A Comparison of Neuroimaging Software and a Contour Inference Method for analysis of Task-fMRI data

by

Alexander Bowring

St Catherine's College

Submitted to the University of Oxford

for the degree of

Doctor of Philosophy

Nuffield Department of Population Health

October 2019

Contents

Acknowledgments

Declarations

I, Alexander Bowring, hereby declare that except where specific reference is made to the work of others, the content of this dissertation is original and has not been submitted in whole or in part for consideration for any other degree or qualification in these, or any other Universities. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

- The work presented in Chapter 3 has been published in the *Human Brain Mapping* journal (Bowring et al., 2019). This work was presented at the *Organization for Human Brain Mapping* (OHBM) Annual Meetings in 2017 and 2018. At the OHBM 2018 Annual Meeting, this work was the recipient of an oral presentation and a Merit Abstract Award.
- The work presented in Chapter 4 has been published in the *NeuroImage* journal (Bowring et al., 2018). This work was presented at the OHBM Annual Meeting in 2017, where it was the recipient of an oral presentation.
- The work presented in Chapter 5 is based on a pre-printed manuscript.

Alexander Bowring
September 2019

Abstract

Over the last three decades, Functional Magnetic Resonance Imaging (fMRI) has rapidly progressed to become the primary tool for human brain mapping. Recently however, considerable attention within the field has been directed towards data-sharing and open science initiatives. This has been driven by a growing apprehension about the reproducibility of findings within the neuroimaging literature, amid concerns that current inference procedures are often misused or misinterpreted such that the overall scientific conclusions become distorted. One aspect specific to neuroimaging pinpointed as a cause for poor reproducibility is the high flexibility of a typical fMRI workflow. In the first part of this thesis, we investigate how the choice of software package used to conduct a statistical analysis can influence the group-level results of a task-fMRI study. We use publicly shared data from three published task-fMRI studies, and reanalyze each study within the three main neuroimaging software packages, AFNI, FSL and SPM, using parameteric and nonparametric inference. All information on how to process, analyze, and model each dataset we obtain from the publications. We use a variety of quantitative and qualitative comparison methods to gauge the scale of variability in our results and assess fundamental differences between each software package. While qualitatively we find broad similarities between packages, we also discover marked differences, such as Dice similarity coefficient values ranging from 0.000 to 0.743 in comparisons of thresholded statistic maps between software. We discuss the challenges involved in our replication attempt, while also utilizing open science tools in an effort to make our own research reproducible. In the second part of this thesis, we extend a contour inference method initially proposed by Sommerfeld, Sain, and Schwartzman (2018) *SSS* to develop spatial confidence sets (CSs) on clusters found in thresholded blood-oxygen-level dependent (BOLD) effect size maps. While traditional inferences based on hypothesis testing indicate where the null, i.e. an effect size of zero, can be rejected, the CSs give statements about where effect sizes exceed a *positive* threshold analogous to confidence intervals simultaneously across the entire brain. We make advancements to theoretical aspects and implementation of contour inference to improve the method's finite-sample performance. We extend the wild bootstrap theory presented in *SSS*, proposing a method based on the t-bootstrap, and recommend that the bootstrapped residuals are multiplied by Rademacher variables instead of Gaussian variables. We also develop a linear interpolation method for computing the topological boundary over which the bootstrap is applied. Notably, we demonstrate that

the framework used in *SSS* for assessing simulations manifests considerable positive bias in the simulation results, and propose our own novel construction to solve this issue. In the final part of this thesis, we make further theoretical developments to contour inference so that the method can operate on the Cohen's *d* and partial R^2 effect sizes commonly reported at the end of a neuroimaging study. For the second and third parts of this thesis, we carry out intensive Monte Carlo simulations on synthetic 3D data to investigate the accuracy of contour inference on signals representative of fMRI activation clusters. We also demonstrate the method on two 'big' fMRI datasets, obtaining confidence sets to localize activation in functional data from the Human Connectome Project and UK Biobank.

CHAPTER 1

Introduction

Since its inception at the end of the twentieth century, functional Magnetic Resonance Imaging (fMRI) has experienced a meteoric rise to become the primary tool for human brain mapping. While many forms of the technique exist, introduction of the particular method based on the Blood Oxygenization Level Dependant (BOLD) effect has ultimately been the catalyst in elevating fMRI to such stature within the neuroimaging community. Taking advantage of the magnetic properties of oxygen-rich red blood cells, BOLD fMRI measures changes in blood oxygenization alongside cerebral blood flow and volume as a proxy to identify brain areas where elevated neuronal activity has occurred in response to a stimulus. While the relationship between the BOLD effect and neuronal activity is complex and remains controversial, it is the unique attributes of BOLD fMRI – in particular, its capacity for non-invasive recording of signals across the entire brain at a high spatial resolution – that set the technique apart from other scanning methods.

However, BOLD fMRI is also a *noisy* process. The MR signals researchers set out to measure during a scanning session are corrupted by artefacts from both the imaging hardware and the physiology of the participant. Examples of scanner noise include inhomogeneities of the magnetic field that can cause spatial distortion or blurring in the MR image, and scanner drift characterized by temporal degradation of the signal. Physiological noise induced by subject motion, respiration, and heartbeat exacerbate the problem.

Because of the low signal-to-noise, researchers must apply a series of statistical techniques to find meaning in the data. This usually entails carrying out a number of preprocessing, modelling and analysis steps that together constitute the fMRI processing pipeline. The fundamental objectives of preprocessing are to standardize brain locations across participants, to apply methods ensuring that the data conform to statistical assumptions required for analysis, and to reduce the influence of the

aforementioned noise artefacts present in the data. This is achieved by conducting a number of steps, including slice-timing correction, motion correction, normalization, registration of the functional data to an anatomical template, and spatial smoothing.

For task-based fMRI, a mass-univariate approach is utilized to model the data. During the scanning session, functional data are acquired in the form of voxels – cubic intensity units that partition the brain comparable to the way in which pixels partition a computer screen. Each voxel's time-series is considered independently within the general linear model framework as a combination of signal components. To evaluate the effect of an experimental task condition relative to a baseline condition, hypothesis testing is performed at each voxel to compute a statistical parametric map of t-statistic values. Here, the behaviour of the signal under the null hypothesis of no activation is estimated using either a parametric approach, appealing to the body of mathematics known as Random Field Theory, or a nonparametric approach, where permutation methods are applied to estimate the null-distribution directly from the data. Finally, the statistical parametric map is thresholded to localize brain function.

While we have provided a brief overview of the fMRI analysis pipeline, it is notable that there is not a general consensus as to how each particular analysis step should be carried out. Consequently, researchers have the freedom to make many choices during an analysis, such as how much smoothing is applied to the data, or how the heamodynamic response of blood flow to active neuronal tissues is modelled. However, this 'methodological plurality' comes with a drawback. While conceptually similar, two different analysis pipelines applied on the same dataset may not produce the same scientific results, and mathematical modelling has shown that the high analytic flexibility associated with fMRI can potentially distort the final scientific findings of an investigation (Ioannidis, 2005). The problem is, with so many statistically valid methodological strategies available, if you try them all you are likely to find *something*. Combined with further issues such as p-hacking and publication bias – where there has been evidence to suggest that studies finding a significant effect are disproportionately represented in the fMRI literature (David et al., 2013; Ioannidis et al., 2014) – these conditions have created the perfect storm: In recent years, many attempts to replicate the results of published fMRI studies have been unsuccessful, in what has been deemed as an ongoing reproducibility crisis within the field (Poldrack et al., 2017; Gorgolewski and Poldrack, 2016; Open Science Collaboration, 2015).

The degree to which varying methodological decisions can lead to discrepancies in observed results has been investigated extensively. Choices for each individual procedure in the analysis pipeline (for example, head-motion regression (Lund et al., 2005), temporal filtering (Skudlarski et al., 1999), and autocorrelation correc-

tion (Woolrich et al., 2001)) alongside the order in which these procedures are conducted (Carp, 2013) can all deeply influence the final determined areas of brain activation. In perhaps the most comprehensive of such studies (Carp, 2012), a single publicly available fMRI dataset was analyzed using over 6,000 unique analysis pipelines, generating 34,560 unique thresholded activation images. These results displayed a substantial degree of flexibility in both the sizes and locations of significant activation.

Alongside issues concerned with the flexibility of the analysis workflow, the statistical procedures carried out for fMRI inference have also come under intense scrutiny. Because statistical tests are conducted at each brain voxel independently, the p-values used to threshold the statistical parametric map are corrected to account for the large number of simultaneous comparisons being carried out and limit the expected number of voxels falsely declared as significant. This is almost always done using a false discovery rate correction procedure (Benjamini and Hochberg, 1995) or a Bonferonni correction to limit the family-wise error rate of making at least one significant finding.

The importance of such statistical correction methods were made prominent within the neuroimaging community using a humorous example, where one author identified significant activation in the brain of a dead salmon after applying inference with uncorrected p-values (Bennett et al., 2009). However, in recent times they have been a source of major controversy. In 2016, a shocking paper by Eklund, Nichols, and Knutsson (2016) discovered that many fMRI software packages were incorrectly carrying out the multiple-correction procedures for clusterwise inference, inflating the false-positive rate to up to 70%. In a damning blow to the field, the implications of this study brought into question the validity of thousands of published fMRI results.

While the relevant software packages have now been patched, deeper conceptual problems have been raised regarding the fMRI approach to inference. Specifically, there is a considerable amount of information that is *not* captured when applying inference using cluster-size. In this setting, a significant p-value only indicates that a cluster is larger than expected by chance, and although a significant cluster may have a large spatial extent, since we can only infer that at least one voxel in the cluster has statistically significant signal, spatial specificity is low (Woo et al., 2014). In addition, this method does not provide a measure of the spatial variation of significant clusters. For illustration, imagine that a single fMRI study is repeated using two varying cohorts of participants; whereas we would expect moderate differences in the size and shape of clusters within each cohort's group-level thresholded map, current statistical results do not characterize this variability.

A more pressing issue stems from an age-old paradox caused by the fallacy of the null hypothesis (Rozeboom, 1960). The paradox is that while the statistical models used for fMRI conventionally assume mean-zero noise, in reality all sources of noise will *never* completely cancel. Therefore, improvements in experimental design will eventually lead to statistically significant results, and the null-hypothesis will, eventually, *always* be rejected (Meehl, 1967). The recent availability of ambitious, large-sample studies (e.g Human Connectome Project (HCP), N = 1,200; UK Biobank, N = 30,000 and counting) have exemplified this problem. Analysis of high-quality fMRI data acquired under optimal noise conditions has been shown to display almost universal activation across the entire brain after hypothesis testing, even with stringent correction (Gonzalez-Castillo et al., 2012). For these reasons, there is an increased urgency for methods that can provide meaningful inference to interpret all significant effects.

