Statistical Models for Neuroimaging Meta-analytic Inference

Gholamreza Salimi-Khorshidi

Analysis Group

Oxford Centre for Functional MRI of the Brain

Department of Clinical Neurology

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Supervised by Prof. Stephen M. Smith and Dr. Thomas Nichols

March 2011

Abstract

A statistical meta-analysis combines the results of several studies that address a set of related research hypotheses, thus increasing the power and reliability of the inference. Meta-analytic methods are over 50 years old and play an important role in science; pooling evidence from many trials to provide answers that any one trial would have insufficient samples to address. On the other hand, the number of neuroimaging studies is growing dramatically, with many of these publications containing conflicting results, or being based on only a small number of subjects. Hence there has been increasing interest in using meta-analysis methods to find consistent results for a specific functional task, or for predicting the results of a study that has not been performed directly. Current state of neuroimaging meta-analysis is limited to coordinate-based meta-analysis (CBMA), i.e., using *only* the coordinates of *activation* peaks that are reported by a group of studies, in order to "localize" the brain regions that respond to a certain type of stimulus. This class of meta-analysis suffers from a series of problems and hence cannot result in as accurate results as desired.

In this research, we describe the problems that existing CBMA methods are suffering from and introduce a hierarchical mixed-effects image-based meta-analysis (IBMA) solution that incorporates the sufficient statistics (i.e., voxel-wise effect size and its associated uncertainty) from each study. In order to improve the statistical-inference stage of our proposed IBMA method, we introduce a nonparametric technique that is capable of adjusting such an inference for spatial nonstationarity. Given that in common practice, neuroimaging studies rarely provide the full image data, in an attempt to improve the existing CBMA techniques we introduce a fully automatic model-based approach that employs Gaussian-process regression (GPR) for estimating the meta-analytic statistic image from its corresponding sparse and noisy observations (i.e., the collected foci). To conclude, we introduce a new way to approach neuroimaging meta-analysis that enables the analysis to result in information such as "functional connectivity" and networks of the brain regions' interactions, rather than *just* localizing the functions.

Acknowledgements

This thesis would not have been possible without help, support, and data from many people. I would particularly like to thank:
- Steve Smith and Tom Nichols for outstanding supervision, for always making time for me and steering me back onto the right track and teaching me how to carry out an academic research;
- Again to Steve Smith for fostering a collaborative and supportive atmosphere at FMRIB (especially in the analysis group) that has helped to make this degree a joy;
- Mark Woolrich for his FLAME paper/code and extremely helpful advice on Gaussian processes and Bayesian inference;
- Irene Tracey and her team (John Keltner, Merle Fairhurst, Siri Leknes, Mike Lee and Chantal Berna), Gwenaëlle Douaud and probably many others at FMRIB for the data sets that formed the foundation of this work;
- My fellow student and great friend Morgan Hough who taught me about Neuroimaging, Star Wars, Family Guy and Jesus;
- Adrian Groves for kindly sharing his DPhil experience, particularly in C++ coding and GP/Bayesian modelling;
- Steven Reece for being a great transfer examiner, especially for his GP idea/advice;
- Alek Petrovic for brainstorming, help and support, and occasional Family-guy chat;
- Matthew Webster for randomise and C++ coding;
- The Guarantors of Brain, Soudafar Foundation, and European Science Foundation for generous travel grants;
- The DHPA and GlaxoSmithKline for funding my studies at Oxford;
- The organisers and tutors of FMRIB Graduate course for preparing me for all the regressions and P-values;
- Amirkabir University of Technology (Tehran, Iran) Biomedical Engineering program for preparing me for anything and opening my eyes to how much more there is yet to learn;
- FMRIB IT, Jalapeno, Skype, Facebook, YouTube, Gmail, iTunes, Torrent, Cisco VPNClient, RadioFarda, OUCS, 0 and 1 for making my Mac extremely useful;

- Mom and Dad for all their love and support, for encouraging curiosity in all things, and for their continuing guidance;
- My darling Mona for being a constant source of inspiration and reassurance, and for her endless patience for me;
- My great friends in Oxford Mo, Neil, Ravina, Ruxandra, Samira, Ana, Pouya, Mehrdad (via Skype), Louisa, Hannah, Joel & Sara, Jingyi, Ricarda and ... for making DPhil fun;
- Thomas Linacre for the sexy-subfusc, support and fun he and his students offered me over the past three years;
- Oxford University for being AWESOME; and
- Summertown for Esporta, Wine cafe, M&S, FindersKeepers, and 117 Frenchay Road.

Contents

1 Introduction **11**

1.1 Magnetic Resonance Imaging 12
1.2 FMRI Study Design and Analysis 13
1.2.1 Activation FMRI 13
1.2.2 Resting-state FMRI 15
1.3 Meta-analysis 16
1.4 Meta-analytic Inference under Nonstationarity 19
1.5 Existing Meta-analytic Problems 21
1.6 Outline of the Thesis 23

2 Meta-analysis of Neuroimaging Data: a Comparison of Image-based and Coordinate-based Pooling of Studies **27**

2.1 Introduction 30
2.2 Materials and Methods 33
2.2.1 IBMA Analyses 33
2.2.1.1 Combining Methods 35
2.2.1.2 Single-level Regression 35
2.2.1.3 Hierarchical Model for Fixed- or Mixed-Effects Inferences 36
2.2.1.4 Hierarchical Model for Image-based Meta-analysis 36
2.2.2 CBMA Analyses 37
2.2.2.1 ALE 37
2.2.2.2 KDA 39
2.2.2.3 MKDA 39
2.2.2.4 Group Comparisons with CBMA Methods 40
2.2.3 Data 40
2.2.4 Map Comparison 42
2.3 Results 44
2.4 Discussion and Conclusions 53

3 Adjusting The Effect Of Nonstationarity In Cluster-based And TFCE Inference **57**
3.1 Introduction . . . 59
3.2 Materials and Methods . . . 61
3.2.1 Smoothness in Random Field Theory . . . 61
3.2.2 RFT Cluster-size Adjustment . . . 63
3.2.3 Empirical Cluster-size Adjustment . . . 64
3.2.4 Empirical TFCE Adjustment . . . 65
3.2.5 Nonstationarity Assessment . . . 66
3.2.6 ROC-based Evaluations . . . 68
3.2.7 Data . . . 69
3.3 Results . . . 72
3.4 Discussion and Conclusions . . . 82

4 Using Gaussian-Process Regression for Meta-analytic Neuroimaging Inference Based on Sparse Observations **88**
4.1 Introduction . . . 90
4.2 Materials and Methods . . . 93
4.2.1 Image-based Neuroimaging Meta-analysis . . . 94
4.2.2 Coordinate-based Neuroimaging Meta-analysis . . . 95
4.2.3 Using GPR for CBMA . . . 95
4.2.3.1 Hyperparameter Estimation with EO . . . 98
4.2.3.2 Prior on length-scale hyperparameter ℓ . . . 99
4.2.3.3 Fixing σ_f^2 to Account for Coordinate Sampling Bias . . . 99
4.2.4 CBMA Notations . . . 100
4.2.4.1 GPR Notations . . . 101
4.2.5 Data . . . 101
4.2.5.1 Simulated Data . . . 101
4.2.5.2 Real Data . . . 104
4.2.6 Map Comparison . . . 106
4.3 Results . . . 107
4.3.1 Simulated Data: 1D . . . 107
4.3.2 Simulated Data: 3D . . . 111
4.3.3 fMRI Data . . . 114
4.4 Discussion and Conclusions . . . 117

5 Identifying Modulatory Network-interactions in the Brain: Correspondence between Activation and Rest **126**
5.1 Introduction . . . 128
5.2 Materials and Methods . . . 130
5.2.1 Data . . . 131
5.2.1.1 Simulated Data . . . 131
5.2.1.2 Resting FMRI Data . . . 133
5.2.1.3 BrainMap Data . . . 134
5.2.2 Functional Nodes . . . 135
5.2.3 Log-linear Graphical Model . . . 137
5.2.4 Correspondence Analysis . . . 139
5.2.5 Odds-ratio Analysis . . . 141
5.3 Results . . . 143
5.4 Conclusions and Discussion . . . 152

6 Conclusions and Future Works **157**
6.1 Summary of Contributions . . . 157
6.2 Problems to Overcome and Future Directions . . . 160
6.3 Final Conclusions . . . 162

A Image-based meta-analysis **163**
A.1 Pseudo-code for ALE Method . . . 163
A.2 Pseudo-code for KDA Method . . . 163
A.3 Pseudo-code for KDA Method . . . 164

B Nonstationarity **165**
B.1 Estimation of Smoothness/Roughness . . . 165
B.1.1 Kiebel's Method . . . 165
B.1.2 Jenkinson's Method . . . 166
B.2 Validity of Empirical Cluster Size Adjustment . . . 167
B.3 Storing the Empirical Statistics . . . 170
B.4 Pseudo-codes . . . 170

C Triplet Analysis **172**
C.1 Analysis Procedure . . . 172

List of Figures

2.1 Illustration of a 4-study, 1-dimensional meta analysis 38
2.2 IBMA map resulting from MFX at second level 45
2.3 Reference image against which CBMA methods are compared 46
2.4 Evaluating CBMA methods for different kernel parameter values 47
2.5 Illustration of a 20-study, 1-dimensional meta analysis 47
2.6 Evaluating the effect of smoothing extent of studies on optimal CBMA methods' kernel parameter 48
2.7 The reference IBMA map 49
2.8 Using a third-level design to answer cognitive pain questions 51
2.9 Using difference contrasts to answer cognitive pain questions 52
2.10 The effect of deactivation information in difference contrasts 53

3.1 Sample slices from the simulated data 71
3.2 Heterogeneity of cluster-size test's false positive risk for simulated null data 75
3.3 Heterogeneity of TFCE false positive risk for simulated null data 76
3.4 Heterogeneity of false positive risk for null fMRI data versus applied image smoothing 77
3.5 Heterogeneity of false positive risk for null VBM data versus applied image smoothing 78
3.6 Joint sensitivity-power evaluation using area under the ROC curve (AUC) with simulated nonstationarity data 79
3.7 VBM cluster-size results (MCI>AD) with smoothness maps 80
3.8 TFCE results for VBM data (MCI>AD) using different adjustment methods 81
3.9 The 33 clusters surviving the cluster-forming threshold of T-stat=3 . . . 83
3.10 Relationship between empirical and RPV measures of cluster size for the VBM data 85

4.1 The effect of GP's hyperparameters on its shape 97
4.2 Comparing the $\text{GPR}_{\text{joint,s}}$ with $\Gamma(1,5)$ prior on ℓ 108
4.3 GPR's flexibility in accommodating the prior knowledge on ℓ 109

4.4 GPR's flexibility in incorporating various foci types 110
4.5 The results of $GPR_{joint,s}$ with $\Gamma(1,5)$ prior on ℓ 111
4.6 IBMA and CBMA results when pooling a one-group set of 3D simulated studies 112
4.7 IBMA and CBMA results when contrasting two groups of 3D simulated studies 113
4.8 Using DC for evaluating the performance of CBMA methods 114
4.9 The foci- and frequency-maps from the fMRI study pool 115
4.10 The effect of the number of foci included in the CBMA on the inferred ℓ 116
4.11 IBMA and CBMA results when pooling a one-group set of fMRI studies . 118
4.12 IBMA and CBMA results when contrasting two groups of fMRI studies . 119
4.13 Using DC for assessing the performance of CBMA methods when applied to pooling real fMRI studies 120
4.14 Using the area under the ROC curve for assessing the performance of CBMA methods when applied to the real fMRI data 120
4.15 The effect of fixed σ_f in $GPR_{joint,s}$'s performance 121

5.1 The network topology fed into the FMRI data simulations and its corresponding connection matrix 132
5.2 Summary diagram of the method, illustrating all the various steps involved in the analysis 142
5.3 Subject-level and pooled $\log_{10}$(OR) and Z-stat for a triplet of nodes with non-additive interaction 144
5.4 The spatial maps of the ICA components 145
5.5 A sample triplet with (a subset of) its associated time-series 147
5.6 The extent of correspondence between the three-way interactions found in BrainMap and rFMRI 148
5.7 Assessing the inter-domain variability of three-way interactions in BrainMap data for a triplet 150
5.8 Assessing the inter-subject variability of three-way interactions in rFMRI data for a triplet 151
5.9 Regions with high frequency of appearance in three-way-interactive triplets 153
5.10 The number of studies in each sub-population of the BrainMap database 155

List of Tables

3.1 The coordinates (x,y,z) of the centre of the Gaussian functions and their standard deviation 70
3.2 Specifications of the clusters formed in real VBM analysis 84

4.1 The list of analyses carried out in this chapter in terms of their underlying model, inputs and outputs 100
4.2 The list of GPR inference scenarios utilized in this chapter 101
4.3 The list of parameters used in 1D and 3D data simulation 104

Chapter 1

Introduction

A statistical meta-analysis combines the results of several studies that address a set of related research hypotheses, thus increasing the power and reliability of the inference (Sutton et al., 2000). Meta-analytic methods are over 50 years old (Fisher, 1948) and play an important role in science; pooling evidence from many trials to provide answers that any one trial would have insufficient samples to address. For example, a recent meta-analysis of clinical trials of depression found that anti-depressants were ineffective on all but sickest patients (Kirsch et al., 2008).

The number of neuroimaging studies is growing dramatically, where a conservative measure of the functional magnetic resonance imaging (FMRI) literature shows a growth from 2 publications in 1993 to 1,970 publications in 2007 and an exponential growth since 2000, predicting a doubling of the yearly publication rate every 3.5 years[1]. However many of these publications contain conflicting results, or are based on only a small number of subjects. Hence there has been increasing interest in using meta-analysis methods to find consistent results for a specific functional task. These same methods can also be used to predict the results of a study that has not been performed directly, by combining studies that intersect on a particular concept.

The typical number of subjects in an FMRI study is low, from 10 to 20 subjects, which results in low power (probability of detecting a true positive) and increasing the chance that the results will not replicate in another group of subjects. For example, Thirion et al.

[1]Based on a per-year PubMed search of "FMRI" in title or abstract.

(2007) investigated the reproducibility of statistical inference over different numbers of randomly-selected subjects from a pool of 80 subjects performing the same task. They showed that the number of subjects needed to give a generalizable result is greater than 20. This suggests that studies in the literature based on small samples are difficult to interpret in isolation and researchers could greatly benefit from pooling evidence from multiple studies.

The standard brain imaging meta-analysis methods are based on just the *x,y,z* locations of local maxima in a statistic image. These locations comprise only a small fraction of all the information in the statistical results, yet the number of publications using these methods continues to grow. Thus the overall goal of the work that follows is to understand the limitations of the current imaging meta-analysis methods, and propose new methods that make better use of the data, and provide more interpretable results. Having investigated such methods, we conclude by applying them to assessment of the correspondence between the "active" and "resting" brain in terms of their underlying functional components and the functional connectivity of such components.

The purpose of this chapter is to provide a brief introduction to both functional MRI and meta-analysis, describe the current state of neuroimaging meta-analysis, and outline the rest of the thesis. First, we describe magnetic resonance imaging (MRI) and its use in functional neuroimaging, design and analysis of neuroimaging studies and a brief introduction to meta-analysis as a statistical tool for pooling studies. Next, we describe the recent advances in neuroimaging and challenges that current neuroimaging meta-analysis is facing. We conclude this chapter with an outline of the thesis, including a short summary of each chapter.

1.1 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a technique commonly used in visualizing the anatomy and function of the human body, which employs three main components in order to generate an image. As the main component, a *strong magnetic field* is used to

align nuclear magnetization of the hydrogen atoms' protons of the water molecules in the body. The resulting magnetization of the protons is in the same direction as the main magnet and thus is undetectable. This justifies the use of a *radio-frequency (RF) pulse* for altering the alignment of this magnetization in order to produce a transverse magnetic field that is detectable by the scanner. The hydrogen protons precess at a known frequency proportional to the strength of the constant magnetic field, known as the Larmor frequency. Thus, a *magnetic field gradient* can be used to encode spatial location in the frequency and (with changing gradients) the phase of the precessing protons to reconstruct an image. An excellent reference for MR physics and image formation is Jezzard et al. (2003).

Functional MRI (FMRI) is a specialized MRI scan that indirectly measures changes in neuronal activity. An increase in local neuronal activity increases oxygen consumption, which increases blood flow and blood volume. The flow and volume increases provide an abundance of oxyhemoglobin, decreasing the proportion of deoxygenated haemoglobin, which results in an increase in signal intensity (Jezzard et al., 2003). While this blood-oxygen-level dependent (BOLD) signal is not quantitative, and percent change in signal is only approximately proportional to blood flow change, this is the most commonly used method in neuroscience imaging studies of human subjects.

1.2 FMRI Study Design and Analysis

An FMRI study consists of imaging the brain while the subject performs a cognitive task or is at rest. In the first part of this section we describe different steps in a generic activation-data analysis, which will be followed by a brief introduction to resting-state data and its analysis.

1.2.1 Activation FMRI

There are two types of experimental designs used in activation FMRI studies: blocked and event-related design (Amaro and Barker, 2006). In a blocked design, subjects should

maintain their cognitive engagement to a series of stimuli of the same condition (presented subsequently) over a period of time (until the beginning of the next condition's stimulus series). The resulting BOLD response corresponds to the whole block and thus has poor temporal resolution, is strong and robust, and results in more powerful statistical inference than with event-related designs, in general (Friston et al., 1999). In an event-related design, however, the BOLD response corresponding to each stimulus (as opposed to each block in blocked design) is detected, which can be analysed in detail, may be more robust to artifacts like head motion (because of its speed), allows for randomization of the order of conditions (and hence more complicated experiments), and allows for analyses of the individual responses to trials (D'Esposito et al., 1999).

After running the experiment, each subject's data consists of a 4D volume corresponding to each voxel's (i.e., an x, y, z coordinate) activity over time. This data must be preprocessed, which includes motion correction (and realignment) (Jenkinson et al., 2002) and spatial smoothing, before voxel-wise modelling of the time series using a standard haemodynamic response (convolution) function (HRF) (Woolrich et al., 2001; Beckmann et al., 2003; Woolrich et al., 2004). The final step of a neuroimaging analysis is the statistical inference (hypothesis testing) using parametric (Worsley et al., 1996; Poline et al., 1997; Cao and Worsley, 2001; Worsley et al., 2002) or nonparametric methods (Nichols and Holmes, 2002), including correcting the P-values for the multiple-comparison problem caused by searching the entire brain for significance (Nichols and Hayasaka, 2003). The result is an 'activation map' for a single subject, which is thresholded and typically shown as a colour overlay on an anatomical reference image. Such a single-subject analysis is referred to as a 'first-level analysis'. In order to combine different subjects' results together their brain-images need to be transformed into a common atlas space (Jenkinson and Smith, 2001; Jenkinson et al., 2002). When multiple subjects are combined it is referred to as a 'second-level analysis'. When a second-level analysis simply pools the intrasubject variance estimates, it produces a 'fixed-effects' (FFX) inference where significance is gauged against within-subject variation

(measurement error) only. When a second-level analysis considers both within- and between-subject variation, it is called a 'random-effects' (RFX) or 'mixed-effects' (MFX) analysis. RFX inferences should generalize to the population from which the units (here, subjects) are sampled to increase the final inference's statistical usefulness.

In order to form the parametric/statistical maps one can use a general linear model (GLM) as follows. Consider the analysis of the data from a group of subjects (i.e., a group analysis) where there are S subjects and that each subject, s, uses a within-subject analysis to estimate the effect-size at voxel k ($k = 1, ..., K$). This subject-specific effect size, $y_{s,k}$, can be shown to be given by:

$$\begin{aligned} y_{s,k} &= \alpha_{s,k} + w_{s,k}, \text{ where } w_{s,k} \sim \mathcal{N}(0, \tau^2_{s,k}). \\ \alpha_{s,k} &= \mu_k + u_{s,k}, \text{ where } u_{s,k} \sim \mathcal{N}(0, \sigma^2_k), \end{aligned} \tag{1.1}$$

where μ_k is the overall population mean effect at voxel k, $\alpha_{s,k}$ denotes the effect for subject s, $\tau^2_{s,k}$ represents the within-subject variances, and σ^2_k is the random-effects variance (or the inter-subject heterogeneity variance) (Woolrich et al., 2004). Combining the two lines in Equation 1.1 gives:

$$y_{s,k} = \mu_k + u_{s,k} + w_{s,k}, \tag{1.2}$$

which implies that group analysis estimates the voxel-wise group-level mean effect size, $\{\mu_k\}_{k=1,...,K}$, and between-subject variance, σ^2_k, using the subject-level summary statistics, $\{\{\alpha_{s,k}\}_{s=1,...,S}\}_{k=1,...,K}$ and $\{\{\tau^2_{s,k}\}_{s=1,...,S}\}_{k=1,...,K}$, e.g., by employing a Bayesian method such as FLAME (FMRIB's Local Analysis of Mixed Effects) (Woolrich et al., 2004).

1.2.2 Resting-state FMRI

Spontaneous or 'resting-state' fluctuations in the BOLD signal, as measured by FMRI, may present a valuable data resource for understanding the human neural functional architecture. Consistent large-scale spatial patterns of coherent signal have been identified in the human brain using both FMRI and positron emission tomography. Techniques

assessing functional connectivity, originally applied to BOLD FMRI data alongside studies of model-driven, task-evoked activation, have also proven useful for resting-state research and have greatly supported and contributed to increasing scientific interest in the spontaneous, or 'default' neural activity of the brain at baseline. As outlined in Cole et al. (2010), these methods provide useful conceptual complements to the inferences made from task-FMRI data, and hence are increasingly being applied across multiple fields of neuroscience, to further inform our understanding of the fundamental organisation of processing systems in the human brain.

The majority of approaches to analyzing resting-state FMRI data have thus far been spatially model-driven, with strong a priori hypotheses regarding the functional connectivity of a small number of brain regions of interest (ROIs) or individual voxel locations of interest. Recently, however, a great deal of attention has been focused on the patterns of connectivity between multiple ROIs within spatially distributed, large-scale "networks", characterized via both model-driven (e.g., seed-based correlation analysis) and data-driven analyses (e.g., independent component analysis (Beckmann et al., 2005)). These patterns have been variously termed 'intrinsic connectivity networks', or 'resting-state networks' (RSNs). They are purported to reflect the intrinsic energy demands of neuron populations that, via firing together with a common functional purpose, have subsequently wired together through synaptic plasticity. RSNs can be reliably and reproducibly detected at individual subject and group levels across a range of analysis techniques. A characteristic set of co-activating functional systems is found consistently across subjects (Beckmann et al., 2005; Smith et al., 2009), stages of cognitive development, and degrees of consciousness (please see Cole et al. (2010) for an extensive review).

1.3 Meta-analysis

In statistical hypothesis testing, there are two types of errors (or incorrect conclusions) that can be made. If a null hypothesis is incorrectly rejected when it is in fact true, this

is called a "Type I" error (also known as a false positive). A "Type II" error (also known as a false negative), however, occurs when a null hypothesis is not rejected despite being false. Small sample size, which happens to be an issue in neuroimaging studies, increases the chance of Type II error and hence reduces the statistical power (i.e., the probability that the test will reject a false null hypothesis) of the analysis (Thirion et al., 2007).

In order to overcome the statistical power problem of individual studies, meta-analysis pools a group of studies by employing appropriate statistical methods, in order to achieve a more reliable and powerful inference. Standard meta-analysis methods are typically based on combining the Z-stats or P-values of a group of studies. As reviewed in Lazar et al. (2002), the two most common methods are the Fisher's P-value combining (based on $-2\sum_{i=1}^{n}\log P_i \sim \chi^2_{2n}$ statistic) and Stouffer's average Z-stat method (based on $\sqrt{n}\bar{Z}$ statistic). These methods test the null hypothesis that all the studies are truly negative, resulting in a fixed effects inference. This type of inference can be driven by a single study and thus does not reflect the consistency of the studies considered.

While a behavioural study can completely report the Z-stats or P-values in one or more tables in a published report, a neuroimaging study has 10's or 100's of thousands of statistic values that can only be partially reported by traditional paper publication. Thus, neuroimaging papers summarize their results by providing the readers with a series of brain images (qualitative results) and a table including the local maxima locations (i.e., (x,y,z) coordinates) in a standard space (quantitative results). In spite of different solutions proposed for sharing the raw data or sufficient summary statistic images, neuroimaging still suffers from the lack of such data-sharing policy. As a result, the input to the standard neuroimaging meta-analysis consists of such coordinate lists, which we refer to as coordinate-based meta-analysis (CBMA). CBMAs harvest coordinates from individual journal papers or from curated databases such as BrainMap[2] (Laird et al., 2005b). There are several limitations to CBMA methods, one being the information loss due to the relative sparseness of such a representation of the image results, and

[2]http://www.brainmap.org

another being that coordinates are very sensitive to methods adopted in the study, from thresholding to report preparation (e.g., how many foci per cluster are reported) (Wager et al., 2007, 2009).

One of the first CBMA approaches was Fox et al.'s functional volumes modelling (FVM) method (Fox et al., 1997), though it lacked a formal statistical framework (Fox et al., 1998). The FVM method assumed a Gaussian spatial distribution of activations (Fox et al., 1999), though subsequent authors relaxed this assumption using non-parametric modelling of the distribution of foci (Nichols and Holmes, 2002; Wager et al., 2007). Currently, there are three widely used CBMA methods: ALE, KDA and MKDA. ALE, or activation likelihood estimation (Turkeltaub et al., 2002), is implemented in software provided by the BrainMap database. In brief, ALE constructs 'likelihood' maps for each activation focus by placing a 3D Gaussian density with specified FWHM at the focus location; these maps are then combined with the addition rule for probabilities, giving the probability that one or more foci are near a given voxel. KDA, or kernel density analysis (Wager et al., 2004) also treats each focus independently, but instead uses a spherical kernel and a simple addition rule to produce a map showing the number of foci within a given radius. MKDA, or multi-level KDA (Wager et al., 2007), does not treat each focus independently, and instead creates a binary map for each study, showing where there is one or more foci within a given radius; these study binary maps are then averaged, giving the proportion of studies having any foci within a given radius from a voxel. Unlike ALE and KDA, MKDA does not treat all foci equally and uses studies as the units of analysis, and thus minimizes the potential for one study with many foci to drive a meta-analytic result.

While authors of meta-analyses rarely have access to the complete original datasets, when they are available, it is natural to perform an image-based meta-analysis (IBMA), which combines whole-brain statistic volumes, rather than just using a summary of them (i.e., a list of local maxima coordinates). To account for both within- and between-study variance, a hierarchical FFX or MFX model is a natural approach. In FMRI a

generic hierarchical modelling framework is often used where, instead of modelling *all* of the data at all levels simultaneously, summary statistics are passed between levels of the hierarchy (Beckmann et al., 2003; Woolrich et al., 2004). It was shown by Beckmann et al. (2003) that a "two-level MFX model with its study-level parameters being estimated from parameter and variance estimates of the subject level" can be made equivalent to a "single complete mixed-effects model whose parameters are estimated directly from all of the original single sessions' time series data" if the (co-)variance at the second level is set equal to the sum of the (co-)variances in the single-level form. This statement is generalizable to fMRI meta-analysis, i.e., IBMA *only* requires the values of the parameter estimates and their (co-)variance from each study, generalizing the well-established "summary statistics" approach to IBMA.

Thus, an essential requirement of such models is that both effect size (contrast of parameter estimate, or COPE) images and their variance (variance of the contrast of parameter estimate, or VARCOPE) images are passed up from one level to the next, allowing subjects with poor precision to be down-weighted relative to high precision subjects, and provide MFX inferences that incorporate both within- and between-subject variation. At the third level, such methods down-weight *studies* with poor precision, and results in inferences that account for both within- and between-*study* variation. If between-study random variation is not of third-level analysis's interest, FFX model can be employed instead. Justification and interpretation of such models are discussed in Higgins et al. (2009).

1.4 Meta-analytic Inference under Nonstationarity

When detecting changes in functional or structural brain image data in a meta-analysis (as well as a study-level analysis), it is necessary to have powerful inference methods that offer precise control of false positive risk. Once a meta-analytic statistic image is created that assesses the evidence of an effect at each voxel, the two most common "thresholding" approaches are voxel-based and cluster-based inference. While voxel-wise

methods use a single threshold to classify signals as "real," cluster-based inference uses two thresholds, an arbitrary cluster-forming threshold followed by a cluster-size threshold to label clusters as "real". Cluster-based inference has a higher sensitivity compared to voxel-intensity-based tests when the signal is spatially extended (Friston et al., 1996; Poline et al., 1997).

An approach closely related to cluster-wise inference, that can be applied to meta-analytic statistic images, is threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009), which removes the dependence on the arbitrary cluster-forming threshold u_c. TFCE attempts to keep the sensitivity benefits of cluster-based inference (by using the cluster size information at a range of possible u_c values), while avoiding an arbitrary choice of a single u_c. The method produces a voxel-wise output image in which each voxel's value represents the accumulative cluster-like local spatial support over a range of cluster-forming thresholds.

When the image noise (i.e., IBMA model's residual) fails to have uniform smoothness it is said to be "nonstationary", as the correlation between neighboring voxels depends on where the voxels are located. Under nonstationarity, the sensitivity and specificity of a (standard, stationary) cluster-size test depend on local smoothness of the image, as bigger clusters are expected in smoother areas. Thus, any improvement in adjusting the statistical inference for nonstationarity can be of both study-level and meta-analytic inference's interest.

To overcome this problem, cluster sizes can be adjusted for nonstationarity with a local smoothness estimate based on RFT (Worsley et al., 1996). Alternatively, the adjusted cluster-size statistic can be assessed with a permutation test to obtain P-values, constituting a semi-parametric approach (Hayasaka et al., 2004): The statistic is derived using parametric RFT to adjust for the impact of spatially-varying smoothness, but non-parametric permutation is used to assign P-values. A comparison of the results from such adjusted cluster-size statistics and the maximal-voxel statistic is presented in Moorhead et al. (2005). TFCE, on the other hand, has been shown to give generally better sensitivity

than other methods over a wide range of test signal shapes and SNR values (Smith and Nichols, 2009; Smith et al., 2008). However, TFCE's performance when nonstationarity is present has not been studied to date.

1.5 Existing Meta-analytic Problems

As previously described, current state of neuroimaging meta-analysis is limited to CBMA, i.e., using *only* the coordinates of *activation* peaks that are reported by a group of studies, in order to "localize" the brain regions that respond to a certain type of stimulus. This class of meta-analysis suffers from a series of problems and hence cannot result in as accurate results as desired. Depending on such problems in the analysis being related to the input, output or underlying model, we call them "input, output and model problems", respectively. Even when carrying out an IBMA using the full COPE-VARCOPE data, a highly recommended approach is to use cluster-related (e.g., standard and threshold-free-enhanced cluster size) statistics, whose results are shown to be very sensitive to the spatial variation in image smoothness and hence require an adjustment for nonstationarity (Hayasaka et al., 2004) (i.e., "nonstationarity problem"). Finally, there are other ways to approach neuroimaging meta-analysis to enable the analysis to result in information such as "functional connectivity" and networks of the brain regions' interactions (Smith et al., 2010), rather than *just* localizing the functions (i.e., "approach problem").

Input problem It is important to appreciate the censoring that takes place by summarizing the full statistic images with a list of (x,y,z) coordinates. Well-established summary statistics approach to hierarchical neuroimaging modelling implies that in order to have the highest level of inference (e.g., a meta-analysis) as accurate as pooling all the lowest-level data (e.g., subject-level BOLD time series) in a single-level model, voxel-wise COPE and VARCOPE at lower levels *must* be passed on to the higher levels. However, in CBMA, COPE-VARCOPE images (each with 100s of thousands of values) are first

converted into a Z-stat image from which, approximately 10 to 20 (x,y,z) coordinates are incorporated in the analysis. It is clear that this is likely to result in a great loss of information, which will influence the meta-analysis result.

Model problem In spite of the censored nature of CBMA input data (i.e., a small group of voxels), existing CBMA models such as ALE and KDA, require their users to further censor the data by excluding the effect-size estimates (e.g., Z-stat) corresponding to the input coordinates. This also makes it impossible for the existing models to jointly incorporate the activation and deactivation information. In order for ALE and KDA to offer a solution that incorporates both activation and deactivation foci, they run two separate analyses: one using activation information, and the other using deactivation information. An additional problem caused by excluding the effect size is automatic estimation of the image smoothness. The only attempt made to infer the meta-analytic images' smoothness was by Eickhoff et al. (2009), by using the coordinate information. However, smoothness represents how similar voxel-wise "effect sizes" are as a function of their spatial distance from each other. Thus, any smoothness measure that does not incorporate effect-size similarity into account, will suffer from an estimation-accuracy problem.

Output problem This category of current CBMA problems addresses the fact that the output of methods such as ALE and KDA is different from what is provided by standard (image-based) analyses. That is, in an ideal world, one would prefer the CBMA results to resemble IBMA results, which is a COPE and VARCOPE (i.e., a mean effect and its standard deviation). Another important strength of IBMA is its RFX model that can account for heterogeneity in the study pool. Since such RFX variance is estimated based on the mean and SD of the study-level "effect sizes", which cannot be incorporated by CBMA, coordinate-based method still cannot yet provide an ideal output.

Nonstationarity problem One solution to avoid the previous problems is to use the study-level sufficient statistics and carry out an IBMA. However, as previously shown Hayasaka et al. (2004), validity of cluster-related inference, as one of the most advocated techniques for image-based inference, depends on stationarity of the statistical image. The problem is caused by the fact that in nonstationary images, clusters tend to be larger in smooth regions than in rough regions *just* by chance. Thus, in order to achieve an accurate meta-level inference, choosing IBMA over CBMA is not enough and requires an extra adjustment in order to cancel this effect out.

Approach problem Finding solutions for the above problems will presumably result in an accurate meta-level localization. However, given the number of studies that are stored in a large neuroinformatic database, one can redefine the meta-analysis problem as "finding the network of the brain-regions' interaction" rather than "finding the regions in charge of a certain class of stimuli". Under this approach, one can use the rich literature of graphical models and functional connectivity in order to address the structure under which the brain regions interact with each other (Smith et al., 2010).

1.6 Outline of the Thesis

The general outline of the thesis is as follows. Although we appreciate that some researchers may not have access to full-image data, for those who have this luxury, we first (in Chapter 2) introduce an IBMA technique and advocate it as the ideal solution. For those employing this solution, we recommend a complementary analysis (in Chapter 3) in order to adjust the (cluster-related) inference for the nonstationarity in the data. Given that most neuroimaging meta-analyses are CBMA and that existing CBMA techniques suffer from multiple weaknesses (e.g., subjective model selection, and excluding the effect size from the analysis), for those who cannot employ our IBMA solution (described in Chapters 2 and 3), we introduce a novel CBMA technique (in Chapter 4) that results in an accurate MFX meta-analysis. We conclude our research (in Chapter 5) by introducing

a new way of analyzing coordinate data (e.g., from BrainMap), where instead of using them for *localizing activation* in the brain, we introduce methods for analyzing them jointly with (resting-state) BOLD time-series, in order to pinpoint the structure of the brain's functional connectivity. This last solution can be useful for addressing cognitive neuroscience questions such as the extent of correspondence between the activation and rest, in terms of their functional connectivity.

In more detail, each of the major research chapters independently addresses one or more problems from the ones described above. That is, apart from the introductory text in Chapter 1, each of the other chapters has its own literature review, materials and methods description and conclusions.

Chapter 2 introduces a hierarchical IBMA technique (addressing CBMA's input problem) and presents the results from the evaluation of CBMA methods; using full image data of a set of studies to directly compare the results with when a sparse representation of the same data is used. By definition CBMA methods retain less information from each study than IBMA methods, but this work attempts to quantify how much information is lost, whether the CBMA methods can capture similar patterns of activations that IBMA methods provide, and how sensitive the CBMA results are to the change in the value of their spatial tuning parameters.

Chapter 3 is about adjusting the image-based inference (such as the one introduced in Chapter 2) for the nonstationarities (addressing the nonstationarity problem). Neuroimaging inference depends on the local image smoothness, as activation areas tend to be larger in smoother regions by chance alone. In order to adjust the inference for such nonstationarities, activation extents can be adjusted according to a local smoothness estimate. In our new model empirically-estimated average cluster size at each voxel is defined as a measurement of image smoothness. This adjustment technique plus the standard approach (using random-field theory) are employed in a nonparametric framework and tested on both simulated and real data; results show improvement in inference accuracy.

Chapter 4 returns to the coordinate-based meta-analysis in an attempt to improve the existing CBMA methods (addressing CBMA's input, output and model problems). In common practice, although IBMA is recommended, neuroimaging studies rarely provide the full image data and instead only report the magnitude and coordinate of their activation peaks in the papers. On the other hand, current CBMA models incorporate minimal information from published studies and do not infer the parameters of their model from the data, and hence result in an inaccurate solution. We solved these problems by employing Gaussian-process regression (GPR) for estimating the underlying statistic image at the location of each voxel given sparse noisy observations from the landscape at some nearby voxels. Our results on both simulated and real data show that GPR outperforms the existing CBMA techniques and is capable of offering an accurate solution to the problems that the current CBMA techniques suffer from.

Chapter 5 employs some of the previously described models in order to extract the functional components of the brain (at both activation and rest) and infer the structure of their multi-way interaction (addressing the approach problem). Neural connections, providing the substrate for functional networks, exist whether or not they are functionally active at any given moment. However, it is not known to what extent brain regions are continuously interacting when the brain is "at rest." In this work, we identify the major explicit activation networks by carrying out an image-based activation network analysis of thousands of separate activation maps derived from the neuroinformatic databases of functional imaging studies, involving nearly 30,000 human subjects. In addition, we extract networks of covariation when the brain is "at rest" by using resting FMRI data of 36 subjects. The results show that the full repertoire of functional networks utilized by the brain in action is continuously and dynamically active even when at rest. Next, we carry out a joint multivariate exploratory analysis of resting-state FMRI and activation data in order to pinpoint the structure of the underlying network of interactions. However, like others we have found that network modelling using a high number of nodes is difficult (at least for the task database), hence, we used a log-linear graphical model to look for

interactions, finding triplets of functional nodes which appear to interact (i.e., when one node modulates the functional connection between the other two).

