{O}_c^{n,\alpha}$. As S_1 is path connected, we can define a path from $\tilde{s}$ to $\tilde{t}$ which does not leave S_1. By the definitions of $(\mathcal{F}_c)^\circ$ and $\mathcal{O}_c^{n,\alpha}$, for all $i \in \alpha$, $g^i(\tilde{s}) > 0$ and $g^i(\tilde{t}) < 0$. Therefore, along the path from $\tilde{s}$ to $\tilde{t}$ there must exist at least one point at which $g^i = 0$ for at least one $i \in \alpha$. Let p be the first such point encountered along the path when moving from $\tilde{s}$ to $\tilde{t}$. This construction is illustrated by Fig. C.12.

By the definition of S_1, $g^j(p) > 0$ for all $j \in \alpha_2 \setminus \alpha$ and, by Statement (C.16), $g^k(p) > 0$ for all $k \in \mathcal{M} \setminus \alpha_2$. Therefore, $g^j(p) > 0$ for all $j \in \mathcal{M} \setminus \alpha$. Furthermore, $p \in \partial\mathcal{F}_c$ as it is the first point along the path from $\tilde{s}$ to $\tilde{t}$ which does not belong to $(\mathcal{F}_c)^\circ$. As $p \in \partial\mathcal{F}_c$ but $g^j(p) > 0$ for all $j \in \mathcal{M} \setminus \alpha$, it follows that $p \in \partial^{\alpha_1}\mathcal{F}_c$ for some $\alpha_1 \subseteq \alpha$. As such a point, p, can be found for any sufficiently small neighbourhood

of s^*, we can define a sequence $\{p_n\}_{n\in\mathbb{N}}$ such that $p_n \to s^*$ as $n \to \infty$ and each $p_n \in \partial^{\alpha_1}\mathcal{F}_c$ for some $\alpha_1 \subseteq \alpha$.

As there are only finitely many $\alpha_1 \subseteq \alpha$, for some fixed $\alpha_1 \subseteq \alpha$, $\{p_n\}_{n\in\mathbb{N}}$ must contain a subsequence $\{\tilde{p}_n\}_{n\in\mathbb{N}}$ which tends to s^* such that all $\tilde{p}_n \in \partial^{\alpha_1}\mathcal{F}_c$. As we have identified a sequence $\{\tilde{p}_n\}_{n\in\mathbb{N}} \in \partial^{\alpha_1}\mathcal{F}_c$ which tends to s^*, it follows that $s^* \in \overline{\partial^{\alpha_1}\mathcal{F}_c}$ for some $\alpha_1 \subseteq \alpha$. As by construction $s^* \notin \overline{\partial^{\alpha}\mathcal{F}_c}$, Statement (C.15) now follows. □

C.2.10 Theorem C.7: Convergence of $\max(W_\alpha^n) - \max(X_\alpha^n)$

Theorem C.7. *Defining W_α^n and X_α^n as in Section C.2.2, it is true that:*

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(W_\alpha^n) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(X_\alpha^n) \xrightarrow{p} 0.$$

Proof. To begin, we note that, by a similar argument to that employed in Section C.2.5, it is true that $W_{\mathcal{M}}^n = X_{\mathcal{M}}^n$ and therefore, $W_{\mathcal{M}}^n - X_{\mathcal{M}}^n \xrightarrow{p} 0$ as $n \to \infty$ trivially. Suppose $\alpha \neq \mathcal{M}$. By Lemma C.4, it is true that:

$$|W_\alpha^n - X_\alpha^n| \leq \sup_{s\in\mathcal{F}_c^{-n,\alpha}} \tau_n^{-1} \left| \min_{i\in\mathcal{M}}(\hat{g}_n^i(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s)) \right|$$

$$= \sup_{s\in\mathcal{F}_c^{-n,\alpha}} \tau_n^{-1} \left| \min\left(\min_{i\in\alpha}(\hat{g}_n^i(s)), \min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s)) \right) - \min_{i\in\alpha}(\hat{g}_n^i(s)) \right|$$

$$= \sup_{s\in\mathcal{F}_c^{-n,\alpha}} \tau_n^{-1} \left| \min\left(0, \min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s)) \right) \right|,$$

where the first equality is due to the fact that, by the definition of $\mathcal{F}_c^{-n,\alpha}$, $g^i(s) = 0$ for all $i \in \alpha$ and $s \in \mathcal{F}_c^{-n,\alpha}$. By repeating the argument of Section C.2.5 and noting that, by definition, for all $s \in \mathcal{F}_c^{-n,\alpha}$ and $j \in \mathcal{M} \setminus \alpha$ it is true that $g^j(s) > \eta_n$, it now

follows that for arbitrary $\epsilon > 0$:

$$\mathbb{P}[|W_\alpha^n - X_\alpha^n| \geq \epsilon] \leq \mathbb{P}\left[\sup_{s\in\mathcal{F}_c^{-n,\alpha}} \tau_n^{-1}\left|\min\left(0, \min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s))\right)\right| \geq \epsilon\right]$$

$$\leq \mathbb{P}\left[\tau_n^{-1} \inf_{s\in\mathcal{F}_c^{-n,\alpha}}\left(\min_{j\in\mathcal{M}\setminus\alpha}(\hat{g}_n^j(s) - g^j(s)) - \min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\right) \leq -\epsilon - \tau_n^{-1}\eta_n\right].$$

Through similar logic to that employed in the proof of Theorem C.2, it now follows that the above probability tends to zero as $n \to \infty$. Therefore, for all $\alpha \in \mathcal{P}^+(\mathcal{M})$, $W_\alpha^n - X_\alpha^n \xrightarrow{p} 0$ as $n \to \infty$. By applying Lemma C.5, the result of the theorem now follows. □

C.2.11 Theorem C.8: Convergence of $\max(X_\alpha^n) - \max(Y_\alpha^n)$

Theorem C.8. *Defining X_α^n and Y_α^n as in Section C.2.2:*

$$\max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(X_\alpha^n) - \max_{\alpha\in\mathcal{P}^+(\mathcal{M})}(Y_\alpha^n) \xrightarrow{p} 0.$$

Proof. By the definitions of X_α^n and Y_α^n it follows that;

$$|X_\alpha^n - Y_\alpha^n| = \left|\sup_{s\in\mathcal{F}_c^{-n,\alpha}}\left|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\right| - \sup_{s\in\partial^\alpha\mathcal{F}_c}\left|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))\right|\right|$$

$$\leq \omega\left(|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))|, \delta_{-n}^\alpha\right),$$

where ω is the modulus of continuity (c.f. Lemma C.7) and δ_{-n}^α is the Hausdorff distance between $\mathcal{F}_c^{-n,\alpha}$ and $\partial^\alpha\mathcal{F}_c$. It therefore follows that, for arbitrary $\epsilon > 0$:

$$\mathbb{P}[|X_\alpha^n - Y_\alpha^n| \geq \epsilon] \leq \mathbb{P}\left[\omega\left(|\tau_n^{-1}\min_{i\in\alpha}(\hat{g}_n^i(s) - g^i(s))|, \delta_{-n}^\alpha\right) \geq \epsilon\right].$$

By applying limits to both sides of the above and noting the result of Lemma C.7, it now follows that $X_\alpha^n - Y_\alpha^n \xrightarrow{p} 0$ if $\delta_{-n}^\alpha \to 0$ as $n \to \infty$.