In this work, we make contributions in two thematic areas currently challenging the field of task-based fMRI: Firstly, the need for further transparency to the degree in which the body of work comprising the fMRI literature is reproducible. Secondly, the need for further statistical methods to improve current inference practices carried out within the field. To end this section, we summarize our main contributions before providing an outline of the organisation of this dissertation:

1. While we have already discussed a number of studies exploring how decisions made at each stage of the analysis pipeline can influence the final scientific results of an fMRI investigation, for all of these studies the fundamental decision of which analysis software package the pipeline was conducted through remained constant. This is despite a vast array of analysis packages that are used throughout the neuroimaging literature, the most popular of which are AFNI, FSL and SPM. Motivated by this, in Chapter 2 we comprehensively assess how each of these software packages can impact analysis results by reanalyzing three published task-fMRI neuroimaging studies and quantifying several aspects of variability between the three package's group-level statistical maps. Our findings suggest that exceedingly weak effects may not generalise across software. We are unaware of any comparable exercise in the literature.

2. In carrying out this software comparison exercise, we implement a range of quantitative methods for the novel application of comparing fMRI statistical maps. These include Dice statistic comparisons, for assessing differences in the determined regions of activation between the three software's thresholded statistical maps, Bland-Altman plots, for assessing differences between the magni-

tude of the t-static values in the unthresholded maps, Euler Characteristics, for assessing differences in the topological properties of each software's activation profile, and Neurosynth analyses, for assessing differences in the anatomical regions associated to each software's activation pattern. We believe these methods are generalizable and hope they may benefit any further comparison of neuroimaging results.

3. In Chapter 3, we develop an inference method originally proposed for application on geospatial data in Sommerfeld, Sain, and Schwartzman (2018) *SSS* to create spatial confidence sets on clusters found in fMRI percentage BOLD effect size maps. While currently used hypothesis testing methods indicate where the null, i.e. an effect size of zero, can be rejected, this form of inference allows for statements about anatomical regions where effect sizes have exceeded, and fallen short of, a *non-zero* threshold, such as areas where a BOLD change of 2.0% has occurred.

4. In developing the inference method proposed by *SSS*, we make theoretical advancements that improve the performance of the confidence sets, particularly for 3D data with moderate sample sizes. We also find that the methods used to assess the empirical coverage for simulations presented in *SSS* are positively biased. We develop our own weighted-interpolation method for assessing empirical coverage, and on using this method, our simulation results validate the asymptotical mathematical theory set out in *SSS*.

5. In Chapter 4, we make further theoretical advancements to the confidence sets for application on the Cohen's d and partial R^2 effect sizes.

This dissertation is organized into five chapters: Chapter 2 is dedicated to presenting the context of this work and providing background on the current methodological procedures carried out for analysis of task-fMRI data, with a particular emphasis on the statistical inference methods relevant to this thesis. In Chapter 3, we assess the analytic variability of group-level task-fMRI results under the choice of software package through which the analysis is conducted. We reanalyze three published task-fMRI studies whose data has been made publicly available, attempting to replicate the original analysis procedure within each software package. We then make a number of comparisons to assess the similarity of our results. In Chapter 4, we develop the inference method originally proposed in *SSS* to create spatial confidence sets on clusters found in fMRI percentage BOLD effect size maps. We summarize the theory in *SSS* before detailing our proposed modifications. We then carry

out intensive Monte Carlo simulations to investigate the accuracy of the confidence sets on synthetic 3D signals representative of clusters found in fMRI effect size maps. Futhermore, we illustrate the method by computing confidence sets on 80 subject's percentage BOLD data from the Human Connectome Project working memory task. In Chapter 5 we make further theoretical developments to the inference method for application on the Cohen's d and partial R^2 effect sizes commonly used in a task-fMRI study. Finally, in Chapter 6 we conclude this dissertation and provide further discussion of possibilities for future work.

CHAPTER 2

Background

In this chapter, we provide the context that forms the basis of our research. We begin by presenting a broad overview of the study of brain function, before narrowing down to the specific field of task-based functional Magnetic Resonance Imaging (t-fMRI) that will be the main focus of study in this thesis. Here, we describe each of the preprocessing and modelling components of a typical t-fMRI analysis pipeline. Finally, we give an in-depth discussion of the state-of-the-art procedures used for subject- and group-level t-fMRI inference that are of particular relevance to the remaining chapters of this work.

2.1 The Study of Brain Function

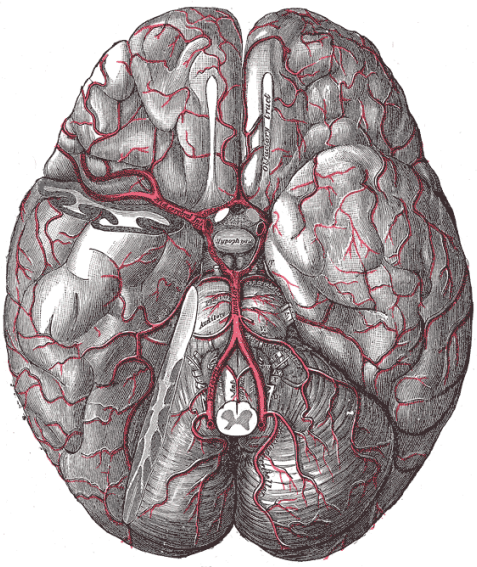

Figure 1: An illustration showing the arteries at the base of the human brain from (Gray, 1918).

The human brain, the central organ of the human nervous system, has been described as one of the most complex structures in the known universe. Made up of approximately 86 billion neurons (Azevedo et al., 2009), where neuronal interaction occurs continuously via trillions of synaptic networks to form intricate and dynamic neural networks, the myriad of processes taking place inside the brain at any given time make the study of brain function an intimidating challenge. Nonetheless, our understanding of this organ has come along way from our ancient Egyptian ancestors, who believed that the heart was at the source of human intelligence, and for whom the practice of drilling a hole into the skull was regarded as a solution to cure a headache (Adelman and Others, 1987; Mohamed, 2014).

Remarkably, much of this progress has come in the last century alone. A number of key developments within this time-frame include: Confirmation of the neuron doctrine, the concept that the nervous system is a collection of discrete individual cells, postulated by Santiago Ramon y Cajal at the end of the 19th century and demonstrated in the 1950s thanks to the development of electron microscopy (López-Muñoz et al., 2006); the first evidence of neuroplasticity, the ability for the brain's structure to change during an individual's lifetime (Diamond et al., 1964; Bennett et al., 1964); and the emergence of neuroimaging techniques such as electroencephalography (EEG), positron emission tomography (PET), and MRI. The toils of this

scientific endeavour are now translating into concrete advancements influencing a wide variety of aspects concerned with population health. Neuroscience research is beginning to find applications in the clinical setting to advance our understanding of neurodevelopmental and neurodegenerative disorders and generate novel therapies to treat and prevent such diseases. Brain imaging has been used to localize the source of neurological impairment for diseases such as epilepsy (Stacey and Litt, 2008), and neuroengineering techniques based on our capability to stimulate neural circuits are implemented to treat Parkinson's disease (Kalia et al., 2013) and dystonia (Fox and Alterman, 2015). Structural- and functional-MRI are being explored to determine biomarkers for diagnosis of Alzheimer's disease *prior* to symptom onset (Sperling et al., 2014; McEvoy et al., 2009), alongside providing information about the role of different brain regions in human behaviour that can contribute to an improved prognosis and patient response to therapy (Matthews et al., 2006).

Modern neuroscience can be dissected into many major branches, each subfield taking a specific slant to studying the nervous system. It is therefore perhaps unsurprising that in isolation, the phrase 'the study of brain function' is rather vague. Brain function can manifest itself in ways that can be observed using a variety of different measurements, whether that be with a molecular, chemical, structural, or functional approach (Hargreaves and Klimas, 2012). Different modalities of MRI are employed to evaluate specific properties that ultimately characterize whichever approach is taken. For instance, looking at brain function from an anatomical perspective, voxel-based morphometry (VBM) could be used to measure differences in local concentrations of brain tissue, to assess, for example, changes in grey matter volume (Mechelli et al., 2005). Additionally, one could apply diffusion tensor imaging (DTI) to instead map white matter tractography in the brain (Alexander et al., 2007; Soares et al., 2013). From a functional outlook, resting state fMRI (rs-fMRI) determines that spatially remote brain areas are functionally connected when each region's BOLD response is temporally correlated in the absence of an explicit task (Lee et al., 2013). On the other hand, task-based fMRI (t-fMRI) measures spatio-temporal changes in the BOLD signal between task-stimulated and control states to find brain regions that are activated in the presence of a stimulus (Glover, 2011).

Each imaging method and modality does not live inside a vacuum, and recent work within the field has provided further insight of the interdependence between different approaches to examining brain function. One example of this is in the study of resting state networks, which explores how distinct sets of brain regions can reveal temporally correlated activation patterns when the brain is at rest. While resting state networks have been most widely investigated using rs-fMRI tech-

niques (e.g. Smith et al., 2009; Lee et al., 2012; Moussa et al., 2012), more recently, the same correlation patterns have been independently detected using EEG and MEG (Brookes et al., 2011; Fomina et al., 2015). This work not only demonstrates how utilization of numerous tools can further our understanding of resting state mechanisms, but also suggests a direct relationship between the electro-physiological signals recorded with MEG and the BOLD fluctuations associated to fMRI. Similarly, other recent efforts have shown that the functional response to a cognitive task measured with t-fMRI may be able to be predicted by connectivity features from the same individual's brain at rest (Parker Jones et al., 2017; Tavor et al., 2016). This research signals towards an innate functional signature that defines our behaviour, while also providing potential clinical solutions to obtain t-fMRI data from patients who are unable to perform the specific task of interest.

In the context of this thesis, we will study brain function from a functional perspective, primarily focussed on task-based fMRI.

2.2 Blood Oxygenation Level Dependant (BOLD) Functional Magnetic Resonance Imagery (fMRI)

Whereas structural MRI is concerned with the anatomy of the brain, functional MRI (fMRI) measures dynamic changes in blood flow in order to ultimately make inference on neuronal activation. This is possible due to the intrinsic relationship between local neuronal activity and subsequent changes in cerebral blood flow (CBF), a biological phenomenon known as neurovascular coupling. An increased supply of oxygen is carried by haemoglobin in red blood cells to provide energy to active neurons, and it is the magnetic properties of the haemoglobin that MRI takes advantage of. Specifically, as deoxygenated haemoglobin is more magnetic (paramagnetic) than oxygenated haemoglobin, MRI uses haemoglobin as an endogenous contrast agent from which to source the signal. Neurovascular coupling induces inhomogeneities in the local magnetic field due to a decreased concentration of deoxygenated haemoglobin, that lead to a detectable change in the MR signal.

