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 – 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.

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 correction (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, 2012a), 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 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. However, in recent times they have been a source of major controversy. In 2016, a shocking paper by *Eklund et. al* discovered that many fMRI software packages were incorrectly carrying out the multiple-correction procedures for clusterwise inferences, 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 (Choong-Wan Woo, Cluster-extent based). 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). 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 magnitude 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*, *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 Blood-Oxygen-Level-Dependent (BOLD) functional Magnetic Resonance Imaging (MRI) that will be the main focus of study in this thesis. Here, we describe each of the preprocessing and modelling components of a typical task-fMRI analysis pipeline. Finally, we give an in-depth discussion of the state-of-the-art procedures used for subject- and group-level fMRI inference that are of particular relevance to the remaining chapters of this work.

2.1 The Study of Brain Function

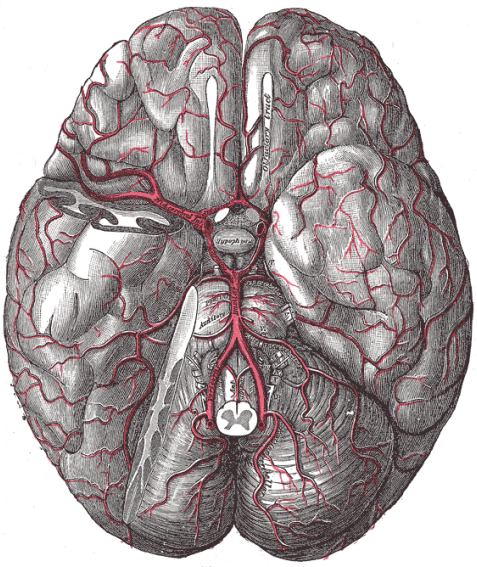

Figure 1: An illustration showing the arteries at the base of the human brain from (Gray, 1918).

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.

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; the first evidence of neuroplasticity, the ability for the brain's structure to change during an individual's lifetime; 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 in-

fluencing 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, and neuroengineering techniques based on our capability to stimulate neural circuits are implemented to treat Parkinson's disease and dystonia. Structural- and functional-MRI are being explored to determine biomarkers for diagnosis of Alzheimer's disease *prior* to symptom onset, 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.

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. 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. Additionally, one could apply diffusion tensor imaging (DTI) to instead map white matter tractography in the brain. 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. On the other hand, task-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.

Crucially, 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 techniques, more recently, the same correlation patterns have been independently detected using EEG and MEG. 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. On a similar concept, another effort has shown that resting state connectivity features may be used to predict task-evoked activation. This research signals towards an innate functional signature that defines our behaviour, while also providing potential clinical solutions to obtain task-fMRI data from patients who are unable to perform a specific task.

2.2 Magnetic Resonance Imagery (MRI)

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

2.3.1 Overview

2.3.2 Pre-processing

2.4 Statistical Analysis: Subject-level

2.4.1 Parametric Methods

2.4.2 Nonparametric Methods

2.5 Statistical Analysis: Group-level

2.5.1 Parametric Methods

2.5.2 Nonparametric Methods

2.6 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

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