Chapter 2

Meta-analysis of Neuroimaging Data: a Comparison of Image-based and Coordinate-based Pooling of Studies

Abstract

With the rapid growth of neuroimaging research and accumulation of neuroinformatic databases the synthesis of consensus findings using meta-analysis is becoming increasingly important. Meta-analyses pool data across many studies to identify reliable experimental effects and characterize the degree of agreement across studies. Coordinate-based meta-analysis (CBMA) methods are the standard approach, where each study entered into the meta-analysis has been summarized using only the (x,y,z) locations of peak activations (with or without activation magnitude) reported in published reports. Image-based meta-analysis (IBMA) methods use the full statistic images, and allow the use of hierarchical mixed effects models that account for differing intra-study variance and modeling of random inter-study variation. The purpose of this work is to compare image-based and coordinate-based meta-analysis methods applied to the same dataset, a group of 15 FMRI studies of pain, and to quantify the information lost by working only with the coordinates of peak activations instead of the full statistic images. We apply a 3-level IBMA mixed model for meta-analysis, highlighting important considerations in the specification of each model and contrast. We compare the IBMA result to three CBMA methods: ALE, KDA and MKDA, for various CBMA smoothing parameters. For the datasets considered, we find that ALE at $\sigma = 15mm$, KDA at $\rho = 25 - 30mm$ and MKDA at $\rho = 15mm$ give the greatest similarity to the IBMA result, and that ALE was the most similar for this particular dataset, though only with a Dice similarity coefficient of 0.45 (Dice measures range from 0 to 1). Based on this poor similarity, and the

greater modeling flexibility afforded by hierarchical mixed models, we suggest that IBMA is preferred over CBMA. To make IBMA analyses practical, however, the neuroimaging field needs to develop an effective mechanism for sharing image data, including whole-brain images of both effect estimates *and* their standard errors[1].

[1]The work in this chapter has appeared as Salimi-Khorshidi et al. (2009a).

2.1 Introduction

The number of neuroimaging studies is growing dramatically, with the FMRI literature having grown from 2 publications in 1993 to 1,970 publications in 2007, and exponential growth since 2000 predicting a doubling of the yearly publication rate every 3.64 years[2]. However many of these publications contain conflicting results, or are based on only a small number of subjects. Hence there has been increasing interest in using meta-analysis methods to find consistent results for a specific functional task. These same methods can also be used to predict the results of a study that has not been performed directly, by combining studies that intersect on a particular concept.

Many neuroimaging studies are under-powered, with the typical number of subjects ranging from 10 to 20 subjects. The main challenge in performing statistical inference over such small sample sizes is the limited power and, related, the chance that results will be reproduced in another group of subjects. For example, Thirion et al. (2007) investigate the reproducibility of statistical inference over different numbers of randomly-selected subjects from a pool of 80 subjects performing the same task. They show that the number of subjects needed to give a generalizable result is greater than 20. This suggests that studies in the literature based on small samples are difficult to interpret in isolation and researchers could greatly benefit pooling evidence from multiple studies.

A statistical meta-analysis combines the results of several studies that address a set of related research hypotheses, thus increasing power and reliability (Sutton et al., 2000). While authors of meta-analyses rarely have the complete original datasets, when they are available, it is natural to perform an image-based meta-analysis (IBMA) which combines whole-brain statistic volumes, rather than just using a summary of them (i.e., a list of local maxima coordinates). Lazar et al. (2002) review a number of ways to combine different subjects' statistic maps, and such methods can equally be applied to combining different studies' maps. In particular, the Fisher's P-value combining method

[2]Based on a per-year PubMed search of "FMRI" in title or abstract.

and Stouffer's average Z method ($\sqrt{n}\bar{Z}$) have frequently been used in traditional meta-analyses. These methods are fixed effects (FFX) methods, however, and their output does not reflect the consistency of the studies considered.

To account for both within- and between-study variance, a hierarchical mixed effects (MFX) model is a natural approach. In FMRI a generic hierarchical modeling framework is often used where, instead of modeling *all* of the data at all levels simultaneously, summary statistics are passed between levels of the hierarchy (Beckmann et al., 2003; Worsley et al., 2002). While this work has generally been used to combine first-level intra-subject FMRI model results into a second-level group FMRI model, it can be equally well used to combine multiple second-level studies into a third-level meta-analysis. An essential component of the work is that both effect size (contrast of parameter estimate) images and their variance (variance of the contrast of parameter estimate) are passed up from one level to the next, allowing subjects with poor precision to be down-weighted relative to high precision subjects, and provide MFX inferences that incorporate both within- and between-subject variation. At the third level, this translates to a method that can down-weight *studies* with poor precision, and inferences that account for both within- and between-*study* variation. At the third level, between-study random variation may not be of interest and so FFX may be used instead. For example, if one only wants to obtain the most sensitive pooling of a group of studies, a FFX inference at the 3rd (study) level would be appropriate. If, on the other hand, one wants to find the areas found most reliably in many studies, then a 3rd level MFX inference would be desired.

In common practice, neuroimaging studies rarely provide the full image data, and instead only activation foci magnitude and location are reported in journal papers, or submitted to results databases such as BrainMap[3] (Laird et al., 2005b). Hence most meta-analysis methods are based only on activation foci in a standard space (e.g. MNI152) and we called this the coordinate-based meta-analysis (CBMA) approach. There are several limitations to CBMA methods, one being the information loss due

[3]http://www.brainmap.org

to the relative sparseness of such a representation of the image results, and another being that coordinates are very sensitive to methods adopted in the study, from thresholding to report preparation (e.g., how many foci per cluster are reported) (Wager et al., 2007). For example, from one single dataset, three different sets of foci could be obtained depending on whether just three local maxima per cluster are reported (the default in SPM) above a corrected threshold, all local maxima are reported above a corrected threshold, or all maxima above an uncorrected threshold are reported. Since there is no universal standard for reporting results, CBMA methods should ideally take account of these differences but rarely do.

CBMA approaches were pioneered by Fox et al.'s functional volumes modeling (FVM) method (Fox et al., 1997), though it lacked a formal statistical framework (Fox et al., 1998). The FVM method assumed a Gaussian spatial distribution of activations (Fox et al., 1999), though subsequent authors relaxed this assumption using non-parametric modeling of the distribution of foci.

Currently, there are three widely used CBMA methods: ALE, KDA and MKDA. ALE, or activation likelihood estimation (Turkeltaub et al., 2002), is implemented in software provided by the BrainMap database. In brief, ALE constructs 'likelihood' maps for each activation focus by placing a 3D Gaussian density with specified FWHM at the focus location; these maps are then combined with the addition rule for probabilities, giving the probability that one or more foci are near a given voxel. KDA, or kernel density analysis (Wager et al., 2004) also treats each focus independently, but instead uses a spherical kernel and a simple addition rule to produce a map showing the number of foci within a given radius. MKDA, or multi-level KDA (Wager et al., 2007), does not treat each focus independently, and instead creates a binary map for each study, showing where there is one or more foci within a given radius; these study binary maps are then averaged, giving the proportion of studies having any foci within a given radius from a voxel. Unlike ALE and KDA, MKDA does not treat all foci equally and uses studies as the units of analysis, and thus minimizes the potential for one study with many foci to

drive a meta-analytic result.

By definition, the CBMA methods retain less information from each individual study than IBMA methods. However, it is an open question as to *how much* information is lost, and whether the CBMA methods can capture similar patterns of activations that IBMA methods provide. Further, the CBMA methods have spatial tuning parameters (Gaussian FWHM for ALE, and sphere radius for KDA and MKDA) which have no objectively-defined optimal setting. Hence the purpose of this work is to compare CBMA results to IBMA results for a variety of CBMA tuning parameter settings, to understand the relative sensitivity of each method and how performance depends on CBMA parameters.

2.2 Materials and Methods

In this section we first describe IBMA methods, reviewing the hierarchical MFX model, itemising practical issues and discussing when a MFX versus FFX model is appropriate. We then describe and compare the three considered CBMA methods. After introducing the collection of pain datasets used, we present the evaluation methods used to compare the different IBMA and CBMA methods.

2.2.1 IBMA Analyses

Several preparations must be made before any IBMA. First, all image data or relevant summary images must be warped into a common atlas space. While this is a fundamental pre-processing step, it is important that the atlas is the same for all subjects and all studies, and that the warping methods are as similar as possible between studies. All images should use the same size kernel smoothing, though if subjects come from different imaging centers a "Smooth to" strategy can be used (Friedman et al., 2006).

Another fundamental issue with IBMA methods is masking. Most standard analysis software will only analyze a voxel if all subjects (or studies) have data. This means that the analysis mask is the intersection of all the masks contributing to the model, which can result in dramatic erosion of the brain volume analysed. In particular, a single subject

with some missing data (e.g. due to motion, or a poor anatomical-functional alignment) can dramatically reduce the final analysis mask.[4] It is important to be aware of such effects and ameliorate these through either careful investigation of each session's data, generous intrasubject mask definition, or the use of statistical methods that allow for missing data.

IBMA methods that are only based on Z or T statistic images are unitless, while methods that use effect magnitude images require great care to ensure compatible units between the design matrices and contrasts in each study. For example, if one study has BOLD regressors with a baseline-to-peak height of 1 and a second study has BOLD regressors with a baseline-to-peak height of 2, then first study will have parameter estimate units twice that of the second study. Similar issues arise with respect to compatibility of contrasts, especially those expressing differences between conditions or groups. The best strategy is to ensure that all intrasubject model predictors have the same scaling, and that all contrasts preserve that scale. To ensure that a contrast preserves the units of the predictors, it is usually sufficient to require that all positive contrast elements sum to 1.0 and all negative elements (if any) sum to -1.0.

There are a variety of possible IBMA analysis methods (described next), but each can be classified as providing either FFX or MFX inferences. A FFX meta-analysis measures evidence for a non-zero effect relative to the inter-subject variability pooled over studies, while a MFX meta-analysis measures an effect relative to the combination of inter-subject and inter-study variability. As meta-analysis is often used to increase power with less concern given to inter-study consistency, a FFX may well be the most appropriate type of inference, whereas a MFX inference should only be required if a strong statement about inter-study consistency is needed.

[4]Note that the FSL 4.0 software introduced a new masking approach that resulted in a smaller mask than in previous versions, sometimes resulting in severely contracted group masks. FSL 4.1 uses a more generous masking scheme and is recommended for IMBA or any study using a large number of subjects.

2.2.1.1 Combining Methods

One approach to IBMA is the generic combining of statistic images, one per study. For a thorough review of this approach see Lazar et al. (2002), which discusses many of the well-known methods in the meta-analysis literature. In this work we consider only Fisher's P-value combining method ($-2 \times \sum_i \ln P_i$, where P_i is the uncorrected $p-value$ of the ith study) and Stouffer's Z-transform test ($\sum_i Z_i/\sqrt{n}$, where Z_i is the z-score for the ith study). These two methods are FFX methods, which provide evidence of one or more studies possessing an effect. One limitation of Fisher's method is that it can give significant results even when the signs of one-sided tests input to it are highly discordant, while conflicting signs will cancel with the Stouffer's method.

2.2.1.2 Single-level Regression

The simplest model for the effect magnitude data is a single regression model for all data. If all first-level time series data are modeled at once, the resulting (giant) regression model would yield FFX inferences. This requires massive computing resources to simultaneously access gigabytes of data, therefore it is not really practical and thus we do not consider it further.

A regression of the study-level data, in contrast, is very practical. It consists of an ordinary least square (OLS) regression, a simple unweighted analysis of mean effect magnitude data (one per study), as is typically done in SPM and as is available in many other packages such as FSL and AFNI. This produces MFX inferences, but does not weight studies according to their sample size or standard errors. Hence we prefer a multi-level model which weights each study according to study-level precision and which can produce either FFX or MFX inferences.

(One may think of an OLS analysis of study-level MFX Z-score data, instead of effect magnitude data. While this provides a kind of MFX inference, as between study variance is considered and the z-scores themselves convey group-level significance, the fitted model

is difficult to interpret as it is modeling average *significance* rather than average effect magnitude. However, we include in our results as "Stouffer's-MFX" for completeness.)

2.2.1.3 Hierarchical Model for Fixed- or Mixed-Effects Inferences

A multi-level hierarchical model (Beckmann et al., 2003; Woolrich et al., 2004; Worsley et al., 2002) fits data of any kind that is grouped within levels, for example time-series data within subjects, or subject data within studies. We first describe it in terms of combining subjects for a single group analysis[5]. First, each subject is modeled individually, producing effect estimates and standard errors. Next these intrasubject effects and standard errors are modeled together, producing group-level effect estimates and separate within- and between-subject variance estimates. For a MFX inference (FLAME-MFX), each subject is individually weighted according to the balance of their within-subject and the between-subject variance, producing an optimal estimate of the population effect. For a FFX inference (FLAME-FFX), the between-subject variance is ignored, but subjects are still individually weighted (unlike OLS) using just the within-subject variance.

2.2.1.4 Hierarchical Model for Image-based Meta-analysis

We use this same multi-level hierarchical framework to fit a three-level meta-analysis model: Level 1 is the intra-subject modeling of each subject's FMRI time series data, level 2 is the inter-subject analysis for each study, and level 3 is the inter-study meta-analysis. For details of the FSL's FLAME method used we refer the reader to the original citations (Beckmann et al., 2003; Woolrich et al., 2004), but in brief: At level 1, temporal autocorrelation is modeled voxel-wise, providing efficient estimates of each subject's effect estimates; at level 2, after alignment into standard space, each subject's effect estimates *and* standard errors are combined to give a mean group effect size estimate and MFX variance; at level 3, the study-level effect sizes and variances are again jointly modeled to

[5]A multi-level hierarchical model is implemented in the FSL software's "FMRIB's local analysis of mixed effects" or FLAME package, `http://www.fmrib.ox.ac.uk/fsl`

provide either MFX or FFX inference. The 3rd level model will typically be very simple (e.g. a column of ones to estimate the mean effect over studies), but can have any form. For example, a 3rd level model could be used to test for differences between studies or account for study-level covariates.

Note that a potential source of confusion is how, at both level 2 and 3, either MFX or FFX inferences can be produced. We are not advocating the use of FFX standard errors at the second (study) level. In both single-study and multiple-study analyses, it is crucial that the second-level standard errors incorporate the between-subject variation. Otherwise the final meta-analysis will not reflect between-subject variation in response magnitudes and will have a very limited interpretation. Hence, the only inference choice is whether to use MFX or FFX at the 3rd level.

2.2.2 CBMA Analyses

While there have been a wide variety of methods proposed for CBMA (Fox et al., 1997, 1998, 1999; Chien et al., 2002; Neumann et al., 2005), we have limited our evaluations to three: ALE, KDA and MKDA. In all three methods a map of the evidence for activations is created based on a set of foci coordinates. A qualitative 1D example is shown in Figure 2.1. All of the methods assess significance using a Monte Carlo resampling approach where, under the null hypothesis of no coherent activation, the foci are randomly distributed across space. At each voxel an uncorrected p-value is obtained by counting the number of Monte Carlo realisations that equal or exceed the original value. Family-wise error corrected p-values can similarly be obtained by counting the number of realisations where the maximal (image-wise) value exceeds the original value.

2.2.2.1 ALE

For each focus, ALE scores each voxel as a function of its distance from that focus using a Gaussian kernel of size σ (Turkeltaub et al., 2002). After this step, each voxel has a vector of "activation likelihood" probability values whose elements correspond to foci (one probability per foci). These values are assumed to be independent (the occurrence

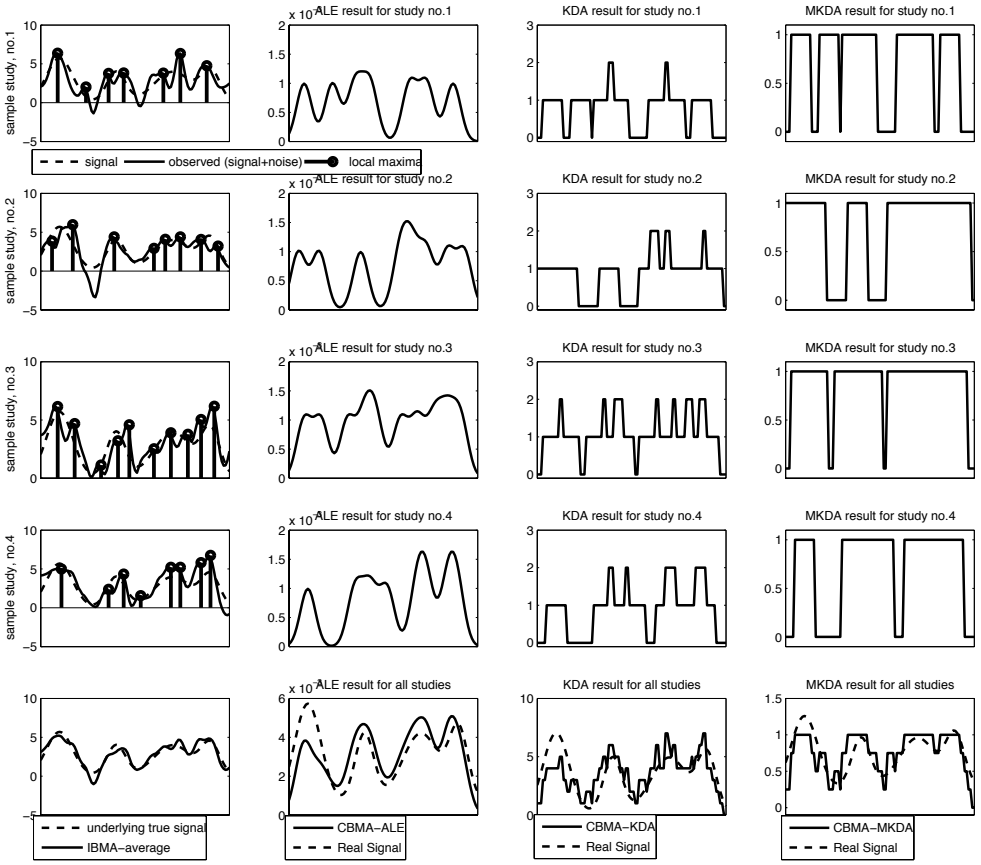


Figure 2.1: Illustration of a 4-study, 1-dimensional meta analysis. A true signal (dashed line) is created and four simulated statistic "images" are created by adding smoothed white noise to the true signal (bold lines in the first column of the first four rows). To apply CBMA to these simulated 1D studies, local maxima (foci) are extracted from each observed signal (circles on the bold lines). Next, the locations of these foci are fed into each CBMA technique. In the last row, the results of each method in reproducing the true signal using the foci are shown. As can be seen, averaging over the complete signals (as IBMA does) yields a better estimate of truth compared to using local maxima (CBMA). ALE results in a smooth estimate (due to its Gaussian kernel) which KDA and MKDA are rougher (due to its spherical kernel). Note that the "true" profile is generated as a sum of Gaussian densities, which is most consistent with the ALE method.

of one focus is assumed to give no information about whether or not the other foci will occur) and then combined with the addition rule for probabilities to yield the final activation likelihood, or ALE statistic value. This statistic indicates the probability of having at least one peak lying in that particular location, based on the Gaussian model for each focus. The procedure is repeated with Monte Carlo realisations of the data (the same number of foci randomly distributed over the brain) building up a null distribution of ALE maps. The significance test formally tests the global null hypothesis of no coherent activation, but rejecting this null hypothesis voxel-wise should provide evidence of consistent activation at a particular location. Pseudo-code for ALE is shown in Appendix A.1 and a 1D simulated example is shown in the second column of Figure 2.1.

2.2.2.2 KDA

KDA is similar to ALE, but uses a different kernel and method for combining the statistic maps. KDA creates maps for each focus with a spherical indicator function "kernel", with a radius ρ (Wager et al., 2004, 2007). A statistic map is created by summing, producing a map of the number of peaks activated within radius ρ. Similarly to ALE, a Monte Carlo test is used to reject the global null hypothesis of no coherent activation. Pseudo-code for KDA is shown in Appendix A.2 and its 1D simulation is shown in the third column of Figure 2.1.

2.2.2.3 MKDA

A clear limitation of both ALE and KDA is the independent treatment of each focus. If one study has 100 foci and another only 10, the first study will have an immense impact on the results, even if the increased number of foci is only due to different thresholding. The ALE and KDA Monte Carlo procedures also independently scramble each focus, even though a null study would be expected to generate some clustering of foci, due to the smoothness of the image data.

MKDA (Wager et al., 2007) attempts to address these limitations with two modifications to KDA. First, the convolved images are summed by study and truncated at

unity, creating study-specific images which indicate the presence of one or more foci within radius ρ. These study images are averaged, creating the mean number of studies that have one or more foci near a given voxel. This provides robustness to possible bias from studies that systematically report more foci per cluster and produces a more interpretable map. Second, the Monte Carlo procedure scrambles foci as clusters, producing realisations that bear greater resemblance to real data (i.e., have clustered foci) but lack any inter-study coherence. Hence MKDA is testing against a more realistic null hypothesis (no study-level coherence) and, since no single study can contribute disproportionately to the result, it is expected to produce more reliable and reproducible activation results. Pseudo-code for MKDA is shown in Appendix A.3 and its 1D simulation is shown in the fourth column of Figure 2.1.

2.2.2.4 Group Comparisons with CBMA Methods

While the CBMA methods don't have the flexibility of the hierarchical modeling framework described above, it is possible to make simple tests between groups of studies. As presented in (Laird et al., 2005a), if two groups of studies are separately analyzed for creating their corresponding whole-brain statistic maps, subtracting these two maps gives a measure of the difference contrast. Statistical significance of this difference map is assessed with respect to a null distribution of no coherence in either maps, created by taking null maps from each analysis and computing the difference. The final result provides evidence for difference in activation, though this approach has several caveats (detailed in the Discussion section).

2.2.3 Data

The aim was to pool results of 15 pain studies to find regions of activation induced by painful stimuli. In total, 163 healthy adult subjects were imaged (age range 20-35 y, mean 26.2 y; 97 male, 66 female) in either a 3T Siemens Trio MRI scanner (using a 12-channel head coil), or 3T Varian MRI scanner (using a 4-channel head coil). All data employed had been collected in accordance with local ethics approval. In spite of some differences,

all studies concentrated on pain as the main effect of interest. In three of these studies, a pain stimulus is combined with some language-related explanatory variables (EVs, or covariates). In two other studies, a painful stimulus is combined with some cues that warn or deceive subjects about an upcoming painful stimulus. Another group of six studies considers the effect of treatment on subjects' pain perception. In the other four studies, a pain stimulus is modulated to obtain different perceived pain levels. All studies have at least one pain EV, which allows us to form a simple "pain" contrast for each subject at the first level (and consequently at second and third levels) (Iannetti et al., 2005).

Despite having a pain covariate in all studies, the pain delivery mechanisms are different across the studies. For example, six of the studies used a mechanical pain stimulus, while the other nine studies used a thermal pain stimulus. We investigate a differential response to the two forms of pain delivery in 3rd level (meta) analysis. The result of this analysis will be areas of the brain that show more or less thermal-induced pain activation relative to mechanical-induced pain.

Processing of functional images at the first level was performed using FSL (Smith et al., 2001). Functional images were motion corrected (Jenkinson et al., 2002) and spatially smoothed (full width half maximum = 5 mm) prior to temporal model fitting (Beckmann et al., 2003; Woolrich et al., 2004, 2001). Co-registration to the MNI152 standard brain space was performed in 2 stages: (1) the FMRI data from a given subject was registered to that subject's T1 structural using linear registration and (2) the subject's structural image was registered to the MNI standard brain using nonlinear registration (Jenkinson and Smith, 2001; Jenkinson et al., 2002).

In the second-level analyses (Woolrich et al., 2004) MFX activation maps corresponding to the main pain effect were created. Third-level cross-study analyses were carried out using all studies, with a one-group model or a two-group model split by mechanical vs. thermal stimulus study type. Both fixed (FLAME-FFX) and mixed (FLAME-MFX) activation maps were created at the third level.

We used the results of the 15 pain studies to create the foci lists for the CBMA

analyses. For each study, the second-level analysis produced a list of foci, the locations of local maxima in the statistic image. A constraint is imposed to find local maxima that are not closer than 8mm to each other, which matches the default behavior of SPM's results. Based on this framework, a list of 231 foci are extracted from all 15 available studies. This foci list is the main input to all following CBMA (ALE, KDA and MKDA).

2.2.4 Map Comparison

We use the results of the IBMA FLAME-FFX model to define a "reference" result against which the other methods are compared. This choice of standard result follows from a sequence of three assessments: IBMA is preferred over CBMA, as the image data are a strict superset of the information in CBMA analyses; FFX is preferred over MFX, as the typical meta-analysis goal is aggregation of evidence for an effect, not MFX's inference on inter-study concordance; and, for the choice of IBMA analysis method, FLAME's hierarchical model is preferred over other traditional meta-analystic measures, due to its statistical optimality and flexibility for dealing with group differences and covariates.

We compare CBMA maps to the IBMA reference image with one symmetric and two asymmetric measures. The Dice similarity measure (DSM) (Dice, 1945) is a symmetric measure of the resemblance of two binary images:

$$DSM = \frac{2|I \cap C|}{|I| + |C|} \tag{2.1}$$

where $|I|$ and $|C|$ are the number of non-zero voxels in a thresholded IBMA (reference) image and a thresholded CBMA image, and $|I \cap C|$ is the number of non-zero voxels in their intersection. DSM ranges from 0 (no overlap), to 1 (perfect overlap).

If the reference is taken as "truth", we can compute the traditional (asymmetric) similarity measures, the true positive rate (TPR), and the false positive rate (FPR):

$$TPR = \frac{|C \cap I|}{|I|} \tag{2.2}$$

$$FPR = \frac{|C \cap (\neg I)|}{|\neg I|} \tag{2.3}$$

where $|\neg I|$ are the number of zeroed voxels in the thresholded reference image. The interpretation of TPR is the probability of a CBMA method correctly labeling a voxel as "active", averaged over all truly active voxels, where "true activation" is defined by a threshold applied to the reference image (see below). Likewise, the FPR is the probability of a CBMA method falsely labelling a voxel as "active", averaged over all truly inactive voxels.

To evaluate CBMA methods with respect to selected IBMA methods, two thresholding schemes are utilized. In the first scheme, three uncorrected p-values (0.001, 0.01 and 0.05) are used to threshold both IBMA and CBMA output maps, providing equal (nominal) false positive rates for each method, and yielding equivalent thresholded images to compute DSM.

In the second thresholding strategy, maps from IBMA and CBMA are each thresholded differently. For IBMA, a 0.05 false discovery rate (FDR) (Nichols and Hayasaka, 2003) corrected threshold is used to create a reference map with high sensitivity. For CBMA, the same set of uncorrected p-values as before (0.001, 0.01 and 0.05) is used. With this strategy, the TPR and FPR measures can be computed, while keeping the reference image fixed (i.e. it doesn't change with the CBMA uncorrected p-value threshold).

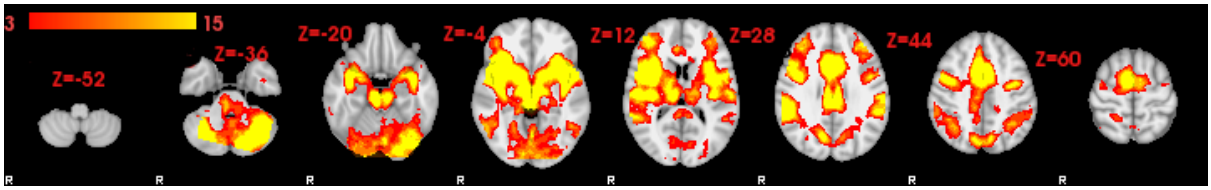
For each thresholding scheme, each CBMA method is tested over a range of kernel parameters. ALE's kernel parameter is the value of the Gaussian kernel's standard deviation (σ), and MKDA/KDA's kernel parameter is its indicator kernel's radius (ρ). The aim is to find the optimal setting for each method (for this dataset and a 5mm FWHM Gaussian smoothing), while comparing CBMA with IBMA. σ values compared are {5, 10, 15, 20, 25, 30, 35, 40, 45, 50}, and ρ values {5, 10, 15, 20, 25, 30, 35, 40, 45, 50}. For each CBMA method, each threshold (0.001, 0.01 and 0.05) and each kernel parameter, the binary resulting map is compared with two binary references (the first and second thresholding scheme), and each comparison yields a single DSM. The larger the DSM, the better the method-threshold-kernel combination.

2.3 Results

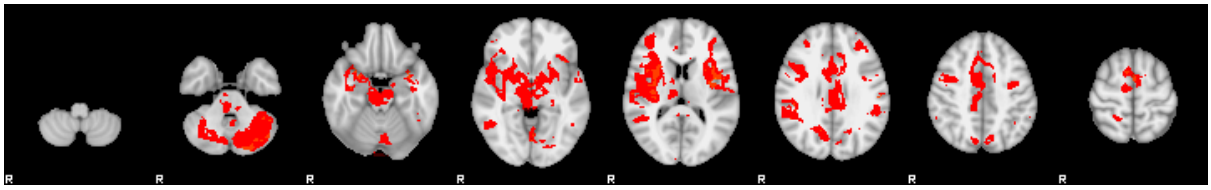
Figure 2.2 shows statistic maps for the six IBMA methods considered, each thresholded at uncorrected p=0.001, to give an indication of the differences between the different FFX and MFX image-based combining methods. The figure shows a clear distinction between the FFX methods (Figure. 2.2a, 2.2d, & 2.2e) and the MFX methods (Figure. 2.2b & c), with the FFX showing considerably more activation. The FFX result based on a hierarchical model (Figure. 2.2a, FLAME-FFX) shows a smoother profile of activation, while the classic meta-analytic statistics (Figure. 2.2d Fisher's, 2.2e Stouffer's) were more irregular, perhaps indicating their greater sensitivity to individual (instead of average) study significance. A complete map of the FLAME-FFX reference is shown in Figure 2.3.

The evaluation of CBMA methods as a function of kernel parameter is shown in Figure 2.4 for the first thresholding scheme (same uncorrected threshold for IBMA and CBMA). Results for the second thresholding scheme had lower DSM scores overall and are qualitatively similar (not shown). For all methods, the best DSM was for the most liberal p-value threshold considered (0.05). For DSM and TPR, the curves generally had the same shape, with an optimal value that was consistent over different p-value thresholds.

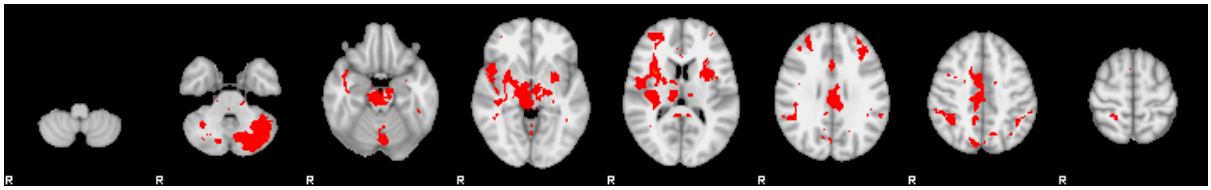
Among all CBMA methods, ALE seems to yield the best results overall, with KDA performing similarly and MKDA being more conservative with respect to our reference image. This conservativeness could be due to MKDA treating studies as independent units, instead of each focus, suggesting that it would require more studies to obtain a similar consistency map to KDA. As can be seen in a 1D illustration in Figure 2.5, increasing the number of studies makes MKDA's statistic more similar to the reference image. The false positive rate for ALE and KDA analyses increases with kernel size, and the "optimal" kernel in both ALE and KDA has a false positive rate close to 0.1. Setting a kernel size of of 5mm for ALE or 20mm for KDA limits their FPR to .05, which puts them on more similar footing in terms of DSM to MKDA. Thus, the MKDA's DSM is



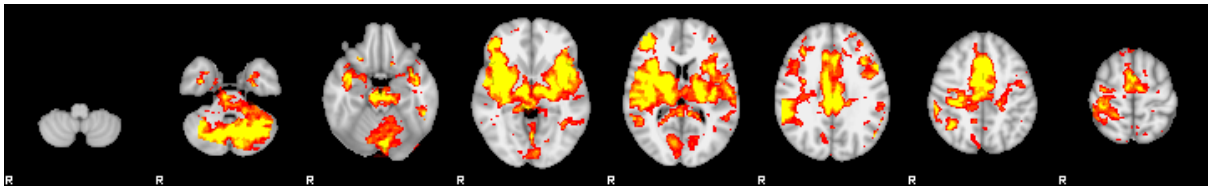
(a) FLAME-FFX



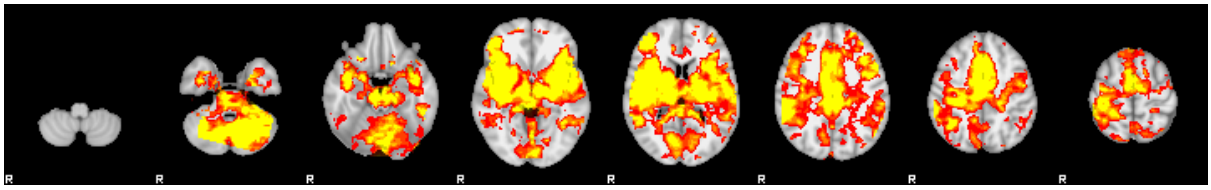
(b) FLAME-MFX



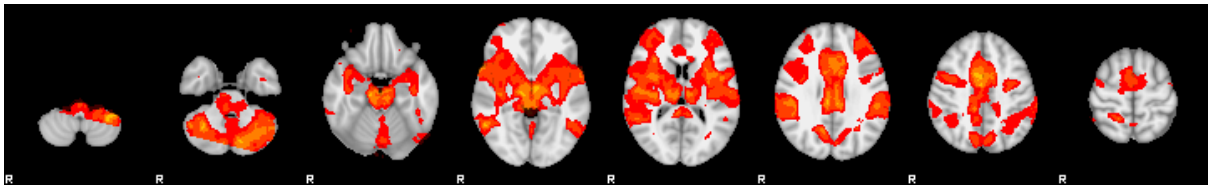
(c) OLS-MFX



(d) Fisher's-FFX



(e) Stouffer's-FFX



(f) Stouffer's-MFX

Figure 2.2: IBMA map resulting from MFX at second level and different IBMA methods at third level. Z-stat maps are converted to their corresponding p-value maps and then, to give clearer visualization, the $-\log_{10} p$ map is shown (with min-max of 3-15). As can be seen, FLAME-MFX and OLS show less extended activations, in areas of consistency across studies. Slice locations are in mm in MNI space.

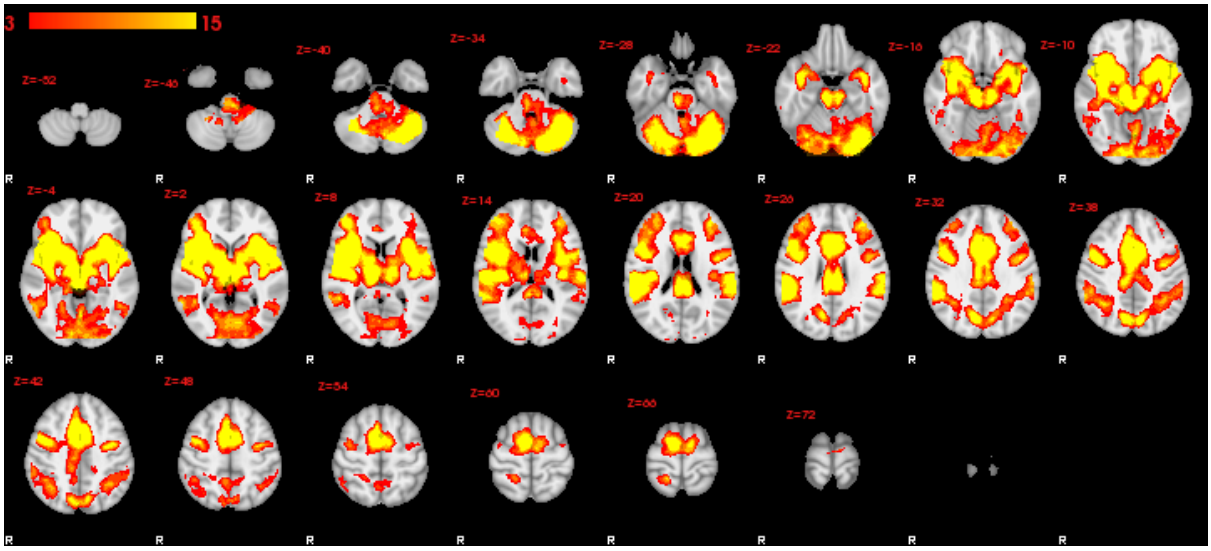


Figure 2.3: Reference image against which CBMA methods are compared. This is the resulting map from a three-level hierarchical analysis, with FLAME-FFX at the third level. Color overlays show $-\log_{10} p$ map (with min-max of 3-15). Slice locations are in mm in MNI space.

possibly lower because it is more conservative and more similar to a MFX analysis.

The optimal kernel parameter values (shown in Figure. 2.4) can depend on the amount of first level smoothing applied to studies from which foci are collected. We investigate this by repeating the entire comparative analysis on the basis of 4, 5, 7, 10 & 15mm FWHM first level smoothing (with the reference map re-defined for each smoothing). The DSM results of these comparisons are shown in Figure 2.6. As these plots show, the optimal kernel parameter is not very sensitive to smoothing extent; particularly when it varies in the range of 4-10mm (which is the typical smoothing range used in FMRI studies).

We find the optimal kernel parameter for each CBMA method to be: $\sigma = 15mm$ for KDA, $\rho = 25mm$ for KDA and $\rho = 15mm$ for MKDA. Figure 2.7 compares the reference IBMA result to the DSM-optimized CBMA results. Note the dramatic difference in the sensitivity and the overlap pattern of detected regions.