The complete chain of events linking neuronal activity to a change in MRI signal is referred to as the Blood Oxygenation Level Dependant (BOLD) effect, and this type of imaging is known as BOLD fMRI. Proof of concept of the BOLD effect was first provided in Ogawa, Lee, Kay, and Tank (1990), and the first use of BOLD fMRI for human brain mapping was carried out in 1992 (Bandettini et al., 1992; Kwong et al., 1992; Ogawa et al., 1992), leading to a large uptake of the method that has continued to this day. Alternative approaches to functional imaging exist, the most

popular of which is functional Arterial Spin Labelling (fASL), that uses magnetically labelled arterial blood water to quantify changes in CSF. While fASL can offer some advantages over fMRI, and changes in CSF measured with the technique are more closely tied to neuronal activation than the BOLD signal, fASL suffers from a much lower signal-to-noise ratio that consequently has made fMRI the preferred imaging modality of choice.

2.2.1 Physiology of the BOLD response

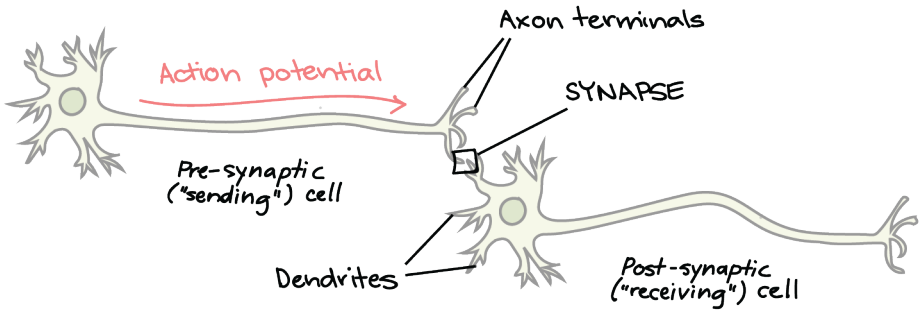

Figure 2: A schematic of the interaction between neurons. Image reused from Khan Academy (CC BY-NC-SA 3.0 US). Note: All Khan Academy content is available for free at (www.khanacademy.org).

Neuronal interaction transpires via a system of electrical and chemical activity. To send out information, an individual neuron – the pre-synaptic cell – emits an electrical signal known as an action potential, for the purpose of stimulating another target neuron – the post-synaptic cell. The action potential travels along the axon of the sending cell, and is transmitted to the receiving cell at the synapse. Information is delivered from the output branches (or, *axon terminals*) of the sending cell across the synapse to the input branches (or, *dendrites*) of the receiving cell, involving the release of chemical *neurotransmitters* alongside a number of other cellular processes. This may stimulate or inhibit the firing of action potentials at the target cell to communicate with other neurons, eventually leading to a configuration of neurons collectively processing and responding to information.

The electrical and chemical processes involved in neuronal activation require energy, which drives the neurovascular coupling. Blood vessels that flow into the capillaries pervading the neuronal tissue dilate and the rate of CBF increases to regulate a greater supply of oxygen and nutrients to localized regions of active neurons. Overall, the increases in CBF and cerebral blood volume (CBV) are many orders of magnitude greater than the increases in oxygen extraction ($CMRO_2$) caused by the neuronal activation. Thus, there is an overall net increase of oxygenated heamoglobin,

and an increase in the BOLD signal. The expected BOLD response generated from a brief stimulus is characterized quantitatively by the Hemodynamic Response Function (HRF), encompassing the individual changes in CBF, CBV and $CMRO_2$ induced by neuronal stimulation.

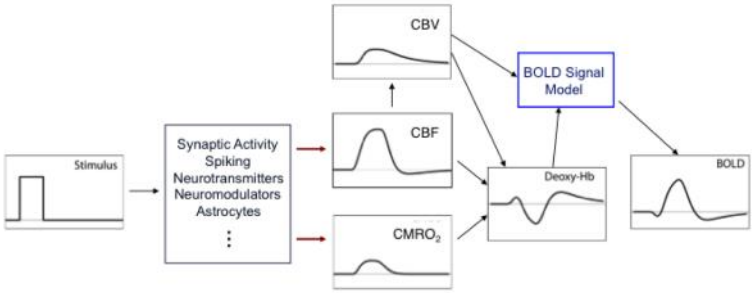

Figure 3: Blah

2.3 Task-based functional Magnetic Resonance Imagery (t-fMRI)

The ultimate goal of a task-based functional Magnetic Resonance Imagery (t-fMRI) experiment is to understand the brain regions that are responsive to a particular task or stimulus the researcher has chosen to investigate. Explicitly, the researcher seeks to detect brain areas whose BOLD time series data is correlated to the task the participant is instructed to perform in the scanner. Researchers can choose from a wide range of possible tasks to explore how the brain processes in a variety of circumstances. For example, a cognitive task may be chosen to gain insight into how the brain processes decision-making or recognition, while a physiological task may be used to see how the brain reacts to a stimulus intended to cause pain or arousal, or how the brain functions when participants are told to hold their breath. In general, the experimenter is only limited in choice of task by the constraints that the task must be able to be conducted within the scanner, and that the task should not involve any sort of head movement which could corrupt the signal.

The MR signal measured in the scanner is noisy, and the heamodynamic response induced by a stimulus only causes fractional changes in the BOLD response, typically of around one percent. Therefore, in order to increase the signal-to-noise ratio (SNR) of the BOLD signal participants repeat the task several times in the scanner. The type of task used, alongside the timings for which the participant is instructed to perform the task inside the scanner, are together known as the *task paradigm* or *experimental design*. Many different task conditions can be investigated within one task paradigm, however, it is fundamental that at least two conditions are included. This is

because BOLD data are not quantitative, insofar that we are unable to interpret the level of neuronal activity from the absolute magnitude of the BOLD response alone. Instead, neuronal activity is inferred by using *contrasts* to measure the difference in the MR signal between two conditions. Commonly, the BOLD response to a task condition is contrasted with a *baseline* condition, where the participant is at rest within the scanner. However, it is equally acceptable to contrast two separate task conditions depending on the aims of the investigation.

An experimental design where the task condition is carried out for an extended period of time is said to have a *block* design (or *boxcar* design). One example of this could be a task paradigm where the participant is instructed to look at an animal photo for five seconds in each task repetition. Alternatively, in an *event-related* design the task or stimulus takes the form of a discrete, rapid event, such as a study where the participant experiences a mild electric shock. A graphical representation of each of these task paradigms is shown in *Figure BLAH*, showing the *onset timing function* along with the anticipated response for both types of stimuli. While block designs have greater statistical power, with a relatively larger BOLD response, the researcher has more control with respect to how the stimuli are delivered in an event-related design.

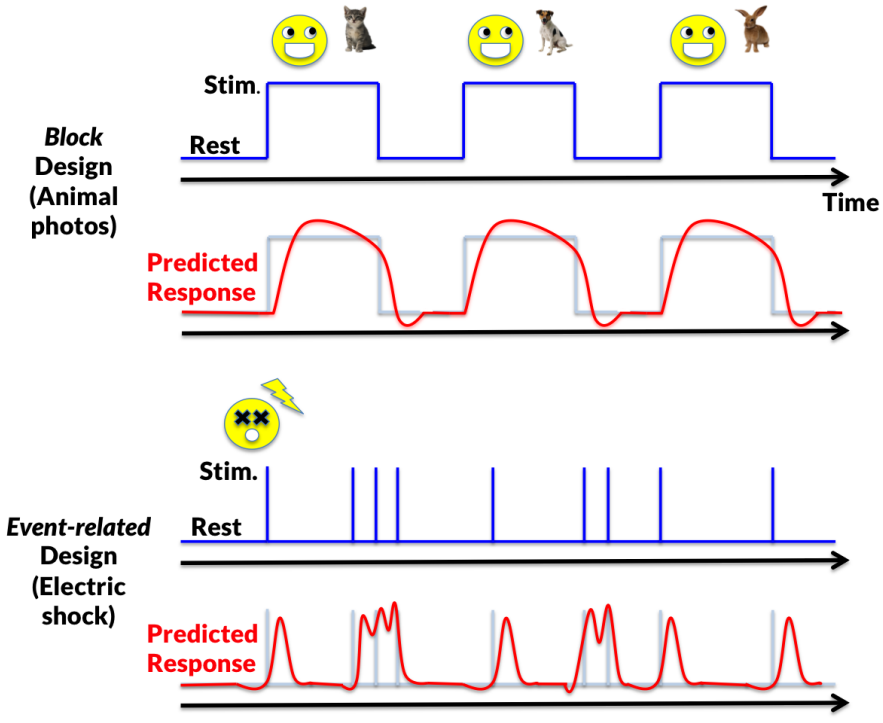

Figure 4: Blah

2.4 Overview of Analysis Pipeline

To analyse voxelwise t-fMRI BOLD data, a series of analysis steps are performed on the data in succession. Together, the complete chain of analysis procedures carried out is known as the analysis *pipeline*. Researchers have great flexibility as to how the analysis pipeline is comprised, with various options and adjustable parameters for each individual analysis step, as well as choices as to the order in which certain procedures are conducted. Nevertheless, a standard analysis pipeline of t-fMRI data can be partitioned into three main stages: preprocessing, modelling, and statistical inference. In the upcoming sections we will describe the individual processing steps that are usually carried out in each of these analysis stages, while here we provide a brief overview.

The main goals of preprocessing are to reduce the severity of noise artefacts present in the raw BOLD fMRI data, and to prepare the data for statistical analysis. At the modelling stage, a mass-univariate approach is adopted, whereby each voxel's functional time series data is considered independently as an instance of the general linear model (GLM) framework. Within the GLM, the contrasts discussed in the previous section are formulated to statistically test the main hypotheses investigated within the study. At the inference stage, a *statistical parametric map* is generated containing statistic values at each voxel for each contrast of interest. For *subject-level* inference, a participant's statistical parametric map is thresholded to display only voxels showing statistically significant results. This type of inference may be of interest in a clinical setting, particularly to aid in the diagnosis of a patient, however in a research study there is usually a greater emphasis placed on finding results that generalize across the larger population. In this case, each participant's statistic map is entered into a second-level model for *group-level* inference, and a thresholded map is computed to localize effects that were consistent across all individuals in the study.

In practice, the analysis pipeline is usually carried out within a neuroimaging *software package*. Various software packages are available, many of which are freely distributed on the internet. The three most popular packages are *AFNI* (Cox, 1996), *FSL* (Jenkinson et al., 2012), and *SPM* (Penny et al., 2011). While there are several differences as to how each software package operates, most packages follow the same fundamental principles to implement the three main stages of the analysis pipeline.