However, it may be seen from the definitions of $\mathcal{F}_c^{-n,\alpha}$ and $\partial^\alpha \mathcal{F}_c$ that the sets $\{\mathcal{F}_c^{-n,\alpha}\}_{n\in\mathbb{N}}$ are strictly increasing with $\cup_{n\in\mathbb{N}}\mathcal{F}_c^{-n,\alpha} = \partial^\alpha \mathcal{F}_c$. From standard properties of the Hausdorff distance, it now follows that $\delta_{-n}^\alpha \to 0$ as $n \to \infty$. All that remains is to apply Lemma C.5 to obtain the result of the theorem. □

C.3 Linear Model Assumptions

In this section, we provide additional assumptions which allow the methods of Section 5.2.3 to be applied in the linear modelling context of Section 5.3.1. The assumptions provided in this section were first described by Sommerfeld et al. (2018) and the reader is referred to this canonical text for further detail. To state the assumptions, we must introduce some additional notation.

We denote the $p-$norm and $\infty-$norm of a matrix A as $\|A\|_p = \sup_{\|x\|_p=1} \|Ax\|_p$ and $\|A\|_\infty = \max_i \sum_j |a_{ij}|$, where $\|x\|_p := (\sum_i |x_i|^p)^{\frac{1}{p}}$ is the standard vector $p-$norm and a_{ij} is the $(i,j)^{th}$ element of the matrix A, respectively. For spatial sets $B \subseteq S$, we let $|B|$ denote the Lebesgue measure of B and, for $s, t \in \mathbb{R}^N$ we define the block indexed by $s = (s_1, ..., s_n)$ and $t = (t_1, ..., t_n)$ as $(s, t] = (s_1, t_1] \times \cdots \times (s_N, t_N] \subset \mathbb{R}^N$. For a stochastic process $\delta(s)$ with index set containing $(s, t]$ we now define the increment of $\delta(s)$ across $(s, t]$ as follows:

$$\delta((s,t]) = \sum_{\kappa_1=0,1} \dots \sum_{\kappa_N=0,1} (-1)^{N-\sum_j \kappa_j} \delta(s_1 + \kappa_1(t_1 - s_1), \cdots, s_N + \kappa_N(t_N - s_N)).$$

For further discussion of the above definition of increment see Bickel and Wichura (1971) and Sommerfeld et al. (2018).

For each study condition $i \in \mathcal{M}$, we now assume that there exists a positive-definite spatial correlation function $\mathfrak{c}^i$: $S \times S \to [-1, 1]$ defined by the following

relationship:

$$\text{cov}\left(\epsilon^i(s), \epsilon^i(t)\right) = \mathfrak{c}^i(s,t)\left(\Sigma^i(s)\Sigma^i(t)\right)^{\frac{1}{2}} \text{ for all } s,t \in S,$$

for some appropriately defined notion of matrix square root. For each $i \in \mathcal{M}$, we also denote $Z^i(s) = \Sigma^i(s)^{-1/2}X^i(X^{i'}\Sigma^i(s)^{-1}X^i)^{-1/2}$. We are now in a position to state the assumptions which allow the theory of Section 5.2.3 to be applied to the linear modelling context described in Section 5.3.1 of the main text.

Assumption C.1. We assume that, for all $i \in \mathcal{M}$:

(a) The noise field ϵ^i has continuous sample paths with probability one. In addition, a centered unit variance Gaussian random field exists which possesses the same spatial correlation function, $\mathfrak{c}^i$, as the noise field ϵ^i and continuous sample paths with probability one.

(b) The function $\Sigma^i(s) : S \to \mathbb{R}^{n\times n}$ is continuous.

(c) The decorrelated error field for study condition i, $\tilde{\epsilon}^i(s) = \Sigma^i(s)^{-\frac{1}{2}}\epsilon^i(s)$, has independent components.

(d) There exists $\delta > 0$ and $K_\delta > 0$ such that $\sup_{s\in S}\mathbb{E}\left[|\tilde{\epsilon}^i_j(s)|^{2+\delta}\right] \leq K_\delta$ (independent of j and n), and $\sup_{s\in S}\left(n\|Z^i(s)\|_\infty^{2+\delta}\right) \to 0$ as $n \to \infty$.

(e) There exists $\gamma \geq 0$ and $\beta > 0$ such that for some constant $K_{(\beta,\gamma)} > 0$, it is true for any block $B \subseteq S$ that $\mathbb{E}\left[\|[Z^i(\cdot)\epsilon^i(\cdot)](B)\|_1^{2+\gamma}\right] \leq K_{(\beta,\gamma)}|B|^{1+\beta}$.

(f) The function $v^i_n(s) = L^{i'}(X^{i'}\Sigma^i(s)^{-1}X^i)^{-\frac{1}{2}}$ satisfies $\tau_n\frac{v^i_n(s)}{\sigma^i(s)} \to v^i(s)$ uniformly in s as $n \to \infty$ for some continuous function $v^i : S \to \mathbb{R}^{1\times p}$.

The above assumptions ensure that the CLT described by Assumption 5.1 is satisfied. A full discussion, justification and proof of sufficiency for the above conditions may be found in Sommerfeld et al. (2018). The additional requirements listed by Assumptions 5.2-5.3 may be satisfied through placing equivalent assumptions of

continuity and differentiability on the estimator and target functions, $L^{i'}\hat{\beta}^i(s)$ and $L^{i'}\beta^i(s)$, respectively.