Using the optimal settings we also tested a contrast between two sub-groupings of the 15 studies. There were 9 studies with thermal pain and 6 studies with mechanical pain. We examined the IBMA and CBMA inference for just thermal (i.e., THERM), just

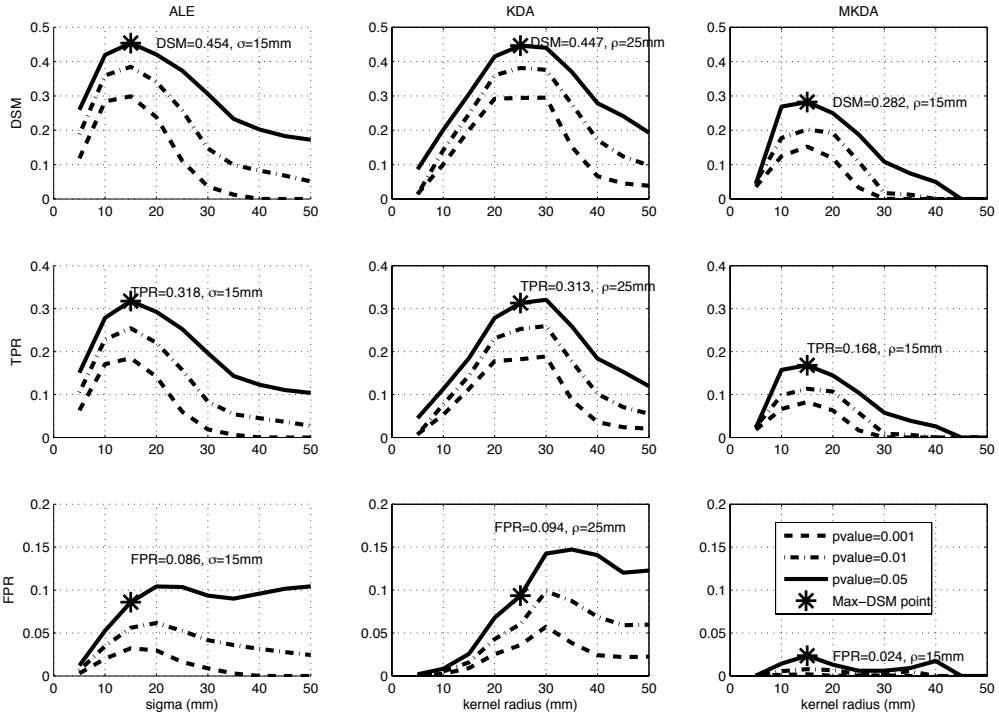


Figure 2.4: Evaluating CBMA methods for different kernel parameter values with respect to the reference map. In the first column, DSM, TPR and FPR of the ALE method are shown. In the second and third columns, the same performance measures are shown for KDA and MKDA. In all plots, the x-axis is the kernel parameter (σ for ALE and ρ for KDA and MKDA). To estimate the DSM value, images are thresholded at different p-values (shown in the legend) and then binary images are compared. Note that in this plot all scores are for the first thresholding scheme. For the second thresholding scheme, plots are very similar, but with smaller DSM scores overall. More liberal thresholding yields higher DSMs ($*$ indicates the coordinate corresponding to maximum DSM).

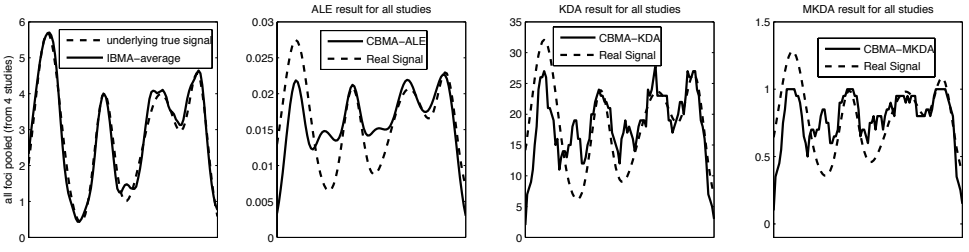


Figure 2.5: Illustration of a 20-study, 1-dimensional meta analysis. Using the same setting as in Figure 2.1 except with 20 studies, MKDA's estimate is more similar to ALE and KDA's estimate.

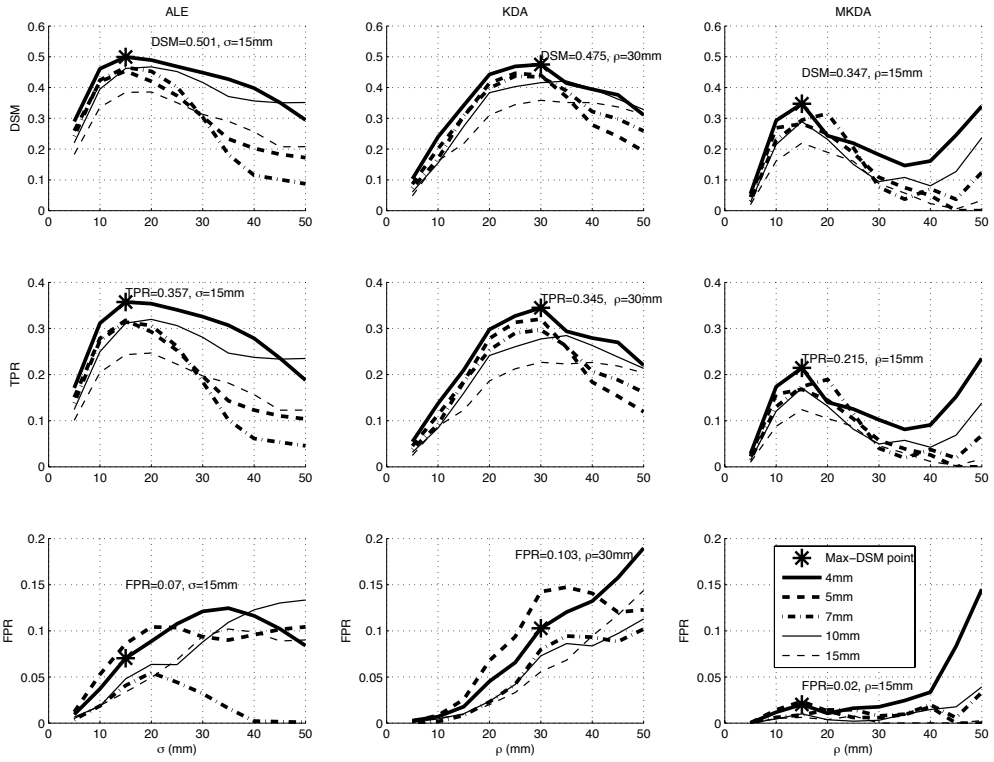
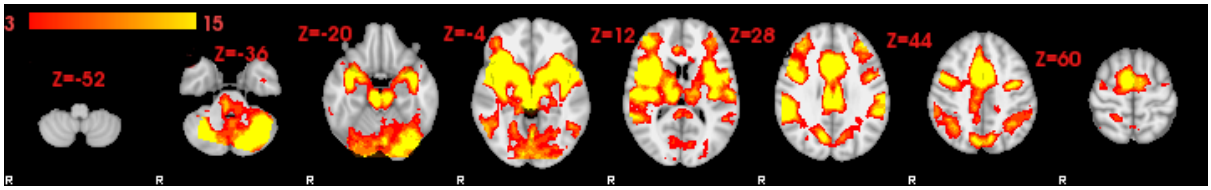
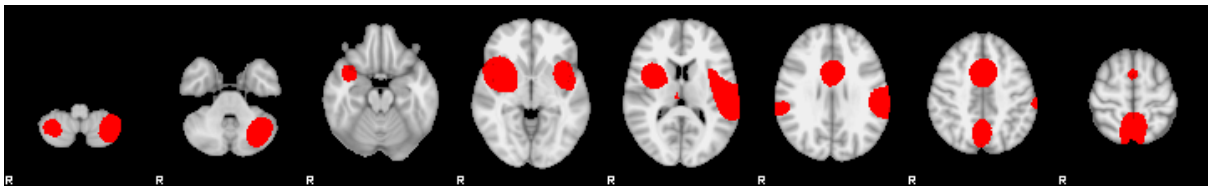


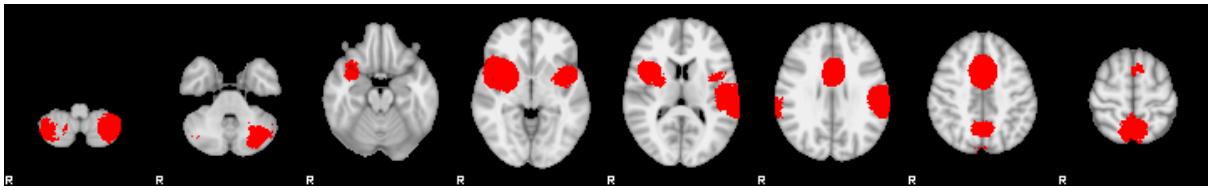
Figure 2.6: Evaluating the effect of smoothing extent of studies on optimal CBMA methods' kernel parameter values with respect to their corresponding reference map. The first, second and third columns show the results for ALE, KDA and MKDA, respectively. The first, second and third row show DSM, TPR and FPR for each CBMA method, respectively ($*$ indicates the coordinate corresponding to maximum DSM). Each line in each subplot corresponds to one FWHM from 4, 5, 7, 10 and 15mm, and all CBMA maps are thresholded at 0.05 ($p_{thresh} = 0.05$).



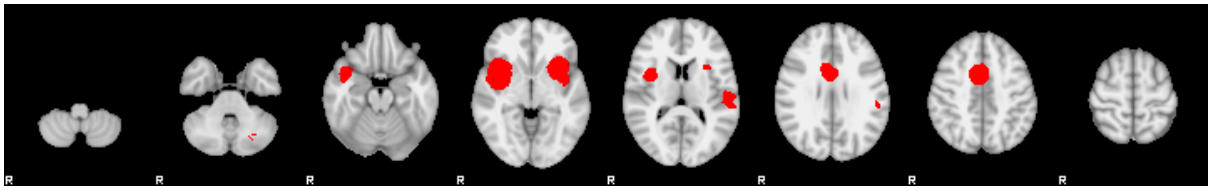
(a) IBMA: FLAME-FFX



(b) CBMA: ALE



(c) CBMA: KDA



(d) CBMA: MKDA

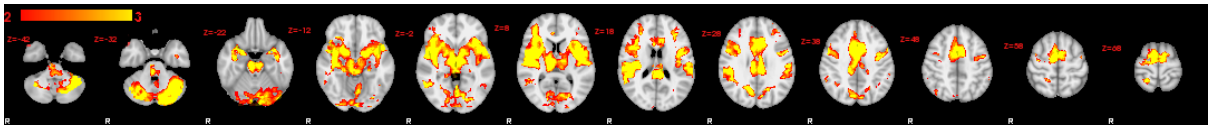
Figure 2.7: The reference IBMA map (panel a) shown with CBMA maps. Colour overlays show $-\log_{10} p$ values (with min-max of 3-15). Maps in panel b,c and d are resulting from ALE with $\sigma = 15mm$, KDA with $\rho = 25mm$ and MKDA with $\rho = 15mm$, respectively. Slice locations are in mm in MNI space.

mechanical (i.e., MECH) and their difference (i.e., THERM>MECH and THERM<MECH), shown in Figure 2.8 a,e and Figure 2.9 a,e. To generate the THERM/MECH contrast image, foci are collected only from those studies using thermal/mechanical pain stimuli. All of the CBMA analyses are performed using ALE, KDA and MKDA with $\sigma = 15mm$, $\rho = 25mm$ and $\rho = 15mm$, respectively. Results from this analysis are shown in Figure 2.8b-d,f-h and Figure 2.9b-d,f-h.

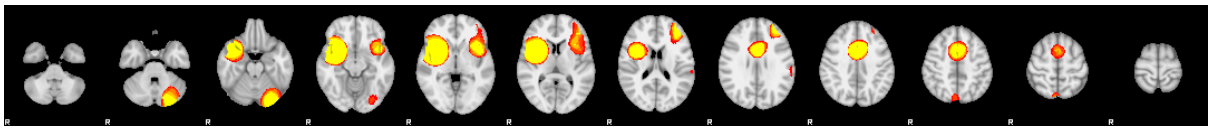
In studies using both thermal and mechanical pain stimuli, activity is widely extended across the cortex. For example, both sets of studies activate the attentional network, including the parietal sulcus. Note that we cannot unambiguously attribute THERM vs. MECH effects to differences in pain perception, as there are multiple confounding factors. For example, in most of the mechanical studies, stimuli were delivered to the right foot, while the thermal stimuli were delivered to the left arm. This confounding effect can be seen clearly in activation maps as a lateralization effect, where the thermal stimuli cause activation in right somatosensory cortex, while mechanical stimuli cause more activation on left somatosensory cortex. Also note that mechanical activations are more medial (Figure. 2.8 e), while the thermal activations are more lateral (Figure. 2.8 b), consistent with typical somatosensory findings Becerra et al. (2006); Borsook et al. (2008).

The other confound arising from the studies' experimental setups are the difference in visual cortex activity. Most thermal studies used a visual analogue scale (VAS), while studies using mechanical stimuli instructed subjects to close their eyes during the experiment (compare visual cortex in Figure. 2.8b & 2.8e).

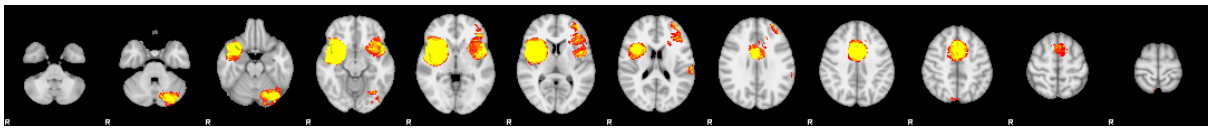
Figure 2.8b-d, f-h and Figure 2.9b-d, f-h show the corresponding results for the coordinate-based methods. Note that the CBMA and IBMA results are more similar for the THERM and MECH effects, and less so for the THERM>MECH and THERM<MECH results. This differential performance is likely due to the lack of information about activation decreases in the coordinate-based data. For example, while THERM>MECH can be significant if THERM activation is greater than MECH, it can also be significant if the MECH deactivation is greater than THERM deactivation, which can't be reconstructed



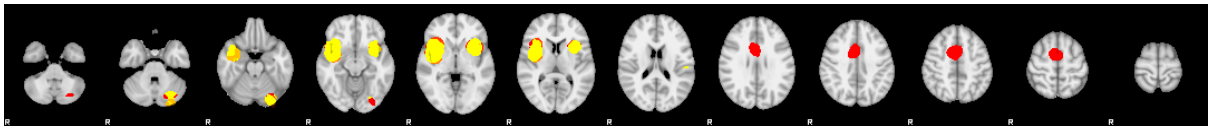
(a) IBMA : Thermal



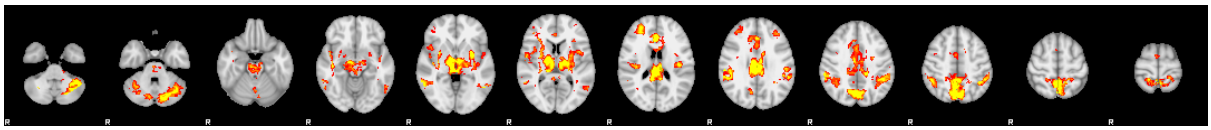
(b) ALE : Thermal



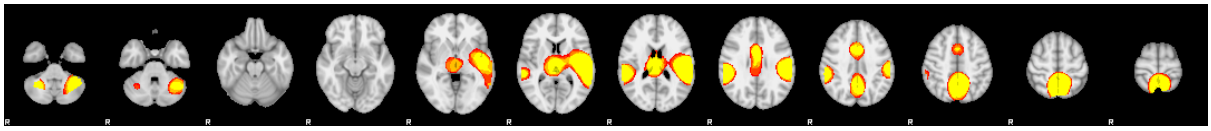
(c) KDA : Thermal



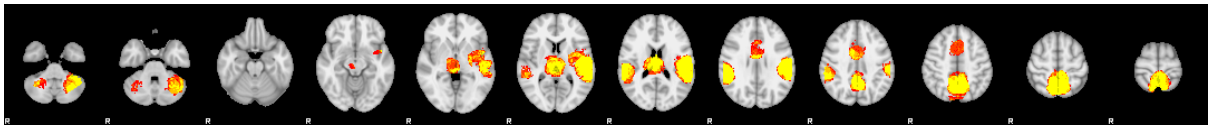
(d) MKDA : Thermal



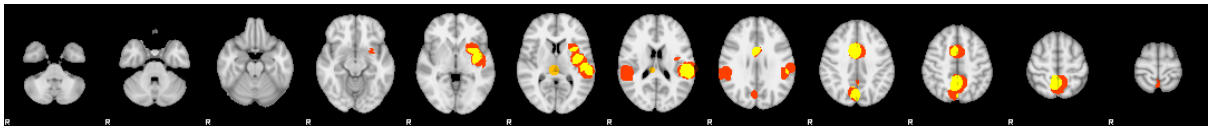
(e) IBMA : Mechanical



(f) ALE : Mechanical

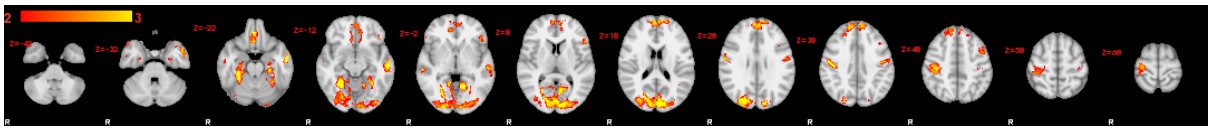


(g) KDA : Mechanical

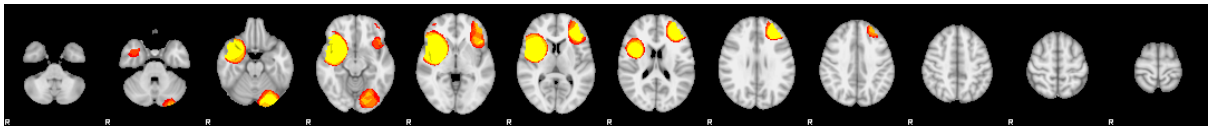


(h) MKDA : Mechanical

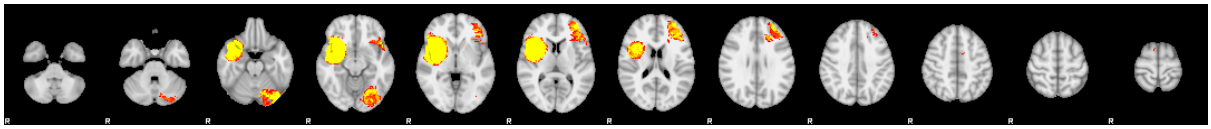
Figure 2.8: Using a third-level design to answer cognitive pain questions both in IBMA and CBMA. $-\log_{10} p$ overlay maps are shown (with min-max of 2-3). Slice locations are in mm in MNI space.



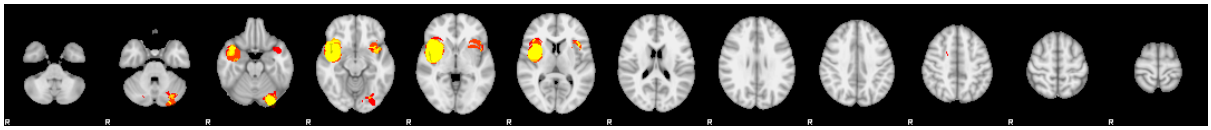
(a) IBMA : Therm>Mech



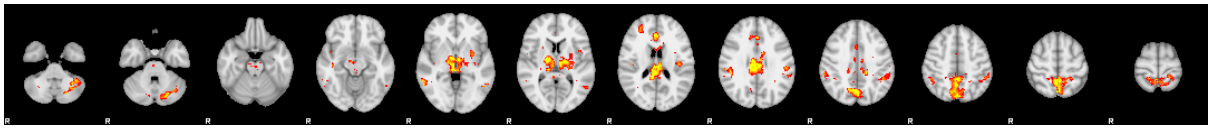
(b) ALE : Therm>Mech



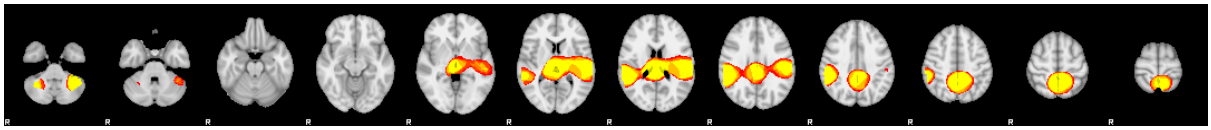
(c) KDA : Therm>Mech



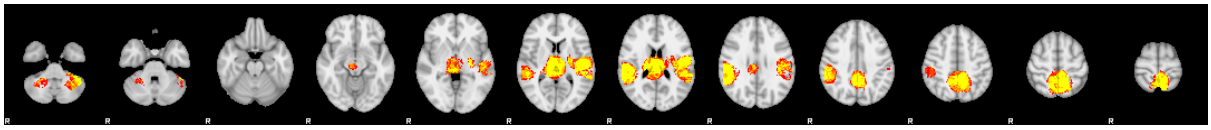
(d) MKDA : Therm>Mech



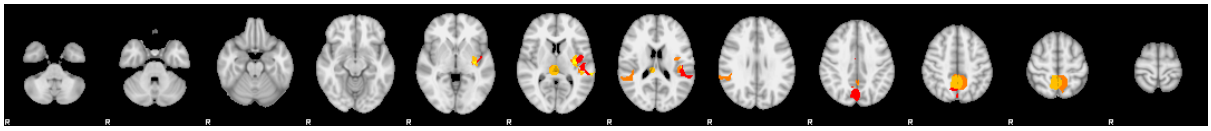
(e) IBMA : Mech>Therm



(f) ALE : Mech>Therm



(g) KDA : Mech>Therm



(h) MKDA : Mech>Therm

Figure 2.9: Using difference contrasts to answer cognitive pain questions both in IBMA and CBMA, for the THERM > MECH and MECH > THERM contrasts. $-\log_{10} p$ maps are shown (with min-max of 2-3). Slice locations are in mm in MNI space.

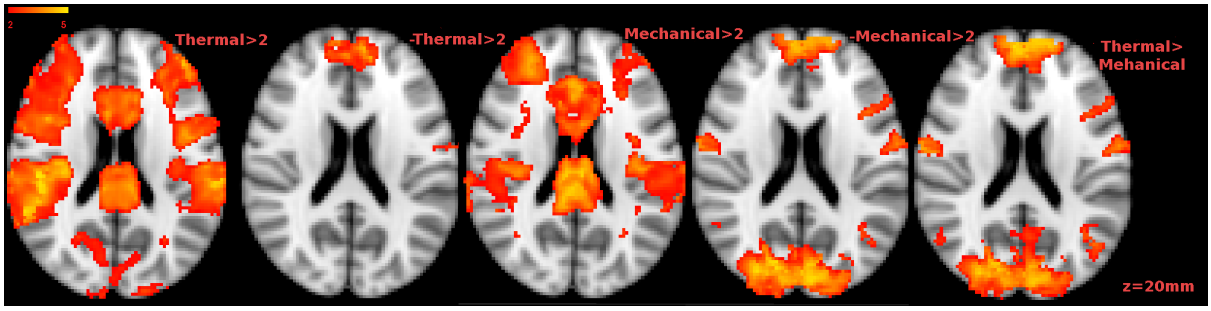


Figure 2.10: The effect of deactivation information in difference contrasts for a sample slice (z=20mm in MNI space). This figure shows z-statistic images ($z > 2$) for THERM, -THERM, MECH, -MECH, and THERM> MECH, from left to right respectively. It can be seen that the reason for having significant areas corresponding to THERM> MECH that are not significant in THERM is associated with deactivation in MECH.

by positive activation foci. This highlights a potential danger of omitting deactivation foci from such a comparison (see Figure. 2.10).

Finally, note the much greater spatial detail available in the IBMA results. These show a greater general richness, as well as greater sensitivity for finding small activation areas, again attributable to IBMA's retention of all data.

2.4 Discussion and Conclusions

In this chapter we have tried to assess the information lost from working only with foci when the voxel-level data are available, as well as highlight the importance of kernel parameters in a typical CBMA. Using a group of 15 pain studies, all analyzed in a similar fashion, we generated a reference map using a three-level hierarchical model, and generated foci to produce the data that would normally be used for a CBMA. While we know of no other work that considers IBMA and CBMA in such a parallel way with a common set of studies, we stress that our findings may depend on the number of studies considered, number of subjects in each, scanning techniques, and chosen foci extraction/reporting style. None-the-less, we believe our collection of pain studies are representative of common practice and are useful for the evaluations considered.

Using DSM plots, it is clear that CBMA methods cannot produce the same map as an IBMA. The best result for CBMA comes from ALE for $\sigma = 15mm$, which is just

DSE=0.45. Using this evidence, it seems a necessary future concern for the neuroimaging field to find a way to share datasets (Toga, 2002; Van Horn et al., 2004). Full data need not be transferred, rather just sufficient statistics, that is effect magnitude estimates and their standard errors. However there are a multitude of issues to be taken into consideration for a successful future sharing policy which are beyond the scope of this chapter. These issues include description of the experimental design (Miller et al., 2001; Liu and Frank, 2004; Liu, 2004; Smith et al., 2007), image acquisition, and analysis techniques adopted in different research groups and institutes (Friedman et al., 2006, 2007; Zou et al., 2005).

Hierarchical GLM models can be applied simply by using summary statistic maps (like contrasts of effect sizes and their variances) (Beckmann et al., 2003; Woolrich et al., 2004). As an appropriate way of doing IBMA, there is no hard rule about whether to use FFX or MFX at the third level. We propose, in general, using FFX at the third level on the basis that individual studies are valid samples from the population and, even if just one is significant, this is valid information to drive a meta-analysis. If there is greater concern about the quality of the constituent studies in a meta-analysis, a MFX approach would be a safer approach, and would only identify consistent results relative to the inter-study variability.

Two important parameters that all three CBMA techniques depend on are kernel parameter and significance threshold. We tried to investigate both of these factors to find the optimal combination of these parameters to maximize the IBMA-CBMA similarity (at least for our datasets). The main reason for adopting a voxel-wise comparison with a fixed, uncorrected p-value threshold is to have comparable thresholds for all methods. For example, using a threshold from FDR would create adaptively-determined thresholds for each result. This is also the reason why we adopted the same protocol in CBMA techniques.

For the other important factor, kernel parameter, we tried to find a good kernel value given typical first-level analyses (FWHM=5mm first-level smoothing). Our recommendation for these parameters is σ=15mm for ALE, ρ=25mm for KDA and

ρ=15mm for MKDA. These values are dependent on our 15-study sample, but can provide a guide for other similar data. Although this raises another weak-point of CBMA—that the most optimal setting can vary from one dataset to another—the same could be said with respect to the effect that first-level smoothing has on IBMA approaches.

Comparing results from analysis techniques using DSM is not necessarily the best way for all such comparisons. DSM is a combination of TPR and FPR and in other comparisons/applications it may not be necessary for all variables to have the same weight. For example, in cases where FPR is the most important variable, a method such as MKDA may then appear to perform relatively better, in spite of having a smaller overall DSM. Another issue that might cause MKDA to be more desirable than ALE and KDA is cases where there is a large difference in the number of foci extracted from each study. In such cases, pooling in ALE or KDA style can be highly biased toward studies with higher numbers of foci. As the MKDA technique looks for consistency over studies (by using studies as input units), outlier studies and foci will have less chance of having noticeable effects on the final result.

The obvious primary weak point of CBMA techniques arises from discarding a huge amount of information, simply by using coordinates of maxima (i.e., a systematic flaw of CBMA). When a comparison is made between different conditions, there will be further loss of accuracy if deactivations are not included. In case of having two contrasts, C_1 and C_2, a difference contrast like $C_1 - C_2$ can be significant if C_1 is more active than C_2, or C_1 is less de-active than C_2. The first case can be *partially* assessed by activation foci, while the second case cannot be assessed in the absence of deactivation foci. In short, the difference between IBMA and CBMA group comparisons can be due to either omission of decreases resulting in CBMA false negatives, or thresholding artifacts resulting in CBMA false positives.

This was highlighted by the differences in our thermal-mechanical comparisons between the CBMA and IBMA results. Based on this, it is strongly recommended to include deactivation foci in CBMA, as well as activation foci, to have a more accurate

and reliable result. However, of course, the greater data reduction implicit in CBMA approaches is considerably more convenient than needing to provide full summary images from all studies; CBMA can even be carried out purely on the basis of activations reported in journal papers.

We have offered a three-part justification on why the IBMA FLAME-MFX analysis should be the reference. As further evidence, if CBMA were a better choice, and in fact IBMA were less sensitive than CMBA, this would be evidenced by CMBA having essentially perfect power relative to IBMA reference image. Instead, CMBA shows quite poor power relative to the IBMA reference, and thus further justifies the choice of reference method.

Finally, we note that the recommended IBMA method (hierarchical linear modeling) depends on comparable contrast (and standard error) images obtained for each subject and each study, unlike the CBMA methods, and other IBMA methods, which are based only on t- or z-statistics that are invariant to design matrix or contrast scaling. All of the IBMA methods can be affected by corrupted masks for one or more subjects, resulting in excessive erosion of the analysis mask. These issues simply highlight the importance of careful quality control of the analysed data to maximize the interpretability of the final results.

Chapter 3

Adjusting The Effect Of Nonstationarity In Cluster-based And TFCE Inference

Abstract

In nonstationary images, cluster inference depends on the local image smoothness, as clusters tend to be larger in smoother regions by chance alone. In order to correct the inference for such a nonstationary, cluster sizes can be adjusted according to a local smoothness estimate. In this study, adjusted cluster sizes are used in a permutation-testing framework for both cluster-based and threshold-free cluster enhancement (TFCE) inference and tested on both simulated and real data. We find TFCE inference is already fairly robust to nonstationarity in the data, while cluster-based inference requires an adjustment to ensure homogeneity. A group of possible multi-level adjustments are introduced and their results on simulated and real data are assessed using a new performance index. We also find that adjusting for local smoothness via a separate resampling procedure is more effective at removing nonstationarity than an adjustment via a random field theory based smoothness estimator[1]. The results in this chapter are recommended as a complimentary analysis to the IBMA method introduced in Chapter 2.

[1]The work in this chapter has appeared as Salimi-Khorshidi et al. (2010).

3.1 Introduction

When detecting changes in functional or structural brain image data, it is necessary to have powerful inference methods that offer precise control of false positive risk. Once a statistic image is created that assesses the evidence of an effect at each voxel, the two most common "thresholding" approaches are voxel-based and cluster-based inference. While voxel-wise methods use a single threshold to classify signals as "real," cluster-based inference uses two thresholds, an arbitrary cluster-forming threshold followed by a cluster-size threshold to label clusters as "real". Cluster-based inference has a higher sensitivity compared to voxel-intensity-based tests when the signal is spatially extended (Friston et al., 1996; Poline et al., 1997).

In the original implementation of cluster-based inference (Roland et al., 1993; Poline and Mazoyer, 1993) a null distribution of cluster sizes from simulated images having the same characteristics (e.g., spatial autocorrelation) as the observed data is generated to assess the significance of the clusters. Further modifications of this approach have been proposed for fMRI (Forman et al., 1995) and PET (Ledberg et al., 1998). The most widely used approaches to cluster-based inference, however, are the ones based on the random field theory (RFT) (Friston et al., 1994; Hayasaka et al., 2004). Like any other parametric tests, several assumptions are required, such as smooth images, a sufficiently high cluster-forming threshold u_c, and the uniform smoothness of images (Worsley et al., 1992; Petersson et al., 1999). An alternative to the RFT-based cluster inference is the permutation test (Holmes et al., 1996; Bullmore et al., 1999; Nichols and Holmes, 2002), which requires almost no assumptions (except the exchangeability of the data under the null hypothesis).

When the image noise fails to have uniform smoothness it is said to be "nonstationary", as the correlation between neighboring voxels depends on where the voxels are located. Under nonstationarity, the sensitivity and specificity of a (standard, stationary) cluster-size test depend on local smoothness of the image, as bigger clusters are

expected in smoother areas. To overcome this problem, cluster sizes can be adjusted for nonstationarity with a local smoothness estimate based on RFT (Worsley et al., 1996). Further assumptions and approximations produce a null distribution for this adjusted cluster-size statistic that accounts for both variation in cluster size and the substantial sampling variability in the estimate of local smoothness, providing valid uncorrected or corrected P-values[2]. Alternatively, the adjusted cluster-size statistic can be assessed with a permutation test to obtain P-values, constituting a semi-parametric approach (Hayasaka et al., 2004): The statistic is derived using parametric RFT to adjust for the impact of spatially-varying smoothness, but non-parametric permutation is used to assign P-values. A comparison of the results from such adjusted cluster-size statistics and the maximal-voxel statistic is presented in Moorhead et al. (2005).

An approach closely related to cluster-wise inference is threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009), which removes the dependence on the arbitrary cluster-forming threshold u_c. While TFCE does introduce two new parameters, in practice they are fixed to values justified by theory and empirical results. TFCE attempts to keep the sensitivity benefits of cluster-based inference (by using the cluster size information at a range of possible u_c values), while avoiding an arbitrary choice of a single u_c. The method produces a voxel-wise output image in which each voxel's value represents the accumulative cluster-like local spatial support at a range of cluster-forming thresholds. TFCE has been shown to give generally better sensitivity than other methods over a wide range of test signal shapes and SNR values (Smith and Nichols, 2009; Smith et al., 2008). However, TFCE's performance when nonstationarity is present has not been studied to date.

In this work we assess the accuracy of the established RFT-based nonstationary cluster-size inference method, and propose a new fully nonparametric approach to nonstationarity adjustment based on the local empirical distribution of the cluster-related

[2]We only consider family-wise error (FWE) corrected P-values, defined as the smallest critical value α such that a cluster can be declared significant controlling for the chance of one or more false positive clusters anywhere in the image.

statistics. Our proposed method uses a resampling-based estimate of nonstationarity, followed by a permutation test that applies an adjustment based on that estimate. Since this entails two successive null-hypothesis resampling procedures, we call it "2-pass". We evaluate this new approach, in the context of both standard cluster-based and TFCE inference, and measure the spatial homogeneity of false-positive risk using the variability of uncorrected P-values. We compare the impact of using no adjustment, RFT-based (i.e., explicit smoothness-estimation-based) adjustment, and our proposed empirical adjustment using various real and simulated data with spatially-varying smoothness.

3.2 Materials and Methods

Our starting point is a general linear model (GLM) fit at each voxel. For a D-dimensional image, at voxel i we have

$$Y_i \;=\; X\beta_i + \epsilon_i \tag{3.1}$$

where Y_i is the observed intensity vector (M×1), β_i is the parameter vector (P×1), X is the design matrix (M×P). We assume $\mathsf{E}(\epsilon_i) = 0$ and $\mathsf{Var}(\epsilon_i) = \tau_i^2$ (mean zero and constant variance errors). An unbiased estimate of $\hat{\beta}_i$ gives residuals

$$\hat{\epsilon}_i \;=\; Y_i - X\hat{\beta}_i, \tag{3.2}$$

which are used to estimate $\hat{\tau}_i^2 = \hat{\epsilon}_i{}^\top \hat{\epsilon}_i/\eta$, where η are the degrees of freedom of the error, and leads to the test statistic $T_i = c\hat{\beta}_i/(\hat{\tau}_i\sqrt{c(X^\top X)^{-1}c^\top})$ for a given contrast c.

3.2.1 Smoothness in Random Field Theory

We first introduce the notions of smoothness and roughness for standard, stationary RFT methods, then generalize to the nonstationary case. A core assumption of standard RFT methods is that the standardized errors $\epsilon_i^* = \epsilon_i/\tau_i$ comprise a sampled version of the mean zero and unit variance "component fields", which are smooth, homogeneous Gaussian processes; "smooth" meaning that the spatial autocorrelation function has two

derivatives at the origin, and "homogeneous" meaning that the autocorrelation function is spatially invariant, i.e., stationary. RFT roughness (inverse smoothness) is parameterized by the $D \times D$ matrix Λ,

$$\Lambda = \mathsf{Var}(\dot{\epsilon}^*) \tag{3.3}$$

where $\dot{\epsilon}^*$ is the D×1 vector of spatial partial derivatives of the component fields. A common assumption is that Λ is diagonal[3], and so we denote λ_d, $d = 1, ...,$D as the diagonal entries.

It is more convenient to work in terms of smoothness instead of roughness, and so the following transformations are used:

$$\sigma_d^2 = \frac{1}{2\lambda_d} \tag{3.4}$$

$$\text{FWHM}_d = (8 \cdot \ln(2))^{1/2} \sigma_d \tag{3.5}$$

where σ_d is the standard deviation of a Gaussian kernel needed to convolve a white noise field to have roughness λ_d and FWHM_d is the full width at half maximum (FWHM) of the same Gaussian kernel. The geometric mean of the FWHM_d values,

$$\text{FWHM} = \left(\prod_{d=1}^{D} \text{FWHM}_d \right)^{1/D} \tag{3.6}$$

is a useful summary measure of smoothness over the D dimensions. Related to FWHM is RESEL, short for RESolution ELement; it is the size of the D-dimensional search volume in units of its smoothness,

$$\text{RESEL} = \frac{V}{\prod_{d=1}^{D} \text{FWHM}_d} \tag{3.7}$$

where V is the number of voxels in the search volume.

Nonstationary RFT methods require voxel-wise estimates of roughness and smoothness. Details of how such local estimates are made are reviewed in Appendix B.1, but we simply use a voxel index i to indicate local versus global estimates (e.g. FWHM_{id}

[3]This is a common assumption in FSL and SPM for low degrees of freedom.

for the FWHM in direction d at voxel i). The "RESEL count" for a single voxel is the size of one voxel in units of RESELs, and has the designation RPV for "RESELs per voxel":

$$\text{RPV}_i \;=\; \frac{1}{\prod_{d=1}^{D} \text{FWHM}_{id}}. \tag{3.8}$$

The RPV of each voxel is crucial for the RFT nonstationary cluster size adjustment described below.

We next describe the RFT adjustment followed by our proposed empirical adjustment. First some common notation: for a given cluster-forming threshold u_c let S be the size of a cluster (i.e., its voxel count), and let $\mathcal{S}$ be the set of voxel indices that define the cluster; let S_i be the size of a cluster that covers voxel i with the statistic value T_i, zero if $T_i < u_c$. Generically, any time we index a cluster statistic with i, we imply the use of a voxel-wise "cluster image", where each voxel takes the value of the cluster statistic of the cluster that covers voxel i, or zero if there is no cluster at i.

3.2.2 RFT Cluster-size Adjustment

In a nonstationary cluster test, the size of a cluster is measured in units of RESELs. While the most straightforward approach might seem to be to compute cluster size in voxels S and then standardize it in some way, instead the RFT method specifies summing up the RPV_i within each cluster. That is, for a given cluster the RPV cluster size is

$$S_{\text{RFT}} \;=\; \sum_{i \in \mathcal{S}} \text{RPV}_i. \tag{3.9}$$

The S_{RFT} can be shown to equal to the size of the cluster in another version of the data that has been spatially distorted such that stationarity holds (Worsley et al., 1999). Also, under stationarity it can be shown that each RPV_i is proportional to the inverse of the expected cluster size (Friston et al., 1994). Thus, while S_{RFT} does not take the form of "S standardized", it can be seen as the "sum of S standardized voxels".