2.5 Preprocessing

In this section, we present each of the analysis steps that are typically conducted within a t-fMRI preprocessing pipeline: brain extraction, distortion correction, slice-timing correction, realignment, coregistration, spatial normalization, spatial smoothing, temporal filtering and intensity normalization. These procedures are carried out to compensate for artefacts present in the data, and to ensure that the data satisfy the assumptions used for modelling and inference.

2.5.1 Brain Extraction

Brain extraction is commonly the first procedure carried out in the analysis pipeline, with the purpose of removing the skull and any other non-brain tissue from a participant's anatomical image. Since the purpose of the analysis is to infer areas of activation *within* the brain, conceptually it is sensible to remove any external structures that are not of interest. However, brain extraction also has a more important role in improving the outcome of subsequent steps in the preprocessing pipeline. In the upcoming sections we describe *coregistration*, where the subject's functional data is spatially realigned to the anatomical image, as well as *spatial normalization*, where each participant's data is registered to a standard space. Brain extraction helps to increase the robustness of both of these registration methods, since differences in non-brain structures can sidetrack the registration algorithms causing an inaccurate alignment of the respective images.

In SPM, brain extraction is carried out by first applying a *segmentation* to the anatomical image, in order to generate probability maps of the gray and white matter tissue in the structural scan. The grey and white matter probability maps are summed and thresholded, creating a binary map containing the brain regions to be included in the analysis. Finally, the anatomical scan is masked with the binary map to remove any non-brain structures. AFNI and FSL both use variants of the Brain Extraction Tool (Smith, 2002) algorithm, implementing an adaptive model that evolves to fit the brain's surface in order to segment brain and non-brain tissue types.

2.5.2 Distortion Correction

Distortion correction is applied to account for signal loss and geometric distortions in the functional data that can manifest due to spatial inhomogeneities in the main static magnetic field during the acquisition. These inhomogeneities arise due to the different magnetic susceptibility properties of each tissue type in the brain, and the

most severely affected regions are those close to air-filled synuses, such as the temporal or frontal lobe. If not corrected, signal dropout and distortion can cause failure in the registration of the functional data to the non-distorted anatomical image.

While it is not possible to recover regions of signal loss, field distortions can be rectified with the use of a *field map*. A field map is obtained as part of the acquisition to estimate the intensity of the static magnetic field. The analysis software uses the field map to calculate the magnitude of the geometric distortions, and then applies spatial transformations to unwarp the functional data. The field map is also used to deweight areas of substantial signal dropout during the registration, and if the signal loss is particularly severe, ignore these locations in the analysis of the data.

2.5.3 Slice timing Correction

While the statistical modelling of fMRI data assumes that the signal is measured over the entire brain simultaneously, in reality functional imaging is usually carried out on a slice-by-slice basis, creating a single 3D volume as a combination of multiple 2D slices. Slices are acquired sequentially from top-to-bottom or bottom-to-top, or by using an interleaved sequence where all odd-numbered slices are collected first, followed by the even slices. Because of this, the BOLD signal is sampled at different points of the HRF. This can create the illusion that the signal peaks earlier for slices that are collected later in the acquisition, even though the underlying response is identical.

Slice timing correction uses temporal interpolation methods to artificially obtain an intensity estimate at each brain voxel at a single time point, shifting the data to recreate the image as if all measurements were obtained collectively. The reference time point is commonly chosen to be halfway through the scanning procedure, and in this case the timings from all slices are corrected to match-up with the timings of the volume collected midway in the acquisition, which acts as the reference slice. The time series data from each voxel is temporally shifted to line-up with the signal response from the reference slice, and the voxel's data between the acquisition time points of the reference slice are re-estimated with interpolation. Commonly, this is done using either sinc or spline interpolation.

It is debatable whether slice timing correction should be conducted before or after realignment, and some practitioners have suggested that slice timing correction should be excluded from the analysis pipeline altogether. One alternative to slice timing correction is to account for timing differences at the modelling stage of the analysis. FSL recommend that temporal derivatives are incorporated as extra regressors into the GLM, effectively making the model flexible to temporal shifts in the

signal response.

2.5.4 Realignment

While a participant is told to remain as still as possible in the scanner, over the course of the acquisition some head movement is inevitable. This is particularly problematic in fMRI, as it can corrupt the functional data in numerous ways. If left uncorrected, head movement may cause a voxel's time series data to contain signal from two different tissue types, and if the voxel is located at the edge of the brain, may lead to a loss of signal altogether. Additionally, the change in signal intensity induced by head motion can be many orders of magnitude greater than the BOLD effect. Therefore, if head motion is elicited by the task the participant performs in the scanner, this can lead to false activations in the statistical results that invalidate the analysis.

Realignment (or *motion correction*) of the functional data is performed to remove any substantial movement throughout the time series. To do this, each volume in the time series is spatially transformed to match a reference volume, usually chosen as the first volume of the data or an average image of all the scans. Specifically, a rigid-body transformation of translations and rotations is applied to superimpose each volume onto the reference image. The transformation is determined to optimize a cost function that quantifies the goodness of alignment between the images, e.g. a least squares (used by default in SPM) or normalised correlation (used by default in FSL) cost function. Finally, the transformed data are spatially interpolated to obtain estimates of the signal response on the same voxel grid as the reference image, usually with spline interpolation.

2.5.5 Coregistration

In order to carry out group analyses, corresponding voxels between each participant's functional data should contain information from the same physical anatomical location. However, prior to normalizing data *between* subjects, *coregistration* is conducted to align a participant's functional time series data with their own anatomical image. Similar to realignment, coregistration is achieved via a rigid-body transformation chosen to minimize an appropriate cost function. However, to account for differences between the blurry, distorted functional data and the high-resolution structural image, scalings are also included as parameters of the rigid-body transformation, and a mutual information cost function is commonly used.

2.5.6 Spatial Normalization

The goal of *spatial normalization* (or *intersubject registration*) is to warp all participants functional time series data into a universal coordinate space, integrating the data between subjects to facilitate for group analyses. To remove structural variability between subjects, each participant's data is spatially transformed onto a standard template brain image. The most commonly used templates are the MNI152 images, created by the Montreal Neurological Institute by combining structural data from 152 healthy adults. The transformation is computed on a participant's structural image; the anatomy is registered to the template with a series of linear and non-linear transformations, permitting for local deformations to change the size and shape of the subject's structural image for a better alignment with the brain standard. Finally, the functional data are warped to standard space by concatenating the transformation from functional to structural space computed during coregistration with this transformation from structural to standard space.

2.5.7 Spatial Smoothing

Prior to statistical analyses, *spatial smoothing* is conducted on the functional data. Although this step may seem unsound, as any smoothing will effectively reduce some of the spatial resolution of the fMRI data, the reasons for spatial smoothing are twofold. First and foremost, the main reason for smoothing is to improve the SNR of the data by filtering out high-frequency regions. Intuitively, this works because averaging should reduce the intensity of noisey areas, while leaving the underlying functional signal of interest relatively unaffected. The second reason for smoothing is as a prerequisite for statistical analysis. Specifically, the Gaussian random field theory used for parametric inference is adaptive in how it corrects for the multiple comparison problem dependant on the smoothness of the data. However, a minimum amount of smoothing is required to obtain accurate control over the false discovery rate of activations in the thresholded statistical results.

In practise, the functional data are convolved with a three-dimensional Gaussian filter, and the amount of smoothing applied is proportional to the full width at half maximum (FWHM) of the kernel function. A suitable degree of smoothing is conditional on many factors, such as the quality of the data, the statistical power required, and the expected size of the final activation clusters. A typical smoothing kernel FWHM is between 6 and 10mm^3, although the preprocessing pipelines for recent high-quality, large-sample fMRI datasets have used a lesser degree of smoothing (e.g. 5mm FWHM for the UK Biobank, 4mm FWHM for the Human Connectome Project).

2.5.8 Temporal Filtering

Temporal filtering is another processing step that aims to increase the SNR of the functional data, by taking advantage of the fact that the BOLD signals fMRI sets out to measure generally have a consistent frequency range. Temporal filtering suppresses or removes frequencies outside of this range, implicitly eliminating any artefactual signals present in the data while leaving the neuronal signals of interest untouched.

A well-known source of noise is slow drifts that occur due to imperfections in the scanning hardware. As components of the scanner heat up, this can induce a gradual change in the MR signal resulting in low frequency trends of less than 0.01 Hz in the data. The expected frequency of the BOLD signal response to a task stimulus is around 0.2Hz. Therefore, a *high-pass filter* can be applied, removing all frequencies below a set threshold to attenuate scanner-related drifts. Other forms of noise are caused by physiological effects such as respiratory and cardiac cycles. These artefacts have a frequency range higher than the expected BOLD response (respiratory frequencies are $\sim$0.3hz, cardiac frequencies $\sim$1.0hz), although they may also manifest in the data as lower frequencies due to the effects of *aliasing*. A *low-pass filter* may be used to cut off higher frequencies and subdue artefacts such as physiological noise. While a high-pass filter is commonly included as part of the preprocessing pipeline, low-pass filters are more controversial as they can cultivate autocorrelation in the signal, violating the assumption of temporal independence made for statistical inference.

2.5.9 Grand Mean Scaling

As touched on in *2.3*, BOLD t-fMRI data are not quantitative. Because of this, the fMRI scanner assigns arbitrary units to the signal intensities during the acquisition, and data can be scaled differently across scanning sessions. *Grand mean scaling* (or *intensity normalization*) is applied to rescale each individual's functional time series to increase the interpretability of the data across the group of participants. This is done by multiplying the functional time series (across all voxels and time points) by a constant so that the mean intensity takes a fixed value of, for example, 100. While grand mean scaling will not affect the statistical inference results, normalizing the data facilitates for comparability of the regression coefficient maps (i.e. *beta* maps) obtained for each task condition at the modelling stage of analysis.

2.6 Modelling of t-fMRI data with the General Linear Model

The *General Linear Model* (GLM) is the most widely used approach to modelling BOLD t-fMRI time series data, and a crucial part of any neuroimaging analysis. The GLM generalises a broad class of models that estimate the observed response as a linear combination of experimental and confounding variables. A key strength of this framework is its flexibility, allowing for analyses of data both within and between individuals, and providing a foundation for which experimental hypotheses can be assessed with a variety of statistical tests, using either parametric or nonparametric statistics. In this section we provide an overview of the GLM in the context of brain imaging, before describing some of the most commonly used statistical tests performed within the GLM for analysing fMRI data.

2.6.1 The GLM Set-up

To analyse voxelwise t-fMRI data, each voxel's time series is independently modelled within the GLM. This is commonly referred to as a *mass-univariate* analysis – the 'mass' term specifies that the same analysis is performed many times, and 'univariate' indicates that each analysis is performed separately at every brain voxel (as opposed to *multivariate*, which considers many locations as part of one analysis).