C.4 Comparison to Naive Intersections

In this section, we illustrate why intersecting the single-'study condition' confidence regions obtained using the methods of Sommerfeld et al. (2018) is not a valid approach for obtaining confidence regions for $\mathcal{F}_c$. To do so, we define the naive intersection sets $\hat{\mathcal{I}}_c^{\pm} := \cap_{i\in\mathcal{M}}\hat{\mathcal{A}}_c^{\pm,i}$ where $\{\hat{\mathcal{A}}_c^{\pm,i}\}_{i\in\mathcal{M}}$ are the single-'study condition' confidence regions obtained using the methods of Sommerfeld et al. (2018). For ease, fix $M = 2$ and $\alpha = 0.05$. By construction, we have that;

$$\lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{A}}_c^{+,1} \subseteq \mathcal{A}_c^1 \subseteq \hat{\mathcal{A}}_c^{-,1}] = 0.95 \quad \text{and} \quad \lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{A}}_c^{+,2} \subseteq \mathcal{A}_c^2 \subseteq \hat{\mathcal{A}}_c^{-,2}] = 0.95. \tag{C.18}$$

In other words, if $\{\hat{\mathcal{A}}_c^{\pm,i}\}_{i\in\{1,2\}}$ were generated many times using different datasets, it would be expected that the statements $\hat{\mathcal{A}}_c^{+,1} \subseteq \mathcal{A}_c^1 \subseteq \hat{\mathcal{A}}_c^{-,1}$ and $\hat{\mathcal{A}}_c^{+,2} \subseteq \mathcal{A}_c^2 \subseteq \hat{\mathcal{A}}_c^{-,2}$ would each be violated approximately 1 in 20 times. However, as illustrated by Fig. C.13, these violations can occur in a variety of different ways. Consider the following three cases. For every 20 runs of the method;

- **Case 1:** Fig. C.13 (a) occurs 18 times, (e) occurs once and (f) occurs once.
- **Case 2:** Fig. C.13 (a) occurs 18 times, (e) occurs once and (c) occurs once.
- **Case 3:** Fig. C.13 (a) occurs 19 times and (d) occurs once.

In each case, it can be seen that Statement (C.18) is satisfied. However, it also follows that;

$$\textbf{Case 1:} \quad \lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{I}}_c^{+} \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{I}}_c^{-}] = 0.90,$$

$$\textbf{Case 2:} \quad \lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{I}}_c^{+} \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{I}}_c^{-}] = 0.95,$$

$$\textbf{Case 3:} \quad \lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{I}}_c^{+} \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{I}}_c^{-}] = 1.00.$$

Figure C.13: Potential violations of the inclusion statements $\hat{\mathcal{A}}_c^{+,1} \subseteq \mathcal{A}_c^1 \subseteq \hat{\mathcal{A}}_c^{-,1}$, $\hat{\mathcal{A}}_c^{+,2} \subseteq \mathcal{A}_c^2 \subseteq \hat{\mathcal{A}}_c^{-,2}$ and $\hat{\mathcal{I}}_c^{+} \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{I}}_c^{-}$. For brevity, only violations involving the outer sets are illustrated.

We note that, by inflating the noise variance surrounding the boundary violations shown in Fig. C.13, datasets can be constructed for which each of the above cases occur. The above logic may be extended to show that, in general, $\mathbb{P}[\hat{\mathcal{I}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{I}}_c^-]$ is bounded as follows;

$$1 - M\alpha \leq \lim_{n\to\infty} \mathbb{P}[\hat{\mathcal{I}}_c^+ \subseteq \mathcal{F}_c \subseteq \hat{\mathcal{I}}_c^-] \leq 1,$$

but can attain any arbitrary value in $[1 - M\alpha, 1]$ depending on the spatial structure of the noise in the data. In other words, the sets $\hat{\mathcal{I}}_c^{\pm}$ do not serve as $(1-\alpha)$ confidence regions for $\mathcal{F}_c$.

Nomenclature

Symbols and Maths (Linear Mixed Models)

β	Fixed effects parameter vector, page 20
ϵ	Random error, page 20
Λ_k	Cholesky factor, page 51
$\mathcal{C}_k$	Constraint matrix, page 65
$\mathcal{D}_m$	Duplication matrix, page 49
$\mathcal{L}_m$	Elimination matrix, page 50
σ^2	Fixed effects variance, page 20
θ	Shorthand for (β, σ^2, D), page 23
D	Random effects covariance matrix, page 22
e	Residual vector, $e = Y - X\beta$, page 20
f_k	The k^{th} random factor, page 23
$K_{m,n}$	Commutation matrix, page 49
l_k	The number of levels possessed by f_k, page 23
n	Number of observations, page 19

N_m Symmetrization matrix, page 49

p Number of fixed effects parameters, page 19

q Second dimension of Z, page 22

q_k The number of random effects grouped by f_k, page 23

r Number of random factors, page 23

V The matrix $I_n + ZDZ'$, page 23

X Fixed effects design matrix, page 19

Y Response vector, page 19

Z Random effects design matrix, page 22

vec Vectorisation operator, page 49

vec_m Generalized vectorisation operator, page 62

vech Half-vectorisation operator, page 49

vecu Unique vectorisation operator, page 64

Symbols and Maths (Confidence Regions)

α Significance level, page 123

$\hat{\mathcal{A}}_c^i$ i^{th} estimated excursion set, page 121

$\hat{\mathcal{F}}_c^{+/-}$ Upper/lower confidence regions, page 122

$\hat{\mathcal{F}}_c$ Estimated conjunction set, page 121

$\hat{\mu}_n^M$ Estimator function for i^{th} study condition, page 119

$\hat{g}_n^i(s)$ The fraction $\frac{\hat{\mu}_n^i(s)-c}{\sigma^i(s)}$, page 126

$\mathcal{A}_c^i$ i^{th} true excursion set, page 121

$\mathcal{F}_c$ True conjunction set, page 121

$\mathcal{M}$ The set $\{1, ..., M\}$, page 121

$\mathcal{P}$ Power set, page 124

$\mathcal{P}^+$ Power set, without the empty set, page 124

μ^i Target function for i^{th} study condition, page 119

∂ Boundary, page 124

$\sigma^i(s)$ Standard deviation across observations, page 126

τ_n Decreasing function of sample size, page 126

$\xrightarrow{d}$ Convergence in distribution, page 126

$G^i(s)$ Limiting variable, with continuous sample paths in S, page 126

$g^i(s)$ The fraction $\frac{\mu^i(s)-c}{\sigma^i(s)}$, page 126

H Exceedance variable, page 128

M Number of study conditions, page 119

S Spatial domain, page 118

Subscript Conventions

(k) Matrix columns corresponding to the k^{th} factor, page 24

(k, j) Matrix columns corresponding to the j^{th} level of the k^{th} factor, page 24