While this S_{RFT} statistic has an approximate parametric null distribution (Hayasaka et al., 2004), for symmetry with other methods we do not rely on these results and only

consider P-values from nonparametric permutation. Even with permutation P-values for S_{RFT}, however, inference with this statistic may not completely eliminate the effects of nonstationarity. That is, RPV may not entirely account for differences in average cluster size, due to insufficient smoothness or highly structured, convoluted patterns of smoothness. And even if RPV captures the differences in cluster size on average, the distribution of S_{RFT} may still vary depending on true smoothness; for example, the effects of discreteness at very low smoothness (since S is an integer) may impact the distribution of S_{RFT} differently in very smooth areas. This motivated the development of alternate approaches, as described next.

3.2.3 Empirical Cluster-size Adjustment

Motivated by the RFT statistic which sums up standardized voxels, we propose an empirical adjustment where we replace the voxel-wise measure of roughness (RPV) with the inverse expected cluster size under the null hypothesis. This method relies on the use of a resampling method (e.g., permutation) to provide null-hypothesis realizations of the data. For permutation k ($k = 1, \ldots, K$), let $S_{i,k}$ be the size of the cluster at voxel i, where $k = 1$ corresponds to the (unpermuted) original data. Then the empirical cluster size per voxel (ECSPV) is

$$\mathrm{ECSPV}_i \quad = \quad \left(\frac{\sum_{k=2}^{K} (S_{i,k})^E}{K_{S_i}} \right)^{1/E} \tag{3.10}$$

where K_{S_i} is the number of permutations $k \in \{2, ..., K\}$ where $S_{i,k} > 0$ (i.e., to normalize the sum of cluster sizes by the count of clusters at voxel i), and E is a normalization parameter. We found that the severe skew of cluster sizes in D=3 dimensions makes the arithmetic mean quite sensitive to outliers, and thus considered $E < 1$ to reduce the skew and provide a better measure of central tendency of cluster size. We consider different values of E (2/3, 1 and 2) in our evaluations (more details below). We exclude the original data from ECSPV calculation to avoid non-null signal and hence biasing this measure.

A "first pass" resampling is needed to estimate the ECSPV, and then a "second pass" of resampling is done as part of a formal cluster size permutation test, with the test statistic

$$S_{\text{Emp}} = \sum_{i \in \mathcal{S}} \frac{1}{\text{ECSPV}_i}, \tag{3.11}$$

which, analogous to S_{RFT}, adjusts the size of each voxel before summing.

It is also possible to jointly use the RFT-based and empirical adjustment, i.e., replacing the S in Eq. 3.10 with S_{RFT}. While we don't expect this to be so useful since S_{RFT} is already adjusted for spatial inhomogeneities, we consider this procedure as well for completeness. Note that the first-pass estimation of ECSPV_i is not a permutation test, but rather a resampling-based estimation of a nuisance parameter. See Appendix B.2 for a detailed justification.

3.2.4 Empirical TFCE Adjustment

The TFCE statistic is defined voxel-wise and based on cluster-wise results. The (unadjusted) TFCE statistic at voxel i is

$$\text{TFCE}_i = \sum_{u=0,du,2du,\ldots,u_{\max}} S_i^{E_{\text{TFCE}}} u^{H_{\text{TFCE}}} \tag{3.12}$$

where E_{TFCE} and H_{TFCE} are tuning parameters set to their recommended values of 0.5 and 2.0 (Smith and Nichols, 2009), respectively, du is discretization step size and $u_{\max}$ is a value larger than the maximum statistic in the image. We define the empirical TFCE per voxel (ETPV) as

$$\text{ETPV}_i = \frac{\sum_{k=1}^{K} \text{TFCE}_{i,k}}{K_{\text{TFCE}_i}} \tag{3.13}$$

where $\text{TFCE}_{i,k}$ is the value of the TFCE statistic at voxel i on permutation k, and K_{TFCE_i} is the number of permutations where $\text{TFCE}_{i,k} > 0$ (TFCE is zero for all voxels where $T_i \leq 0$). The voxel-wise normalization is

$$\text{TFCE}_{\text{Emp}_i} = \frac{\text{TFCE}_i}{\text{ETPV}_i}. \tag{3.14}$$

We also define a RPV-adjusted version of TFCE, by replacing S with S_{RFT}in Eq. (3.12) and hence creating $\mathrm{TFCE}_{\mathrm{RFT}}$.

Note that so far we introduced two cluster-related statistics in the first run (i.e., unadjusted): TFCE and S whose adjustment with RFT method results in S_{RFT} and $\mathrm{TFCE}_{\mathrm{RFT}}$, respectively. Both of these statistics can be adjusted empirically via a second run in order to result in S_{Emp} and $\mathrm{TFCE}_{\mathrm{Emp}}$. Note that *if* the heterogeneity in the distribution of cluster-size were completely explained by local variation in RPV/FWHM, an empirical method could do no better than an RFT-based solution (Hayasaka et al., 2004). However, the accuracy of RFT has been shown to break down in many settings; hence we propose empirical approaches that do not assume a theoretical relationship between RPV/FWHM and cluster size. As there is no unique solution to this problem we consider a range of possible empirical adjustments.

3.2.5 Nonstationarity Assessment

A formal permutation test for nonstationarity is not possible, since stationarity does not imply exchangeability (i.e., spatial correlation is still present under stationarity). However, a null hypothesis of stationarity does justify a global pooling of permutation distributions. In fact, some authors (Bullmore et al., 1999) use such globally pooled distributions to reduce the number of permutations needed. Pooling also allows an evaluation of nonstationarity in the following way. Under stationarity, a global permutation distribution will produce P-values that are valid and homogeneous over the brain. Under nonstationarity, however, these P-values are very significant in smooth areas and very insignificant in rough areas. This spatial effect will be consistent, and seen even in permuted versions of the data. Hence our assessment of nonstationarity is based on the *spatial* standard deviation of uncorrected P-values based on a globally pooled permutation distribution.

We now precisely define this process for unadjusted cluster size S, but the very same method is also applied to S_{RFT}, S_{Emp}, TFCE, $\mathrm{TFCE}_{\mathrm{RFT}}$ and $\mathrm{TFCE}_{\mathrm{Emp}}$. Let $\{S^*_{k'}\}$ be the

pooled cluster-size permutation distribution (more details in Appendix B.3), the set of all cluster sizes observed over space *and* over permutations; as each of the K permutations will generally contribute multiple clusters, the size of the pooled distribution K_{S^*} will be very large, $K_{S^*} \gg K$. The pooled uncorrected P-value for cluster size S is the proportion of pooled statistics that equal or exceed the observed cluster size:

$$P_S = \frac{\#_{k'}\{S^*_{k'} \geq S\}}{K_{S^*}} \tag{3.15}$$

where $\#'_k\{S^*_{k'} \geq S\}$ is the number elements in the pooled distribution that equal or exceed the observed statistic, $k' = 1, \ldots, K_{S^*}$.

Further, each element $S^*_{k'}$ of the permutation distribution can also be transformed to an uncorrected P-value, $P_{S^*_{k'}}$, by applying Eq. (3.15) to $S^*_{k'}$ instead of S. Then, considering the corresponding "cluster images", the pooled uncorrected P-value for a cluster observed at voxel i on permutation k is

$$P_{S_{i,k}} = \frac{\#_{k'}\{S^*_{k'} \geq S_{i,k}\}}{K_{S_i}}. \tag{3.16}$$

Valid uncorrected P-values are uniformly distributed ($U(0,1)$ with the mean of 0.5) under the null hypothesis and stationarity, while invalid or conservative P-values induced by nonstationarity will bias P-values up or down. In particular, we are concerned with spuriously small P-values in relatively smoother regions. Working instead with $-\log_{10}$ P-values to emphasize very significant P-values, one can show that $-\log_{10}$ P-values are exponentially distributed (under the null) with mean and standard deviation of $1/\ln(10) = 0.4343$. Hence, we quantify the heterogeneity of false-positive risk with null data by computing $-\log_{10}$ P-values, averaging them over the K permutations at each voxel (since the nonstationarity effects will be consistent over permutation), and computing the resulting image's standard deviation (SD).

Note that such voxel-wise averaging of $-\log_{10}$ P-values across permutations is sensible for any voxel-wise statistic, such as TFCE. For cluster-based statistics however, there is a bias in that large clusters are more likely to hit a given voxel than small clusters. As

an extreme example, suppose that 99% of clusters have size of 1 voxel, and 1% of clusters have size $V/2$, equal to half the search volume; for a given voxel i, the relative chance that it will ever see a 1-voxel cluster is $0.99 \times (1/V)$, while the chance it will observe the half-volume cluster is $0.01 \times (1/2)$. Therefore in order to correct this bias a weighted mean is used to compute the average $-\log_{10}$ P-value at each voxel,

$$M_i = \frac{\sum_{k=1}^{K} -\log_{10}(P_{S_{i,k}})\, w_{i,k} \mathrm{I}_{S_{i,k}>0}}{\sum_{k=1}^{K} w_{i,k} \mathrm{I}_{S_{i,k}>0}}, \tag{3.17}$$

where $w_{i,k} = 1/S_{i,k}$ is the weight, proportional to the chance of a cluster of size $S_{i,k}$ hitting voxel i, and I is the indicator function.

All permutation tests use 1000 permutations, resulting in a minimum possible P-value of 0.001 ($-\log_{10}(P)$ of 3). Also, for null-data cluster-based analyses, the cluster-forming threshold of $u_c = 2.5$ and 3 for simulated and real data, respectively, are applied to T-statistic images, and the optimal parameter set of $E_{\text{tfce}} = 0.5$ and $H_{\text{tfce}} = 2$ (as proposed in Smith and Nichols (2009) for T-statistic images) is used for TFCE inference.

3.2.6 ROC-based Evaluations

To summarize each method's performance on a simulated signal+noise data, we use a receiver-operator characteristic (ROC) curve, where the free parameter is the critical threshold on S (or which ever method is being evaluated). We use the area under the curve (AUC) as a single-valued summary of an ROC curve; the higher the AUC, the better.

An ROC curve is strictly defined only for a single detection task, repeated many times with different thresholds. In our setting, following Smith and Nichols (2009), with thousands of tests (one per voxel) a free-response (FR) ROC curve is more appropriate (Bunch et al., 1978). An FR-ROC curve replaces either FPR or TPR (or both) with measures that aggregate true or false positives over voxels. We chose to use the FWE-FPR (the chance of one or more false positives anywhere) on the x-axis instead of a marginal (uncorrected) FPR, as practitioners are usually interested in FWE-corrected

inferences, and FW-TPR to measure the chance of one or more detections. Finally, since only performance with low false positive risk is of interest, we computed the AUC only for FW-FPR<0.05, scaling by 1/0.05 to renormalize the AUC to the range $[0, 1]$.

3.2.7 Data

Each method's performance is assessed on both simulated and real data. For the stationary null data simulation, two subject-groups of size 20, standard Gaussian noise images, dimension $150 \times 150 \times 150$, are generated and smoothed with a Gaussian smoothing kernel, with σ=2, 3, 4 and 5 voxels. To avoid generating nonstationarity at the edge, the outer 30 voxels are excluded, and the remaining central $90 \times 90 \times 90$ cube is analysed for a between-group difference contrast using the permutation testing in FSL's (FMRIB Software Library[4]) randomise.

To simulate the nonstationary null data, we again start with two groups of 20 $150 \times 150 \times 150$ Gaussian noise images. The images are then smoothed with three different 3D Gaussian kernels, producing three images with low, medium, and high smoothness. These images are combined in a way that an outer layer smoothed with σ_1 encloses a middle layer smoothed with σ_2, which encircles a core smoothed with σ_3 (we denote the nonstationary configuration as $\sigma_1/\sigma_2/\sigma_3$). The core is $30 \times 30 \times 30$ voxels, centred within a $60 \times 60 \times 60$ voxel middle layer, which itself was centred in a $90 \times 90 \times 90$ volume. The combined image is smoothed again with a 3D Gaussian filter with σ =1.5 voxels to eliminate discontinuities at the borders of different smoothness, and, finally, to avoid the edge effects the outer 30 voxels are discarded (see Algorithm 2 of Appendix B.4 and Figure 3.1).

To generate realizations with true signal we simulate a two-group analysis with expected between-group differences in the form of scaled Gaussian probability density functions (PDFs) located at (x,y,z) coordinates of (45,45,45), (55,55,55), (60,60,60), (70,70,70), (90,60,50), (50,70,100), (90, 100, 70), (65, 80, 90), (105,65,40), (110,60,110),

[4]http://www.fmrib.ox.ac.uk/fsl

Table 3.1: The coordinates (x,y,z) of the centre of the Gaussian functions and their standard deviation (σ). These Gaussian blobs simulate a 3D signal for one group of subjects.

Index	X	Y	Z	σ
1	45	45	45	6
2	55	55	55	2
3	60	60	60	5
4	70	70	70	4
5	90	60	50	3
6	50	70	100	4
7	90	100	70	2
8	65	80	90	6
9	105	65	40	4
10	110	60	110	3
11	90	60	90	4
12	90	90	80	3

(90, 60, 90) and (90, 90, 80) inside a 150x150x150 volume, with corresponding σ = 6, 2, 5, 4, 3, 4, 2, 6, 4, 3, 4, and 3, respectively. Each Gaussian function is scaled to have unit peak intensity and, after summing the 12 foci together, the resulting signal image is also scaled to have unit peak intensity and, finally, all values below 0.25 are set to zero. This signal image is used as described in Algorithm 2 of Appendix B.4 for an ROC analysis.

To assess each method's performance on real data, we also considered null fMRI and VBM datasets. All data employed had been collected in accordance with local ethics approval. The fMRI dataset is a pain study with 16 healthy subjects (also used in Chapter 2). Processing of the functional images at the first level is performed using FSL (Smith et al., 2001). Functional images were motion corrected (Jenkinson et al., 2002) and spatially smoothed (σ = 1, 1.5, 2, 3 and 5 mm) prior to temporal model fitting including modelling of autocorrelation (Woolrich et al., 2001). Registration to the MNI152 standard brain space was performed in two stages: (1) the fMRI data from a given subject was registered to that subject's T1 structural using linear registration and (2) the subject's structural image was registered to the MNI standard brain using linear registration (Jenkinson and Smith, 2001; Jenkinson et al., 2002). Dividing these subjects arbitrarily into two groups is a null-data analysis as there is no expected difference.

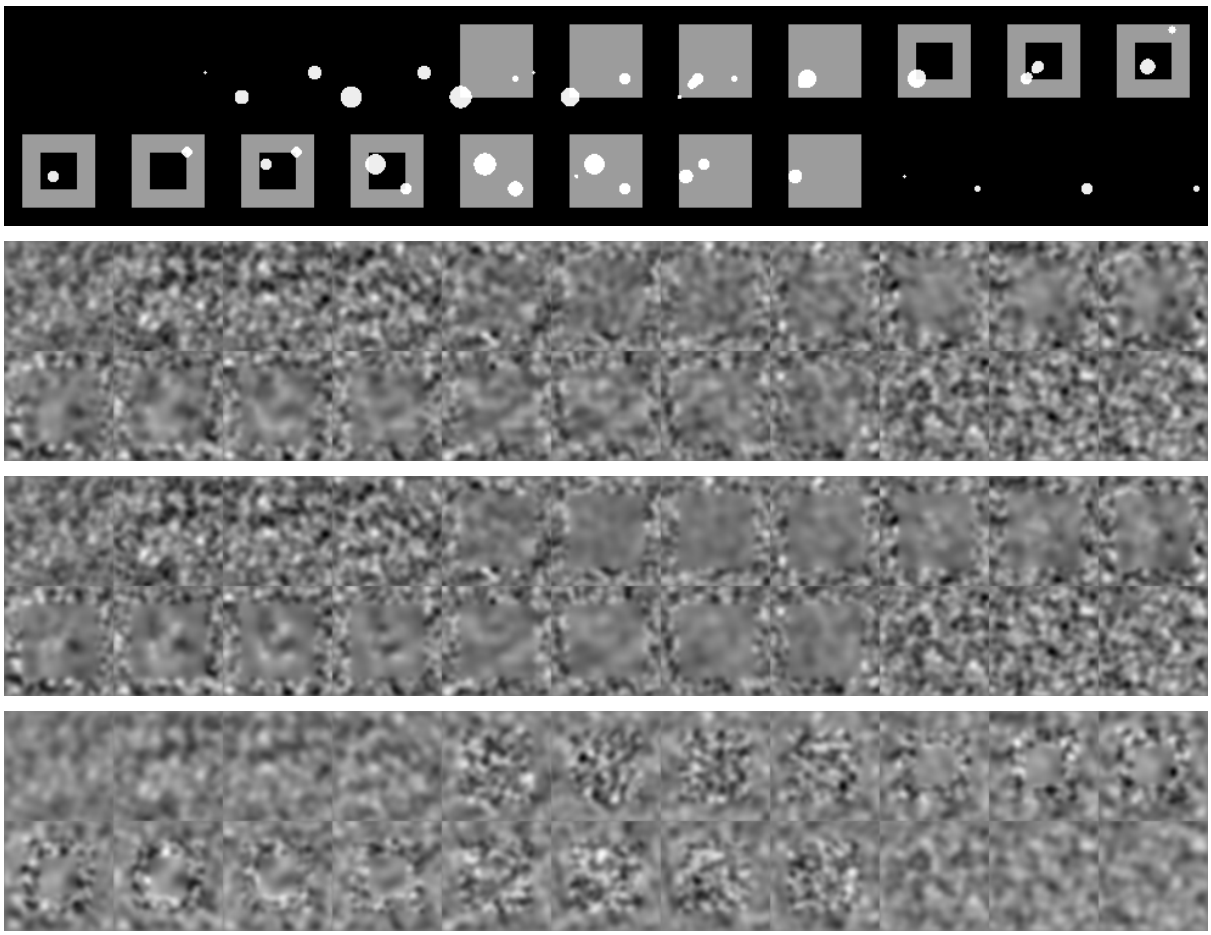


Figure 3.1: Sample slices from the simulated data. The top image shows slices from the signal image (white blobs) overlaid on the mask corresponding to the middle-layer smoothness region (σ_2 region, with gray color). As this figure shows, there are signals at the boundaries as well within each smoothing region, which helps assessing different methods' performance at each region. In other rows, nonstationary *noise* images for σ=2/3/4, 2/4/3 and 3/2/4 (from top to bottom, respectively) are shown to illustrate the effect of variation in local smoothness of the image. The displayed slices are selected from z=1 to z=90, including every fourth slice.

The VBM data includes one null dataset composed of structural gray-matter images of 35 healthy control subjects (age range 51-86 y, mean 70.1 y; 20 male, 15 female), and one three-group dataset composed of structural gray-matter images of 46 healthy control (age range 51-86 y, mean 71.2 y; 28 male, 18 female), 50 AD (Alzheimer's disease; age range 49-89 y, mean 74.1 y; 31 male, 19 female) and 57 MCI (mild cognitive impairment; age range 50-84 y, mean 70.1 y; 32 male, 25 female). For VBM analysis of the null data, subjects in the one-group data are randomly assigned a group label and divided into two groups (hence no group difference is expected). In order to assess the inter-group differences, an optimized VBM protocol is carried out using FSL-VBM (Douaud et al., 2007), which localizes the brain regions with significant grey-matter-volume differences between the groups. In this protocol, first a left-right-symmetric study-specific grey-matter template is built from the healthy cohort grey matter, i.e., segmented native images. The gray-matter images are nonlinearly registered to the ICBM-152 grey matter template using FNIRT (FMRIB's Non-linear Image Registration Tool) [5], flipped along the x-axis and then averaged. This step is followed by nonlinear normalization of grey-matter images onto this study-specific template. The method also introduces a compensation (or "modulation') for the contraction/enlargement due to the nonlinear component of the transformation by dividing each voxel of each registered grey matter image by the Jacobian of the warp field. Finally, modulated registered grey-matter-volume images are smoothed at $\sigma = 2$, 3, 4 and 5 mm for the null data and $\sigma = 7$ mm for the three-group data.

3.3 Results

In the following figures we label unadjusted methods, TFCE and S, "1-pass"; we label results with RFT-based adjustment, TFCE_{RFT} and S_{RFT}, "1-pass, rpv". When using an empirical adjustment, TFCE_{Emp} and S_{Emp}, we use the label "2-pass", and when empirical adjustment is applied to RFT-based adjusted statistic, we label it "2-pass,

[5] http://fsl.fmrib.ox.ac.uk/fsl/fnirt/

rpv". Also, when excluding the outlier observations from the calculations (as described in Appendix B.1 and B.1.2), we use "robust" in the label. We first review the results for null simulated and real data. Figure 3.2 shows the heterogeneity of uncorrected cluster size P-values, measured as the standard deviation (SD) of $-\log_{10}(\mathrm{P})$ over space (see Section 3.2.5), for stationary and nonstationary simulated null data.

As expected, Figure 3.2 shows that all of the methods show good homogeneity with stationary data, but the unadjusted (1-pass through the permutation testing) method has dramatically increased SD with nonstationary data. RPV adjustment reduces SD to the level of stationary data, as does the robust RPV-adjustment; as the simulated data of course do not have outliers, it is notable that the robust method does not appear to suffer from using less data. The 2-pass empirical adjustment has the smallest SD of all methods considered. An empirical adjustment following RPV adjustment ("2-pass, rpv") is nearly as bad as no adjustment at all, possibly because the empirical adjustment does a poor job capturing the small amount of nonstationarity remaining after RFT-based adjustment.

Note that if the data is stationary we do not expect to see dramatic differences when comparing one method with another. However, in the so called "stationary" null data in Figure 3.2, there still can be some nonstationarities caused by "edge effects", i.e., voxels closer to the edges have a smaller chance for being hit by a large cluster than those voxels at central regions, and hence a smaller ECSPV. This cannot be evaluated using RFT-based RPV calculations, and thus, can be counted as a strength of the empirical method. Also, it is not the absolute standard deviation that is important, but rather the relevant comparison between the right and left side of Figure 3.2 is the actual performance index for each method; this cross-data comparison of methods indicates that two-pass methods are similar while one-pass has degraded performance under nonstationarity.

As mentioned in Section 3.2.5, the cluster-forming threshold for the analysis of null data is u_c=2.5. In many papers, however, higher values such as u_c=3 are also used in order to avoid violating the RFT assumptions for parametric cluster-based inference (Poline

et al., 1997). As the choice of u_c is arbitrary and hence has no gold standard value, in order to address this concern we carried out the same analysis with u_c=3, which did not improve the results. Thus, we used u_c=2.5 and as we do not carry out the parametric cluster inference this should not be a concern.

Figure 3.3 shows equivalent results for TFCE inference. Note that the absolute range of SD's is an order of magnitude smaller than the SDs for cluster inference. Using TFCE_{RFT} (i.e., 1-pass, rpv) does not provide good correction for nonstationarity, while TFCE_{Emp} (i.e., 2-pass and 2-pass, rpv) has dramatically reduced the SD.

Real fMRI and VBM null data analyses show similar reductions in uncorrected P-value heterogeneity with the use of adjustment methods; 2-pass empirical methods always produce the most homogeneous inference (Figures 3.4 & 3.5).

Several variations in the methods did not show noticeable differences in performance and are omitted from the plotted results. Kiebel's smoothness estimation method was indistinguishable from Jenkinson's method (see the related formula in Appendix B.1), and so only results using the latter are displayed. Use of the empirical adjustment normalization factor $E = 2/3$ always performed slightly better than $E = 1$ and 2, and thus only the $E = 2/3$ results are shown (simulation results exploring different values of E can be found in Salimi-Khorshidi et al. (2009a)).

Using simulated-data analyses with signal present and AUC as a measure of power, the optimal method depended on the configuration of the nonstationarity and the signal-to-noise ratio (SNR). Figure 3.6 shows AUC at representative samples of the patterns of nonstationarity and SNR levels considered. Over all, cluster-based inference appears to improve with either of the two adjustment techniques (i.e., RPV and empirical), depending on the configuration of nonstationarity. TFCE, however, does not necessarily show an improvement when adjusted, which implies its robustness to nonstationarities.

For the real-data VBM analysis, testing for MCI grey matter greater than AD grey matter, clusters were found in the areas typical of AD atrophy: medial temporal lobe, posterior cingulate, and frontal lobe (see Karas et al. (2004) and Risacher et al. (2009)).

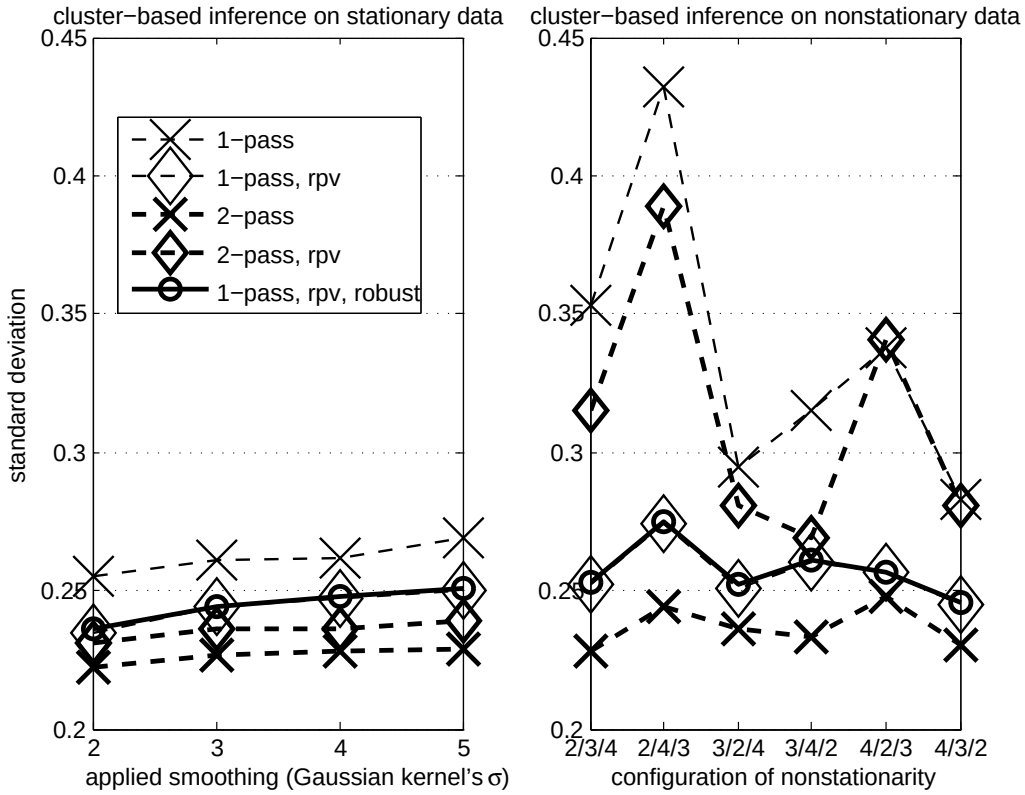


Figure 3.2: Heterogeneity of cluster-size test's false positive risk for simulated null data, as measured by the spatial standard deviation of uncorrected cluster-wise $-\log_{10}(\mathrm{P}-\mathrm{values})$. For stationary null data (left) the SD is plotted against applied smoothing (the σ of the Gaussian kernel); there is low heterogeneity for all the measures considered. For nonstationary null data (right) $-\log_{10}(\mathrm{P}-\mathrm{values})$ standard deviation is plotted against different configurations of nonstationarity; here the standard deviation is higher over all, and is much greater for 2/4/3 in particular. The 1-pass P-values are always the most heterogeneous, and the 2-pass empirically-adjusted P-values are always the most homogeneous, with RPV-based adjusted P-values in-between.

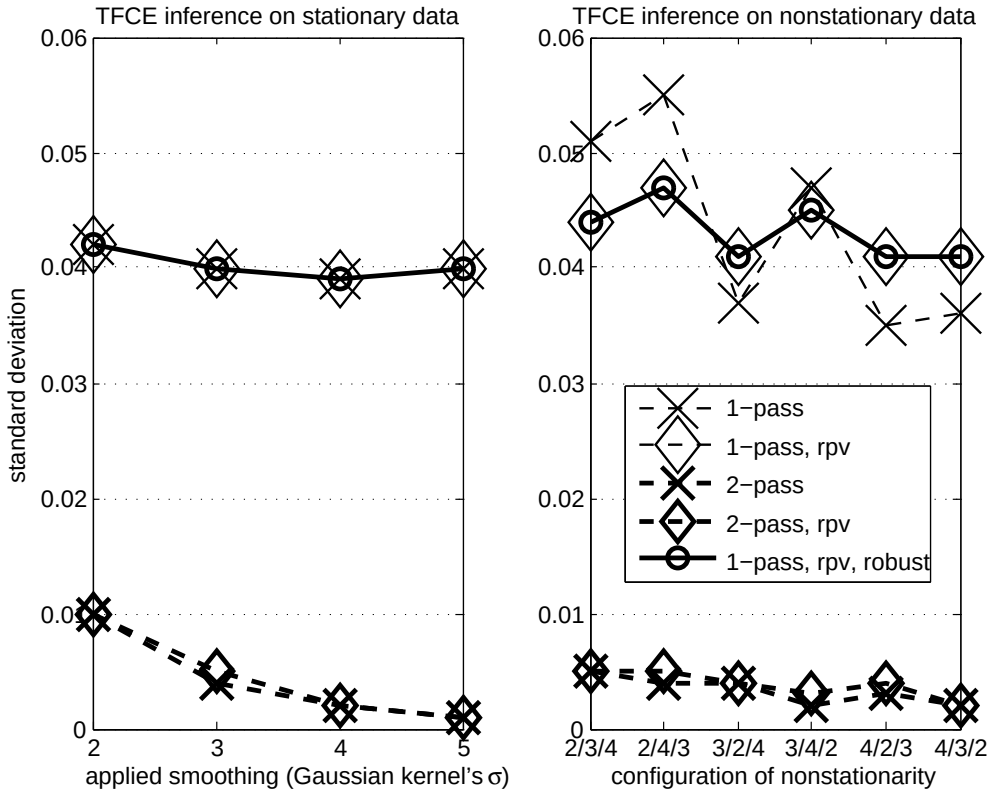


Figure 3.3: Heterogeneity of TFCE false positive risk for simulated null data, as measured by the spatial standard deviation of uncorrected TFCE $-\log_{10}(\mathrm{P}-\mathrm{values})$. See Fig. 3.2 caption for descriptions of abscissa values. For stationary null data (left), there is low heterogeneity for either standard (1-pass) TFCE or RPV-adjusted TFCE, and ultra-low heterogeneity for 2-pass TFCE (compare with Fig. 3.2). For nonstationary null data (right), the standard deviation is not appreciably higher overall, indicating TFCE's robustness with respect to nonstationarity. While standard TFCE shows higher standard deviation for some configurations of nonstationarity (2/4/3 in particular), the RPV-adjusted TFCE is more stable. Again, the 2-pass TFCE has exceptionally low P-value standard deviation (i.e., high homogeneity).

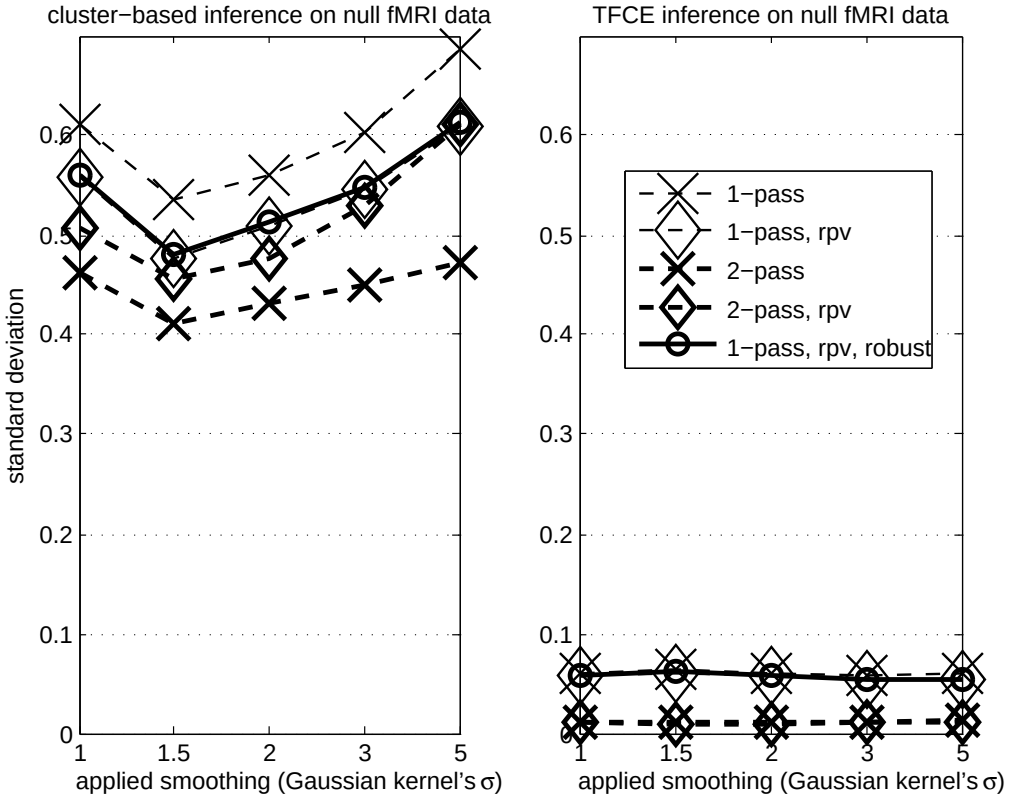


Figure 3.4: Heterogeneity of false positive risk for null fMRI data versus applied image smoothing, as measured by the spatial standard deviation of uncorrected cluster and TFCE $-\log_{10}(\mathrm{P}-\mathrm{values})$. For cluster size inference (left) heterogeneity increases with increasing smoothing, but 1-pass P-values are always the most heterogeneous, and 2-pass empirically adjusted P-values always the most homogeneous. For TFCE inference (right) P-values are much more homogeneous and are generally unaffected by image smoothness. Again, the 2-pass TFCE has very low heterogeneity.

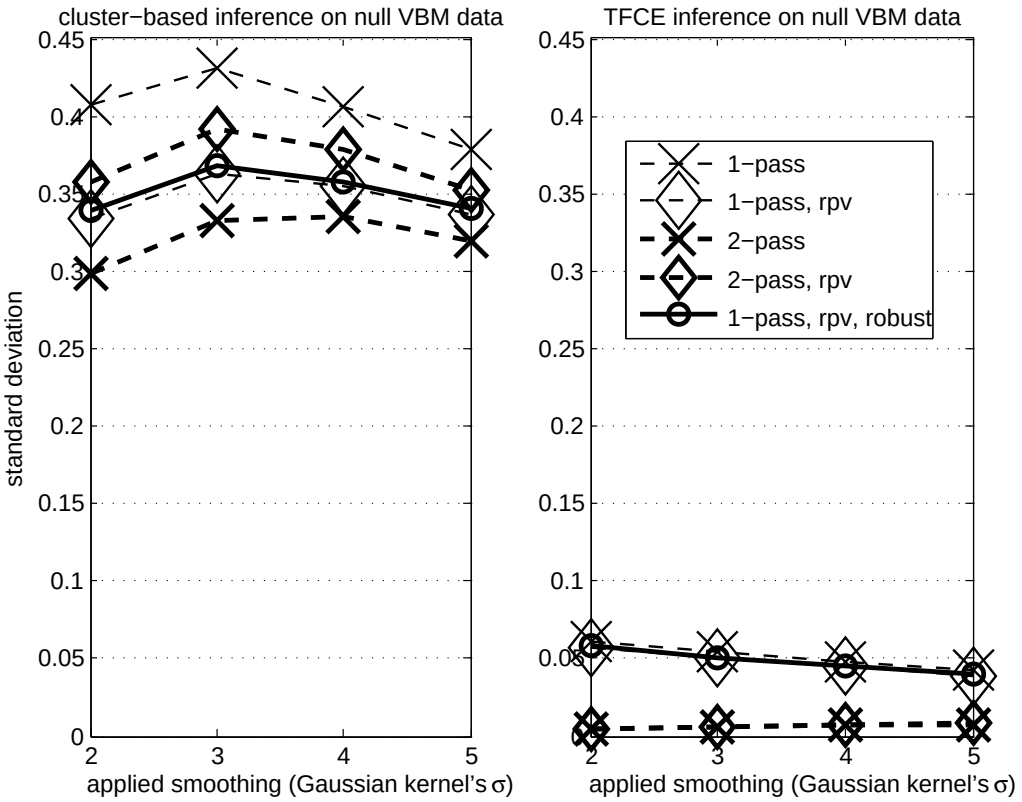


Figure 3.5: Heterogeneity of false positive risk for null VBM data versus applied image smoothing, as measured by the spatial standard deviation of uncorrected cluster and TFCE $-\log_{10}(\mathrm{P}-\mathrm{values})$. For cluster size inference (left) heterogeneity tends to decreases with increasing smoothing, but 1-pass P-values are always the most heterogeneous, and 2-pass empirically adjusted P-values always the most homogeneous. For TFCE inference (right), the 1-pass TFCE P-values show decreasing heterogeneity with increasing image smoothness, just like cluster-based inference. However, the 2-pass TFCE again has very low heterogeneity.