Mathematically, for a compact domain $S \subset \mathcal{R}^D$ (in fMRI, D = 3 and S is the brain mask), the GLM at location (or brain voxel) $\boldsymbol{s} \in S$ is expressed as

$$\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}), \tag{2.1}$$

where $\boldsymbol{Y}(\boldsymbol{s})$ is an $N \times 1$ vector of observations at $\boldsymbol{s}$, $\boldsymbol{X}$ is an $N \times p$ design matrix containing explanatory variables linking the observations in $\boldsymbol{Y}(\boldsymbol{s})$ to the effect sizes in $\boldsymbol{\beta}(\boldsymbol{s})$, $\boldsymbol{\beta}(\boldsymbol{s})$ is an $p \times 1$ vector of the unknown parameters, and $\boldsymbol{\epsilon}(\boldsymbol{s})$ is an $N \times 1$ vector of error terms. It is assumed that the errors are independently distributed conditional on $\boldsymbol{X}$ by a Gaussian distribution with mean zero.

The aim of the regression is to find parameter estimates $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ that best fit the model to the data. The goodness of fit is determined by a method of *least squares*, depending on additional constraints added to the model. The parameter estimates are then used at the inference stage to test hypotheses about the data expressed in terms of the unknown parameters contained in $\boldsymbol{\beta}(\boldsymbol{s})$.

2.6.2 Estimating the Parameters with Ordinary Least Squares (OLS)

OLS is used to solve 2.1 with the assumption that the errors are *spherical*, which means that there is no autocorrelation and that each error term has constant variance. Combined with the normality assumption stated in the previous section, this means

$$\boldsymbol{\epsilon}(\boldsymbol{s}) \mid \boldsymbol{X} \sim \mathcal{N}(0, \sigma^2(\boldsymbol{s})\boldsymbol{I}_N), \tag{2.2}$$

where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. OLS solves the GLM by minimizing the *sum of squares* cost function S given by

$$S(\boldsymbol{\beta}(s)) = \|\boldsymbol{Y}(\boldsymbol{s}) - \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s})\|^2 = \sum_{i=1}^{N} |Y_i(\boldsymbol{s}) - \sum_{j=1}^{p} X_{ij}\beta_j(\boldsymbol{s})|^2. \tag{2.3}$$

This gives the OLS estimates

$$\hat{\boldsymbol{\beta}}(\boldsymbol{s}) = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}(\boldsymbol{s})) = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{Y}(\boldsymbol{s}). \tag{2.4}$$

By the Gauss-Markov Theorem, it can be shown that the OLS estimates are the Best Linear Unbiased Estimates (BLUE) of $\boldsymbol{\beta}(\boldsymbol{s})$ providing all the assumptions are satisfied.

2.6.3 Prewhitening

The key assumption of OLS is that the errors are spherical, however, this is often violated for fMRI data. As discussed in 2.5.8, functional data are characterized by slow drifts which induce temporal autocorrelation in the MR signal. While high-pass filtering can be applied in an attempt to remove the majority of low frequency components, another strategy is to estimate the autocorrelation directly and then remove it by *prewhitening* the data. This can be more efficient than filtering for event-related designs (Woolrich et al., 2001).

If the data are correlated, the error terms have marginal distribution

$$\boldsymbol{\epsilon}(\boldsymbol{s}) \mid \boldsymbol{X} \sim \mathcal{N}(0, \sigma^2(\boldsymbol{s})\boldsymbol{V}(\boldsymbol{s})), \tag{2.5}$$

where $\boldsymbol{V}(\boldsymbol{s})$ is the correlation matrix. Since $\boldsymbol{V}(\boldsymbol{s})$ is symmetric and positive-definite, $\boldsymbol{V}(\boldsymbol{s})$ satisfies the assumptions for the Cholesky decomposition, which means there exists a lower triangular matrix $\boldsymbol{K}(\boldsymbol{s})$ such that $\boldsymbol{V}^{-1}(\boldsymbol{s}) = \boldsymbol{K}^\intercal(\boldsymbol{s})\boldsymbol{K}(\boldsymbol{s})$. Providing that $\boldsymbol{K}(\boldsymbol{s})$ can be accurately determined, the idea is to update the model by multiplying both sides of the GLM by $\boldsymbol{K}(\boldsymbol{s})$ so that the error terms are spherical. Denoting $\boldsymbol{Y}^*(\boldsymbol{s}) = \boldsymbol{K}(\boldsymbol{s})\boldsymbol{Y}(\boldsymbol{s})$, and defining $\boldsymbol{X}^*(\boldsymbol{s})$ and $\boldsymbol{\epsilon}^*(\boldsymbol{s})$ similarly, then for the updated

GLM

$$\boldsymbol{Y}^*(\boldsymbol{s}) = \boldsymbol{X}^*(\boldsymbol{s}) + \boldsymbol{\epsilon}^*(\boldsymbol{s}), \tag{2.6}$$

the conditional covariance of $\boldsymbol{\epsilon}^*(\boldsymbol{s})$ is

$$\mathrm{Cov}(\boldsymbol{\epsilon}^*(\boldsymbol{s}) \mid \boldsymbol{X}) = \boldsymbol{K}(\boldsymbol{s})\mathrm{Cov}(\boldsymbol{\epsilon}(\boldsymbol{s}) \mid \boldsymbol{X})\boldsymbol{K}^\intercal(\boldsymbol{s}) = \sigma^2(\boldsymbol{s})\boldsymbol{I}_N. \tag{2.7}$$

Therefore, the sphericity assumption is satisfied for the updated model, and OLS can be applied to obtain the BLUE of $\boldsymbol{\beta}(\boldsymbol{s})$.

2.6.4 Estimating the Variance

For statistical inference, the variance of the errors $\sigma^2(\boldsymbol{s})$ needs to be estimated. This can be done using the OLS estimates. The *fitted values* given by the OLS estimates are $\hat{\boldsymbol{Y}}(\boldsymbol{s}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{s})$. The differences between the observed data points and the fitted values are known as the *residuals*, denoted by $\hat{\epsilon}(\boldsymbol{s}) = \boldsymbol{Y}(\boldsymbol{s}) - \hat{\boldsymbol{Y}}(\boldsymbol{s})$. The variance of the errors is estimated as the sum of squares of the residuals divided by the *degrees of freedom* of the model

$$\hat{\sigma}^2(\boldsymbol{s}) = \frac{\boldsymbol{\epsilon}^\intercal(\boldsymbol{s})\boldsymbol{\epsilon}(\boldsymbol{s})}{N - p}; \tag{2.8}$$

The degrees of freedom are $N - p$, since there are N observations $Y_1(\boldsymbol{s}), ..., Y_N(\boldsymbol{s})$, and p parameters $\beta_1(\boldsymbol{s}), ..., \beta_p(\boldsymbol{s})$ to estimate.

2.6.5 Inference with Null-Hypothesis Significance Testing

The pay off from obtaining the parameter estimates using OLS is that it enables us to statistically test hypotheses about the unknown effect sizes. This form of inference is known as *Null-Hypothesis Significance Testing* (NHST). Hypotheses are expressed using a contrast vector $\boldsymbol{c}$ to define a linear combination of the parameters. The null hypothesis is always expressed in the form $H_0 : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = 0$, although this covers a wide variety of tests. For example, in a GLM with two parameters, $\boldsymbol{\beta}(\boldsymbol{s}) = (\beta_1(\boldsymbol{s}), \beta_2(\boldsymbol{s}))$, the contrast $\boldsymbol{c} = (1, 0)$ would lead to the null hypothesis $H_0 : \beta_1(\boldsymbol{s}) = 0$, establishing a test to determine if the first parameter $\beta_1(\boldsymbol{s})$ is significantly different from zero. However, one could also test for significant differences between the two parameters by choosing the contrast vector $\boldsymbol{c} = (1, -1)$ to form the null hypothesis $H_0 : \beta_1(\boldsymbol{s}) = \beta_2(\boldsymbol{s})$.

If the sphericity assumption is satisfied, then the distribution of $\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ is

$$\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}) \sim \mathcal{N}(\boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}), \sigma^2(\boldsymbol{s})\boldsymbol{c}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{c}). \tag{2.9}$$

Therefore, hypotheses about a contrast of the model parameters $\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ can be assessed with a t-test, using the knowledge that

$$\frac{\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}) - \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s})}{\sqrt{\hat{\sigma}^2(\boldsymbol{s})\boldsymbol{c}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim t_{N-p}, \tag{2.10}$$

where t_{N-p} is a Student's t-distribution with $N - p$ degrees of freedom. In practice, the hypothesis $H_0 : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = 0$ is tested by computing the t-statistic

$$T(\boldsymbol{s}) = \frac{\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})}{\sqrt{\hat{\sigma}^2(\boldsymbol{s})\boldsymbol{c}^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{c}}} \tag{2.11}$$

and then obtaining a p-*value* by comparing $T(\boldsymbol{s})$ to a t-distribution with $N - p$ degrees of freedom. For a one-sided hypothesis test where the alternative hypothesis is given by $H_A : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) > 0$, the p-value is computed as $p = \Pr(t_{N-p} \geq t)$. In fMRI, the *statistical parametric map* (or *unthresholded statistic map*) is an image containing the p-values computed at every voxel. A p-value is said to be *statistically significant* when $p < \alpha$, where α is a predetermined *significance level* set according to inference standards appropriate for the study (typically for fMRI, α is set at 5% *before* correction for multiple-comparisons). In this case, the conclusion of the test is that there is sufficient evidence to reject the null hypothesis in favour of the alternative. Thus, for the null $H_0 : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = 0$, a statistically significant p-value would suggest a non-zero effect size at location $\boldsymbol{s}$. A *thresholded statistic map* is obtained by masking the statistical parametric map to show only voxels with a statistically significant p-value.

In addition to testing a single contrast, one may also wish to test multiple contrasts at once. For example, in the two-parameter GLM described above, the null hypothesis $H_0 : \beta_1(\boldsymbol{s}) = \beta_2(\boldsymbol{s}) = 0$ could be chosen to test for a significant effect size in either of the parameters. In this case, the contrast $\boldsymbol{c}$ is given as a matrix

$$\boldsymbol{c} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{2.12}$$

where each row corresponds to each of the hypotheses being tested (for this example, $\beta_1(\boldsymbol{s}) = 0$ and $\beta_2(\boldsymbol{s}) = 0$ respectively). This time, inference is carried out using an F-*test*. The F-*statistic* is computed as

$$F(\boldsymbol{s}) = (\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}))^\intercal[r\boldsymbol{c}^\intercal\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}(\boldsymbol{s}))\boldsymbol{c}]^{-1}(\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})), \tag{2.13}$$

where r is the rank of $\boldsymbol{c}$, and a p-value is obtained by comparing $F(\boldsymbol{s})$ to an F-distribution with r numerator and $N - p$ denominator degrees of freedom.