$[X, Y]$ Matrix indexing, page 48

c Threshold, page 40

k Factor number, page 23

n Number of observations used to generate estimate, page 119

s Iteration number, page 50

v Voxel number, page 29

Superscript Conventions

i Study condition, page 121

h/f/c Half/Full/Cholesky representation, page 51

Abbreviations

BOLD Blood Oxygen Level Dependent, page 14

GLS Generalized Least Squares, page 56

HRF Haemodynamic Response Function, page 14

LM Linear Model, page 19

ML Maximum Likelihood, page 23

OLS Ordinary Least Squares, page 20

ReML Restricted Maximum Likelihood, page 23

References

Adler R (1981) The geometry of random fields. Wiley series in probability and mathematical statistics. Probability and mathematical statistics, J. Wiley

Adler R, Taylor J (2009) Random Fields and Geometry. Springer Monographs in Mathematics, Springer New York

Adler RJ (1977) Hausdorff Dimension and Gaussian Fields. The Annals of Probability 5(1):145 – 151, DOI 10.1214/aop/1176995900

Adler RJ, Bartz K, Kou SC, Monod A (2017) Estimating thresholding levels for random fields via euler characteristics. 1704.08562

Alagapan S, Lustenberger E Caroline; Hadar, Shin HW, Fröhlich F (2019) Low-frequency direct cortical stimulation of left superior frontal gyrus enhances working memory performance. NeuroImage 184, DOI 10.1016/j.neuroimage.2018.09.064

Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, Gallacher J, Green J, Matthews P, Pell J, Sprosen T, Collins R (2012) Uk biobank: Current status and what it means for epidemiology. Health Policy and Technology 1(3):123 – 126, DOI https://doi.org/10.1016/j.hlpt.2012.07.003

Azas JM, Wschebor M (2009) Level sets and extrema of random processes and fields. Level Sets and Extrema of Random Processes and Fields DOI 10.1002/9780470434642

Baayen R, Davidson D, Bates D (2008) Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59(4):390 – 412, DOI https://doi.org/10.1016/j.jml.2007.12.005, special Issue: Emerging Data Analysis

Bandettini PA (2012) Twenty years of functional mri: The science and the stories. NeuroImage 62(2):575–588, DOI https://doi.org/10.1016/j.neuroimage.2012.04.026

Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, Glasser MF, Curtiss S, Dixit S, Feldt C, Nolan D, Bryant E, Hartley T, Footer O, Bjork JM, Poldrack R, Smith S, Johansen-Berg H, Snyder AZ, Van Essen DC (2013) Function in the human connectome: Task-fmri and individual differences in behavior. NeuroImage 80:169–189, DOI https://doi.org/10.1016/j.neuroimage.2013.05.033

Barnett S (1990) Matrices: Methods and Applications. Oxford applied mathematics and computing science series, Clarendon Press

Bates D (2006) lmer, p-values and all that. `https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html`, accessed: 2020-12-07

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software, Articles 67(1):1–48, DOI 10.18637/jss.v067.i01

Bates DM, DebRoy S (2004) Linear mixed models and penalized least squares. Journal of Multivariate Analysis 91, DOI 10.1016/j.jmva.2004.04.013

Bear M, Connors B, Paradiso M (2020) Neuroscience: Exploring the Brain, Enhanced Edition: Exploring the Brain, Enhanced Edition. Jones & Bartlett Learning

Bearden CE, Thompson PM (2017) Emerging global initiatives in neurogenetics: The enhancing neuroimaging genetics through meta-analysis (enigma) consortium. Neuron 94(2):232 – 236, DOI https://doi.org/10.1016/j.neuron.2017.03.033

Beckmann C, Jenkinson M, Smith S (2003) General multilevel linear modeling for group analysis in fmri. NeuroImage 20:1052–63, DOI 10.1016/S1053-8119(03)00435-X

Belliveau JW, Kennedy DN Jr, McKinstry RC, Buchbinder BR, Weisskoff RM, Cohen MS, Vevea JM, Brady TJ, Rosen BR (1991) Functional mapping of the human visual cortex by magnetic resonance imaging. Science 254(5032):716–719

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300, DOI https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berger H (1929) Über das elektrenkephalogramm des menschen. Archiv für Psychiatrie und Nervenkrankheiten 108:407–431

Berger RL (1997) Likelihood Ratio Tests and Intersection-Union Tests, Birkhäuser Boston, Boston, MA, pp 225–237. DOI 10.1007/978-1-4612-2308-5_15

Berger RL, Hsu JC (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science 11(4):283 – 319, DOI 10.1214/ss/1032280304

Bernal-Rusiel JL, Greve DN, Reuter M, Fischl B, Sabuncu MR (2013a) Statistical analysis of longitudinal neuroimage data with linear mixed effects models. NeuroImage 66:249 – 260, DOI https://doi.org/10.1016/j.neuroimage.2012.10.065

Bernal-Rusiel JL, Reuter M, Greve DN, Fischl B, Sabuncu MR (2013b) Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. NeuroImage 81:358 – 370, DOI https://doi.org/10.1016/j.neuroimage.2013.05.049

Bhattacharyya KB (2016) Godfrey newbold hounsfield (1919-2004): The man who revolutionized neuroimaging. Ann Indian Acad Neurol 19(4):448–450

Bickel PJ, Wichura MJ (1971) Convergence Criteria for Multiparameter Stochastic Processes and Some Applications. The Annals of Mathematical Statistics 42(5):1656 – 1670, DOI 10.1214/aoms/1177693164, URL https://doi.org/10.1214/aoms/1177693164

Boisgueheneuc Fd, Levy R, Volle E, Seassau M, Duffau H, Kinkingnehun S, Samson Y, Zhang S, Dubois B (2006) Functions of the left superior frontal gyrus in humans: a lesion study. Brain 129(12):3315–3328, DOI 10.1093/brain/awl244, https://academic.oup.com/brain/article-pdf/129/12/3315/741983/awl244.pdf

Bolin D, Lindgren F (2015) Excursion and contour uncertainty regions for latent gaussian models. Journal of the Royal Statistical Society Series B (Statistical Methodology) 77(1):85–106, URL http://www.jstor.org/stable/24774726

Bowring A, Telschow F, Schwartzman A, Nichols TE (2019) Spatial confidence sets for raw effect size images. NeuroImage 203:116187, DOI https://doi.org/10.1016/j.neuroimage.2019.116187

Bowring A, Telschow FJ, Schwartzman A, Nichols TE (2021) Confidence sets for cohen's d effect size images. NeuroImage 226:117477, DOI https://doi.org/10.1016/j.neuroimage.2020.117477