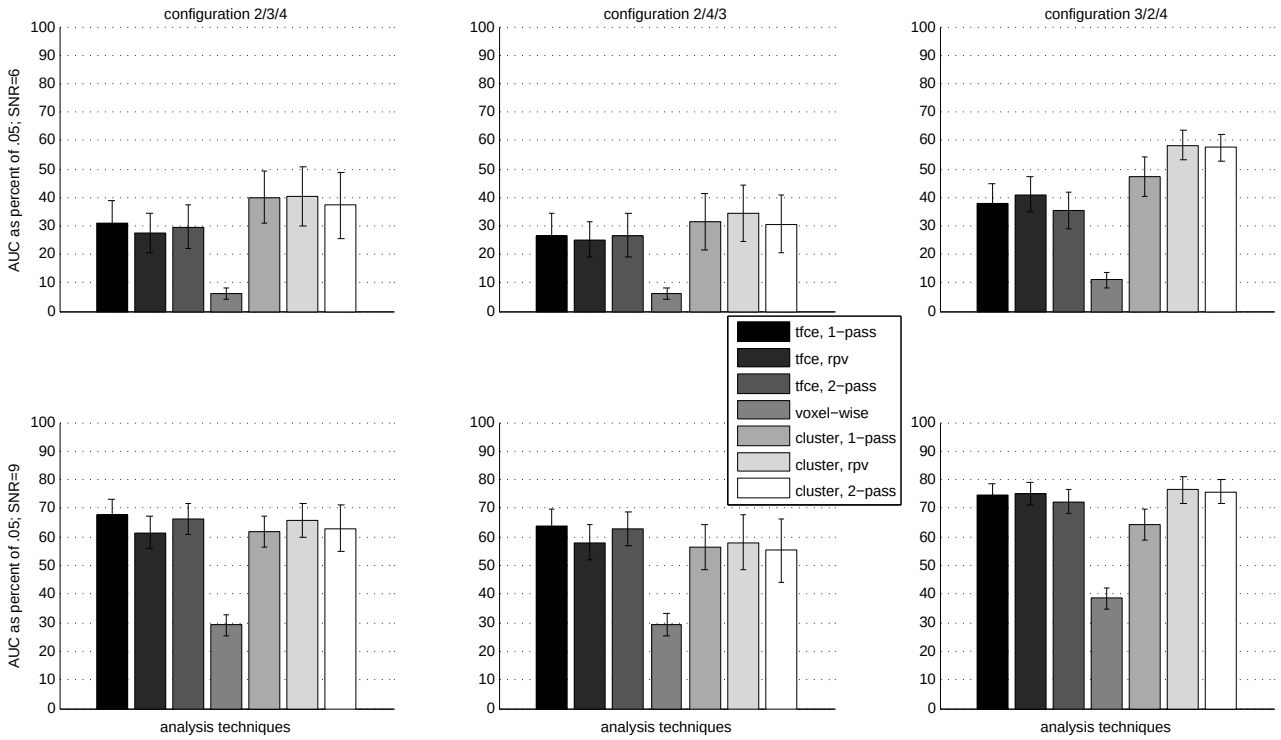


Figure 3.6: Joint sensitivity-power evaluation using area under the ROC curve (AUC) with simulated nonstationarity data. From a wide range of simulations, six representative results are shown: Upper row shows peak SNR=6, and lower row shows peak SNR=9; columns show different configurations of nonstationarity, from left to right: 2/3/4, 2/4/3, and 3/2/4. Highest AUC depends on exact setting, but is usually an adjusted cluster method or a TFCE method.

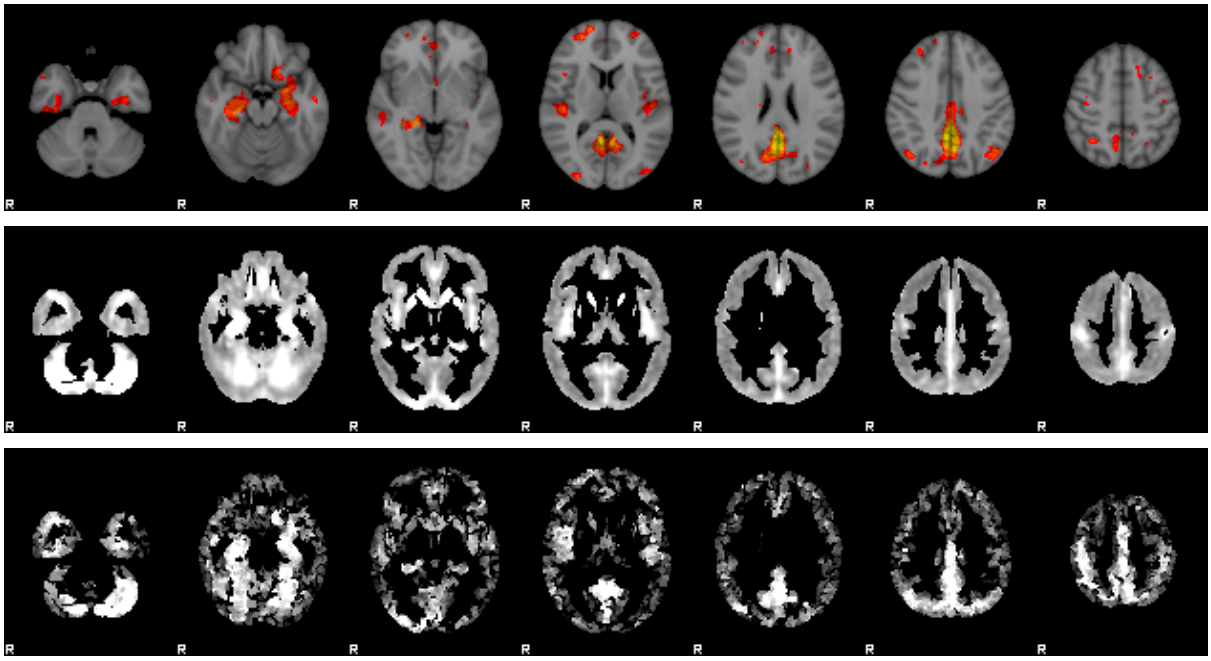
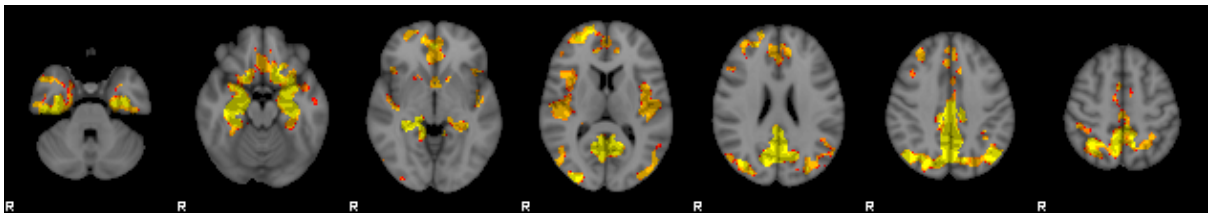


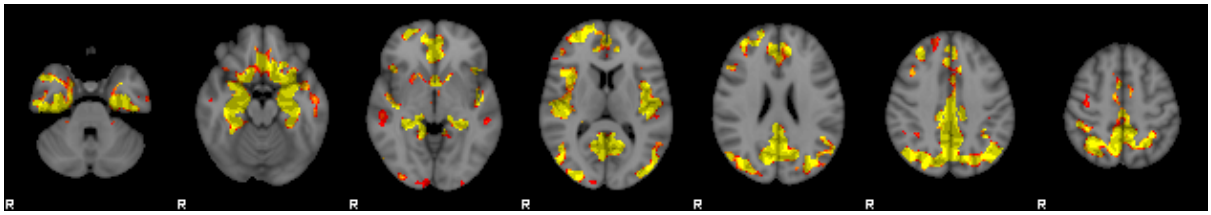
Figure 3.7: VBM cluster-size results (MCI>AD) with smoothness maps. Top row shows clusters obtained with a T-stat=3 cluster-forming threshold. Middle row shows FWHM-based smoothness image (intensity range 0-8), and bottom row shows ECSPV$^{1/3}$ image (intensity range 0-15). Note not only how the largest clusters appear in smoothest areas (by either measure), but also the differences between the two smoothness measures; FWHM is greatest in the cerebellum, while ECPSV is greatest in the posterior cingulate area. Displayed slices are selected from z=-32mm to z=38mm, every 14 millimeters in MNI coordinates.

Figure 3.7 shows these clusters, along with maps of the two measures of local smoothness, FWHM (middle) and ECSPV$^{1/3}$. It is notable that some of the areas of highest ECSPV are exactly where the clusters have been observed, suggesting that the clusters may be false positives *or* these areas of signal simply coincide with areas where there is exceptionally structured noise. Figure 3.8 shows the MCI>AD results using TFCE. The 2-pass adjustment has reduced sensitivity, but precisely in the areas where the ECSPV is large; thus the adjustment is doing what it is supposed to do. The 2-pass RPV, however, is almost as good as the 2-pass, suggesting that, in this case, the empirical adjustment was not harming the accuracy of the the method.

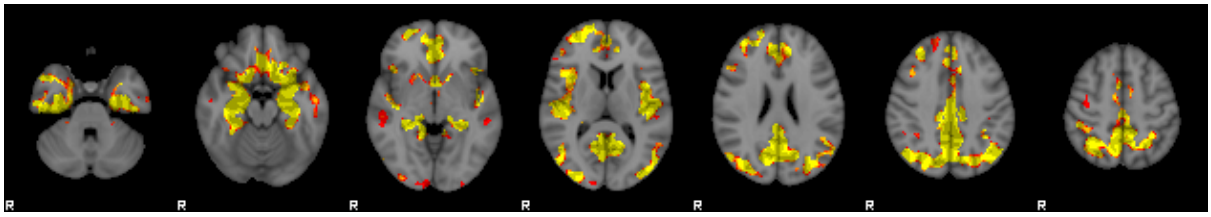
To further understand the difference between the RPV-based and ECSPV-based adjusted inferences, we directly compared smoothness estimates and the change in P-value after adjustment within each of the 33 clusters in the real VBM data (Figure 3.9).



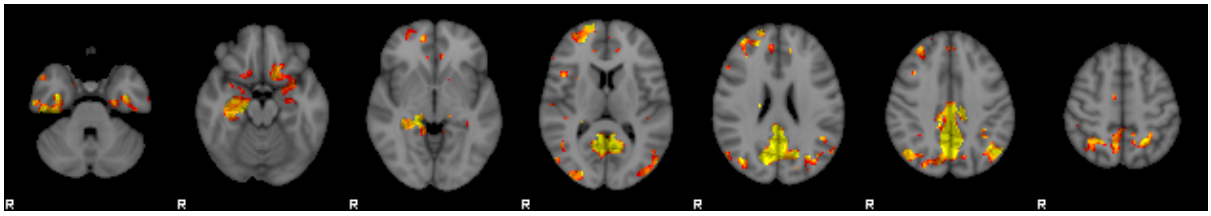
(a) 1-pass (unadjusted)



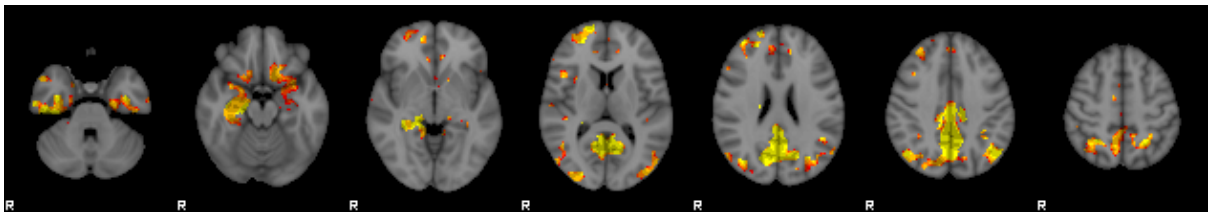
(b) 1-pass RPV



(c) 1-pass RPV, Robust



(d) 2-pass



(e) 2-pass RPV

Figure 3.8: TFCE results for VBM data (MCI>AD) using different adjustment methods. The standard TFCE result (a) is very similar to RPV-adjusted TFCE (b). RPV-based adjustment by using a robust RPV calculation (c) shows similar posterior cingulate effects, but finds reduced parietal, frontal or insular differences. 2-pass adjustment (d) and RPV-adjustment with 2-pass adjustment (e) show patterns suggesting that the significance of the parietal, frontal and insular areas may have been over-estimated with the standard TFCE or RPV-adjusted TFCE. Colour (red-yellow) shows 1-P-value with range 0.95-1, and the displayed slices are selected from z=-32mm to z=38mm, every 14 millimeters in MNI coordinates.

These clusters' specifications are shown in the table in Table 3.2. The table specifies how smooth/rough each cluster is (on average) using different measures introduced in the chapter, and also, what the effect of adjustment was on its significance. As an example of this information's usage, the plots in Figure 3.10 confirm that FWHM shows a positive correlation with ECSPV$^{1/3}$, but there are some exceptions. In particular the posterior cingulate cluster had the largest ECSPV but only moderate FWHM. The effect of adjustment, measured by $-\log_{10}(P^{\text{FWE}}_{1-\text{pass}}) - (-\log_{10}(P^{\text{FWE}}_{\text{adjusted}}))$, shows the expected pattern of increased significance for clusters in relative rough regions, and decreased significance for clusters in relatively smooth regions. Note that since the right panel of Figure 3.10 compares the effect of adjustment versus FWHM, RFT's measure of nonstationarity, it is not surprising that RFT shows the expected pattern. More important is that the empirical adjustment shows the general pattern expected, indicating that it is making similar corrections as RFT.

3.4 Discussion and Conclusions

Cluster-based inference must account for nonstationarity in VBM and other types of data with highly variable smoothness. Parametric RFT methods that ignore nonstationarity are invalid, and result in inflated false positive risk in smooth areas. Nonparametric cluster-size tests based on the maximum distribution (i.e., FWE corrected inferences) are valid, but will have non-uniform sensitivity related to the local smoothness. For example, in a very rough region, the mean null cluster size might be 2.5, and clusters as large as 10 voxels might never occur by chance; hence we might regard a 50 voxel cluster in this region as very unusual. But since the maximum distribution considers all regions, including smooth regions where, say, the mean null cluster size is 1000, the maximum-based FWE significance will never find a 50 voxel cluster to be significant. If (as may often be the case) the likelihood of finding extended signal is affected by nonstationarity in a similar manner to the effect of the nonstationarity on the null distribution, then correction of nonstationarity could hope to improve both sensitivity and specificity

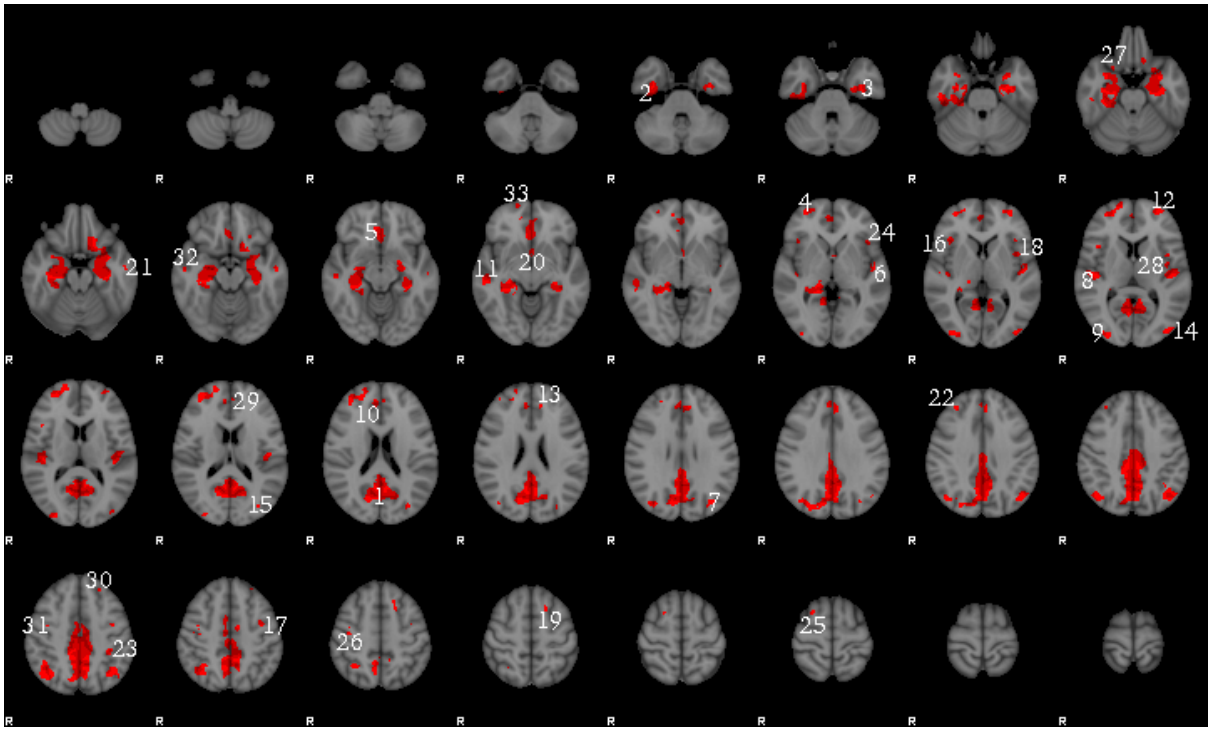


Figure 3.9: The 33 clusters surviving the cluster-forming threshold of T-stat=3 with their indices (the same indices as shown in Table 3.2). Clusters with a size less than 10 voxels are excluded from the analysis. In this image colour is just to show cluster location, and the displayed slices are selected from z=-52mm to z=72mm, every 4 millimeters in MNI coordinates.

Table 3.2: Specifications of the clusters formed in real VBM analysis (same as those in Figures 3.9 and 3.10); columns 1-7 are clusters' indices as are as shown in Figure 3.9 and their size (in voxel count), mean FWHM, mean ECSPV$^{1/3}$, pre-adjustment $-\log_{10} P$, empirically-adjusted $-\log_{10} P$ and RPV-adjusted $-\log_{10} P$.

Index	Voxel Count	<FWHM>	<ECSPV$^{1/3}$ >	$-\log_{10} P^{\mathrm{FWE}}$	$-\log_{10} P^{\mathrm{FWE}}_{emp}$	$-\log_{10} P^{\mathrm{FWE}}_{rpv}$
01	5035	6.16	13.35	3.00	3.00	3.00
02	1723	7.43	9.55	2.52	2.70	2.40
03	1262	7.89	8.91	2.30	2.40	2.30
04	451	4.58	6.33	1.60	2.40	2.40
05	348	7.81	6.02	1.46	1.80	1.16
06	252	7.52	6.01	1.24	0.49	0.80
07	241	4.73	5.40	1.21	1.44	1.77
08	164	7.23	5.26	1.00	0.42	0.54
09	128	5.00	4.83	0.89	0.88	1.10
10	113	4.90	4.40	0.82	1.01	1.02
11	108	6.38	6.43	0.80	0.03	0.49
12	80	5.01	3.90	0.67	1.06	0.71
13	79	4.50	4.06	0.64	0.82	0.95
14	77	5.23	4.06	0.62	0.66	0.60
15	57	5.09	5.00	0.47	0.03	0.51
16	50	6.31	4.58	0.41	0.04	0.29
17	35	4.53	3.12	0.30	0.33	0.42
18	32	6.71	4.12	0.23	0.02	0.08
19	32	7.15	5.77	0.27	0.01	0.06
20	32	4.23	2.95	0.27	0.50	0.45
21	31	4.96	3.26	0.26	0.07	0.24
22	23	4.52	3.15	0.18	0.05	0.25
23	22	4.87	3.20	0.17	0.04	0.16
24	21	6.55	4.54	0.17	0.01	0.05
25	17	4.51	3.96	0.13	0.06	0.22
26	17	4.71	4.16	0.13	0.01	0.14
27	15	5.02	2.97	0.10	0.03	0.09
28	15	7.99	2.51	0.10	0.16	0.01
29	13	5.52	6.30	0.09	0.01	0.04
30	13	4.65	2.50	0.09	0.05	0.11
31	13	4.77	2.67	0.09	0.04	0.09
32	13	4.26	2.82	0.10	0.03	0.14
33	13	4.48	2.66	0.09	0.05	0.12

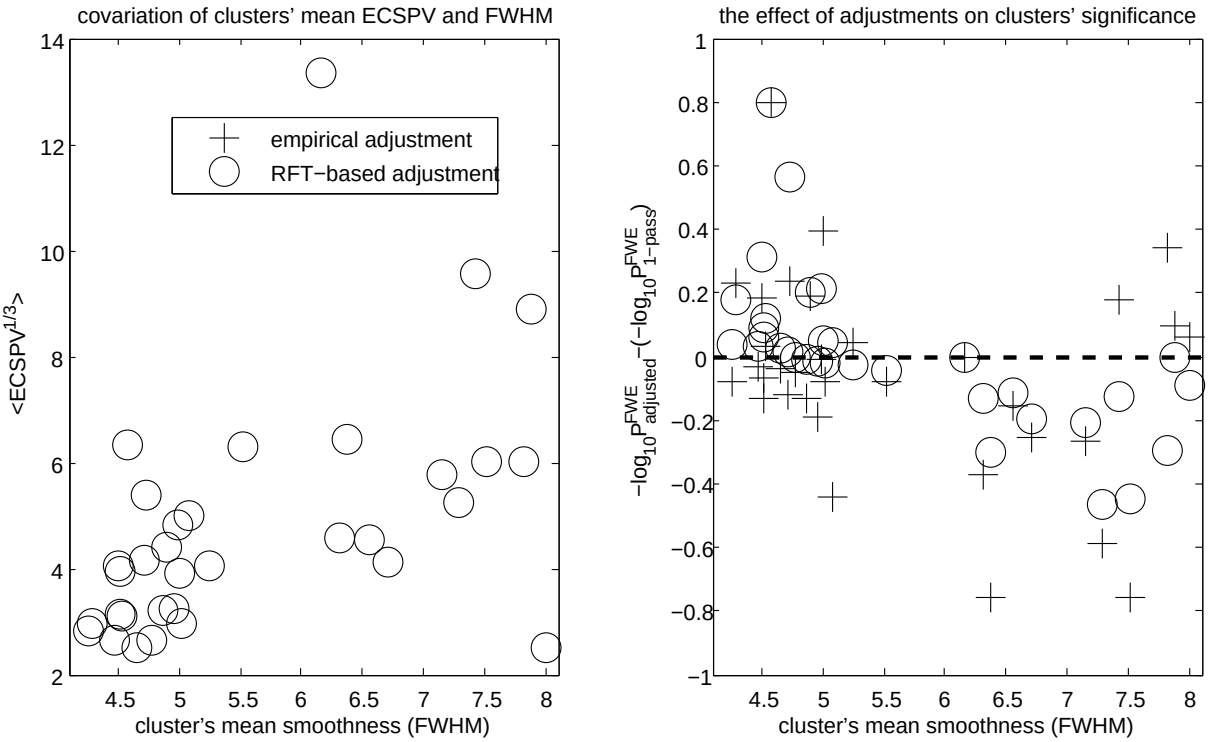


Figure 3.10: Relationship between empirical and RPV measures of cluster size for the VBM data (MCI>AD). Left shows the 33 clusters found with a $t = 3$ cluster-forming threshold, plotting ECSPV$^{1/3}$ versus FWHM (each measure averaged within a cluster). There is a roughly linear relationship, with some outliers. Right shows the relationship between the change in $-\log_{10} P^{FWE}$ (between no adjustment and adjustment with either empirical, i.e., 2-pass, or RPV-based method) and FWHM. In general, adjustment causes clusters in rough areas to increase in significance and clusters in smooth areas to decrease in significance, as would be expected.

In this work we have shown that the RFT-based method to account for nonstationarity does not completely eliminate spatial heterogeneity, and we have proposed an empirical measure that produces more uniform false positive risk over the image. The success of the RFT-based RPV approach to nonstationarity-adjustment depends on local smoothness accurately accounting for differences in cluster size over the image, and does not account for edge effects where, all things being equal, clusters near boundaries will always be smaller than clusters in the middle of the analysis volume. Our empirical ECSPV measure directly estimates local cluster size under the null hypothesis, and forms a cluster size statistic that should be uniform with respect to both smoothness and cluster position relative to the analysis boundary.

The primary finding of our work is that cluster inference with our proposed ECSPV measure has minimal heterogeneity of false positive risk over the image, as measured in a diverse collection of real and simulated null datasets (Figures. 3.2, 3.4 & 3.5). We obtained the best homogeneity when the cluster sizes were raised to the $E = 2/3$ power before averaging. We did not find any appreciable difference between Kiebel's vs. Jenkinson's method for RPV estimation.

The uniformity of our empirically-adjusted cluster-size method, however, comes at the expense of a somewhat noisier measure of local smoothness (Figure. 3.7, middle vs. bottom), and results in lower sensitivity (Figure. 3.6). TFCE, a cluster-informed voxel-wise inference method, was found to be much less affected by nonstationarity (even without any adjustment) than cluster-size-based thresholding (Figure. 3.3) and to have good power.

On the real VBM data analysis, the unadjusted method detected 5 significant clusters, the RPV-adjusted analysis detected 5, and the ECSPV-adjusted analysis detected 6 clusters. The apparent overall reduction in sensitivity (in both significant and insignificant clusters) is due to the ECSPV finding evidence of large clusters in the posterior cingulate area even under permutation. While AD declines in that region are of course typical for the disease, we find that the strength of evidence based on spatial

extent is over-stated by the RFT nonstationary method.

In conclusion, while our proposed method does not dominate in terms of power, whenever severe nonstationarity is considered we find it to be the method of choice to ensure false positive risk is optimally invariant to image smoothness. However, when compared with cluster-based inference, TFCE appears to be a very safe approach in order to minimize the effect of nonstationarity in the inference.

Chapter 4

Using Gaussian-Process Regression for Meta-analytic Neuroimaging Inference Based on Sparse Observations

Abstract

The purpose of neuroimaging meta-analysis is to localize the brain regions that are activated consistently in response to a certain intervention. As a commonly used technique, current coordinate-based meta-analyses (CBMA) of neuroimaging studies utilize relatively sparse information from published studies, typically only using (x,y,z) coordinates of the activation peaks. Such CBMA methods have several limitations. First, there is no way to jointly incorporate deactivation information when available, which has been shown to result in an inaccurate statistic image when assessing a difference contrast. Second, the scale of a kernel reflecting spatial uncertainty must be set without taking the effect size (e.g., Z-stat) into account. To address these problems, we employ Gaussian-process regression (GPR), explicitly estimating the unobserved statistic image given the sparse peak activation "coordinate" and "standardized effect-size estimate" data. In particular, our model allows estimation of effect size at each voxel, something existing CBMA methods cannot produce. Our results show that GPR outperforms existing CBMA techniques and is capable of more accurately reproducing the (usually unavailable) full-image analysis results[1].

[1]The work in this chapter is under review by IEEE Transactions on Medical Imaging.

4.1 Introduction

A statistical meta-analysis combines the results of studies that address a set of related research hypotheses, thus increasing the power and reliability of the inference (Sutton et al., 2000). It is becoming more popular in the field of neuroimaging as the number of studies in the field is increasing and many of the conducted studies either contain conflicting results or are based on only a small number of subjects. The small sample size makes the studies statistically under-powered (i.e., increases the chance that their results will not be reproduced in another group of subjects) and hence emphasizes the need for a meta-analysis.

Neuroimaging meta-analyses are either coordinate-based meta-analysis (CBMA) or image-based meta-analysis (IBMA); while existing CBMA methods are based only on activation foci in a standard space (i.e., a minimal summary of each study that can be found in journal papers), IBMA methods combine whole-brain statistic volumes. Although authors of neuroimaging meta-analyses rarely have access to the complete original datasets, when they are available, it is natural to perform an IBMA. Lazar et al. (2002) review a number of ways to combine different subjects' statistic maps, which also applies to combining different studies' maps. Among these solutions are the Fisher's method for combining P-values and Stouffer's method for combining Z-stats ($\sqrt{n}\bar{Z}$) that have been frequently used in traditional meta-analyses (Lazar et al., 2002). As an alternative to such simple fixed effects (FFX) models, we proposed a flexible hierarchical mixed effects (MFX) model in Chapter 2 to account for both within- and between-study variance; instead of modeling *all* of the data at all levels simultaneously, this model passes the summary statistics between the levels of the hierarchy (Beckmann et al., 2003; Woolrich et al., 2004).

In common practice, however, although IBMA is preferable over CBMA, neuroimaging studies rarely make public the full image data, and instead only report the magnitude and coordinates of their activation peaks in the papers; a collection of such information can

be found in databases such as BrainMap[2] (Laird et al., 2005b). With coordinate-based data, there are two widely used meta-analysis methods: activation likelihood estimation (ALE) (Turkeltaub et al., 2002) and kernel density approximation (KDA) (Wager et al., 2004); more details and a review can be found in Wager et al. (2007). In both ALE and KDA methods, the stereotactic coordinates of activation peaks are the "units" of analyses (they are treated/processed independent from each other and then pooled). In rough terms, they assess the consistency across studies by convolving an impulse at each peak activation location, combining the convolved images into test statistic images, and comparing the observed statistic images to null-hypothesis images. In KDA, the smoothing kernel is spherical with radius ρ, while in ALE it is Gaussian with standard deviation of σ.

Given the nature of the inputs to any given coordinate-based analysis, the result is not expected to perfectly resemble the IBMA results. For instance, in Chapter 2, we found a loss of sensitivity in CBMA when compared with IBMA, which is mostly due to the fact that study images are summarised by a list of sparsely-located coordinates. Also, as addressed by (Wager et al., 2007), the fact that each study employs a different approach (e.g., which of Z- or T-stat images is used; at what value the statistic image is thresholded; how many supra-threshold coordinates are reported in the paper), will influence the input and hence the result of coordinate-based analyses. In addition to such problems that any coordinate-based technique will suffer from, existing coordinate-based techniques such as ALE and KDA suffer from their strong dependency on their arbitrarily-selected kernel size. Investigating this issue, Eickhoff et al. (2009) modeled ALE's kernel size as the parameter that "should reflect the uncertainty of the reported spatial location due to between-template and between-subject variance". However, as the extent of this kernel is a property of the meta-analysis result (i.e., its smoothness), it should ideally be "inferred" from the study specifications, and spatial arrangement of the foci with respect to each other (in terms of *both* their coordinates *and* effect size).

[2]http://www.brainmap.org

When solely using "activation" foci, deactivated and neutral regions are represented similarly (i.e., both regions have no representative foci in the CBMA input), which can result in inaccurate results when inferring difference contrasts. For example, imagine a region in which studies in Group1 report effect sizes around 0 and studies in Group2 report very negative effect sizes; using full-image information is likely to result in a Group1-Group2 that is significantly bigger than zero in this region. However, there is no evidence (i.e., foci) supporting a nonzero Group1-Group2 difference when carrying out CBMA, which is the discrepancy between IBMA and CBMA in difference contrasts (shown in Chapter 2). Thus, deactivation foci are important and it would be preferable if CBMA could "jointly" incorporate the activation and deactivation information. Lastly, no matter how similar the studies included in the meta-analysis are, there is always heterogeneity in the study pool (e.g., caused by slight variations in the study design, or different numbers of subjects taking part in each study), the extent of which cannot be easily assessed by ALE and KDA, as they only use the coordinates, not the effect size.

Regarding some of the aforementioned shortcomings, Neumann et al. (2008) propose a post-ALE hierarchical clustering, which models the foci as samples from a mixture of clusters with various prior shapes located across space. A modification of the core ALE method is also introduced by Eickhoff et al. (2009), which performs a MFX by pooling the study-level activation-likelihood (AL) maps and testing the resulting map against the distribution under the H0 of "there is no consistency across individual studies' AL maps". A similar MFX version of KDA (known as multi-level KDA or MKDA) was introduced by Wager et al. (2007), where each study is represented by a binary map (with voxels being 1 if the study reports a foci in its vicinity and is 0 otherwise). This approach alleviates the inconsistency across studies in terms of the number of foci they report, but still does not address the extent of uncertainty in the meta-level effect size. In a very different approach, Costafreda et al. (2009) use a parametric CBMA by modelling the activation coordinates as random samples drawn from a Poisson point process; assuming different underlying Poisson processes for each group in this model can result in a MFX

meta-analytic solution. However, even with such modifications, CBMA would still benefit from an approach capable of *incorporating all available information* (e.g., deactivation information and foci's Z-stat) and resulting in *effect-size images*.

In this chapter, we employ Gaussian-process (GP) regression (GPR) (Rasmussen and Williams, 2006) in order to tackle these problems. GPR is very similar to Kriging (Stein, 1999), which computes the best linear unbiased estimator of the field at a target point in D-dimensional space from observations of that field at nearby locations with a stochastic model of the spatial dependence. This space is also known as feature space with points being D-dimensional feature vectors that under the field are mapped into their corresponding scalar target values. By using GPR we assume that "points close in the feature space will be close in the target space" holds for the random field (i.e., the statistic image) under consideration. Assuming that the data is generated by a multivariate Gaussian process results in the following advantages: (1) This makes the joint usage of activation and deactivation possible; (2) It enables GPR to incorporate the effect size and the location information (i.e., coordinates) rather than location alone; (3) It provides a solution for estimating the random-effects variance/heterogeneity, which provides the meta-analysis with both FFX and RFX options; (4) It enables the analysis to predict the effect size at voxels that have no observation associated with them (i.e., regression); and (5) the scale of the spatial covariance, playing an equivalent role to the kernel size in CMBA is estimated from the data instead of set arbitrarily or by a rule-of-thumb.

4.2 Materials and Methods

In this section, we first describe the full IBMA model, followed by a mathematical description of the neuroimaging CBMA. The GPR model is introduced next, for application to coordinate-based meta-analysis. Since GPR is applied to two types of data (simulated and real fMRI data), the simulation procedure and general analysis routine and data preparation before the GP meta-analysis are also described. This is followed by

a brief introduction to Dice coefficient (DC) and receiver operating characteristic (ROC) curves as performance indices.

4.2.1 Image-based Neuroimaging Meta-analysis

In order to achieve improved statistical power and more generalizable conclusions, one approach is to pool the results from a group of neuroimaging studies using IBMA. In essence, pooling the study-level statistics in IBMA is very similar to pooling the subject-level statistics in group-level analysis and hence can utilize similar models, i.e., a two-level MFX model. It was previously shown by Beckmann et al. (2003) that a "two-level MFX model with its study-level parameters being estimated from parameter and variance estimates of the subject level" can be made equivalent to a "single complete mixed-effects model whose parameters are estimated directly from all of the original single sessions' time series data" if the (co-)variance at the second level is set equal to the sum of the (co-)variances in the single-level form. This statement is generalizable to fMRI meta-analysis, i.e., IBMA *only* requires the values of the parameter estimates and their (co-)variance from each study, generalizing the well-established "summary statistics" approach to IBMA.

Consider a meta-analysis where there are S studies and that each study, s, uses a within-study analysis to estimate the effect-size at voxel k ($k = 1, ..., K$). This study-specific effect size, $y_{s,k}$, can be shown to be given by:

$$\begin{aligned} y_{s,k} &= \alpha_{s,k} + w_{s,k}, \text{ where } w_{s,k} \sim \mathcal{N}(0, \tau^2_{s,k}). \\ \alpha_{s,k} &= \mu_k + u_{s,k}, \text{ where } u_{s,k} \sim \mathcal{N}(0, \sigma^2_k), \end{aligned} \tag{4.1}$$

where μ_k is the overall population mean effect at voxel k, $\alpha_{s,k}$ denotes the effect for study s, $\tau^2_{s,k}$ represents the within-study variances, and σ^2_k is the random-effects variance (or the inter-study heterogeneity variance) (Woolrich et al., 2004). Combining the two lines in Equation 4.1 gives:

$$y_{s,k} = \mu_k + u_{s,k} + w_{s,k}, \tag{4.2}$$

which implies that IBMA estimates the meta-level mean effect size, $\{\mu_k\}_{k=1,\ldots,K}$, and between-study variance, σ_k^2, using the study-level summary statistics, $\{\{\alpha_{s,k}\}_{s=1,\ldots,S}\}_{k=1,\ldots,K}$ and $\{\{\tau_{s,k}^2\}_{s=1,\ldots,S}\}_{k=1,\ldots,K}$ (e.g., by employing a Bayesian method such as FLAME (FMRIB's Local Analysis of Mixed Effects) Woolrich et al. (2004)).

4.2.2 Coordinate-based Neuroimaging Meta-analysis

As each neuroimaging study consists of hundreds of thousands of hypothesis-tests (i.e., large K), it is almost impossible to report the result of all these tests (i.e., all voxels considered) in one paper. Therefore, in practice, neuroimaging meta-analysis uses a summary from each study, i.e., a set of coordinates corresponding to the location (and often Z-stat magnitude) of activation peaks. Therefore, the model in Equation 4.2 requires two modifications, which we will now address: (1) it needs to offer a solution for sparsity of observations, and (2) it needs to accommodate the use of standardized effect-sizes (i.e., Z-stats) instead of $\alpha_{s,k}$.

Typically CBMA does not have access to both $\alpha_{s,k}$ and $\tau_{s,k}^2$ information; instead it has access to $z_{s,k} = \alpha_{s,k}/\tau_{s,k}$. Therefore, the model in Equation 4.2 changes into

$$z_{s,k} = \mu_k/\tau_{s,k} + e_{s,k}, \text{ where } e_{s,k} \sim \mathcal{N}(0, \frac{\sigma_k^2 + \tau_{s,k}^2}{\tau_{s,k}^2}). \tag{4.3}$$

As an approximation, we assume that for every study, $\tau_{s,k} = \tau_k$ (i.e., studies are similarly reliable in their effect-size estimates), Equation 4.3 can then be rewritten as

$$z_{s,k} = \tilde{\mu}_k + e_{s,k}, \text{ where } e_{s,k} \sim \mathcal{N}(0, 1 + \tilde{\sigma}_k^2), \tag{4.4}$$

where $\tilde{\mu}_k = \mu_k/\tau_k$ and $\tilde{\sigma}_k^2 = \sigma_k^2/\tau_k^2$. This model will produce FFX inferences if we assume $\tilde{\sigma}_k^2 = 0$ (i.e., no RFX variance).

4.2.3 Using GPR for CBMA

Even though CBMA only has access to $n \times 1$ vector $\mathbf{z}=\{z_k\}_{k=1}^n$ at n sparsely-located voxels with voxel coordinates $\mathbf{V}=\{\mathbf{v}_k\}_{k=1}^n$ (i.e., an $n \times 3$ matrix), we can employ GPR to model those voxels' (unobserved) standardized mean effect size $\tilde{\boldsymbol{\mu}} = \{\tilde{\mu}_k\}_{k=1}^n$ ($n \times 1$

vector). Under GPR, $\tilde{\boldsymbol{\mu}}$ is assumed to be a sample from a Gaussian process (i.e., a distribution over functions), i.e.,

$$\tilde{\boldsymbol{\mu}} \sim \mathcal{GP}(\mathbf{m}, \mathbf{C}), \tag{4.5}$$

where $\mathbf{m}$ and $\mathbf{C}$ denote the mean and covariance matrix of the process. We set $\mathbf{m}$=0 (expressing a prior belief of no activation in the absence of foci), and use a squared exponential (SE) covariance function to model the smoothness of $\tilde{\boldsymbol{\mu}}$, i.e.,

$$\mathbf{C}(k, k') = \sigma_f^2 \exp\left(-\frac{d(\mathbf{v}_k, \mathbf{v}_{k'})^2}{2\ell^2}\right), \tag{4.6}$$

where $d(\mathbf{v}_k, \mathbf{v}'_k)$ is the Euclidian distance between kth and k'th voxel, and σ_f and ℓ denote the model's parameters for signal-variance and length-scale, respectively. Thus, $\mathbf{C}$ depends on σ_f and ℓ, and given that they are the prior distribution's parameters, in the rest of this chapter we call them our model's "hyperparameters". Figure 4.1 illustrates how the shape of the GP varies as a function of σ_f and ℓ.