2.6.6 First-Level (Subject-Level) Analysis

In a *first-level* (or *subject-level*) analysis, the GLM set-up in 2.6.1 is used to analyze and test hypotheses related to the t-fMRI data obtained from an individual in a single scanning session.

At each brain voxel s, the $N \times 1$ observations vector $\boldsymbol{Y}(s)$ contains the BOLD signal response data recorded by the scanner across all N sampled time-points during the session. The columns of the design matrix $\boldsymbol{X}(s)$ comprise of task-related and nuisance regressors to model the response in $\boldsymbol{Y}(s)$. The number of task-related regressors is dependent on the task paradigm and the statistical hypotheses the researcher wishes to test. For example, in the animal photo paradigm described at the end of 2.3, to test for activations when the participant was looking at any of the animal photos, a single regressor could be used to model the change in BOLD signal attributable to a photo being displayed. However, if instead the researcher wanted to test for changes in activation when the participant looked at photos of dogs compared to photos of cats, then multiple regressors would need to be used for pictures of each animal type.

The predicted response of each task-related regressor is estimated by convolution of the onset timing function for the stimulus with a HRF. *Figure 5* shows how the predicted response is obtained for the block design used in the animal photo task. There are a variety of ways to model the HRF: by default FSL and SPM use a single canonical HRF, however alternative methods include use of a basis set of smooth functions (Friston et al., 1998) or a more flexible *finite impulse response* basis set (Goutte et al., 2000). For each task-related condition, temporal derivatives may be included as an additional regressor to account for differences between the actual and modelled HRF. See Lindquist et al. (2009) for a comparison of different HRF models.

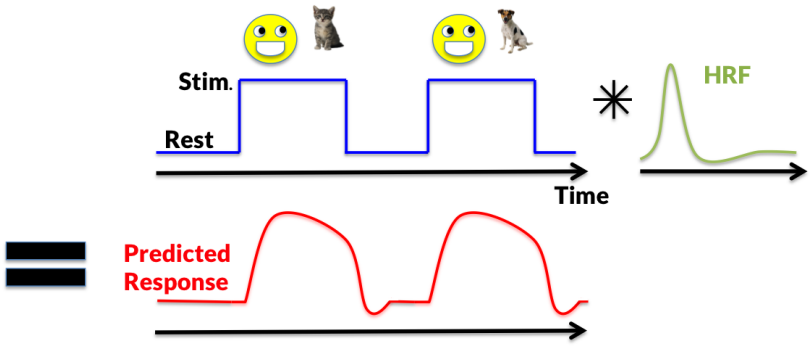

Figure 5: Blah

The time series of motion-related parameters (e.g. translations and rotations) used for realignment are commonly added to the GLM as nuisance regressors to compensate for any left-over motion artefacts in the signal after preprocessing. Further nuisance regressors may include cardiac and respiratory recordings to account for fluctuations in the BOLD signal caused by changes in heart rate and breathing patterns during the scan.

To remove temporal autocorrelation in the BOLD signal the data are whitened (as in 2.6.3), after which unbiased estimators of the parameters in $\boldsymbol{\beta}(\boldsymbol{s})$ can be computed via OLS (as in 2.6.2). Finally, changes in neuronal activity between the task conditions is tested by contrasting the task-related regressors with an appropriate contrast vector $\boldsymbol{c}$ under the null hypothesis $H_0 : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = 0$, using the NHST procedure described in 2.6.5.

2.6.7 Second-Level (Group-Level) Analysis

The contrast parameter estimates obtained for each individual during the first-level analysis are combined in a *second-level* (or *group-level*) analysis to test whether results can be generalized to the larger population. The second-level model aims to account for variation in quantity estimates within each individual's data, as well as variability *between* the participants stemming from biological differences between each subject. This is fundamental from a research perspective as it means that conclusions drawn from a second-level analysis can relate to the whole population from which the participants came from, rather than just the the specific cohort of individuals involved in the study. In this section, we will describe a simplified second-level model for a one-sample *t*-test used to test for a consistent response in a first-level contrast effect across all individuals. However, within the GLM framework it is also possible to conduct a *between*-groups analysis with a *two-sample t-test* to assess activation differences between two distinct groups (e.g. a group of patients and a group of controls), and to incorporate other design types such as a *paired t-test* or *one-way analysis of variance*.

For the GLM set-up in 2.6.1, in a group-level analysis N represents the total number of participants in the study cohort. Theoretically, for a pre-specified contrast, the $N \times 1$ observations vector $\boldsymbol{Y}(\boldsymbol{s})$ contains each subject's first-level contrast effect, i.e. $\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = [\boldsymbol{c}^\intercal\beta_1(\boldsymbol{s}), ..., \boldsymbol{c}^\intercal\beta_N(\boldsymbol{s})]^\intercal$. This gives the instance of the GLM

$$\boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{\beta}_G(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}), \tag{2.14}$$

where the group-level parameters are notated with $\boldsymbol{\beta}_G(\boldsymbol{s})$ to distinguish from the subject-level parameters $\boldsymbol{\beta}(\boldsymbol{s})$. For a one-sample *t*-test the first column of the design matrix $\boldsymbol{X}$ is an intercept where all elements are set equal to 1. Other columns of the design are covariates to be considered for the analysis, such as the age of each participant or a score to describe each individuals performance in the task carried out during the scanning session. For simplicity we will assume no covariates are included, so $\boldsymbol{X}(\boldsymbol{s})$ is a column vector of 1's. $\boldsymbol{\epsilon}(\boldsymbol{s})$ is the vector of group-level errors terms, assumed to have distribution

$$\boldsymbol{\epsilon}(\boldsymbol{s}) \sim \mathcal{N}(0, \sigma^2_G(\boldsymbol{s})\boldsymbol{I}_N), \tag{2.15}$$

where $\sigma^2_G(\boldsymbol{s})$ is the between-subject variance for the group.

Of course, in any practical analysis scenario the true subject-level contrast effects $\boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s})$ are unknown, and therefore the contrast estimates obtained in the first-level analyses $\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ must be used instead. This leads to the GLM formulation

$$\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}) = \boldsymbol{X}\boldsymbol{\beta}_G(\boldsymbol{s}) + \boldsymbol{\epsilon}^*(\boldsymbol{s}), \tag{2.16}$$

where the errors $\boldsymbol{\epsilon}^*(\boldsymbol{s})$ must now account for within-subject variation caused by estimating each participant's parameter effect size as well as the between-subject variance. Rearranging 2.14 and 2.16,

$$\boldsymbol{\epsilon}^*(\boldsymbol{s}) = \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) - \boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}), \tag{2.17}$$

and therefore

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\epsilon}^*(\boldsymbol{s})) &= \mathrm{Cov}(\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})) + \sigma^2_G(\boldsymbol{s})\boldsymbol{I}_N \\ &= \underbrace{\begin{pmatrix}\sigma^2_1(\boldsymbol{s}) & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma^2_N(\boldsymbol{s})\end{pmatrix}}_{\text{within-subject variance}} + \underbrace{\sigma^2_G(\boldsymbol{s})\boldsymbol{I}_N}_{\text{between-subject variance}},\end{aligned} \tag{2.18}$$

where $\sigma^2_i(\boldsymbol{s})$ is the within-subject variance for the ith subject.

From this point on, two approaches are frequently used in the fMRI literature. The first approach assumes that the within-subject variances are equal across individuals so that the errors are *homoscedastic*, meaning that each error term also has equal variance. The second approach relaxes this assumption, allowing the within-subject variances to differ between individuals. In this case, the variance of the error

terms also differ, meaning that the error terms are *hetroscedastic*.

2.6.8 Solving the Second-Level GLM with Homoscedastic Errors

Assuming that the within-subject variance terms are equal, $\sigma_1^2(\boldsymbol{s}) = ... = \sigma_N^2(\boldsymbol{s})$, 2.18 reduces to

$$\mathrm{Cov}(\boldsymbol{\epsilon}^*(\boldsymbol{s})) = (\sigma_S^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s}))\boldsymbol{I}_N \tag{2.19}$$

where $\sigma_S^2(\boldsymbol{s})$ is the common within-subject variance. Since this means the errors are spherical, the group-level effect estimates can be obtained using OLS as described in 2.6.2. For the one-sample t-test where $\boldsymbol{X}$ is a column vector with all elements equal to one, this leads to the parameter estimates

$$\hat{\boldsymbol{\beta}}_G(\boldsymbol{s}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i(\boldsymbol{s}). \tag{2.20}$$

2.6.9 Solving the Second-Level GLM with Hetroscedastic Errors

Assuming that the within-subject variance terms are not equal, then the errors are hetroscedastic, and thus the sphericity assumption required for OLS is violated. In this case, a *weighted least square* (WLS) approach is used to solve the GLM instead. Conceptually, the idea of WLS is to deweight the most variable subject-level parameter estimates contained in $\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s})$. In practise, this is done by constructing a weight matrix $\boldsymbol{W}(\boldsymbol{s})$ equal to the inverse of the covariance matrix of the observations. In the context of the second-level analysis, 2.18 leads to the weight matrix

$$\boldsymbol{W}(\boldsymbol{s}) = \begin{pmatrix} (\sigma_1^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s}))^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & (\sigma_N^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s}))^{-1} \end{pmatrix}. \tag{2.21}$$

The WLS parameter estimates are then computed as

$$\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{s}) = (\boldsymbol{X}^\intercal\boldsymbol{W}(\boldsymbol{s})\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{W}(\boldsymbol{s})\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}(\boldsymbol{s}), \tag{2.22}$$

which for the one-sample t-test example reduces to

$$\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{s}) = \left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s})}\right)^{-1}\sum_{j=1}^{N}\frac{\boldsymbol{c}^\intercal\hat{\boldsymbol{\beta}}_j(\boldsymbol{s})}{\sigma_j^2(\boldsymbol{s}) + \sigma_G^2(\boldsymbol{s})}. \tag{2.23}$$

In the WLS approach to solving the GLM, the weight matrix $\boldsymbol{W}(\boldsymbol{s})$ effectively

applies a whitening transformation to the data. In fact, it can be advantageous to simplify WLS using a similar procedure as described for Prewhitening in 2.6.3. Using the same notation as 2.6.3, the error terms in the group-level GLM can be whitened using the diagonal matrix $\boldsymbol{K}(\boldsymbol{s})$ where $K_{ii}(\boldsymbol{s}) = \sqrt{W_{ii}(\boldsymbol{s})}$. The updated model

$$\boldsymbol{K}(\boldsymbol{s})\boldsymbol{c}^{\intercal}\hat{\boldsymbol{\beta}}(\boldsymbol{s}) = \boldsymbol{K}(\boldsymbol{s})\boldsymbol{X}\boldsymbol{\beta}_G(\boldsymbol{s}) + \boldsymbol{K}(\boldsymbol{s})\boldsymbol{\epsilon}^*(\boldsymbol{s}) \tag{2.24}$$

satisfies the sphericity assumption, and can therefore be solved using OLS. The parameter estimates obtained with this method are equivalent to solving the original model directly with WLS.