Bréchet L, Grivaz P, Gauthier B, Blanke O (2018) Common recruitment of angular gyrus in episodic autobiographical memory and bodily self-consciousness. Frontiers in Behavioral Neuroscience 12, DOI 10.3389/fnbeh.2018.00270

Brett M, Penny W, Kiebel S (2004) An introduction to random field theory. Transactions on Rough Sets DOI 10.1016/B978-012264841-0/50046-9

Brown RE (2019) Why study the history of neuroscience? Front Behav Neurosci 13:82

Buxton RB (2012) Dynamic models of bold contrast. NeuroImage 62(2):953–961, DOI https://doi.org/10.1016/j.neuroimage.2012.01.012

Buxton RB, Uludağ K, Dubowitz DJ, Liu TT (2004) Modeling the hemodynamic response to brain activation. NeuroImage 23:S220–S233, DOI https://doi.org/10.1016/j.neuroimage.2004.07.013

Bzdok D, Yeo BT (2017) Inference in the age of big data: Future perspectives on neuroscience. NeuroImage 155:549–564, DOI https://doi.org/10.1016/j.neuroimage.2017.04.061

Bzdok D, Nichols T, Smith S (2019) Towards algorithmic analytics for large-scale datasets. Nature Machine Intelligence

Ramón y Cajal S (1911) Histologie du système nerveux de l'homme & des vertébrés., vol v. 1. Paris :Maloine,, URL `https://www.biodiversitylibrary.org/item/103261`

Cao J (1999) The size of the connected components of excursion sets of $\chi 2$, t and f fields. Advances in Applied Probability 31(3):579–595, DOI 10.1239/aap/1029955192

Carlson N (2013) Physiology of Behavior. Always learning, Pearson

Casey B, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, Soules ME, Teslovich T, Dellarco DV, Garavan H, Orr CA, Wager TD, Banich MT, Speer NK, Sutherland MT, Riedel MC, Dick AS, Bjork JM, Thomas KM, Chaarani B, Mejia MH, Hagler DJ, Daniela Cornejo M, Sicat CS, Harms MP, Dosenbach NU, Rosenberg M, Earl E, Bartsch H, Watts R, Polimeni JR, Kuperman JM, Fair DA, Dale AM (2018) The adolescent brain cognitive development (abcd) study: Imaging acquisition across 21 sites. Developmental Cognitive Neuroscience 32:43 – 54, DOI https://doi.org/10.1016/j.dcn.2018.03.001

Chang C, Lin X, Ogden RT (2017) Simultaneous confidence bands for functional regression models. Journal of Statistical Planning and Inference 188:67–81, DOI https://doi.org/10.1016/j.jspi.2017.03.002

Chen G, Saad Z, Britton J, Pine D, Cox R (2013) Linear mixed-effects modeling approach to fmri group analysis. NeuroImage 73, DOI 10.1016/j.neuroimage.2013.01.047

Chernozhukov V, Chetverikov D, Kato K (2013) Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. The Annals of Statistics 41(6), DOI 10.1214/13-aos1161

Cohen J (2013) Statistical Power Analysis for the Behavioral Sciences. Elsevier Science

Courtney SM, Ungerleider LG (1997) What fMRI has taught us about human vision. Curr Opin Neurobiol 7(4):554–561

Cowan WM, Harter DH, Kandel ER (2000) The emergence of modern neuroscience: Some implications for neurology and psychiatry. Annual Review of Neuroscience 23(1):343–391, DOI 10.1146/annurev.neuro.23.1.343

Dale HH (1914) The action of certain esters of choline, and their relation to muscarine. Journal of Pharmacology and Experimental Therapeutics 6(2):147–190

Dandy WE (1919) Röntgenography of the brain after the injection of air into the spinal canal. Ann Surg 70(4):397–403

Demidenko E (2013) Mixed Models: Theory and Applications with R. Wiley Series in Probability and Statistics, Wiley

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39(1):1–38

Dempster AP, Rubin DB, Tsutakawa RK (1981) Estimation in covariance components models. Journal of the American Statistical Association 76(374):341–353, DOI 10.1080/01621459.1981.10477653

Dijkstra N, van Gaal S, Geerligs L, Bosch SE, van Gerven M (2021) No overlap between unconscious and imagined representations. DOI 10.31234/osf.io/ctdmk

Dunn OJ (1961) Multiple comparisons among means. Journal of the American Statistical Association 56(293):52–64, DOI 10.1080/01621459.1961.10482090

Efron B, Tibshirani R (1994) An Introduction to the Bootstrap, 1st edn. Chapman and Hall/CRC

Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A 113(28):7900–7905

Fai AHT, Cornelius PL (1996) Approximate f-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. Journal of Statistical Computation and Simulation 54(4):363–378, DOI 10.1080/00949659608811740

Feinberg DA, Yacoub E (2012) The rapid development of high speed, resolution and precision in fMRI. Neuroimage 62(2):720–725

Ferreira RA, Vinson D, Dijkstra T, Vigliocco G (2021) Word learning in two languages: Neural overlap and representational differences. Neuropsychologia 150:107703, DOI https://doi.org/10.1016/j.neuropsychologia.2020.107703

FIL Methods Group (2020) SPM12 Manual. UCL Queen Square Institute of Neurology, 12 Queen Square, London

Flandin G, Friston KJ (2019) Analysis of family-wise error rates in statistical parametric mapping using random field theory. Hum Brain Mapp 40(7):2052–2054

Francq BG, Lin D, Hoyer W (2019) Confidence, prediction, and tolerance in linear mixed models. Statistics in Medicine 38(30):5603–5622, DOI https://doi.org/10.1002/sim.8386

Frégnac Y (2017) Big data and the industrialization of neuroscience: A safe roadmap for understanding the brain? Science 358(6362):470–477

Friston K, Glaser D, Henson R, Kiebel S, Phillips C, Ashburner J (2002) Classical and bayesian inference in neuroimaging: Applications. NeuroImage 16(2):484 – 512, DOI https://doi.org/10.1006/nimg.2002.1091

Friston K, Stephan K, Lund T, Morcom A, Kiebel S (2005) Mixed-effects and fmri studies. NeuroImage 24:244–52, DOI 10.1016/j.neuroimage.2004.08.055

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, Evans AC (1994) Assessing the significance of focal activations using their spatial extent. Human Brain Mapping 1(3):210–220, DOI https://doi.org/10.1002/hbm.460010306