Assuming that $\mathbf{z}$ is sampled from $\tilde{\boldsymbol{\mu}}$ with Gaussian noise $\mathcal{N}(0, \sigma_n^2 I)$, this results in $z_k \sim \mathcal{N}(\tilde{\mu}_k, \sigma_n^2)$ at the location of kth voxel, or $\mathbf{z} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \sigma_n^2 I)$ for the n observations. This resembles CBMA's generative model in Equation 4.4 (where $\forall k\ \sigma_n^2 = 1 + \tilde{\sigma}_k^2$), which, together with the model in Equation 4.5, provides the CBMA with a solution for predicting the $n_* \times 1$ vector of standardized effect-size $\tilde{\boldsymbol{\mu}}_*$ at a new set of n_* given voxels $\mathbf{V}_*$(i.e., $n_* \times 3$ matrix of prediction points that can have overlapping voxels with $\mathbf{V}$ as well). In the first step of this solution (inference phase), the model's hyperparameters $\Theta = \{\sigma_f, \sigma_n, \ell\}$ are estimated with "evidence optimization" (EO). In the next step (prediction phase), given the data ($\{\mathbf{V},\mathbf{z}\}$) and estimated hyperparameters $\Theta = \hat{\Theta}$, GPR uses its key predictive formula (Rasmussen and Williams, 2006)

$$p(\tilde{\boldsymbol{\mu}}_* | \mathbf{V}, \mathbf{z}, \mathbf{V}_*) = \mathcal{N}(\bar{\boldsymbol{\mu}}_*, \mathrm{cov}(\tilde{\boldsymbol{\mu}}_*)), \tag{4.7}$$

where

$$\begin{aligned} \mathbb{E}[\tilde{\boldsymbol{\mu}}_*] &= \bar{\boldsymbol{\mu}}_* = \mathbf{C}(\mathbf{V}_*, \mathbf{V})[\mathbf{C}(\mathbf{V}, \mathbf{V}) + \sigma_n^2 I]^{-1}\mathbf{z} \\ \mathrm{cov}(\tilde{\boldsymbol{\mu}}_*) &= \mathbf{C}(\mathbf{V}_*, \mathbf{V}_*) - \mathbf{C}(\mathbf{V}_*, \mathbf{V})[\mathbf{C}(\mathbf{V}, \mathbf{V}) + \sigma_n^2 I]^{-1}\mathbf{C}(\mathbf{V}, \mathbf{V}_*), \end{aligned} \tag{4.8}$$

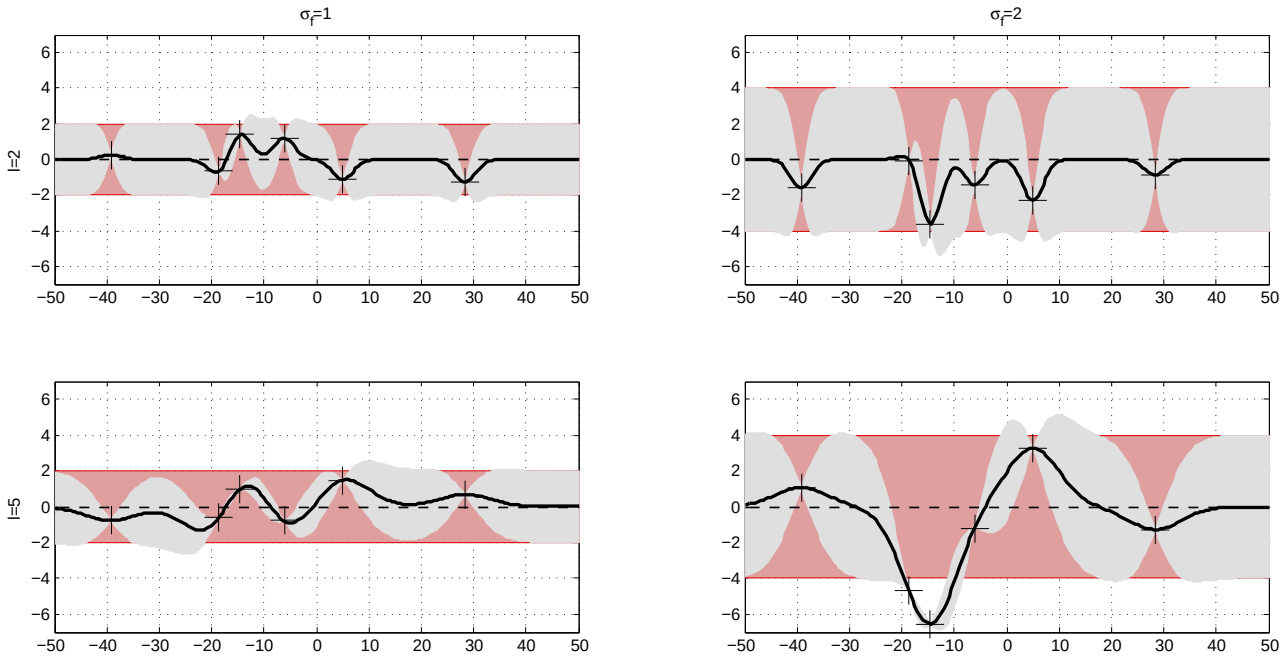


Figure 4.1: The effect of GP's hyperparameters on its shape. Data is generated from a GP with hyperparameters ℓ =2 and 5 (shown at the left side of each row), σ_f =1 and 2 (shown on top of each column), and σ_n =0 (i.e., no obervation noise), as shown by the + symbols. Using Gaussian process prediction with these hyperparameters we obtain a 95% confidence region for the underlying function ($\sim \mathcal{GP}$) (shown in grey). The red band on the background of each subplot shows the the corresponding GP if no data is given (with its mean shown as a black thin dashed line, and its 95% confidence region shown in red). This figure shows how higher ℓ corresponds to higher smoothness (bottom row), and how higher σ_f translates to stronger signal (right column). It also shows how our observation can reduce the uncertainty in our prior (i.e., the red region).

where $\mathbf{C}(\mathbf{V}_*,\mathbf{V}_*)$ is the $n_* \times n_*$ prediction covariance matrix of the GP function at n_* prediction voxels, $\mathbf{C}(\mathbf{V}_*,\mathbf{V})$ is the $n_* \times n$ matrix of covariance between n training and n_* prediction points, and $\mathbf{C}(\mathbf{V},\mathbf{V}_*)= \mathbf{C}(\mathbf{V}_*,\mathbf{V})^T$. Note that the predictive formula is the conditional normal distribution, i.e., $p(\tilde{\boldsymbol{\mu}}_*|\tilde{\boldsymbol{\mu}})$, where $\tilde{\boldsymbol{\mu}}_*$ and $\tilde{\boldsymbol{\mu}}$ jointly form a multivariate normal distribution. According to Equation 4.8, the estimated uncertainty is the difference between two terms: the first term is simply the GP prior covariance from which is subtracted a (positive) term representing the information the observations give us about the function.

4.2.3.1 Hyperparameter Estimation with EO

EO estimates the model's hyperparameters from observations $\{\mathbf{V},\mathbf{z}\}$. Evidence, or "marginal likelihood", is the integral of the likelihood times the prior, where the term marginal refers to the marginalization over $\tilde{\boldsymbol{\mu}}$ (i.e., the GP function)

$$p(\mathbf{z}|\mathbf{V}) = \int_{\tilde{\boldsymbol{\mu}}} p(\mathbf{z}|\tilde{\boldsymbol{\mu}},\mathbf{V})p(\tilde{\boldsymbol{\mu}}|\mathbf{V})d\tilde{\boldsymbol{\mu}}. \tag{4.9}$$

This is straightforward to evaluate since $\tilde{\boldsymbol{\mu}}|\mathbf{V} \sim \mathcal{N}(0, \mathbf{C}(\mathbf{V},\mathbf{V}))$

$$\log p(\tilde{\boldsymbol{\mu}}|\mathbf{V}) = -\frac{1}{2}\tilde{\boldsymbol{\mu}}^T\mathbf{C}(\mathbf{V},\mathbf{V})^{-1}\tilde{\boldsymbol{\mu}} - \frac{1}{2}\log|\mathbf{C}(\mathbf{V},\mathbf{V})| - \frac{n}{2}\log 2\pi. \tag{4.10}$$

and $\mathbf{z}|\tilde{\boldsymbol{\mu}} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \sigma_n^2 I)$, and the log marginal likelihood is

$$\log p(\mathbf{z}|\mathbf{V}) = -\frac{1}{2}\mathbf{z}^T(\mathbf{C}(\mathbf{V},\mathbf{V}) + \sigma_n^2 I)^{-1}\mathbf{z} - \frac{1}{2}\log|\mathbf{C}(\mathbf{V},\mathbf{V}) + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \tag{4.11}$$

where $\mathbf{C}(\mathbf{V},\mathbf{V})$ is the $n \times n$ covariance matrices of the GP function at n training voxels. Note that this can also be obtained directly by observing that $\mathbf{z} \sim \mathcal{N}(0, \mathbf{C}+\sigma_n^2 I)$. The EO results in the $\Theta = \{\sigma_f, \sigma_n, \ell\}$ that satisfies $\hat{\Theta} = \arg\max_{\Theta}(\log p(\mathbf{z}|\mathbf{V}))$. We employed the conjugate gradient method (an algorithm for the numerical solution of particular systems of linear equations, namely those whose matrix is symmetric and positive-definite) for solving EO's optimization problem (Avriel, 2003). As detailed below, we consider different configurations for estimating Θ, however, in practice we only advocate fixing σ_f, using a Gamma prior on ℓ and estimating σ_n and ℓ by using EO.

4.2.3.2 Prior on length-scale hyperparameter ℓ

To improve the accuracy/plausiblity of the estimation of ℓ, one can add prior information on the scale (ℓ) with a Gamma prior distribution

$$\ell \sim \text{Gamma}(\ell;\ \eta,\ \nu) = \frac{1}{\nu^{\eta}\Gamma(\eta)} x^{\eta-1} e^{-\ell/\nu}, \tag{4.12}$$

with mean $\eta\nu$ and variance $\eta\nu^2$. We consider the use of a mean 1 voxel, variance 5 voxel^2 prior, denoted as $\Gamma(1, 5)$, for all our analyses (note that our real fMRI image is 2 mm$\times$2 mm$\times$2 mm). We must choose and fix the prior mean and variance, but our results should be less sensitive to these choices relative to if we had just fixed ℓ to a particular value. Note that, placing a prior on ℓ helps to regularize its estimation to a plausible range. For example, in a CBMA we know that the smoothness extent is normally from a limited range, i.e., values from outside this range are not likely to be the smoothness of an FMRI data (e.g., we will not have ℓ=100 mm in reality).

4.2.3.3 Fixing σ_f^2 to Account for Coordinate Sampling Bias

The basic GPR model as described so far is suitable for problems where $\mathbf{z}$ are *random* samples drawn from a Gaussian random field plus a noise. In a CBMA application, however, samples are not randomly drawn from Z-stat images; rather they are the location of the peaks surviving a threshold, e.g., Z-stat>3. Therefore, it is necessary to reflect this in the model's hyperparameters for a more reliable/accurate inference and prediction. Instead of using EO to estimate $\hat{\sigma_f}$ based on the characteristics of the data, we fix it to force the desired behavior of $\tilde{\boldsymbol{\mu}}_*$, when away from sampled points $\mathbf{V}$.

When there is no sample z_k in the local vicinity of voxel k, the GP will tend towards a voxel-wise marginal distribution such as $\tilde{\boldsymbol{\mu}}_* \sim \mathcal{N}(0, \sigma_f^2 I)$. However, depending on the values of σ_f, σ_n, ℓ, even if k is distant from sampled points $\mathbf{V}$, the predicted value of $\tilde{\boldsymbol{\mu}}_*$ may be far from zero (as determined by Equation 4.8). To ensure that predictions in such regions stay near zero away from sampled data, we fix σ_f (Groves et al., 2009).

Consider a prediction $\tilde{\mu}_*$ at a randomly selected location, with no influence from sampled data, i.e., $\tilde{\mu}_* \sim \mathcal{N}(0, \sigma_f^2 I)$. In this case, if we assume the null hypothesis is true

Table 4.1: The list of analyses carried out in this chapter in terms of their underlying model, inputs and outputs. Note that the $*$ in the second subscript (e.g., such as the one in $\text{GPR}_{\text{joint},*}$) can denote different possibilities (i.e., s, l and n) shown in Table 4.2.

Notation	Model	Input	Output
$\text{ALE}_{\text{activation}}$	ALE	(x,y,z) of activation foci	Positive Z-stat map
$\text{ALE}_{\text{deactivation}}$	ALE	(x,y,z) of deactivation foci	Negative Z-stat map
$\text{KDA}_{\text{activation}}$	KDA	(x,y,z) of activation foci	Positive Z-stat map
$\text{KDA}_{\text{deactivation}}$	KDA	(x,y,z) of deactivation foci	Negative Z-stat map
$\text{GPR}_{\text{activation},*}$	GPR	(x,y,z) and Z-stat of activation foci	Full Z-stat map
$\text{GPR}_{\text{deactivation},*}$	GPR	(x,y,z) and Z-stat of deactivation foci	Full Z-stat map
$\text{GPR}_{\text{joint},*}$	GPR	(x,y,z) and Z-stat of activation and deactivation foci	Full Z-stat map
FLAME-MFX	GPR	voxel-wise COPE and VARCOPE	Full Z-stat map
FLAME-MFX	GPR	voxel-wise COPE and VARCOPE	Full Z-stat map

then $\sigma_f = 1$ ($\tilde{\sigma}^2 = 0$, Equation 4.4). However, as we are working with peaks, even null peak data will have variance in excess of unity. Further, since we believe many studies to be under-powered, there are false negative foci that should have been detected (i.e., true $\tilde{\mu}_* \geq t$, for analysis threshold t) but are not ($z = \tilde{\mu}_* + e < t$). Thus we choose $\sigma_f > 1$ to reflect the omitted foci. We investigate the sensitivity of our choice of σ_f below.

4.2.4 CBMA Notations

It is important to note the differences between various methods we employed for coordinate- and image-based meta-analysis in this chapter. A summary of all these differences (in terms of their underlying model, inputs and outputs) can be found in Table 4.1. We addressed the problem that arises when comparing a "full Z-stat map" such as $\text{GPR}_{\text{joint},*}$ results with uni-sign maps such as $\text{ALE}_{\text{activation}}$ and $\text{ALE}_{\text{deactivation}}$ results, by comparing positive $\text{ALE}_{\text{activation}}$ with positive $\text{GPR}_{\text{joint},*}$ and negative $\text{ALE}_{\text{deactivation}}$ with negative $\text{GPR}_{\text{joint},*}$.

Table 4.2: The list of GPR inference scenarios utilized in this chapter. In all these scenarios, evidence optimization is employed and the $*$ notation in the subscript can be any one of activation, deactivation or joint cases. For instance, $\text{GPR}_{\text{joint},s}$ denotes a case that is provided with both activation and deactivation foci, where σ_f=3 and a Gamma prior on ℓ exists, and σ_n and ℓ are inferred using evidence optimization.

Notation	fixed	prior	EO infers
$\text{GPR}_{*,\text{s}}$ (fixed σ_f)	σ_f=3	Gamma on ℓ	ℓ and σ_n
$\text{GPR}_{*,\text{l}}$ (fixed ℓ)	ℓ	-	σ_f and σ_n
$\text{GPR}_{*,\text{n}}$ (no fixed hyperparameter)	-	-	ℓ, σ_f and σ_n

4.2.4.1 GPR Notations

Given various possibilities for the GPR analysis, e.g., adding priors on hyperparameters and/or fixing hyperparameters' values, we adopted a subscript-representation (shown in Table 4.2) that denotes each case. In the rest of the chapter, we will use these notations.

4.2.5 Data

In this study, meta-analytic methods are applied to contrast of parameter estimates (COPEs) and their variances (VARCOPEs) that are either simulated or are from real fMRI experiments.

4.2.5.1 Simulated Data

In order to compare the methods in a scenario where the underlying ground truth is known, using Algorithm 1 a group of subject-level COPE and VARCOPE images are simulated (i.e., the end-result of the simulation is subject-level COPEs and VARCOPEs, not the functional time-series). Simulation starts with a binary image (with 1s at the location of foci in the list $\mathcal{F}$ and 0s elsewhere) that if smoothed will result in a "meta-level gold standard" signal (i.e., COPE), $\mathcal{G}$. Assuming that each study comes from its own meta-level underlying signal $\mathcal{G}_s$ (i.e., from different populations due to different interventions and designs that each study employs), we next move the foci in $\mathcal{F}$ slightly,

in order to generate $\mathcal{F}_s$. Smoothing the binary image corresponding to $\mathcal{F}_s$ results in $\mathcal{G}_s$, meta-level COPE corresponding to sth study.

In the 1D simulation, $\mathcal{F}$ consists of 12 foci located at coordinates {55, 120, 200, 250, 400, 470, 550, 600, 660, 770, 810, 900}, representing the centres of blobs of σ={12, 15, 8, 13, 9, 10, 12, 12, 9, 12, 11, 15} voxels. In the 3D simulation, $\mathcal{F}$ consists of 12 foci located at coordinates {(20,20,5), (7,7,7), (14,14,14), (20,20,20), (30,30,30), (5,25,35), (10,10,30), (10,30,10), (30,8,28), (10,40,25), (22,11,33), (30,25,15)}, representing the centres of blobs of $\sigma = 3$ voxels. It is worth noting that, given ALE's Gaussian and GP's SE kernels' nature, such simulation might slightly favour them, when compared to KDA, that uses an indicator function as its kernel. However, given the high number of foci that are included in the CBMA, KDA's statistic image will end up having a smoother shape, i.e., representing a Gaussian-blob-type statistic image, which can be an indication of this issue being less important in CBMA of large number of input foci.

Using $\mathcal{G}_s$ as the meta-level COPE in the GLM (described in Section 4.2.1) results in the study-level COPE images, $\mathcal{Y}_s$. We assume that the number of subjects in each study is a random number between 10 and 20 (with uniform probability). For the ith subject of study s, COPE image $\mathcal{Y}_{s,i}$ is generated by feeding the GLM with $\mathcal{Y}_s$ while using the parameters in Table 4.3. Such univariate (i.e., voxel-wise) hierarchical GLM is utilized in many neuroimaging (Beckmann et al., 2003) and general meta-analysis (Copas and Shi, 2001) papers. In order to carry out a study-level analysis on subject-level COPE images, we next simulated VARCOPE images by drawing random samples from $\mathcal{N}(0,1)$ at each voxel (i.e., $\mathcal{Y}_{s,i}(v) \pm \eta_{s,i}(v)$, where $\eta_{s,i}(v) \sim \mathcal{N}(0,1)$).

IBMA of simulated images, i.e., subject-level COPEs and VARCOPEs, is carried out using FSL (Smith et al., 2001). In the second-level analyses, each study's MFX map, corresponding to the average of its subjects' activation maps was created using FLAME (FMRIB's Local Analysis of Mixed Effects) (Woolrich et al., 2004). Third-level analyses were carried out using all studies, with a one- or two-group model (depending on the case) in order to create FLAME-MFX and FLAME-FFX maps. For the CBMA, local maxima

Algorithm 1 Hierarchical Data Simulation

Require: $\mathcal{S},\ \mathcal{F},\ \delta,\ \Sigma,\ \mathcal{V},\ var_{ma}^{hetro},\ var_{ma}^{FFX},\ var_{st}^{hetro},\ var_{st}^{FFX}$

```
for s = 1 to S do
    F_s ← ∅
    for f=1 to length( F) do
        f_new,x ← F(f, x) + random_number_from([-δ δ])
        f_new,y ← F(f, y) + random_number_from([-δ δ])
        f_new,z ← F(f, z) + random_number_from([-δ δ])
        add( F_s, f_new,x, f_new,y, f_new,z)
    end for
    G_s ← 0
    for f = 1 to length(F) do
        T ← 0
        T(F_s(f)) ← 1
        T ← smooth(T, Σ(f))
        T ← T/max( T)
        T ← zero_values_under_0.2(T)
        G_s ← G_s + T
    end for
    I_s ← random_integer_from([10 20])
    for v=1 to length(V) do
        Y_s(v) ~ N(G_s(v), {var_ma^hetro(v), var_ma^FFX(v)})
        for i=1 to I_s do
            Y_s,i(v) ~ N(Y_s(v), {var_st^hetro(v), var_st^FFX(v)})
        end for
    end for
end for
```

Table 4.3: The list of parameters used in 1D and 3D data simulation (see Algorithm 1 for more details). Columns 'Value I', 'Value II', 'Value III' and 'Value IV' show the parameter values in one-group 1D, two-group 1D, one-group 3D and two-group 3D simulations, respectively (see Section 4.3.1 and 4.3.2 for more details).

Notation	Description	Value I	Value II	Value III	Value IV
$\mathcal{S}$	Number of studies	20	20	20	20
$\mathcal{F}$	Coordinates of the Gaussian blobs' centres (gold standard)	-	-	-	-
$\mathcal{F}_s$	Coordinates of the Gaussian blobs' centres for study s (slightly displaced $\mathcal{F}$)	-	-	-	-
$\mathcal{I}_s$	Number of subjects in study s	-	-	-	-
$\mathcal{Y}_s$	COPE image of study s	-	-	-	-
$\mathcal{Y}_{s,i}$	COPE image for subject i in study s	-	-	-	-
$\mathcal{T}$	Temporary image	-	-	-	-
δ	Displacement of the coordinates	5	5	5	5
Σ	SD of the Gaussian blobs	-	-	-	-
$\mathcal{V}$	List of voxels	mask	mask	mask	mask
var_{ma}^{hetro}	Meta-level RFX variance	0	0	0	0
var_{ma}^{FFX}	Meta-level FFX variance	2	2	2	2
var_{st}^{hetro}	Study-level RFX variance	10	10	10	10
var_{st}^{FFX}	Study-level FFX variance	5	5	5	5

and minima along with their Z-stats (positive for maxima and negative for minima) are extracted from study-level FLAME-MFX Z-stat maps. A constraint is imposed to 3D (resp. 1D) cases: only report the local maxima that are not closer than 10 (20) voxels to each other and are in clusters bigger than 25 (20) voxels that survive a cluster-forming threshold of Z-stat=$\pm$2.5. These constraints are to make the simulated scenario more similar to the way papers typically report their foci in neuroimaging, e.g., reported foci of each cluster are not closer than 8 mm when using SPM[3] reports.

4.2.5.2 Real Data

For the real data, the results of 20 different pain studies (collected in accordance with local ethics approval) are pooled in order to find regions of activation induced by painful stimuli. In total, 212 healthy adult subjects were imaged (age range 20-35 y, mean 26.8 y; 127 male, 85 female) in either a 3T Siemens Trio MRI scanner (using a 12-channel

[3]http://www.fil.ion.ucl.ac.uk/spm/

head coil), or 3T Varian MRI scanner (using a 4-channel head coil). Note that the 15 studies employed in Chapter 2 is a subset of these 20 studies. This data In spite of some differences, all studies concentrated on pain as the main effect of interest (the same studies as those described in Chapter 2). Eight of the studies used a mechanical pain stimulus, while the other 12 studies used a thermal pain stimulus. We investigate a differential response to the two forms of pain delivery in the third-level analysis. The result of this analysis extracts the areas of the brain that show more or less thermally-induced pain activation relative to mechanically-induced pain.

Processing of functional images at the first level was performed using FSL (Smith et al., 2001). Functional images were motion corrected (Jenkinson et al., 2002) and spatially smoothed at FWHM (full width half maximum) of 5 mm prior to temporal model fitting (Beckmann et al., 2003; Woolrich et al., 2004, 2001). Co-registration to the MNI152 standard brain space was performed in 2 stages: (1) the fMRI data from a given subject was registered to that subject's T1 structural using linear registration (Jenkinson and Smith, 2001; Jenkinson et al., 2002) and (2) the subject's structural image was registered to the MNI standard brain using nonlinear registration. In the second-level analyses (Woolrich et al., 2004) FLAME-MFX activation maps corresponding to the main pain-effect were created. Third-level cross-study analyses were carried out using all studies, with a one-group model or a two-group model (split by mechanical vs. thermal stimulus study type). Both FLAME-FFX and FLAME-MFX maps were created at the third level.

Using the results of the 20 pain studies, the foci lists are created for the "standard" CBMA analyses; again, one activation (i.e., local maxima) list and one deactivation (i.e., local minima) list. For each study, this foci-list consists of foci not closer than 8 mm to each other (which matches the default behaviour of SPM) that belong to clusters surviving the cluster-forming threshold of Z-stat=$\pm$2.5.

4.2.6 Map Comparison

While the gold-standard reference for the simulated data is available, for the real fMRI data we use the FLAME-MFX IBMA results to define the "gold-standard" reference to which the other methods are compared. This choice of reference is based on three assumptions: IBMA is preferred over CBMA (if the relevant information is available), as the image data are a superset of the information in CBMA analyses; MFX is preferred over FFX, as the goal of the meta-analysis is to estimate the effect size consistently reported in the study pool, which might consist of different populations (particularly when the pool gets larger in terms of the number of studies); and, for the choice of IBMA analysis method, FLAME's hierarchical model is preferred over other traditional meta-analystic measures due to its statistical optimality and flexibility for dealing with group differences and covariates (Woolrich et al., 2004).

We first compare CBMA maps to the IBMA gold-standard with the Dice Coefficient (DC) (Dice, 1945), which is a symmetric measure of the similarity of two binary images:

$$DC = \frac{2|I \cap C|}{|I| + |C|} \tag{4.13}$$

where $|I|$ and $|C|$ are the number of non-zero voxels in a thresholded reference image and a thresholded CBMA image, and $|I \cap C|$ is the number of non-zero voxels in their intersection. DC ranges from 0 (no overlap), to 1 (perfect overlap). CBMA methods are compared to the reference image after thresholding, e.g., binarisation at Z-stat=$\pm$2.5. CBMA methods are tested over a range of kernel parameters, i.e., the Gaussian kernel's standard deviation (σ) for ALE and indicator kernel's radius (ρ) for KDA, in order to find the optimal setting for each method given the data. In fMRI meta-analysis, σ values are {5, 10, 15, 20, 25, 30}mm and ρ values are {15, 20, 25, 30, 35, 40}mm; for simulated 3D data these values are {1,2,4,6,8,10}voxels and {2,4,6,8,10,12}voxels for σ and ρ, respectively.

In order to summarize each method's performance on a range of possible thresholds, we use a receiver-operator characteristic (ROC) curve, where the free parameter is the

binarization threshold. We use the area under the curve (AUC) as a single-valued summary of an ROC curve; the higher the AUC, the better. An ROC curve is strictly defined only for a single detection task, repeated many times with different thresholds. In our setting, following Smith and Nichols (2009), with thousands of hypothesis-tests (one per voxel) a free-response (FR) ROC curve is more appropriate (Bunch et al., 1978). An FR-ROC curve replaces either FPR or TPR (or both) with measures that aggregate true or false positives over voxels. At a given threshold, we define TPR as "the number of voxels that are above the threshold in both CBMA and our gold standard image (e.g., IBMA) divided by the total number of voxels that are above the threshold in our gold standard image" and FPR as "the number of voxels that are incorrectly above the threshold in CBMA divided by the total number of voxels that are below the threshold in our gold standard image".

4.3 Results

Given various data types utilized for the assessment of GPR CBMA, in this section, each subsection is dedicated to one type of input data (e.g., 1D simulated, 3D simulated and real fMRI data).

4.3.1 Simulated Data: 1D

We first illustrate the results when meta-analyzing 20 simulated 1D studies in Figure 4.2. The one-group 1D simulation parameters are shown in column 'Value I' of Table 4.3. As described earlier, meta-analysis of the simulated data starts from pooling the subject-level COPEs and VARCOPEs of each study, which results in "FLAME-MFX study-level" COPEs, VARCOPEs and Z-stat images. There is *one* underlying signal for the meta-analysis ($\mathcal{G}$) whose variants are the underlying signals of the populations that each study is sampled from ($\mathcal{G}_s$).

Figure 4.3 shows a case where GPR accommodates our prior belief on its hyperparameters' values ($\text{GPR}_{\text{joint,l}}$ with ℓ=25 voxels), together with a case where GPR

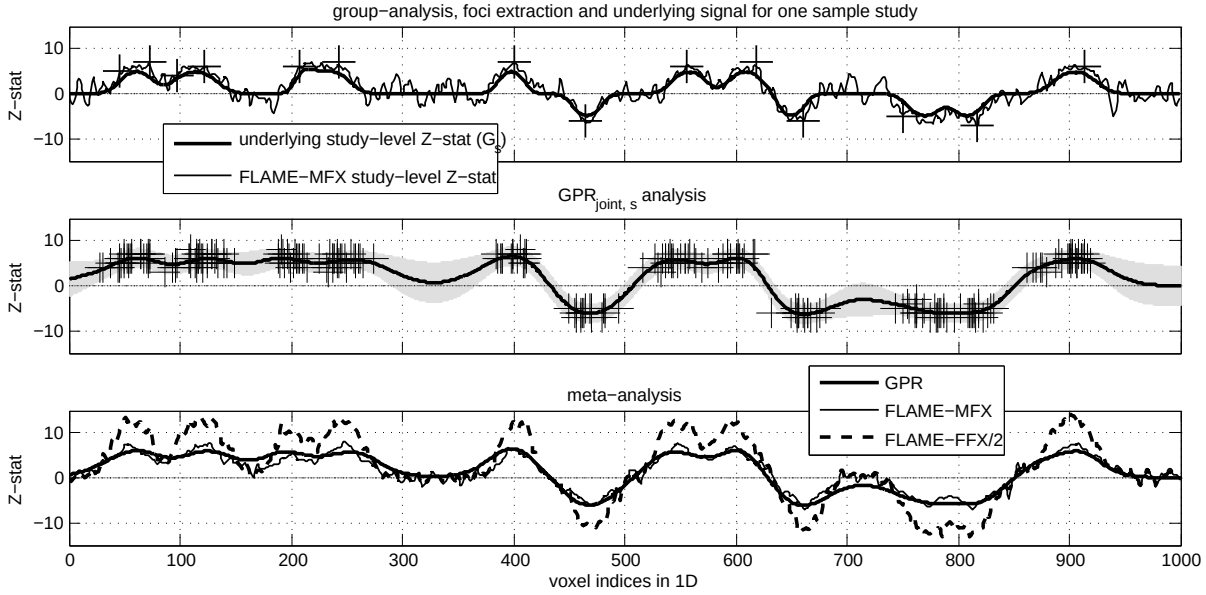


Figure 4.2: Comparing the $\text{GPR}_{\text{joint,s}}$ with $\Gamma(1,5)$ prior on ℓ (EO inferred ℓ=25.6 voxels and σ_n=1) with FLAME when applied to 1D simulated data. In the top row, one sample study's underlying signal (the bold line) is shown together with its FLAME-MFX Z-stats (the thin line). The foci extracted from this FLAME-MFX study-level Z-stat image (i.e., the '+' markers) are collected for $\text{GPR}_{\text{joint,s}}$ whose result (i.e., mean±SD) is shown in the middle row (notice the shading corresponding to the estimation errors, i.e., 90% confidence interval). In this row, +'s represent the foci from the 20 studies included in the CBMA. The bottom row displays the Z-stat images of FLAME-MFX and FLAME-FFX together with the Z-stat resulting from $\text{GPR}_{\text{joint,s}}$. The goodness of this fit is assessed by using the coefficient of determination or R^2, which is 91%, i.e., 91% of the signal variance is explained by the $\text{GPR}_{\text{joint,s}}$ which indicates a very accurate result. Note that the FLAME-FFX results are scaled by a 1/2 factor for the purpose of visualization.

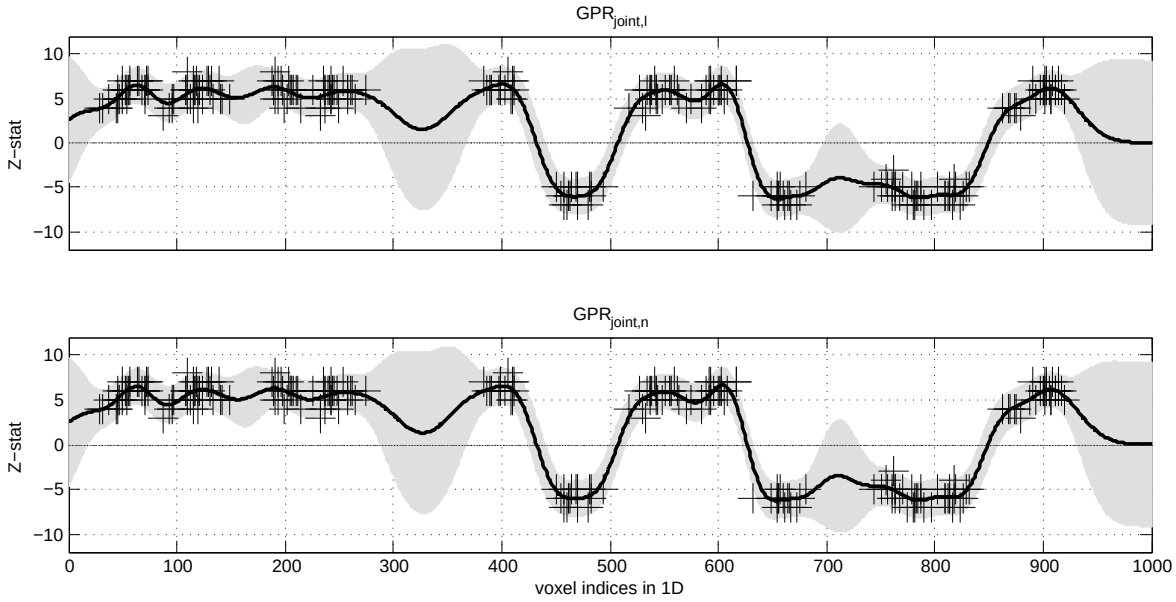


Figure 4.3: GPR's flexibility in accommodating the prior knowledge on ℓ by fixing its value or assuming a probability distribution for it. This figure shows $\text{GPR}_{\text{joint,l}}$ when ℓ=25, and $\text{GPR}_{\text{joint,n}}$ in the first and second rows, respectively. The foci are collected from the same 20 simulated studies described in Figure 4.2 and the shadings correspond to the estimation errors, i.e., 90% confidence interval.

does not incorporate any prior information (i.e., $\text{GPR}_{\text{joint,n}}$). Figure 4.4 on the other hand, displays the variation in GPR in terms of the types of its input by showing the results for $\text{GPR}_{\text{joint,s}}$, $\text{GPR}_{\text{activation,s}}$ and $\text{GPR}_{\text{deactivation,s}}$, all with a $\Gamma(1,5)$ prior on ℓ. Note that, in this chapter we advocate the use of $\text{GPR}_{*,\text{s}}$ and other alternatives are just demonstrated for comparison.

One of the strengths of GPR is its ability to incorporate the deactivation information and hence its better estimation of the underlying statistical landscape in difference contrasts. Figure 4.5 displays the difference-contrast meta-analysis (e.g., similar to comparing two groups of studies: one assessing an intervention on subjects having a Disease and one on Control subjects using the Disease-Control contrast) where two main groups of 1D studies are simulated in order to produce a signal for the difference contrast that is different from zero. The two-group 1D simulation parameters are shown in column 'Value II' of Table 4.3. The results can resemble the results from IBMA methods

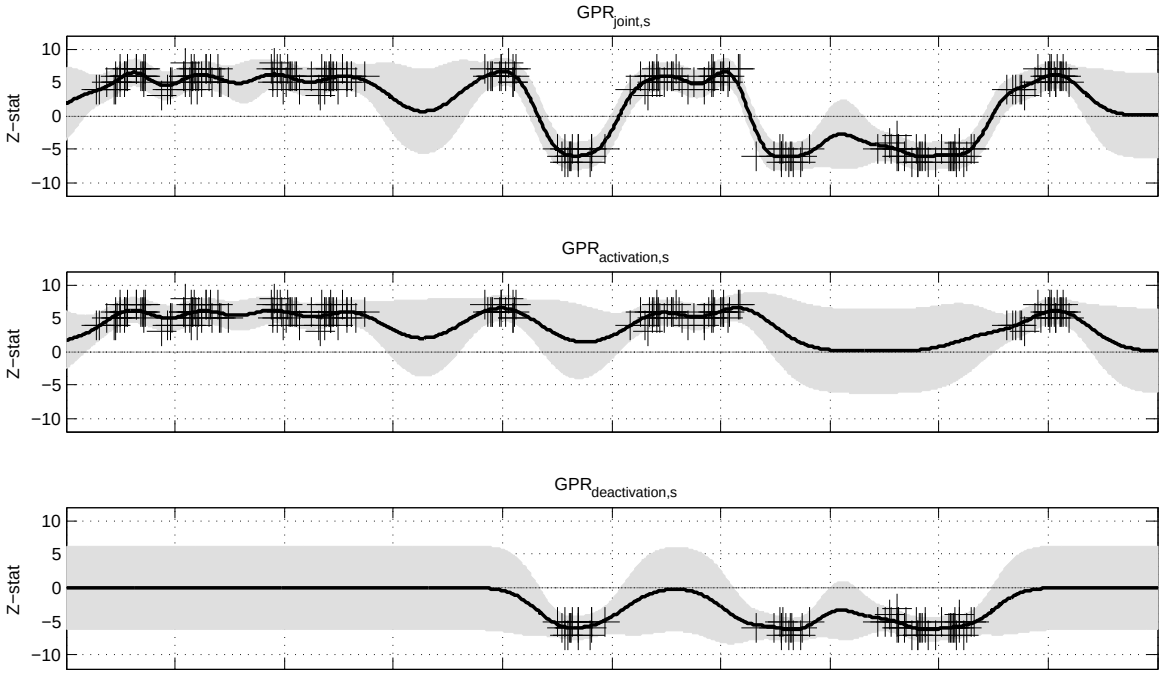


Figure 4.4: GPR's flexibility in incorporating various foci types (e.g., activation only, deactivation only or joint activation and deactivation). In rows 1 to 3 we have $\text{GPR}_{\text{joint,s}}$, $\text{GPR}_{\text{activation,s}}$ and $\text{GPR}_{\text{deactivation,s}}$ with a $\Gamma(1{,}5)$ prior on ℓ. If there are no foci at the location of a voxel, the prior on σ_f pushes the estimated landscape toward an $\mathcal{N}(0{,}\sigma_f^2)$ distribution. The foci are collected from the same 20 simulated studies described in Figure 4.2 and the shadings correspond to the estimation errors, i.e., 90% confidence interval.