In this presentation, we have assumed that the within- and between-subject variance components are known, when in practise they must be estimated. Usually, the within-subject variance estimates obtained from the first-level analyses are also used in the group-level model. There are many proposed methods for estimating the between-subject variance, several of which use an iterative procedure based on OLS or residual maximum likelihood estimates. For more, see (Searle et al., 2009; Woolrich et al., 2004; Worsley et al., 2000).

2.7 The Multiple Comparisons Problem

A fundamental issue with the mass-univariate approach for fMRI inference is the *multiple comparisons problem*. Because the GLM is applied at each voxel independently, across the entire brain mask this means thousands of statistical tests are performed simultaneously. The significance level for each test can be described as the probability of wrongly determining a 'discovery' when the null-hypothesis of no activation is in fact true. In statistical terms, this is the probability of making a *type I error*. The problem is, while a significance level of 5% may be appropriate for one test, as more inferences are carried out the probability of an erroneous inference also increases. To highlight this point, if 100 tests are performed independently, each with a significance level of 5%, then the probability of at least one false discovery is 99.4%. In the context of fMRI, where a typical brain mask contains over 100,000 voxels, this problem is especially severe – if the voxelwise significance level of 5% is not corrected, we can expect over 5,000 brain voxels to falsely be determined as active in the thresholded statistical results.

In the task-fMRI literature, two different approaches are commonly used to correct for the multiple comparisons problem. The *false discovery rate* (FDR) procedure corrects the significance level to control the expected proportion of type I errors across all detected voxels in the thresholded statistical results. For example, if

a 0.05 FDR procedure is applied for subject-level inference on 20 subjects data, in any individual's thresholded statistical image we expect 5% of activations to be false-positives. The second approach is to employ a *familywise error* (FWE) correction, used to control the expected frequency that *any* type I errors are made across the whole brain. This is clearly a more stringent form of correction – in the 20-subject example, using a 0.05 FWE procedure we expect only one individual's thresholded image to contain any false-positives.

In making a choice between FDR or FWE correction there is a trade-off. While FDR is more statistically powerful than FWE, the drawback is a greater risk of false activations. This has lead to criticism concerning the spatial specificity of FDR-corrected results; particularly, since the location of false activations in an FDR-thresholded map are unknown, it is not possible to say with certainty that any given voxel is activated. On the other hand, FWE correction has also come under fire for being too conservative.

Multiple comparison correction procedures are carried out in fMRI at either the voxelwise or clusterwise level. Voxelwise inference is intuitive, once a corrected threshold has been determined via the FDR or FWE procedure, the thresholded statistical results are computed as all voxels whose *t*-statistic value exceeds the corrected threshold. Clusterwise inference involves a two-step procedure. First, a primary voxelwise threshold c is chosen, usually in correspondence with an uncorrected significance level such as $\alpha = 0.005$. Thresholding the statistic map at c creates groups of contiguous voxels above c, or 'clusters'. For this reason, c is commonly referred to as the *cluster-forming threshold*. Subsequently, a *cluster-extent threshold* k is determined based on the distribution of cluster sizes obtained under the null-hypothesis of no activation. The final thresholded results are computed as all suprathreshold clusters with a spatial extent larger than k.

FDR and FWE correction procedures can be applied for both voxelwise and clusterwise inference, with the key difference that voxelwise corrections are based on the sampling distribution of voxel intensities under the null-hypothesis of no signal, while clusterwise corrections use the sampling distribution of cluster size. In the remainder of this section, we consider two methods for obtaining FWE-corrected results at the voxelwise level. First, we present how the body of mathematics called *Random Field Theory* (RFT) is used to treat FWE correction with a parametric approach. Then, we show how FWE correction may also be obtained with permutation methods that make weaker assumptions about the data. While we will not consider FDR here, for more information a reader may refer to Benjamini and Hochberg (1995) where the method was originally proposed, or Genovese, Lazar, and Nichols

(2002) for a presentation of FDR in the context of functional neuroimaging.

2.7.1 Random Field Theory for Voxelwise FWE Correction

Voxelwise FWE control is established using an intrinsic relationship linking the probability of making a false discovery with the distribution of the maximum t-statistic over the brain. For a threshold c,

$$P(\text{Reject } H_0 | H_0 \text{ true}) = P\Big(\max_{\boldsymbol{s} \in S} T(\boldsymbol{s}) > c \Big) \tag{2.25}$$

where $T(\boldsymbol{s})$ is the t-statistic map given in 2.11 and H_0 is the null hypothesis of no activation.

The intuition of RFT is that under the *global null* of no signal *anywhere*, $T(\boldsymbol{s})$ can be modelled by a stationary continuous Gaussian random field $Z(\boldsymbol{s})$ of mean zero and unit variance over the same domain S. Note that this imposes assumptions that the data are stationary and that the discretely sampled statistic image $T(\boldsymbol{s})$ is sufficiently smooth enough to be approximated by a continuous field. It is because of the latter that spatial smoothing of the data must be carried out during preprocessing. Supposing that the model is valid, then the right-hand side of 2.25 is approximated by $P\Big(\max_{\boldsymbol{s} \in S} Z(\boldsymbol{s}) > c \Big)$. Remarkably, this probability can be obtained with use of the *Euler characteristic* (EC), a measure originating from algebraic topology that provides information about a shape's fundamental structure, regardless of ways in which the shape is distorted or deformed.

For a threshold c, defining the *excursion set* A_c as the set of voxels where the random field exceeds c, i.e. $A_c = \Big\{ \boldsymbol{s} \in S : Z(\boldsymbol{s}) > c \Big\}$, the EC χ_c can be characterized as counting the number of clusters minus the number of 'handles' plus the number of 'holes' in A_c. If c is chosen large enough then we expect the handles and holes to disappear, so the EC provides an approximation of the number of clusters. This relates back to the maximum distribution because if $\max_{\boldsymbol{s} \in S} Z(\boldsymbol{s}) > c$, then clearly there must be at least one suprathreshold cluster in the excursion. Putting all this information together,

$$\begin{aligned} P\Big(\max_{\boldsymbol{s} \in S} T(\boldsymbol{s}) > c \Big) &\approx P\Big(\max_{\boldsymbol{s} \in S} Z(\boldsymbol{s}) > c \Big) \\ &\approx P\Big(\chi_c > 0 \Big). \end{aligned} \tag{2.26}$$

Finally, further increasing of the threshold c will result in fewer voxels contained in the excursion set until χ_c will almost certainly take the value of 1 (if the

excursion set is made up of one suprathreshold cluster) or 0 (if the excursion set is empty). In this case,

$$P\Big(\chi_c > 0\Big) = \mathbb{E}\Big(\chi_c\Big). \tag{2.27}$$

In practice, with a parametric approach the FWE-corrected p-value is always approximated using the expectation of the EC. When $S \subset \mathcal{R}^3$, e.g. S is the brain mask in a neuroimaging application, the expected EC has the closed-form solution

$$\mathbb{E}\Big(\chi_c\Big) \approx \lambda(S)|\Lambda|^{\frac{1}{2}}(c^2-1)\exp\left(-c^2/2\right)/(2\pi)^2, \tag{2.28}$$

where $\lambda(S)$ is the Lebesgue measure (i.e. the volume) of the brain mask, and $|\Lambda|$ is the determinant of the covariance matrix of the gradient of $Z(\boldsymbol{s})$,

$$|\Lambda| = \left|\mathrm{Cov}\left(\left[\frac{\partial}{\partial x}Z(\boldsymbol{s}), \frac{\partial}{\partial y}Z(\boldsymbol{s}), \frac{\partial}{\partial z}Z(\boldsymbol{s})\right]\right)\right|. \tag{2.29}$$

Essentially, $|\Lambda|$ provides a measure of the smoothness of the random field $Z(\boldsymbol{s})$; for a less smooth process $Z(\boldsymbol{s})$, the determinant is larger.

2.7.2 Permutation Testing for Voxelwise FWE Correction

In the previous section, we showed that the crux of estimating the FWE-corrected p-value is to approximate the maximum distribution of $T(\boldsymbol{s})$. We demonstrated how this was carried out with a parametric approach by assuming that under the global null, $T(\boldsymbol{s})$ can be modelled by a Gaussian random field. However, this imposed strong assumptions on the data which in practice are seldom fulfilled. In particular, it has been shown that RFT estimates of the FWE p-value are conservative unless the data are extremely smooth with high degrees of freedom. To remediate these problems, nonparametric methods have been proposed as an alternative.

The principle idea of permutation testing for FWE correction is that if the global null hypothesis is true, and there is really no signal anywhere, then the labels of each observation are arbitrary and the data is *exchangeable*. Therefore, the maximum distribution of $T(\boldsymbol{s})$ under the global null can be constructed empirically by creating a large number of surrogate realizations of the data, where on each realization the data points are permuted randomly and the maximum value is obtained from the corresponding statistic map.

In the first-level analysis model described in 2.6.6, where the observations are fMRI time-series data from one individual, permutation testing is usually inappropriate; due to temporal correlation in the data, the exchangeability assumption is vio-

lated. However, permutation testing is perfectly viable for the group-level model in 2.6.7, where the observations are each individual's first-level *contrast of parameter estimates* (cope) map $\boldsymbol{c}^\intercal\beta_1(\boldsymbol{s}), ..., \boldsymbol{c}^\intercal\beta_N(\boldsymbol{s})$.

In a two-sample t-test, where cope maps are obtained from two groups, e.g. a group of patients and a group of controls, permutation testing would be conducted by exchanging the labels between the patients and controls. For a one-sample t-test, where there are no group labels to swap, the principle of exchangability is replaced by an assumption of symmetry; if the null-hypothesis $H_0 : \boldsymbol{c}^\intercal\boldsymbol{\beta}(\boldsymbol{s}) = 0$ is true, then it should not matter if a change of sign is applied to all readings in any individual contrast image. In this case, a permutation test is carried out where each surrogate realization is established by randomly multiplying each individual contrast image $\boldsymbol{c}^\intercal\beta_i(\boldsymbol{s})$ by 1 or -1. Clearly, the assumption of symmetry for the noise distribution is much weaker than Gaussianity required for the parametric approach.