Garrison DH (2015) Vesalius, the China root epistle : a new translation and critical edition / Andreas Vesalius ; translated and edited by Daniel H. Garrison ; with added illustrations from the 1543 and 1555 De humani corporis fabrica. Cambridge University Press, Cambridge

Gebregziabher M, Eckert M, Vaden K, Johnson T, Lawson A (2017) Methods for the analysis of missing data in fmri studies. Journal of Biometrics & Biostatistics 08, DOI 10.4172/2155-6180.1000335

Giles MB (2008) Collected matrix derivative results for forward and reverse mode algorithmic differentiation. Advances in Automatic Differentiation pp 35–44

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC, Jenkinson M (2013) The minimal preprocessing pipelines for the human connectome project. NeuroImage 80:105–124, DOI https://doi.org/10.1016/j.neuroimage.2013.04.127

Good P (2013) Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer Series in Statistics, Springer New York

Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, Sochat VV, Nichols TE, Poldrack RA, Poline JB, Yarkoni T, Margulies DS (2015) Neurovault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Frontiers in Neuroinformatics 9:8, DOI 10.3389/fninf.2015.00008

Hall CN, Howarth C, Kurth-Nelson Z, Mishra A (2016) Interpreting bold: towards a dialogue between cognitive and cellular neuroscience. Philosophical Transactions of the Royal Society B: Biological Sciences 371(1705):20150348, DOI 10.1098/rstb.2015.0348

Hedges LV (1981) Distribution theory for glass's estimator of effect size and related estimators. Journal of Educational Statistics 6(2):107–128

Helmholtz HF (1875) On the sensations of tone as a physiological basis for the theory of music. Nature 12(308):449–452, DOI 10.1038/012449a0

Henderson CR, Kempthorne O, Searle SR, von Krosigk CM (1959) The estimation of environmental and genetic trends from records subject to culling. Biometrics 15(2):192–218

Herculano-Houzel S (2009) The human brain in numbers: a linearly scaled-up primate brain. Front Hum Neurosci 3:31

Hesterberg T (2014) What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. 1411.5279

Hong G, Raudenbush SW (2008) Causal inference for time-varying instructional treatments. Journal of Educational and Behavioral Statistics 33(3):333–362, DOI 10.3102/1076998607307355

Horien C, Noble S, Greene A, Lee K, Barron D, Gao S, O'Connor D, Salehi M, Dadashkarimi J, Shen X, Lake E, Constable R, Scheinost D (2020) A hitchhiker's guide to working with large, open-source neuroimaging datasets. Nature Human Behaviour 5, DOI 10.1038/s41562-020-01005-4

IBM Corp (2015) IBM SPSS Advanced Statistics 23. Armonk, NY: IBM Corp.

Jenkinson M, Chappell M (2018) Introduction to Neuroimaging Analysis. Oxford neuroimaging primers, Oxford University Press

Jennrich RI, Schluchter MD (1986) Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42(4):805–820

Keselman H, Algina J, Kowalchuk RK, Wolfinger RD (1999) The analysis of repeated measurements: a comparison of mixed-model satterthwaite f tests and a nonpooled adjusted degrees of freedom multivariate test. Communications in Statistics - Theory and Methods 28(12):2967–2999, DOI 10.1080/03610929908832460

Khoshnevisan D (2002) Multiparameter Processes: An Introduction to Random Fields, 1st edn. Springer Monographs in Mathematics, Springer-Verlag New York

Koretsky AP (2012) Early development of arterial spin labeling to measure regional brain blood flow by MRI. Neuroimage 62(2):602–607

Kuznetsova A, Brockhoff P, Christensen R (2017) lmertest package: Tests in linear mixed effects models. Journal of Statistical Software, Articles 82(13):1–26, DOI 10.18637/jss.v082.i13

Laird N, Lange N, Stram D (1987) Maximum likelihood computations with repeated measures: Application of the em algorithm. Journal of the American Statistical Association 82(397):97–105, DOI 10.1080/01621459.1987.10478395

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38(4):963–974

Landhuis E (2017) Neuroscience: Big brain, big data. Nature 541(7638):559–561

Lawlor D, Lawlor D, Mishra G (2009) Family Matters: Designing, Analysing and Understanding Family Based Studies in Life Course Epidemiology. Life Course Approach to Adult Health, OUP Oxford

Lawlor GR (2020) l'hôpital's rule for multivariable functions. The American Mathematical Monthly 127(8):717–725, DOI 10.1080/00029890.2020.1793635

Lee T (2010) Western political thought in dialogue with asia by takashi shogimen; cary j. nederman. Perspectives on Politics 8:1220–1222, DOI 10.2307/40984330

Li X, Guo N, Li Q (2019) Functional neuroimaging in the new era of big data. Genomics, Proteomics & Bioinformatics 17(4):393 – 401, DOI https://doi.org/10.1016/j.gpb.2018.11.005

Lindquist MA, Spicer J, Asllani I, Wager TD (2012) Estimating and testing variance components in a multi-level GLM. Neuroimage 59(1):490–501

Lindstrom MJ, Bates DM (1988) Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association 83(404):1014–1022

Lorusso L, Piccolino M, Motta S, Gasparello A, Barbara JG, Bossi-Régnier L, Shepherd GM, Swanson L, Magistretti P, Everitt B, Molnár Z, Brown RE (2018) Neuroscience without borders: Preserving the history of neuroscience. European Journal of Neuroscience 48(5):2099–2109, DOI https://doi.org/10.1111/ejn.14101

Lu H, Hua J, van Zijl PCM (2013) Noninvasive functional imaging of cerebral blood volume with vascular-space-occupancy (VASO) MRI. NMR Biomed 26(8):932–948

Luke S (2017) Evaluating significance in linear mixed-effects models in r. Behavior Research Methods 49:1494–1502

Madhyastha T, Peverill M, Koh N, McCabe C, Flournoy J, Mills K, King K, Pfeifer J, McLaughlin KA (2018) Current methods and limitations for longitudinal fmri analysis across development. Developmental Cognitive Neuroscience 33:118 – 128, DOI https://doi.org/10.1016/j.dcn.2017.11.006

Magnus JR, Neudecker H (1980) The elimination matrix: Some lemmas and applications. SIAM Journal on Algebraic Discrete Methods 1(4):422–449, DOI 10.1137/0601049

Magnus JR, Neudecker H (1986) Symmetry, 0-1 matrices, and jacobians : a review. Econometric Theory p 46

Magnus JR, Neudecker H (1999) Matrix differential calculus with applications in statistics and econometrics, rev. ed edn. Wiley Series in Probability and Statistics, John Wiley