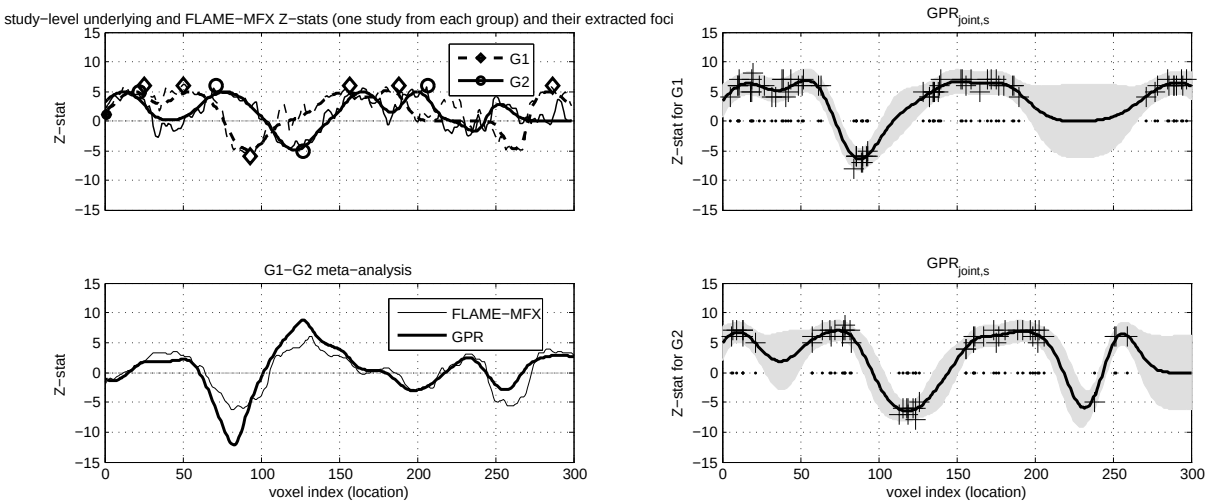


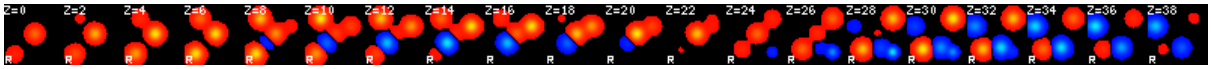
Figure 4.5: The results of $\text{GPR}_{\text{joint,s}}$ with $\Gamma(1,5)$ prior on ℓ when applied to two-group simulated 1D data in order to localize their difference. Top-left displays one sample study from each group whose underlying signals and estimated study-level Z-stats are shown by bold and thin lines, respectively. The foci extracted from the study-level Z-stat images are collected for GPR-CBMA whose results are shown in the right column (top and bottom for G1 and G2, respectively); the shading corresponding to the estimation errors, i.e., 90% confidence interval and +'s represent the foci included in each group's CBMA. Bottom-left displays the Z-stat images of FLAME-MFX together with the Z-stat resulting from $\text{GPR}_{\text{joint,s}}$.

very accurately. Note that the difference Z-stat corresponds to a T-test for the differences between two means (using each group's mean and SD as calculated in Equation 4.8).

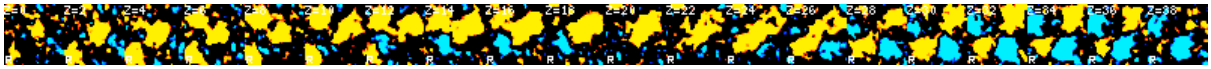
The activation in the meta-analytic result at around voxel 125 corresponds to the deactivation foci reported consistently by studies in G2 in contrast to no foci reported by studies in G1. Traditional CBMA using activation foci only would miss this effect completely.

4.3.2 Simulated Data: 3D

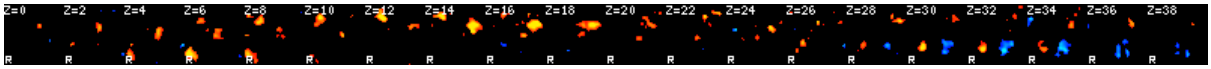
The one-group and two-group 3D simulation parameters are shown in column 'Value III' and 'Value IV' of Table 4.3, respectively. The ground truth signal and the results from various meta-analysis techniques for the 3D simulations can be seen in Figures 4.6 and 4.7. According to these results, GPR performs similarly to the IBMA methods such as FLAME-FFX.



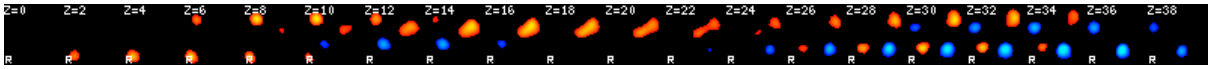
(a) One-group case's gold standard



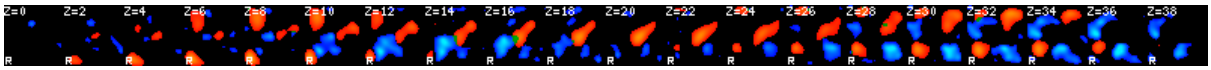
(b) FLAME-FFX



(c) FLAME-MFX



(d) $\text{GPR}_{\text{joint,s}}$

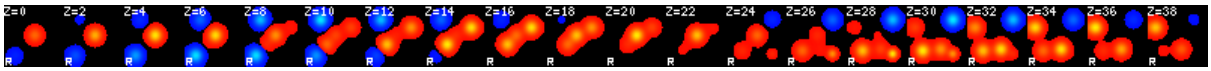


(e) $\text{GPR}_{\text{activation,s}}$ (red) & $\text{GPR}_{\text{deactivation,s}}$ (blue)

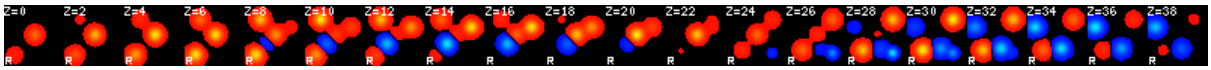
(f) $\text{KDA}_{\text{activation}}$ (red) & $\text{KDA}_{\text{deactivation}}$ (blue)

(g) $\text{ALE}_{\text{activation}}$ (red) & $\text{ALE}_{\text{deactivation}}$ (blue)

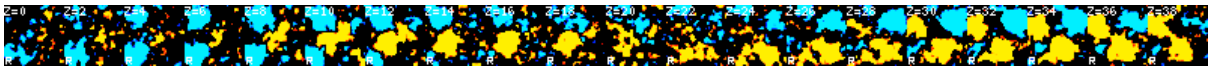
Figure 4.6: IBMA and CBMA results when pooling a one-group set of 3D simulated studies. The underlying signal for the one-group simulation is shown in (a) with the results from pooling these studies under a mean/average contrast using FLAME-FFX, FLAME-MFX, $\text{GPR}_{\text{joint,s}}$ with $\Gamma(1,5)$ prior on ℓ, $\text{GPR}_{\text{activation,s}}$ & $\text{GPR}_{\text{deactivation,s}}$ with $\Gamma(1,5)$ prior on ℓ, $\text{ALE}_{\text{activation}}$ & $\text{ALE}_{\text{deactivation}}$ with σ=4 voxels, and $\text{KDA}_{\text{activation}}$ & $\text{KDA}_{\text{deactivation}}$ with ρ=6 voxels shown in (b)-(g), respectively. Note that the extent of the ALE and KDA are those at which they have their best performance when compared with the gold standard (see Figure 4.8). In this figure, red-yellow and blue colours show Z-stat values with range [2, 4] and [-2, -4], respectively.



(a) Group 1 gold standard

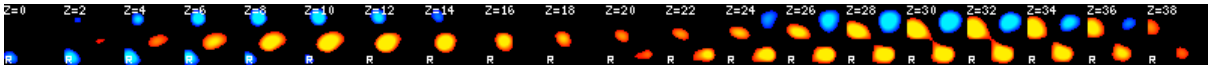


(b) Group 2 gold standard

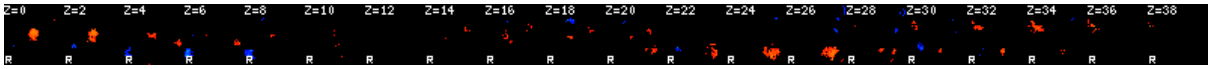


(c) FLAME-FFX: G1-G2

(d) FLAME-MFX: G1-G2



(e) $\text{GPR}_{\text{joint,s}}$: G1-G2



(f) $\text{KDA}_{\text{activation}}$: G1-G2

(g) $\text{ALE}_{\text{activation}}$: G1-G2

Figure 4.7: IBMA and CBMA results when contrasting two groups of 3D simulated studies against each other. The underlying signals for G1 and G2 are shown in (a) and (b), respectively. The results for G1-G2 contrast from FLAME-FFX and FLAME-MFX are shown in (c) and (d), respectively. The CBMA results are for $\text{GPR}_{\text{joint,s}}$ with $\Gamma(1,5)$ prior on ℓ (e), $\text{KDA}_{\text{activation}}$ with ρ=6 voxels (f) and $\text{ALE}_{\text{activation}}$ with σ=4 voxels (g). In this figure, red-yellow and blue colors show Z-stat values with range [2, 4] and [-2, -4], respectively.

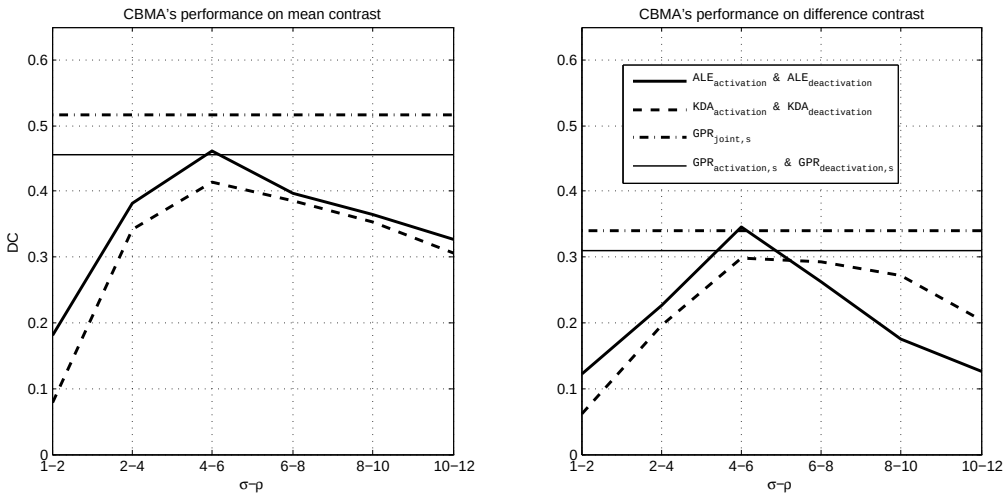
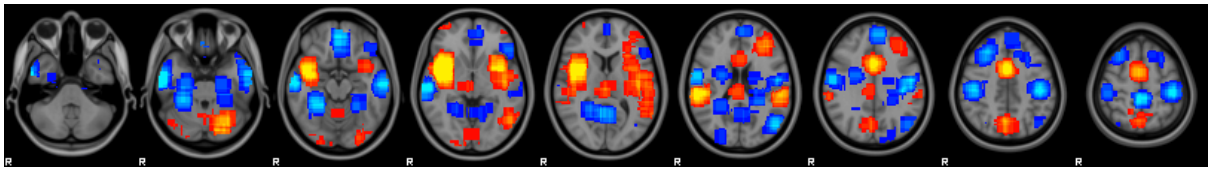


Figure 4.8: Using DC for evaluating the performance of CBMA methods when applied to the simulated 3D data (dash-separated σ and ρ on the x-axis corresponds to ALE and KDA's kernel sizes, respectively). The left panel compares ALE and KDA against our advocated method (i.e., $\text{GPR}_{\text{joint,s}}$ with $\Gamma(1,5)$ prior on ℓ) and $\text{GPR}_{\text{activation,s}}$ & $\text{GPR}_{\text{deactivation,s}}$ with a $\Gamma(1,5)$ prior on ℓ, when applied to mean (left panel) and difference (right panel) contrasts. The result indicates the $\text{GPR}_{\text{joint,s}}$ performs better than ALE and KDA over a range of their kernel sizes.

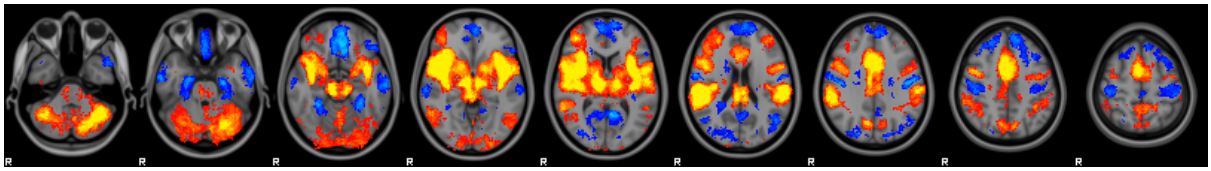
We compare CBMA results with the gold standard using the DC as the measure of overlap. For ALE and KDA, as there is not a best kernel-extent that is estimated by taking into account *both* coordinates and effect-sizes, a fairly large range of possible values are tested (which will include the value proposed in Eickhoff et al. (2009) as the optimal extent of uncertainty of coordinates). The results from the DC analysis are shown in Figure 4.8, where ALE and KDA are compared with $\text{GPR}_{\text{joint,s}}$, the method we advocate for neuroimaging CBMA.

4.3.3 fMRI Data

We first illustrate how foci from the study pool are spatially distributed in Figure 4.9; it also displays how frequently each voxel is reported as (de)active. The foci-map is the sum of the foci, signed (i.e., + for activation and − for deactivation foci) binary maps; one map per foci with +1/-1s at voxels closer than 8 mm to the focus if activation/deactivation and 0s otherwise (e.g., a value of 10 in the foci map means that there are 10 more



(a) Indicator map ($\rho = 8mm$)



(b) Frequency map of the activation/deactivation in the study pool

Figure 4.9: The foci- and frequency-maps from the fMRI study pool, with red and blue spots displaying the positive and negative areas, respectively. This figure is useful for displaying the spatial distribution of foci and the agreement between studies in each area's activation. The indicator and frequency maps are both thresholded at ±(2-10). Red-yellow and blue colours show Z-stat values with range [2, 15] and [-2, -15], respectively (with green regions being their overlap), and the displayed slices are selected from z=-40 mm to z=40 mm, every 12 mm in MNI coordinates.

activation foci than deactivation foci in the 8 mm radius of that voxel). Note that this is the subtraction of the deactivation foci's KDA statistic-image from the activation foci's KDA statistic-image. In order to generate the frequency map, $\sum_{study}[I(z_{study} > 2.5) - I(z_{study} < -2.5]$ (where I(.) returns 1 when its input argument is true and 0 otherwise) is estimated at each voxel (e.g., a value of 5 in the frequency map means that there are 5 more studies that report that voxel as active than studies that report that voxel as deactive). These two images are *only* useful for seeing how foci are scattered in the brain.

As CBMA incorporates spatially-sparse samples with similar target values (e.g., 2.5< $|y_i|$ <5), EO estimates large ℓ values, particularly if samples are only activation foci. This is due to the fact that according to the observations, target values of sample feature vectors are very similar even if they are not spatially close, which in theory is expected to happen under large ℓ. Thus, we require a prior belief of ℓ to blance this phenomenon. In Figure 4.10, each point on the x-axis (i.e., 300, 350, 400, 450 and 500) denote the number of foci selected from the fMRI foci pool, and the y-axis is the mean±SD of estimated ℓ. For instance, the fMRI foci pool consists of 530 foci, we randomly select 300 foci from

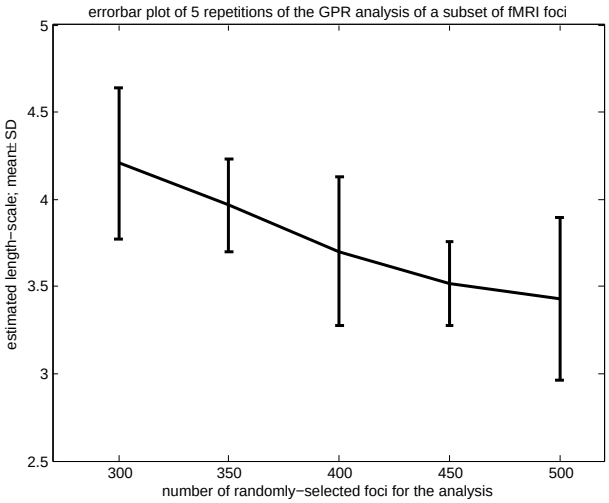


Figure 4.10: The effect of the number of foci included in the CBMA on the inferred ℓ when using the $\text{GPR}_{\text{joint,s}}$ with a $\Gamma(1,5)$ prior on ℓ. For the value x on the x-axis, we randomly select x sample foci from the pool and infer the ℓ; this is repeated 5 times. The mean and SD of the resulting 5 ℓ values are shown on the y-axis. The EO estimate of ℓ increases when the number of foci in the CBMA decreases. This also demonstrates GPR's flexibility in accommodating both prior belief and data in its estimation of hyperparameters, i.e., even though we have a prior on ℓ, there is still flexibility in its estimation depending on the data.

this pool and repeat this process for 5 times (with replacement); each of these 5 repeats results in an estimate for ℓ whose mean and SD are shown on the y-axis for x-axis value of 300. This figure shows that even though we have a prior on ℓ, there is still flexibility in its estimated value depending on the data.

Using the foci in Figure 4.9 for a mean-contrast CBMA results in the maps in Figure 4.11. As in the simulated 3D data, in order for ALE and KDA to use both activation and deactivation foci they are run twice: once with activation (i.e., $\text{ALE}_{\text{activation}}$) and once with deactivation ($\text{ALE}_{\text{deactivation}}$) foci. This figure displays FLAME-FFX and FLAME-MFX together with ALE, KDA and GPR. The lack of sensitivity is an obvious outcome of using the minimal peak information in the CBMA, when compared to IBMA that uses the whole image from each study. In order to assess the CBMA methods in a difference contrast, the results for the Therm-Mech contrast (whose negated image corresponds to the Mech-Therm contrast) are shown in

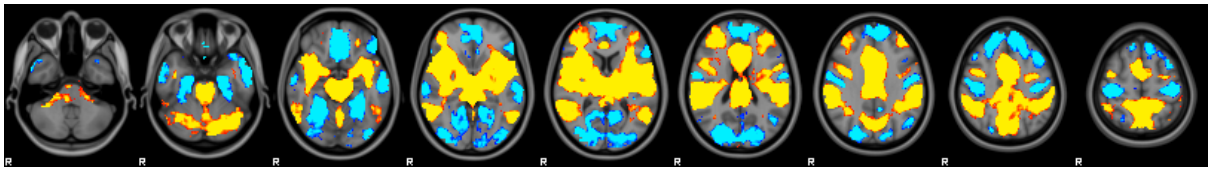
Figure. 4.12. This is expected to display the areas of the brain where thermal stimulus (THERM) causes a different level of (de)activation when compared to mechanical stimulus (MECH).

Following this qualitative summary of each method's performance, in order to quantitatively measure various methods' accuracy when applied to real fMRI data, we use DC with FLAME-MFX as the gold-standard. Both gold-standard and CBMA maps are thresholded at Z-stat=2.5 (i.e., P-value=0.012) in order to result in the DC binary maps; DC scores are plotted in Figure 4.13. As DC correspond to a single binarisation threshold, which may not necessarily be consistent across different CBMA methods, we also used AUC of the ROC curve; this assesses the CBMA methods' performance over a wider range of thresholds (results are shown in Figure 4.14).

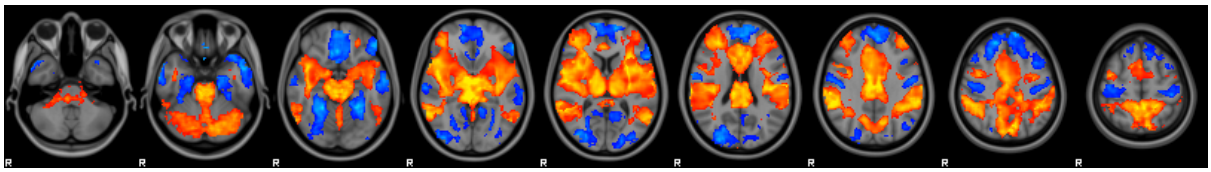
Finally, it is necessary to assess the dependence of $\text{GPR}_{\text{joint,s}}$ CBMA on the value that σ_f is fixed at. For this assessment, DC of $\text{GPR}_{\text{joint,s}}$ is measured when σ_f is fixed at 1,2,3,4 and 5. The results, shown in Figure 4.15, demonstrate how robust the performance is with respect to variations in σ_f, as long as it is larger than 2. This result justifies our choice of σ_f=3 in this chapter.

4.4 Discussion and Conclusions

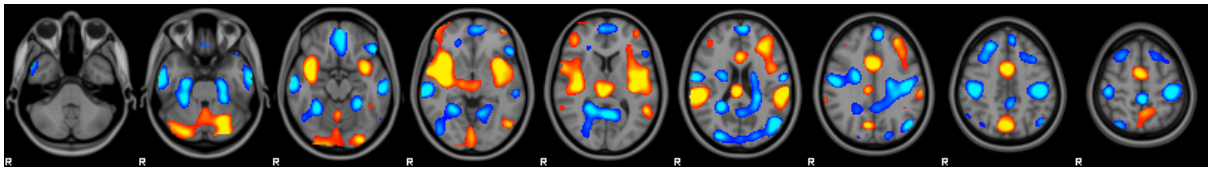
In this study, we reviewed the problems that the current coordinate-based meta-analysis techniques suffer from and offered a framework for tackling them. Our new method employs GPR, a nonparametric approach to regression, using a group of observations/samples from a Gaussian random field (i.e., statistic image) in order to estimate its value at various locations (i.e., voxels). This is the first important difference between GPR and other CBMA approaches; while existing CBMA methods try to localize the consistency, GPR estimates the pooled effect size (i.e., Z-stat in this particular application) as well. Thus, not only does GPR use the (x,y,z) coordinates of the foci, it also incorporates these foci' Z-stats. The most similar method to ours is Costafreda



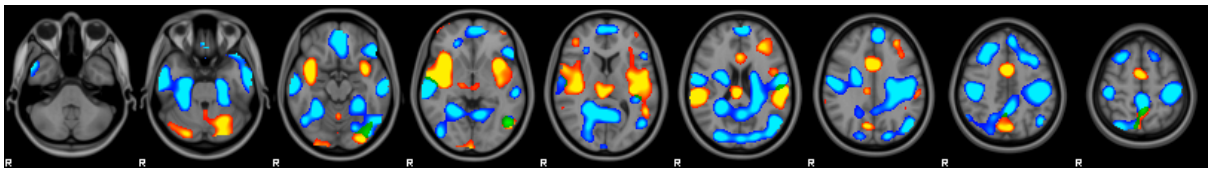
(a) FLAME-FFX



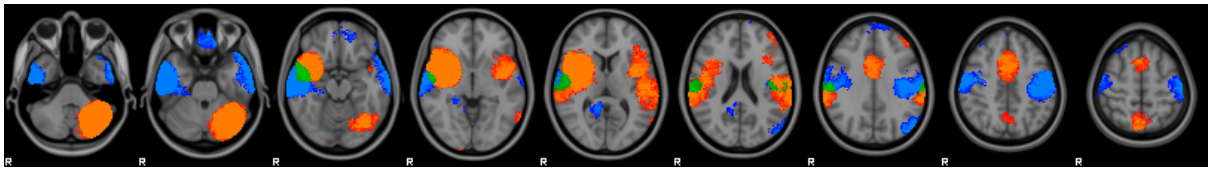
(b) FLAME-MFX



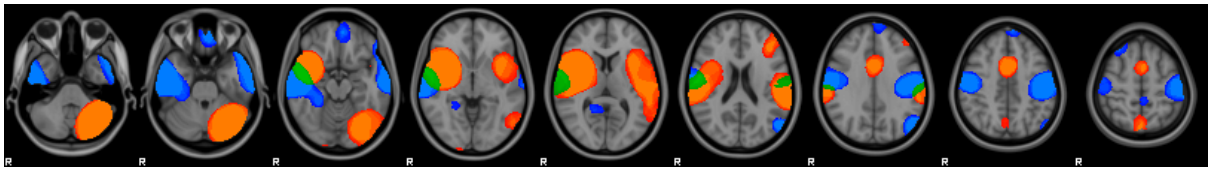
(c) $\text{GPR}_{\text{joint,s}}$



(d) $\text{GPR}_{\text{activation,s}}$ (red) & $\text{GPR}_{\text{deactivation,s}}$ (blue)

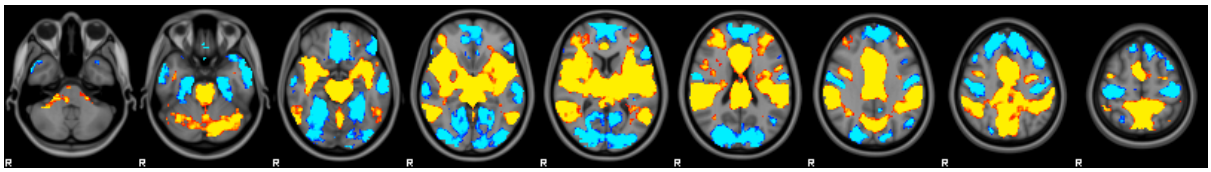


(e) $\text{KDA}_{\text{activation}}$ (red) & $\text{KDA}_{\text{deactivation}}$ (blue)

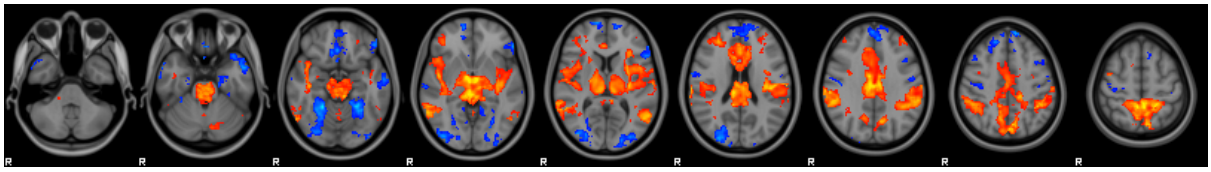


(f) $\text{ALE}_{\text{activation}}$ (red) & $\text{ALE}_{\text{deactivation}}$ (blue)

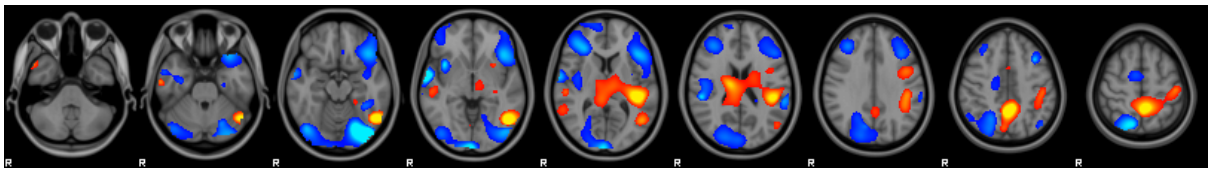
Figure 4.11: IBMA and CBMA results when pooling a one-group set of fMRI studies. Z-stats from FLAME-FFX and FLAME-MFX are displayed in (a) and (b), respectively. The CBMA results are for $\text{GPR}_{\text{joint,s}}$ with $\Gamma(1, 5)$ prior on ℓ, $\text{GPR}_{\text{activation,s}}$ & $\text{GPR}_{\text{deactivation,s}}$ with Γ(1,5) prior on ℓ, $\text{KDA}_{\text{activation}}$ & $\text{KDA}_{\text{deactivation}}$ with ρ=25, and $\text{ALE}_{\text{activation}}$ & $\text{ALE}_{\text{deactivation}}$ with σ=15, shown in (d)-(g), respectively. The false positives in GPR when compared with FLAME are due to the foci existing in those false-positive regions (see Figure 4.9), which also causes ALE and KDA false positives. Red-yellow and blue colours show values with range [2, 4] and [-2, -4], respectively, and the displayed slices are selected from z=-40 mm to z=40 mm, every 12 mm in MNI coordinates.



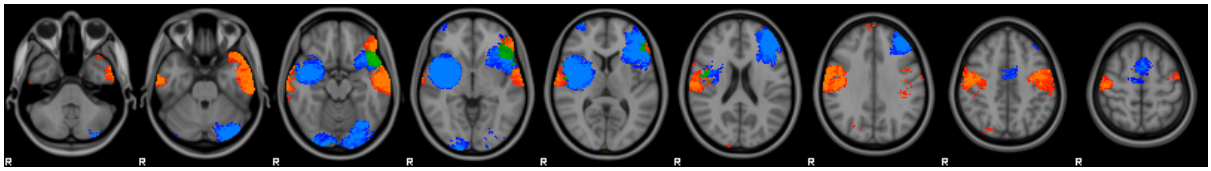
(a) FLAME-FFX



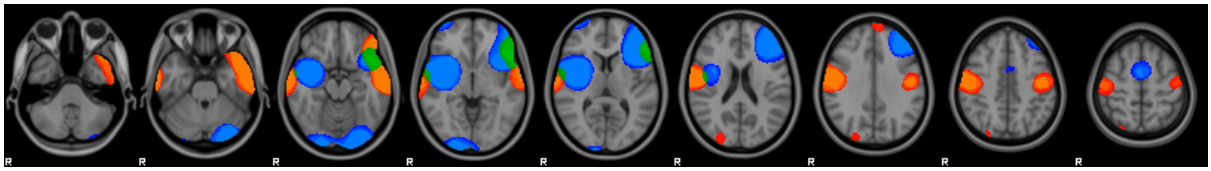
(b) FLAME-MFX



(c) $\text{GPR}_{\text{joint,s}}$



(d) $\text{KDA}_{\text{activation}}$



(e) $\text{ALE}_{\text{activation}}$

Figure 4.12: IBMA and CBMA results when contrasting two groups of fMRI studies (Therm-Mech). Z-stats from FLAME-FFX and FLAME-MFX are displayed in (a) and (b), respectively. The CBMA results are for $\text{GPR}_{\text{joint,s}}$ with $\Gamma(1,5)$ prior on ℓ, $\text{KDA}_{\text{activation}}$ with $\rho = 25$ mm, and $\text{ALE}_{\text{activation}}$ with $\sigma = 15$ mm, shown in (d)-(e), respectively. Red-yellow and blue colours show Z-stat values with range [2, 4] and [-2, -4], respectively, and the displayed slices are selected from z=-40 mm to z=40 mm, every 12 mm in MNI coordinates.

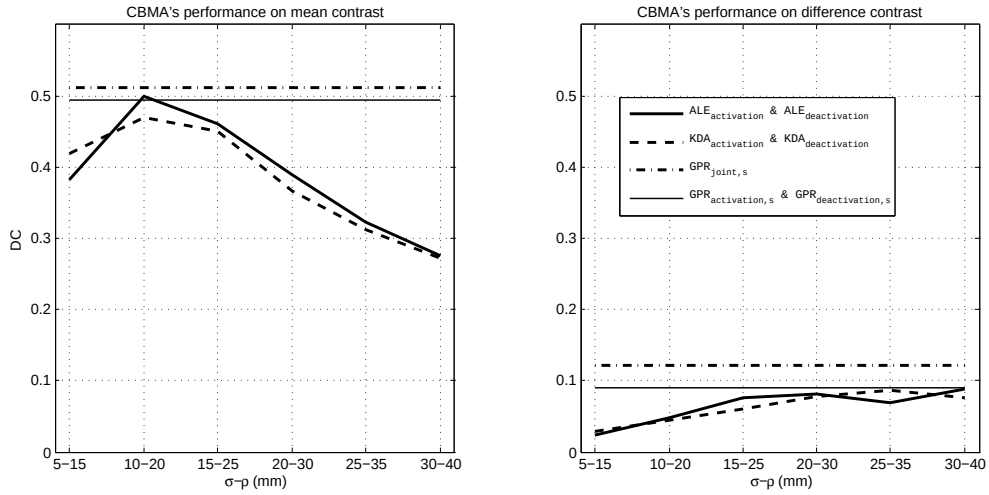


Figure 4.13: Using DC for assessing the performance of CBMA methods when applied to pooling real fMRI studies (dash-separated σ and ρ on the x-axis corresponds to ALE and KDA's kernel sizes, respectively). Having both CBMA and IBMA images binarised at Z-stat=2.5, DC shows that $\text{GPR}_{\text{joint,s}}$ with $\Gamma(1,5)$ prior on ℓ outperforms ALE, KDA and $\text{GPR}_{\text{activation},*}$ & $\text{GPR}_{\text{deactivation},*}$ with $\Gamma(1,5)$ prior on ℓ. The CBMA results are worse for the difference contrast than the mean contrast.

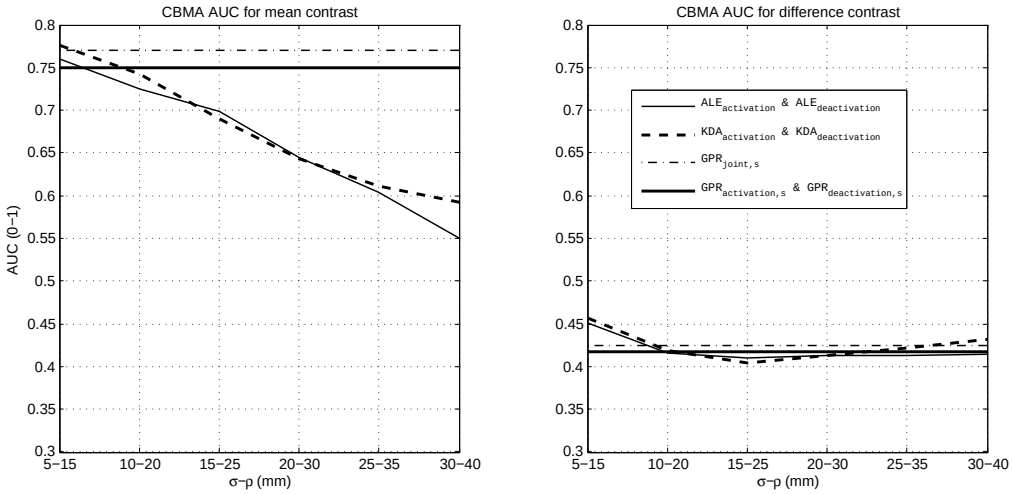


Figure 4.14: Using the area under the ROC curve for assessing the performance of CBMA methods when applied to the real fMRI data (see the text for more details on the notations). AUC shows $\text{GPR}_{\text{joint,s}}$ (with $\Gamma(1,5)$ prior on ℓ) outperforms ALE, KDA and $\text{GPR}_{\text{activation,s}}$ & $\text{GPR}_{\text{deactivation,s}}$ with $\Gamma(1,5)$ prior on ℓ over a range of possible binarisation thresholds for the mean contrasts, and gives a similar results on the difference contrast (note that, dash-separated σ and ρ on the x-axis corresponds to ALE and KDA's kernel sizes, respectively).

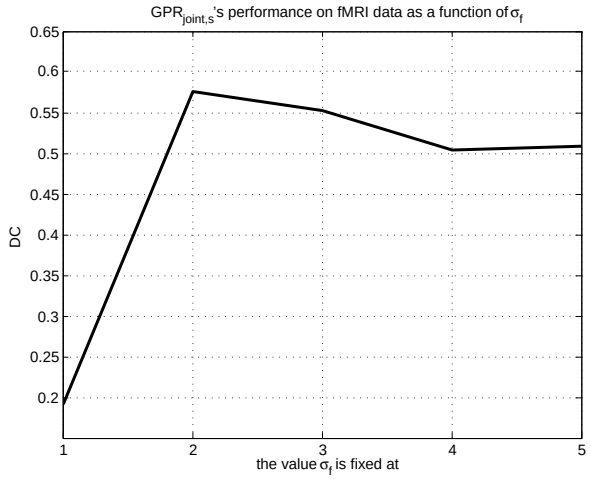


Figure 4.15: The effect of fixed σ_f in $\text{GPR}_{\text{joint,s}}$'s performance. In this figure, the x-axis shows the values σ_f is fixed at and y-axis is the DC score after thresholding the $\text{GPR}_{\text{joint,s}}$ Z-stat maps at 2.5 and comparing against the fMRI gold standard. This plots implies that $\text{GPR}_{\text{joint,s}}$ is fairly robust to varying the σ_f over the 2-5 range.

et al. (2009), which does not utilize the effect sizes and only provides an analytic model for the coordinates.

Existing CBMA methods incorporate only the coordinates of the activation foci and have no solution for jointly-incorporating the deactivation information if available. Therefore, in order to estimate the deactivation map, traditional CBMA needs a separate analysis using the deactivation foci. This is the *second* difference between GPR and the former CBMA approaches, which not only makes the traditional methods less accurate, but also reduces their accuracy when assessing difference contrasts. This is caused by the fact that when solely using the activation foci, regions that are neutral to one group while showing deactivation in another group are represented similarly (i.e., both have no representative foci in the input). This is shown in Chapter 2 for finding the regions showing different activation-deactivation patterns when caused by THERMAL painful stimuli than when caused by MECHANICAL.

Having collected the input information, existing CBMA approaches need the researcher to choose the model parameters (the extent of the kernel). Although smoothing kernels utilized by studies might provide insights for the appropriate CBMA kernel

size, this arbitrariness is a weakness of the method. In an attempt to address this issue, Eickhoff et al. (2009) modeled ALE's kernel size as the parameter that "should reflect the uncertainty of the reported spatial location due to between-template and between-subject variance". However, as the extent of this kernel is a property of the meta-analysis result (i.e., its smoothness), it should be "inferred" from the spatial arrangement of the foci with respect to each other (in terms of *both* their coordinates *and* effect size). As the *third* main difference, GPR has an evidence-optimization step before the regression that automatically determines the optimal values for the model *hyperparameters*, i.e., ℓ, σ_f and σ_n. Additionally, when there is enough information *a priori* about the parameters, GPR offers a flexible framework for incorporating that information in the form of a fixed hyperparameters (e.g., ℓ or σ_f) or a probability distribution over a hyperparameters (e.g., a Gamma distribution for ℓ).