2.8 Reproducibility of fMRI Results

CHAPTER 3

Exploring the Impact of Analysis Software on Task-fMRI Results

3.1 Data and Analysis Methods

3.1.1 Study Description and Data Source

3.1.2 Data Analyses

3.1.3 Comparison Methods

3.1.4 Permutation Test Methods

3.2 Results

3.2.1 Cross-Software Variability for Parametric Inference

3.2.2 Cross-Software Varaibility for Non-Parametric Inference

3.2.3 Intra-Software Variability, Parametric vs Non-Parametric

3.3 Reproducibility

3.3.1 Scripting of Analysis and Figures

3.3.2 Results Sharing

3.4 Discussion

3.4.1 Limitations

3.5 Conclusion

CHAPTER 4

Spatial Confidence Sets for Task-fMRI Inference

CHAPTER 5

Contour Inference for Cohen's *d*

5.1 Theory

5.1.1 Transforming the Residual Field

5.2 Method

5.2.1 2D Simulations

5.2.2 3D Simulations

5.2.3 Application to UK Biobank Data

5.3 Results

5.3.1 2D Simulations

5.3.2 3D Simulations

5.3.3 UK Biobank Data

5.3.4 Comparison to Traditional Inference Procedures

5.4 Discussion

5.4.1 Limitations

5.5 Conclusion

CHAPTER 6

Conclusion and Future Work

Bibliography

George Adelman and Others. *Encyclopedia of neuroscience*. Birkhäuser, 1987.

Andrew L Alexander, Jee Eun Lee, Mariana Lazar, and Aaron S Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, July 2007.

Frederico A C Azevedo, Ludmila R B Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata E L Ferretti, Renata E P Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.*, 513(5):532–541, 2009.

P A Bandettini, E C Wong, R S Hinks, R S Tikofsky, and J S Hyde. Time course EPI of human brain function during task activation. *Magn. Reson. Med.*, 25(2):390–397, June 1992.

Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 1995.

Craig M Bennett, Michael B Miller, and George L Wolford. Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for multiple comparisons correction. *Neuroimage*, 47(Suppl 1):S125, 2009.

E L Bennett, M C Diamond, D Krech, and M R Rosenzweig. CHEMICAL AND ANATOMICAL PLASTICITY BRAIN. *Science*, 146(3644):610–619, October 1964.

Matthew J Brookes, Mark Woolrich, Henry Luckhoo, Darren Price, Joanne R Hale, Mary C Stephenson, Gareth R Barnes, Stephen M Smith, and Peter G Morris. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. *Proc. Natl. Acad. Sci. U. S. A.*, 108(40):16783–16788, October 2011.

Joshua Carp. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.*, 6:149, October 2012.

Joshua Carp. Optimizing the order of operations for movement scrubbing: Comment on power et al. *Neuroimage*, 76:436–438, August 2013.

R W Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.*, 29(3):162–173, June 1996.

Sean P David, Jennifer J Ware, Isabella M Chu, Pooja D Loftus, Paolo Fusar-Poli, Joaquim Radua, Marcus R Munafò, and John P A Ioannidis. Potential reporting bias in fMRI studies of the brain. *PLoS One*, 8(7):e70104, July 2013.

M C Diamond, D Krech, and M R Rosenzweig. THE EFFECTS OF AN ENRICHED ENVIRONMENT ON THE HISTOLOGY OF THE RAT CEREBRAL CORTEX. *J. Comp. Neurol.*, 123:111–120, August 1964.

Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U. S. A.*, 113(28):7900–7905, July 2016.

Tatiana Fomina, Matthias Hohmann, Bernhard Scholkopf, and Moritz Grosse-Wentrup. Identification of the default mode network with electroencephalography. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2015:7566–7569, 2015.

Michael D Fox and Ron L Alterman. Brain stimulation for torsion dystonia. *JAMA Neurol.*, 72(6):713–719, June 2015.

K J Friston, P Fletcher, O Josephs, A Holmes, M D Rugg, and R Turner. Event-related fMRI: characterizing differential responses. *Neuroimage*, 7(1):30–40, January 1998.

Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, April 2002.

Gary H Glover. Overview of functional magnetic resonance imaging. *Neurosurg. Clin. N. Am.*, 22(2):133–9, vii, April 2011.

Javier Gonzalez-Castillo, Ziad S Saad, Daniel A Handwerker, Souheil J Inati, Noah Brenowitz, and Peter A Bandettini. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proc. Natl. Acad. Sci. U. S. A.*, 109(14):5487–5492, April 2012.

Krzysztof J Gorgolewski and Russell A Poldrack. A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS Biol.*, 14(7): e1002506, July 2016.

C Goutte, F A Nielsen, and K H Hansen. Modeling the hemodynamic response in fMRI using smooth FIR filters. *IEEE Trans. Med. Imaging*, 19(12):1188–1201, December 2000.

Richard J Hargreaves and Michael Klimas. Imaging in drug development. In *Principles of Clinical Pharmacology*, pages 327–341. Elsevier, 2012.

John P A Ioannidis. Why most published research findings are false. *PLoS Med.*, 2(8): e124, August 2005.

John P A Ioannidis, Marcus R Munafò, Paolo Fusar-Poli, Brian A Nosek, and Sean P David. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn. Sci.*, 18(5):235–241, May 2014.

Mark Jenkinson, Christian F Beckmann, Timothy E J Behrens, Mark W Woolrich, and Stephen M Smith. FSL. *Neuroimage*, 62(2):782–790, August 2012.

Suneil K Kalia, Tejas Sankar, and Andres M Lozano. Deep brain stimulation for parkinson's disease and other movement disorders. *Curr. Opin. Neurol.*, 26(4):374–380, August 2013.

K K Kwong, J W Belliveau, D A Chesler, I E Goldberg, R M Weisskoff, B P Poncelet, D N Kennedy, B E Hoppel, M S Cohen, and R Turner. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. U. S. A.*, 89(12):5675–5679, June 1992.

M H Lee, C D Smyser, and J S Shimony. Resting-state fMRI: a review of methods and clinical applications. *AJNR Am. J. Neuroradiol.*, 34(10):1866–1872, October 2013.

Megan H Lee, Carl D Hacker, Abraham Z Snyder, Maurizio Corbetta, Dongyang Zhang, Eric C Leuthardt, and Joshua S Shimony. Clustering of resting state networks. *PLoS One*, 7(7):e40370, July 2012.

Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage*, 45(1 Suppl):S187–98, March 2009.

Francisco López-Muñoz, Jesús Boya, and Cecilio Alamo. Neuron theory, the cornerstone of neuroscience, on the centenary of the nobel prize award to santiago ramón y cajal. *Brain Res. Bull.*, 70(4-6):391–405, October 2006.

Torben E Lund, Minna D Nørgaard, Egill Rostrup, James B Rowe, and Olaf B Paulson. Motion or activity: their role in intra- and inter-subject variation in fMRI. *Neuroimage*, 26(3):960–964, July 2005.

Paul M Matthews, Garry D Honey, and Edward T Bullmore. Neuroimaging: Applications of fMRI in translational medicine and clinical practice. *Nat. Rev. Neurosci.*, 7 (9):732, 2006.

Linda K McEvoy, Christine Fennema-Notestine, J Cooper Roddey, Donald J Hagler, Dominic Holland, David S Karow, Christopher J Pung, James B Brewer, and Anders M Dale. Alzheimer disease: Quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment, 2009.

Andrea Mechelli, Cathy J Price, Karl J Friston, and John Ashburner. Voxel-Based morphometry of the human brain: Methods and applications. *Curr. Med. Imaging Rev.*, 1 (2):105–113, 2005.

Paul E Meehl. Theory-Testing in psychology and physics: A methodological paradox. *Philos. Sci.*, 34(2):103–115, June 1967.

W Mohamed. The edwin smith surgical papyrus: Neuroscience in ancient egypt. *IBRO History of Neuroscience*, 2014.

Malaak N Moussa, Matthew R Steen, Paul J Laurienti, and Satoru Hayasaka. Consistency of network modules in resting-state FMRI connectome data. *PLoS One*, 7(8): e44428, August 2012.

S Ogawa, T M Lee, A R Kay, and D W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. U. S. A.*, 87(24): 9868–9872, December 1990.

S Ogawa, D W Tank, R Menon, J M Ellermann, S G Kim, H Merkle, and K Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. U. S. A.*, 89(13):5951–5955, July 1992.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, August 2015.

O Parker Jones, N L Voets, J E Adcock, R Stacey, and S Jbabdi. Resting connectivity predicts task activation in pre-surgical populations. *NeuroImage: Clinical*, 13:378–385, January 2017.

William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier, April 2011.

Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.*, 18(2):115–126, February 2017.

William W Rozeboom. The fallacy of the null-hypothesis significance test. *Psychol. Bull.*, 57(5):416–428, 1960.

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance Components*. John Wiley & Sons, September 2009.

P Skudlarski, R T Constable, and J C Gore. ROC analysis of statistical methods used in functional MRI: individual subjects. *Neuroimage*, 9(3):311–329, March 1999.

Stephen M Smith. Fast robust automated brain extraction. *Hum. Brain Mapp.*, 17(3): 143–155, November 2002.

Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, and Christian F Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U. S. A.*, 106(31):13040–13045, August 2009.

José M Soares, Paulo Marques, Victor Alves, and Nuno Sousa. A hitchhiker's guide to diffusion tensor imaging. *Front. Neurosci.*, 7:31, March 2013.

Max Sommerfeld, Stephan Sain, and Armin Schwartzman. Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. *J. Am. Stat. Assoc.*, 113(523):1327–1340, July 2018.

Reisa A Sperling, Dorene M Rentz, Keith A Johnson, Jason Karlawish, Michael Donohue, David P Salmon, and Paul Aisen. The A4 study: stopping AD before symptoms begin? *Sci. Transl. Med.*, 6(228):228fs13, March 2014.

William C Stacey and Brian Litt. Technology insight: neuroengineering and epilepsy-designing devices for seizure control. *Nat. Clin. Pract. Neurol.*, 4(4):190–201, April 2008.

I Tavor, O Parker Jones, R B Mars, S M Smith, T E Behrens, and S Jbabdi. Task-free MRI predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, April 2016.

Choong-Wan Woo, Anjali Krishnan, and Tor D Wager. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage*, 91:412–419, May 2014.

M W Woolrich, B D Ripley, M Brady, and S M Smith. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, 14(6):1370–1386, December 2001.

Mark W Woolrich, Timothy E J Behrens, Christian F Beckmann, Mark Jenkinson, and Stephen M Smith. Multilevel linear modelling for FMRI group analysis using bayesian inference. *Neuroimage*, 21(4):1732–1747, April 2004.

K J Worsley, C Liao, M Grabove, V Petre, B Ha, and A C Evans. A general statistical analysis for fMRI data, 2000.