Manor O, Zucker DM (2004) Small sample inference for the fixed effects in the mixed linear model. Computational Statistics & Data Analysis 46(4):801 – 817, DOI https://doi.org/10.1016/j.csda.2003.10.005

Maullin-Sapey T, Nichols TE (2021) Fisher scoring for crossed factor linear mixed models. Statistics and Computing 31(5):53, DOI 10.1007/s11222-021-10026-6

Mejia AF, Yue YR, Bolin D, Lindgren F, Lindquist MA (2020) A bayesian general linear modeling approach to cortical surface fmri data analysis. Journal of the American Statistical Association 115(530):501–520, DOI 10.1080/01621459.2019.1611582, pMID: 33060871

Meyer M (1911) The Fundamental Laws of Human Behavior: Lectures on the Foundations of Any Mental Or Social Science. Human personality series, R. G. Badger

Milardi D, Quartarone A, Bramanti A, Anastasi G, Bertino S, Basile GA, Buonasera P, Pilone G, Celeste G, Rizzo G, Bruschetta D, Cacciola A (2019) The cortico-basal ganglia-cerebellar network: Past, present and future perspectives. Frontiers in Systems Neuroscience 13:61, DOI 10.3389/fnsys.2019.00061

Mumford J, Nichols T (2006) Modeling and inference of multisubject fmri data. IEEE Engineering in Medicine and Biology Magazine 25(2):42–51, DOI 10.1109/MEMB.2006.1607668

Mumford JA, Nichols T (2009) Simple group fMRI modeling and inference. Neuroimage 47(4):1469–1475

Mumford JA, Poldrack RA (2007) Modeling group fMRI data. Social Cognitive and Affective Neuroscience 2(3):251–257, DOI 10.1093/scan/nsm019

Neudecker H, Wansbeek T (1983) Some results on commutation matrices, with statistical applications. Canadian Journal of Statistics 11(3):221–231, DOI 10.2307/3314625

Nichols T, Hayasaka S (2003) Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat Methods Med Res 12(5):419–446

Payne L (2005) Robert l. martensen. the brain takes shape: An early history. Journal of The History of Medicine and Allied Sciences - J HIST MED ALLIED SCI 61:222–223

Pinheiro J, Bates D (2009) Mixed-Effects Models in S and S-PLUS. Statistics and Computing, Springer

Pinheiro JC, Bates DM (1996) Unconstrained parametrizations for variance-covariance matrices. Statistics and Computing 6:289–296

Plis SM, Sarwate AD, Wood D, Dieringer C, Landis D, Reed C, Panta SR, Turner JA, Shoemaker JM, Carter KW, Thompson P, Hutchison K, Calhoun VD (2016) Coinstac: A privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. Frontiers in Neuroscience 10:365, DOI 10.3389/fnins.2016.00365

Poldrack R, Mumford J, Nichols T (2011) Handbook of Functional MRI Data Analysis. Cambridge University Press

Poldrack R, Barch D, Mitchell J, Wager T, Wagner A, Devlin J, Cumba C, Koyejo O, Milham M (2013) Toward open sharing of task-based fmri data: the openfmri project. Frontiers in Neuroinformatics 7:12, DOI 10.3389/fninf.2013.00012

Powell M (2009) The bobyqa algorithm for bound constrained optimization without derivatives. Technical Report, Department of Applied Mathematics and Theoretical Physics

Powell MJD (1964) An efficient method for finding the minimum of a function of several variables without calculating derivatives. The Computer Journal 7(2):155–162, DOI 10.1093/comjnl/7.2.155

Pranav P, van de Weygaert R, Vegter G, Jones BJT, Adler RJ, Feldbrugge J, Park C, Buchert T, Kerber M (2019) Topology and geometry of gaussian random fields i: on betti numbers, euler characteristic, and minkowski functionals. Monthly Notices of the Royal Astronomical Society 485(3):4167–4208, DOI 10.1093/mnras/stz541

PubMed statistics (2022) PubMed search results by year for keyword 'fmri'. `https://pubmed.ncbi.nlm.nih.gov/?term=fmri&timeline=expanded`, accessed: 2022-01-08

Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL (2001) A default mode of brain function. Proc Natl Acad Sci U S A 98(2):676–682

Rao C, Mitra SK (1972) Generalized Inverse of Matrices and Its Applications. Probability and Statistics Series, Wiley

Raudenbush SW, Bryk AS (2002) Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd edn. Advanced Quantitative Techniques in the Social Sciences 1, SAGE Publications

Rehman A, Khalili Y (2019) Neuroanatomy, occipital lobe. In: Neuroanatomy, Occipital Lobe

Rocklin M (2015) Dask: Parallel computation with blocked algorithms and task scheduling. In: Huff K, Bergstra J (eds) Proceedings of the 14th Python in Science Conference, pp 130 – 136

Rosenblatt JD, Finos L, Weeda WD, Solari A, Goeman JJ (2018) All-Resolutions inference for brain imaging. Neuroimage 181:786–796

Sandrone S, Bacigaluppi M, Galloni MR, Cappa SF, Moro A, Catani M, Filippi M, Monti MM, Perani D, Martino G (2013) Weighing brain activity with the balance: Angelo Mosso's original manuscripts come to light. Brain 137(2):621–633, DOI 10.1093/brain/awt091

Santoro G, Wood L Mark D; Merlo, Anastasi GP, Tomasello F, Germanò A (2009) The anatomic location of the soul from the heart, through the brain, to the whole body, and beyond. Neurosurgery vol 65 iss 4 65, DOI 10.1227/01.neu.0000349750.22332.6a

SAS Institute Inc (2015) SAS/STAT® 14.1 User's Guide The MIXED Procedure. Springer Berlin Heidelberg, Cary, NC: SAS Institute Inc.

Satterthwaite FE (1946) An approximate distribution of estimates of variance components. Biometrics Bulletin 2(6):110–114

Schaalje GB, McBride JB, Fellingham GW (2002) Adequacy of approximations to distributions of test statistics in complex mixed linear models. Journal of Agricultural, Biological, and Environmental Statistics 7(4):512–524

Scheipl F, Greven S, Küchenhoff H (2008) Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Computational Statistics & Data Analysis 52(7):3283–3299, DOI https://doi.org/10.1016/j.csda.2007.10.022

Schwartzman A, Gavrilov Y, Adler RJ (2011) Multiple testing of local maxima for detection of peaks in 1d. Ann Stat 39(6):3290–3319

Seghier ML (2013) The angular gyrus: multiple functions and multiple subdivisions. The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry 19(1):43–61, DOI 10.1177/1073858412440596

Simpson EH (1951) The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society Series B (Methodological) 13(2):238–241

Smee A (1850) Instinct and Reason: Deduced from Electro-biology. Reeve and Benham

Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44(1):83–98

Smith SM, Nichols TE (2018) Statistical challenges in "big data" human neuroimaging. Neuron 97(2):263 – 268, DOI https://doi.org/10.1016/j.neuron.2017.12.018

Smith SM, Beckmann CF, Ramnani N, Woolrich MW, Bannister PR, Jenkinson M, Matthews PM, McGonigle DJ (2005) Variability in fMRI: a re-examination of inter-session differences. Hum Brain Mapp 24(3):248–257

Sommerfeld M, Sain S, Schwartzman A (2018) Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. Journal of the American Statistical Association 113(523):1327–1340, DOI 10.1080/01621459.2017.1341838

Swazey JP (1970) Action propre and action commune: the localization of cerebral function. J Hist Biol 3(2):213–234

Szucs D, Ioannidis JP (2020) Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. NeuroImage 221:117164, DOI https://doi.org/10.1016/j.neuroimage.2020.117164

T Vu A, Jamison K, Glasser MF, Smith SM, Coalson T, Moeller S, Auerbach EJ, Uğurbil K, Yacoub E (2017) Tradeoffs in pushing the spatial resolution of fmri for the 7t human connectome project. NeuroImage 154:23–32, DOI https://doi.org/10.1016/j.neuroimage.2016.11.049

Telschow F, Schwartzman A (2021) Simultaneous confidence bands for functional data using the gaussian kinematic formula. Journal of Statistical Planning and Inference 216, DOI 10.1016/j.jspi.2021.05.008

Tibaldi FS, Verbeke G, Molenberghs G, Renard D, Van den Noortgate W, de Boeck P (2007) Conditional mixed models with crossed random effects. British Journal of Mathematical and Statistical Psychology 60(2):351–365, DOI 10.1348/000711006X110562

Torres S (1994) Topological Analysis of COBE-DMR Cosmic Microwave Background Maps. apjl 423:L9, DOI 10.1086/187223

Turkington DA (2013) Generalized Vectorization, Cross-Products, and Matrix Calculus. Cambridge University Press, DOI 10.1017/CBO9781139424400

Turnbull BJ, Welsh ME, Heid CA, Davis W, Ratnofsky AC (1999) The longitudinal evaluation of school change and performance (lescp) in title i schools. interim report to congress. In: The Longitudinal Evaluation of School Change and Performance (LESCP) in Title I Schools. Interim Report to Congress.

Uğurbil K, Xu J, Auerbach EJ, Moeller S, Vu AT, Duarte-Carvajalino JM, Lenglet C, Wu X, Schmitter S, Van de Moortele PF, Strupp J, Sapiro G, De Martino F, Wang D, Harel N, Garwood M, Chen L, Feinberg DA, Smith SM, Miller KL, Sotiropoulos SN, Jbabdi S, Andersson JL, Behrens TE, Glasser MF, Van Essen DC, Yacoub E (2013) Pushing spatial and temporal resolution for functional and diffusion mri in the human connectome project. NeuroImage 80:80–104, DOI 10.1016/j.neuroimage.2013.05.012

Vaden KI, Gebregziabher M, Kuchinsky S, Eckert M (2012) Multiple imputation of missing fmri data in whole brain analysis. NeuroImage 60:1843–1855

Van Essen D, Smith S, Barch D, Behrens T, Yacoub E, Ugurbil K (2013) The wu-minn human connectome project: an overview. NeuroImage 80, DOI 10.1016/j.neuroimage.2013.05.041

Van Horn JD, Toga AW (2014) Human neuroimaging as a “big data” science. Brain Imaging Behav 8(2):323–331

Van Laere J (1993) Vesalius and the nervous system. Verh K Acad Geneeskd Belg 55(6):533–576

Verbeke G, Molenberghs G (2001) Linear Mixed Models for Longitudinal Data. Springer Series in Statistics, Springer New York

Vogel T, Smieskova R, Schmidt A André; Walter, Harrisberger F, Eckert A, Lang UE, Riecher-Rössler A, Graf M, Borgwardt S (2016) Increased superior frontal gyrus activation during working memory processing in psychosis: Significant relation to cumulative antipsychotic medication and to negative symptoms. Schizophrenia Research DOI 10.1016/j.schres.2016.03.033

Webb-Vargas Y, Chen S, Fisher A, Mejia A, Xu Y, Crainiceanu C, Caffo B, Lindquist MA (2017) Big data and neuroimaging. Stat Biosci 9(2):543–558

Welch BL (1947) The generalization of ‘student’s’ problem when several different population variances are involved. Biometrika 34(1/2):28–35

West B, Welch K, Galecki A (2014) Linear Mixed Models: A Practical Guide Using Statistical Software. CRC Press

Winkler AM, Webster MA, Vidaurre D, Nichols TE, Smith SM (2015) Multi-level block permutation. NeuroImage 123:253 – 268, DOI https://doi.org/10.1016/j.neuroimage.2015.05.092

Wolfinger R (1996) Heterogeneous variance: Covariance structures for repeated measures. Journal of Agricultural, Biological, and Environmental Statistics 1:205, DOI 10.2307/1400366

Wolfinger R, Tobias R, Sall J (1994) Computing gaussian likelihoods and their derivatives for general linear mixed models. Siam Journal on Scientific Computing 15, DOI 10.1137/0915079

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 – 1747, DOI https://doi.org/10.1016/j.neuroimage.2003.12.023

Worsley KJ (1996) The geometry of random images. CHANCE 9(1):27–40, DOI 10.1080/09332480.1996.10542483

Wu CFJ (1986) Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. The Annals of Statistics 14(4):1261 – 1295, DOI 10.1214/aos/1176350142

Yeung AWK (2018) An updated survey on statistical thresholding and sample size of fmri studies. Frontiers in Human Neuroscience 12, DOI 10.3389/fnhum.2018.00016

Zhang H, Nichols TE, Johnson TD (2009) Cluster mass inference via random field theory. NeuroImage 44(1):51–61, DOI 10.1016/j.neuroimage.2008.08.017

Zhang X, Li L, Huang G, Zhang L, Liang Z, Shi L, Zhang Z (2021) A multisensory fmri investigation of nociceptive-preferential cortical regions and responses. Frontiers in neuroscience

Zhu S, Wathen AJ (2018) Essential formulae for restricted maximum likelihood and its derivatives associated with the linear mixed models. `1805.05188`