According to the results when fixing σ_f while using a Gamma prior on ℓ, GPR yields more accurate results over a wide range of scenarios. The reason we recommend this approach is that it reflects: (1) when no foci are observed, the Z-stat maps could have most likely had a value in [-$Z_{threshold}$, +$Z_{threshold}$] that did not survive the thresholding (at $Z_{threshold}$) prior to foci extraction, and (2) depending on the data, we have a prior belief about a range of values of ℓ that can model the spatial dependency; Gamma PDF models this belief. This PDF is recommended to have a mean ℓ close to the smoothness of interest with a reasonably small SD.

Note that, in our advocated GP model, we fixed the value of σ_f at 3, and only infer the values for ℓ and σ_n. The choice of σ_f=3, i.e., accommodating our prior knowledge of the data in the GP model, is already shown by (Groves et al., 2009) to be useful in neuroimaging models. This value of 3 quantifies our belief on the amount of uncertainty we have around the 0 mean of the GP. This prior belief, when mixed with the data/observations, will show itself as in Equation 4.8, where C($\mathbf{V}_*$,$\mathbf{V}_*$) is our prior belief of uncertainty, from which is subtracted a positive term. That is, new observations reduce our uncertainty about the underlying function. As shown by Figure 4.15, our

results are fairly robust in the range of values that we tested and hence we recommended the ones with better performances (i.e., 2 and 3).

When the covariance, i.e., smoothness of the data, is known *a priori* (similar to ALE and KDA) GPR offers a solution with fixed ℓ (shown in Figure 4.3). Also, ℓ is more interpretable and has a nice correspondence to the statistical image's smoothness, which makes it easier to choose a value for it (if so desired) than for ALE's σ (for example). That is, the kernel in ALE and KDA models the uncertainty associated with the location of the reported foci, which indirectly depends on the smoothness while the value of ℓ is itself the smoothness of the field and hence is in a clearer correspondence to original image's smoothness.

An important strength of GPR-CBMA is its underlying model that resembles the traditional hierarchical GLM-based meta-analytic model that is the fundamental model in many IBMA methods (e.g., Woolrich et al. (2004)). With the model's parameters inferred, CBMA can become nearly as informative as IBMA, e.g., it has the power to offer a FFX as well as an RFX meta-analysis. This model can accommodate various flexibilities (such as "RFX meta-regression") in GPR analysis. However, due to the nature of coordinate inputs, GPR-CBMA is not as accurate as its IBMA alternatives. Figure 4.9 shows this phenomenon by displaying the *discrepancy* that exists between coordinate and full-image (de)activation maps. For instance, a decent number of deactivation foci are reported (top row), which, given how deactivation is not frequently reported by studies (bottom row), implies that a limited number of studies contribute to this foci set.

We have presented a solid mathematical framework for the CBMA, and hence there are many possible extensions to this model. One of the interesting outcomes of the GPR analysis is its estimates for each voxel's effect size with a corresponding uncertainty. Utilizing such voxel-wise pairs of COPEs and VARCOPEs was shown to be useful in producing graphical tools for assessing the bias and heterogeneity in the meta-analytic pool (Salimi-Khorshidi et al., 2009b; Nielsen, 2009). For the traditional CBMA, however,

the lack of such pairs made it almost impossible to harvest this rich graphical meta-analytic literature.

Moving in the direction of incorporating more information from neuroimaging papers is expected to increase the information intake from the literature and hence enable CBMA results to better resemble those from IBMA. While incorporating the deactivation information and Z-stats, as demonstrated in this chapter, is a step in this direction, there is still more information in journal papers that is not taken into account. One important issue in CBMA is that not all studies report the effect sizes corresponding to the reported coordinates. Although this case will have no influence on ALE and KDA (a point of strength for these methods), the current version of GPR CBMA will not be able to accommodate those foci in the analysis. However, in such cases, depending on the other related information in the paper(s), one can use GPR with imputation.

In statistics, imputation is the substitution of some value for a missing data point or a missing component of a data point. Once all missing values have been imputed, the dataset can then be analysed using standard techniques for complete data. For instance, consider a study that reports m coordinates (i.e., (x, y, z) triplets) that have no Z-stats associated with them. In such cases, one can randomly draw m samples from $\mathcal{N}(u_c, 1)$, where u_c reflects the Z-stat at which the study has thresholded its statistic image. In order to have a more robust inference in such a case, however, one can employ "multiple imputation" (i.e., repeating this process for multiple times and pooling the predictions/results) and assess the stability of the result. Also, in cluster-wise-inference papers, one can use assumptions such as those that come from random-field theory, where, given the cluster size, cluster-forming threshold and an estimation of the smoothness, the Z-stat at the location of the peak can be inferred."

Given the complexity of neuroimaging meta-analysis, the ideal model is the one that incorporates all our prior (e.g., anatomical) knowledge in modelling the observations/data. For instance, apart from incorporating our knowledge of plausible ℓ and σ_f, we could limit GPR CBMA to a grey-matter mask, which translates to our

anatomical knowledge. The current state of GPR (as well as ALE and KDA) CBMA cannot employ other anatomical information in its inference and prediction, such as those used by Phan et al. (2002), which makes Phan et al. (2002) a good pre-/post-processing to GPR, ALE, and KDA.

Also, it is important to note that the result of the meta-analysis will depend on the "inclusion criteria", i.e., a set of rules according to which the data from a particular study is chosen to be fed into the meta-analysis. One way of assessing the studies is by implementing the the semantics of the experimental description in the model (Nielsen and Hansen, 2002, 2004). This relates to cases where one wants to conduct a literature search prior to the mathematical meta-analysis, which was not the case in this study. However, meta-analysis researchers are strongly advised to take into account issues such as bias and heterogeneity in their study selection, model selection and results interpretation (Salimi-Khorshidi et al., 2009b).

Chapter 5

Identifying Modulatory Network-interactions in the Brain: Correspondence between Activation and Rest

Abstract

The brain's functional networks can be modelled by graph representations of activity, where the vertices represent functional/anatomical regions and the edges are their functional connectivity. In recent studies the BrainMap task activation database has been used to map such networks of coactivation, whose functional units ("nodes") are shown to be in correspondence with those extracted from the resting brain. However, *correspondences in these nodes' non-additive (e.g., modulatory) interactions* are yet to be assessed. In this work we carried out a joint multivariate exploratory analysis of resting-state FMRI (rFMRI) and BrainMap data in order to (1) extract the spatial maps of these nodes and (2) pinpoint the network structure underlying these nodes' interactions. However, like others we have found that network modelling using a high number of nodes is difficult (at least for the task database), hence, we used a log-linear graphical model (LLGM) to look for interactions, finding triplets of functional nodes which appear to interact (i.e., when one node modulates the functional connection between the other two). Correspondence was found between the set of significant modulatory rFMRI and BrainMap triplets, which extends the component-matching result reported previously to the components' non-additive interactions.

5.1 Introduction

Extracting functional networks describing connectivity between different brain regions is an emerging area in the field of neuroimaging, with 950 papers listed in PubMed from a "functional connectivity" AND FMRI search. These studies differ in the types of data (e.g., activation data (Patel et al., 2006), resting-state data (Zhang et al., 2010) or imaging studies' databases such as BrainMap[1] (Toro et al., 2008)) and their mathematical models (e.g., confirmatory (McIntosh and Gonzalez-Lima, 1994) or exploratory (Burge et al., 2009)).

One of the major data modalities for such studies is resting FMRI (rFMRI), the spontaneous fluctuations in the resting brain, which shows temporal correlation between any part of the brain and other parts of the same functional network. Finding such functional networks, in addition to offering information about the structure and function of the healthy brain, is shown to be of great potential clinical value (e.g., Veer et al. (2010)). Several networks of correlated temporal patterns in the "resting brain" have been identified. These distinct patterns can be separated from each other from a single resting FMRI dataset, because, although each has relatively consistent time courses across its set of involved regions, the different networks have different temporal characteristics from each other.

Although such "resting state networks" (RSNs), and related networks of deactivation under task, have also been investigated in other modalities such as electroencephalography (EEG) and positron emission tomography (PET), the majority of the research to date has used FMRI. Although there has been concern that some patterns of spatially extended spontaneous signals may be of non-neural physiological origin, these concerns are increasingly being addressed, and it has been posited that RSNs do reflect functional networks (see an excellent review in (Fox and Raichle, 2007)).

It was shown by Smith et al. (2009) that the full repertoire of major functional networks utilized by the brain in externally instigated action is continuously and

[1] www.brainmap.org

dynamically active even when "at rest". The activation networks in Smith et al. (2009) are identified by carrying out an analysis of thousands of separate activation maps derived from the BrainMap database of functional imaging (co)activation studies.

BrainMap is currently the largest database of FMRI and PET brain activation studies, including the results from thousands of journal articles, each with several different task conditions and contrasts between these. What makes BrainMap an interesting source of data for brain research is in having inferences from a large number of task studies; any aggregate inference on this data is hence generalizable to the wider population. In other words, the result from employing a meta-analytic approach for determining interdependencies between brain regions is inferring on the most general level possible (Salimi-Khorshidi et al., 2009a). However, (Ramsey et al., 2010) argue that the results obtained as such cannot be interpreted as representations of conditional dependence relations among localized neural activities; specifically, directed pathways in such graphical results may be artifacts of the manner in which distinct kinds of studies are combined in the meta-analytic procedure.

Given the correspondence found between functional modules of the brain in activation and rest (Smith et al., 2009), our study searches for the existence and extent of such correspondence in brain regions' multi-way (or non-additive) interactions. We are specifically interested in looking for modulatory interactions, where one region's "activity" modulates the strength of the connection between two other regions. Thus far, various network-modelling methods have been employed for mining the brain regions' connections (readers are referred to Smith et al. (2010) for an extensive review and comparison of these methods), with little work to date looking for non-additive effects (an obvious exception being the non-exploratory use of dynamic causal modelling (Friston et al., 2003) which allows for the prediction and evaluation of modulatory effects). However, like others (Neumann et al., 2009), we have found that network modelling using a high number of nodes is difficult in the case of the BrainMap database dataset, due to the fact that this contains such sparse representations of the original data (just the activation

peak coordinates) that the more sophisticated network models are not provided with sufficiently rich data to robustly achieve their ambitious goals. Instead, here we use a log-linear graphical model (LLGM) to look for "small-scale" interactions, considering only 3 functional nodes at any one time.

The LLGM in this study is a Poisson regression (PR) model, which is a special case of the generalized linear model (GLZ), a flexible generalization of ordinary least squares regression). The GLZ generalizes linear regression by allowing the linear model to be related to the response variable via a link function (McCullagh and Nelder, 1999); using the logarithm link function is a special form of GLZ known as PR, used to model count data and contingency tables. Under this model, in order to assess an N-way interaction, N time-series are first binarised in order to form an N-way contingency table whose cell-counts are then modelled using PR in order to accept or reject the null hypothesis (H0): "there is not an N-way interaction among the variables".

The small-scale graphical model in this study attempts to find triplets (N=3) of functional "nodes" which appear to interact, e.g., when one node modulates the functional connection between the other two. In order to assess such three-way interactions (if found) further, the binarised time-series of an interactive triplet are fed into an odds-ratio (OR) analysis, one of a range of statistics used to assess the chance of a particular outcome if a certain factor (or exposure) is present. The OR is a relative measure of chance, which pinpoints the directionality of such interactions by telling us how much more likely it is that one region that is exposed to another region's activation will develop the outcome as compared to when it is not exposed. Section 5.2 describes the details of the analysis whose results are then shown in Section 5.3.

5.2 Materials and Methods

In this section, we first introduce the datasets used in this study, consisting of one rFMRI dataset and a group of studies collected from the BrainMap database. Next, the exploratory approach employed for extracting the spatial map and time series of each

functional node is described. At the end of this section LLGM and the post-hoc OR analysis are introduced.

5.2.1 Data

As described earlier, this research aims to find the existence and extent of correspondence between active and resting brain in their small-scale networks of multi-way interactions. Therefore, two groups of data are required: one for assessing the interaction mechanisms under activation, and one for assessing the same thing "at rest".

5.2.1.1 Simulated Data

In order to test our method on a dataset where the truth is known *a priori*, a network of 10 nodes is used to simulate rich, realistic BOLD time-series (see Figure 5.1). These nodes' time-series are simulated so that they resemble the variations that are observed in reality from "functional nodes" (see Section 5.2.2 for the description) that are extracted from the data. The simulations were based upon the dynamic causal modelling (DCM) (Friston et al., 2003) FMRI forward model, which uses the nonlinear balloon model (Buxton et al., 1998) for the vascular dynamics, sitting on top of a neural network model. The fundamentals of our simulation are the same as those carried out by (Smith et al., 2010) for simulating the brain networks.

Each node of this network has an external input that is binary ("up" or "down") and generated based on a Poisson process that controls the likelihood of switching state. Neural noise/variability of standard deviation 1/20 of the difference in height between the two states is added. The mean durations of the states were 2.5 s (up) and 10 s (down), with the asymmetry representing longer average "rest" than "firing" durations; the final results did not depend strongly on these choices (for example, reducing these durations by a factor of 3 made almost no difference to the final results). These external inputs into each node can be viewed equivalently as either a signal feeding directly into each node, or as noise appearing at the neural level. The neural signals propagate around the

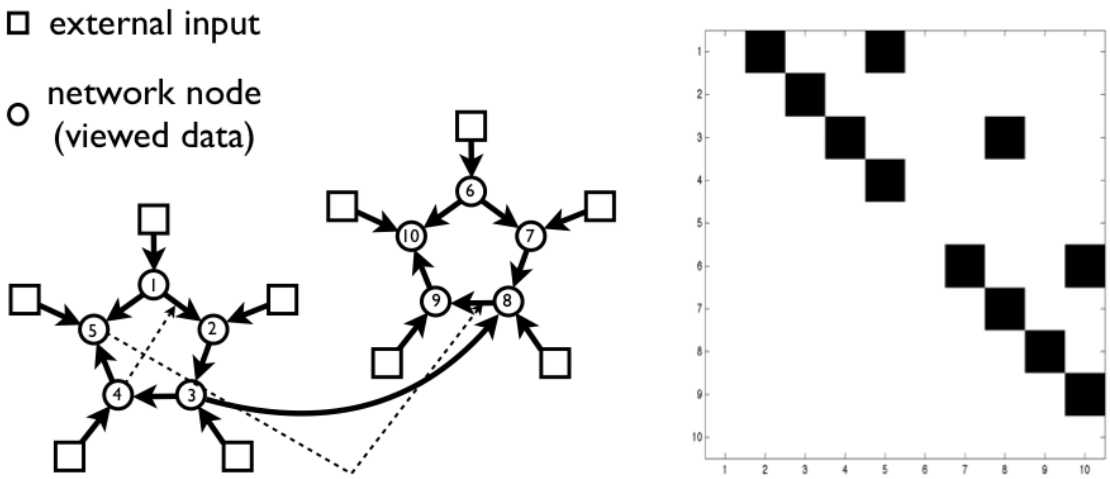


Figure 5.1: The network topology fed into the FMRI data simulations and its corresponding connection matrix. An element in the upper diagonal of the connection matrix implies a directed connection from a lower-numbered node to a higher-numbered one. The nonlinear interaction is added such that activity of node 4 modulates the strength of the connection between nodes 1 and 2, and activity of node 5 modulates the strength of the connection between nodes 8 and 9. Note that the modulatory edges are shown as dashed lines in the topology.

network using the DCM neural network model, as defined by:

$$\dot{z} = (A + \sum_{j=1}^{m} u_j B^j)z + Cu \tag{5.1}$$

where z is the neural time-series, $\dot{z}$ is its rate of change, A denotes the "fixed" connectivity matrix, the $\sum$ term denotes the connectivity "modulation", u are the external inputs and C the weights controlling how the external inputs feed into the network (i.e., the identity matrix in this application). The off-diagonal terms in A determine the network connections between nodes, and the diagonal elements are all set to -1, to model within-node temporal decay. Each node's neural time-series was then fed through the nonlinear balloon model for vascular dynamics responding to changing neural demand. The amplitude of the neural time-series were set so that the amount of nonlinearity (nonlinearity here being potentially with respect both to changing neural amplitude and duration) matched what is seen in typical 3T FMRI data, and BOLD % signal change amplitudes of approximately 4% resulted (relative to mean intensity of simulated

timecourses). The balloon model parameters were in general set according to the prior means in DCM.

However, it is known that the haemodynamic processes vary across brain areas and subjects, resulting in different lags between the neural processes and the BOLD data, with variations of up to at least 1s (Chang et al., 2008). We therefore added randomness into the balloon model parameters at each node, resulting in variations in haemodynamic response function (HRF) delay of standard deviation 0.5s. Finally, thermal white noise of standard deviation 0.11% (of mean signal level) was added. The BOLD data was sampled with a TR of 3 s and the simulations comprised 50 separate realisations (or "subjects"), all using the same simulation parameters, except for having randomly different external input time-series, randomly different HRF parameters at each node (as described above) and (slightly) randomly different connection strengths. Each "subject's" data was a 10-min FMRI session (200 time-points) in this simulation.

5.2.1.2 Resting FMRI Data

The rFMRI dataset and its analysis followed a fairly standard protocol. Thirty-six healthy adult subjects were imaged (age range 20-35 y, mean 28.5 y; 21 male, 15 female) in a 3T Siemens Trio MRI scanner, using a 12-channel head coil. All data employed had been collected in accordance with local ethics approval. Structural brain images were acquired using a T1-weighted 3D MPRAGE sequence, resolution 1×1×1 mm. These were used purely to aid the registration of the functional data into a common standard brain coordinate system (MNI152).

Resting FMRI BOLD (blood-oxygenation-level dependent) data were acquired with a standard gradient echo echo-planar-imaging (EPI) acquisition, TR 2 s, TE 28 ms, flip angle 89°, resolution 3×3×3.5 mm, whole-head coverage except for the lowest parts of the cerebellum in some subjects. The resting FMRI scan lasted 6 min, during which ambient light was minimized, and the subjects were instructed to lie with eyes open, think of nothing in particular, and not to fall asleep.

Data preprocessing was carried out with FSL tools (Smith et al., 2004; Woolrich et al., 2009). The following pre-statistics processing was applied for each subject: head motion correction using MCFLIRT (Jenkinson et al., 2002); non-brain removal using BET (Smith, 2002); spatial smoothing by using a Gaussian kernel of FWHM (full-width at half maximum)=5 mm; grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor; high-pass temporal filtering (subtraction of Gaussian-weighted least-squares straight-line fitting, with σ =50.0 s). Registration of each subject's FMRI data to that subject's high-resolution structural image was carried out using FLIRT (Jenkinson et al., 2002), with the affine transform then refined using BBR (boundary-based registration) (Greve and Fischl, 2009). Registration from the high-resolution structurals to MNI152 standard space was achieved using FLIRT affine registration and then further refined using FNIRT nonlinear registration (Andersson et al., 2007).

This dataset is not related to any experiments contained in BrainMap and the resulting functional modules are expected to represent group-averaged networks of brain regions with BOLD FMRI signals that are temporally correlated, i.e., the most representative networks of covariation when the brain is at rest. This is the same rFMRI data as that used in Smith et al. (2009).

5.2.1.3 BrainMap Data

The BrainMap database contains the results of a large number of brain activation studies; at the time of our analyses, it contained the results from 1,687 journal articles. Each study can involve multiple "conditions", for example, comparing finger tapping with rest and comparing different rates of tapping with each other. The 1,687 studies resulted in 7,342 separate activation/contrast images and involved 29,671 human subjects.

In addition, a large amount of study information is included in the database, including carefully structured, rich descriptive text detailing the experimental paradigm. Each paradigm is also categorized under one or more of 66 behavioural domain classifications;

these provide a more simplistic summary of the experimental tasks but are immediately quantitatively useful.

In the BrainMap database, the spatial distributions of these activations are represented via the coordinate locations of statistically-significant local maxima in the activation images. All coordinates are in a standard brain "space," in the case of BrainMap, the Talairach coordinate system (Talairach and Tournoux, 1987). For each activation result, we recreated a 2-mm-resolution standard space pseudo-activation image by filling an empty image with points corresponding to the activation coordinates and then convolving this with a Gaussian kernel of FWHM=10 mm. Although the actual spatial extent of the original activation has not been preserved, this smoothing extent is a reasonably close match to that applied as data preprocessing in many FMRI activation studies and is close to the spatial variability in database coordinate locations as carefully investigated by Eickhoff et al. (2009). We tested other spatial smoothing extents from 8- to 15-mm FWHM and found the results not significantly sensitive to the exact extent.

5.2.2 Functional Nodes

As shown by Smith et al. (2009), when fed into a probabilistic ICA analysis (Damoiseaux et al., 2006; Beckmann et al., 2005), rFMRI and BrainMap datasets result in very similar networks of covariation. Hence, these two types of data are compatible for a joint analysis in order for the extraction of their underlying functional modules using ICA, i.e., they share virtually the same networks of covariation and hence BrainMap can be pooled with a group of subjects' rFMRI data for a single, integrated ICA decomposition. By applying a *high* dimensional ICA to this combined dataset (in our case ending up with 125 non-artefactual components), the components are sufficiently highly-split that it is more appropriate to consider each component as a functional *node* than a functional *network*.

Therefore, the 7,342 coactivation images extracted from BrainMap (as described above) are concatenated together to produce a 2D dataset, where the first dimension is space (the 3 spatial dimensions are unwrapped onto 1 dimension of size

91×109×91=902,629 voxels) and the second is experiment ID (1:7,342; this dimension is often referred to as "time" in our text, as a convenient shorthand, and in analogy to the temporal dimension in the rFMRI data). Next, the 4D FMRI time series of all subjects are concatenated to produce a 2D dataset with its first dimension being space (of size 902,629 voxels) and the second dimension being time (1:6336; 6336=176×36, is the result of voxel-wise concatenation of 36 subjects' time series). These two datasets are then scaled to have equivalent spatiotemporal variance (estimated through a singular value decomposition analysis of the eigenvalues), and concatenated in their second dimension (i.e., "time"), which results in the "joint" 2D data.

The ICA step was carried out using MELODIC at dimensionality of 150 (i.e., decomposing the joint data into 150 components of covariation). From the resulting 150 maps, 25 artifactual maps were excluded after a visual investigation of the spatial, temporal and (temporal) spectral characteristics, leaving the remaining 125 maps as the representative covarying functional nodes (or "sub-networks") across a large sample of normal subjects in both activation and rest. The individual subject (i.e., 37 subjects: 36 rFMRI subjects and 1 BrainMap "subject") analysis of the joint data was carried out using a regression technique (dual regression (Beckmann et al., 2009)) that allows for voxel-wise comparisons of functional connectivity by identifying subject-specific temporal dynamics of each component. This involves using the full set of group-ICA spatial maps in a linear model fit (spatial regression) against the separate data sets, resulting in matrices describing temporal dynamics for each component and subject.

The result of this stage is a group of 125 spatial maps from the ICA decomposition (i.e., network "nodes") each with thirty-seven time series consisting of thirty-six subject-specific rFMRI time-series (each with 176 time points) and one BrainMap time-series (with 7342 "time" points) from the dual-regression analysis. Given the existing 125 nodes, there are 125!/(3!×122!)=317750 possible ways to choose a triplet of nodes, whose thirty-seven "subjects"' of time series are analyzed next (using the LLGM described in Section 5.2.3) in order to test the null hypothesis of "no multi-way interaction".

5.2.3 Log-linear Graphical Model

As previously described, we are interested in understanding the non-additive multi-way interactions in a small scale while using a relatively interpretable model, i.e., studying triplets of nodes instead of extracting the full structure and parameters of large networks. The model employed in this study is an example of generalized linear models (GLZ) (McCullagh and Nelder, 1999), used here for the analysis of count data and contingency tables. The GLZ allows the independent or explanatory variable (EV) to be linearly related to the response variable via a link function and hence unifies various models such as linear regression (Draper and Smith, 1998), logistic regression (Agresti, 2002) and Poisson regression (PR) (Cameron and Trivedi, 1998).

More precisely, we study the application of PR models to the analysis of contingency tables resulting from binarization of the time-series of the previously-described functional nodes. In this stage, each node's rFMRI and BrainMap time series are first binarised at their 90th percentile (i.e., the "binarization threshold") in order to generate a series of binary events (i.e., ON and OFF). The reason for choosing 90th percentile is its reliability in assessing the correspondence between rFMRI and BrainMap (discussed in Section 5.3). Next, these three binary time-series form 8-cell (i.e., C=8) contingency tables that are the basis of the PR analysis, which is a GLZ with Poisson error (i.e., the stochastic component of the model)

$$\begin{aligned} \Pr(\mathbf{Y} = \mathbf{y}_t|\boldsymbol{\mu}_t) &= p(\mathbf{y}_t|\boldsymbol{\mu}_t) \\ &= \frac{e^{-\boldsymbol{\mu}_t}\boldsymbol{\mu}_t^{\mathbf{y}_t}}{\mathbf{y}_t!}, \end{aligned} \tag{5.2}$$

and natural logarithm link function (i.e., the systematic component)

$$\begin{aligned} \ln(\mathbf{y}_t) &= \mathbf{X}^T\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t \\ &= \ln(\boldsymbol{\mu}_t) + \boldsymbol{\epsilon}_t, \end{aligned} \tag{5.3}$$

where $\mathbf{Y}$ is the random variable representing the C×1 count vector (with its entries being the counts in contingency tables' cells), $\mathbf{y}_t$ is an observed count vector (C×1), $\boldsymbol{\mu}_t$ is the

parameter vector (P×1), $\mathbf{X}^T\boldsymbol{\beta}_t$ is a linear combination of predictors (with $\mathbf{X}^T$ as the P×C design matrix), and $\boldsymbol{\epsilon}_t$ is the Poisson noise for triplet t (note that the factorial is element-wise). A key premise in this model is that the events constituting each count are independent; otherwise the Poisson model does not apply.

In order to see how $\mathbf{y}_t$s are calculated, consider a simple 2×2 contingency table for instance, where each of n independent binary observations are classified in one of its cells and treated as realizations of independent Poisson random variables. With n_{ij} as the number of observations (e.g., time points) and $\{X_1{=}i,\ X_2{=}j\}$, the contingency table

	$X_1 = 1$	$X_1 = 2$
$X_2 = 1$	n_{11}	n_{21}
$X_2 = 2$	n_{12}	n_{22}

is formed with $n = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}$ as the total number of events (i.e., the length of time-series).

Under independence (H0), count vectors $\mathbf{y}_t$ can be modelled/explained by three covariates: a constant mean, X_1 and X_2 (i.e., "simple model"). In case of a dependency or interaction between X_1 and X_2 in order to explain the observed counts, the model requires an X_1X_2 regressor to be introduced as well (i.e., the "saturated" or "full" model). In order to test the no-interaction H0, one can use the simple model and evaluate the goodness of its fit (after estimating the model parameters using a maximum likelihood estimation (MLE) (McCullagh and Nelder, 1999)) by calculating the deviance ($\mathcal{D}$). Deviance is $-2(L_{sim} - L_{sat})$ with L_{sim} and L_{sat} being the likelihood maximized for the simple and saturated models, respectively; if $\mathcal{D}$ is bigger than expected by chance only, the no-interaction hypothesis is rejected, i.e., X_1 and X_2 have a non-additive interaction. Under the H0, $\mathcal{D}$ has a χ^2_ν distribution; ν is the difference between the number of parameters in the saturated and simple models (i.e., ν=1 when simple and saturated models differ in only one covariate).

In this study, however, we assess the three-way interactions where instead of two variables there are three binary variables (X_1, X_2 and X_3) and hence a three-way contingency table. The saturated and simple models are different in that the simple

model does not have the $X_1X_2X_3$ explanatory variable. The rest of the analysis is similar to the two-way tables, i.e., assessing the no-three-way-interaction H0 using $\mathcal{D} \sim \chi^2_\nu$, where $\nu = 1$ as we only lack $X_1X_2X_3$ in the simple model. Rejecting the H0 in this model means there are variations in X_1, X_2 and X_3 that cannot be described by X_1, X_2, X_3, X_1X_2, X_2X_3 and X_1X_3 alone, but also requires $X_1X_2X_3$.

Using such an LLGM analysis for assessing the H0 for each triplet results in 37 Z-stats in total, corresponding to that triplet's 37 chunks of trivariate time series. Pooling the 36 rFMRI Z-stats, using a one-sample T-test, results in a group-level Z-stat indicating the strength of evidence against H0 for a given triplet (in the rFMRI dataset as a whole) and concludes the triplet's assessment with two Z-stats (i.e., one for rFMRI and one for BrainMap). In other words, the result of the LLGM analysis is $\{\mathbf{z}_{\text{rFMRI}}, \mathbf{z}_{\text{BrainMap}}\}$, 317750 pairs of $\{z^t_{\text{rFMRI}},\ z^t_{\text{BrainMap}}\}_{t=1}^{317750}$, each corresponding to the strength of the evidence for rejecting H0 in one triplet (see Algorithm 3 in Appendix C.1 for a pseudo-code summary). Note that the "Z-stat threshold" for triplet t refers to thresholding z^t_{BrainMap} and z^t_{rFMRI}.

5.2.4 Correspondence Analysis

In order to assess the correspondence between the triplets of non-additive interaction between the active and resting brain, we first need to exclude from the existing 317750 triplets those that do not demonstrate any functional connections. That is, a triplet is included in the correspondence analysis if the estimated graph from a "global" network analysis (with all the 125 nodes) has at least one edge connecting a pair of its nodes, and excluded otherwise.

If we consider a continuum of methods, with one extreme being pure pairwise methods (e.g., correlation) and the other extreme global network modelling approaches (e.g., Bayesian networks), there are methods such as partial correlation and sparse Inverse COVariance Estimation (ICOV) that sit somewhere in the middle (still using all nodes' data in their calculations), with ICOV a little closer to the more global modelling extreme. In (Smith et al., 2010) it was shown that such pseudo-global approaches in general (and

ICOV in particular) result in fairly accurate network inference even when fed with a relatively small number of observations (Huang et al., 2010; Smith et al., 2010). Thus, in order to find the triplets with no intra-triplet functional connectivity prior to the correspondence analysis, we use the binarised connectivity (inverse covariance) matrix from ICOV.

The inverse of the covariance matrix is an efficient way to estimate the full set of partial correlations (Marrelec et al., 2006). Regularisation can be applied under the constraint that this matrix is expected to be sparse (e.g., using the Lasso method (Friedman et al., 2008)), which shrinks entries that are close to zero more than those that are not. For a mathematical description, suppose that there are p brain regions to be modelled, i.e., $\{X_1,...,X_p\}$, each with their corresponding time-series that we can assume follow a multivariate normal distribution. Given the measurement data of the brain regions, ICOV finds an estimate for the inverse covariance of the brain regions by solving the following optimization:

$$\hat{\Theta} = \arg\max_{\Theta>0} \ \log(\det(\Theta)) - \mathrm{tr}(S\Theta) - \lambda||\Theta||_1, \tag{5.4}$$

where Θ and $\hat{\Theta}$ denote the inverse covariance and its estimate, S is the sample covariance matrix, $det(.)$ and $tr(.)$ denote the determinant and trace of a matrix, $||.||_1$ denotes the sum of absolute values of all the entries in a matrix, and λ is a pre-selected regularization parameter. We use an implementation of ICOV referred to as L1precision[2], which requires the setting of the regularisation-controlling parameter λ. Following the results in Smith et al. (2010) we set λ to be 10 (higher λ gives greater regularisation).

The result of the previous analysis is a group of triplets (<317750) across which the resting and active brain's correspondence, is to be assessed. As, the estimated extent of such a correspondence depends on the chosen binarization threshold, we carry out an investigation on a range of possible binarization thresholds before choosing one. If there is no systematic overlap between rFMRI and BrainMap (i.e., independence, H0), and

[2]www.cs.ubc.ca/~schmidtm/Software/L1precision.html

having N_{rFMRI} and $N_{BrainMap}$ interactive triplets in rFMRI and BrainMap, respectively, the expected overlap is $N_{expected}$=$N_{total} \times p_1 \times p_2 \times p_3$, where N_{total} is the total number of triplets and p_1 is the chance of a given triplet having at least one edge, p_2 is the chance of the triplet being interactive in the rFMRI data (N_{rFMRI}/N_{total}), and p_3 is the chance of the triplet being interactive in the BrainMap data ($N_{BrainMap}/N_{total}$). Probability p_1 is calculated as:

$$p \;=\; 1 - P(\text{no ICOV edges}) \tag{5.5}$$

where P(no ICOV edges) is the ratio of the triplets that have no ICOV edge to all possible triplets.

We use the robustness of the observed- to expected-overlap ratio (i.e., $N_{observed}/N_{expected}$) (when varying the binarization and Z-stat thresholds) as a performance index for finding the optimal thresholds (described above) on half of the data. Using this threshold for finding the extent of the correspondence on the other half of the data helps us avoid the over-fitting problem when finding the best binarization threshold. Figure 5.2 shows the diagram of the whole analysis (from time-series data to correspondence analysis).

5.2.5 Odds-ratio Analysis

The outcome of the Poisson regression described in Section 5.2.3 is only acceptance/rejection of the H0: whether or not the three-way interaction exists among X_1, X_2 and X_3 given their binary observations, which gives no information about the directionality of such interactions (if found). Having found a three-way interaction it is a natural next-step to question the underlying structure in more detail, e.g., if activation of X_1 increases the interaction between X_2 and X_3. Therefore, the odds ratio (OR), a descriptive statistic for the strength of association or non-independence among binary data values, is used for further analysis.

In statistics, odds are a way of presenting probabilities; the odds of an event happening is the probability that the event will happen divided by the probability that the event

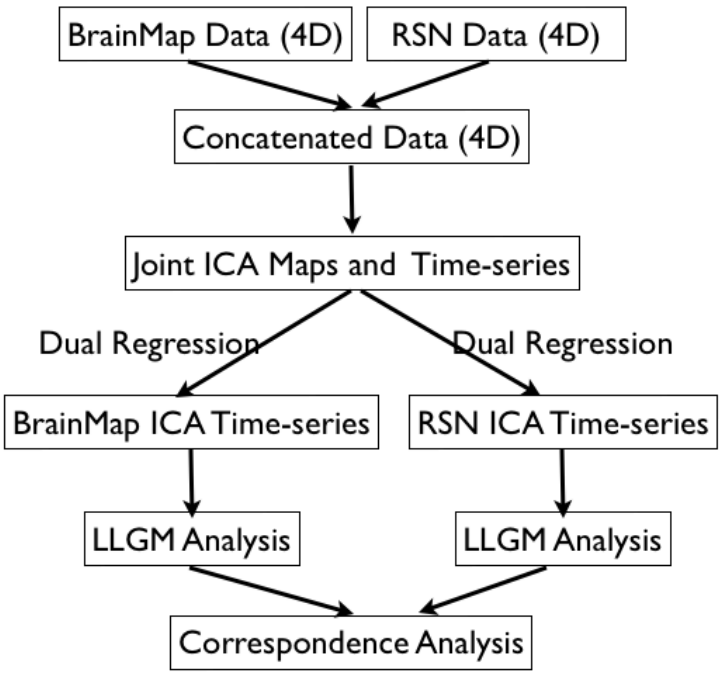


Figure 5.2: Summary diagram of the method, illustrating all the various steps involved in the analysis. Please see Appendix C.1 for more details description of some of the blocks.

will not happen. For example, the odds that a single throw of a die will produce a six are 1 to 5, or 0.2. The odds ratio (OR), however, as the name implies, is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. In recent years ORs have become widely used in medical reports as they provide an estimate (with confidence interval) for the relationship between two binary variables and enable researchers to examine the effects of other variables on that relationship, using logistic regression.

Using OR as a way of comparing whether the probability of a certain event is the same for two groups (i.e., a 2×2 table), an odds ratio of 1 implies that the event is equally likely in both groups; an odds ratio greater than one implies that the event is more likely in the first group while an odds ratio less than one implies that the event is less likely in the first group. For example, using the contingency table from Section 5.2.3 results in

$$\mathrm{OR}_{\mathrm{X}_1,\mathrm{X}_2} = \frac{n_{11}/(n_{11}+n_{12})}{n_{12}/(n_{11}+n_{12})} \Big/ \frac{n_{21}/(n_{21}+n_{22})}{n_{22}/(n_{21}+n_{22})} = n_{11}n_{22}/n_{12}n_{21}, \tag{5.6}$$

which is a symmetric measure, i.e., $\mathrm{OR}_{\mathrm{X_1,X_2}} = \mathrm{OR}_{\mathrm{X_2,X_1}}$. In a three-way table, as OR is defined for two variables, dependencies can be defined by $\mathrm{OR}_{\mathrm{X_1,X_2|X_3=OFF}}$, $\mathrm{OR}_{\mathrm{X_1,X_2|X_3=ON}}$ and so forth. Therefore, for each triplet identified as having a non-additive interaction, the underlying mechanism can be extracted by assessing how the change in the state of one variable can increase or decrease the interaction between the other two. In order to assess the three-way interaction, we defined $\mathrm{OR}_{\mathrm{X_1,X_2,X_3}} = \mathrm{OR}_{\mathrm{X_1,X_2|X_3=ON}}/\mathrm{OR}_{\mathrm{X_1,X_2|X_3=OFF}}$, which tests for existence of a three-way interaction.

5.3 Results

We first illustrate the results from the analysis of the simulated data. As we know *a priori*, there are two three-way interactions in the simulated network: node 4 modulating the strength of the 1-2 edge and node 5 modulating the strength of the 8-9 edge. Having each node's time-series binarised at the 90th percentile for each subject, calculating the P-values corresponding to existence of nonlinear interactions for every possible triplet (pooled across subjects), and keeping the ones that survive a false discovery rate (FDR) (Genovese et al., 2002) threshold of p <0.05, results in exact the two correct triplets of non-additive interaction. The subject-wise and pooled $\log_{10}(\mathrm{OR})$ and Z-stat results are shown in Figure 5.3, for one of the two interacting triplets. This figure shows that (1) how LLGM can find the non-additive interactions, and (2) how odds-ratio analysis can provide more details about the underlying interaction.

In the real data, the 125 components that are included in this analysis lie within cortical and subcortical grey matter, and have temporal/spectral characteristics typical for resting state networks. Figure 5.4 displays the spatial maps corresponding to these components with their indices that we will use in the rest of this section. Each of the components displayed in Figure 5.4 have 37 time-series associated with them: one for BrainMap and thirty-six for the thirty-six subjects' rFMRI. Therefore, for a given triplet of nodes, the acceptance or rejection of H0 (i.e., no three-way interaction) requires 37 LLGM analyses.

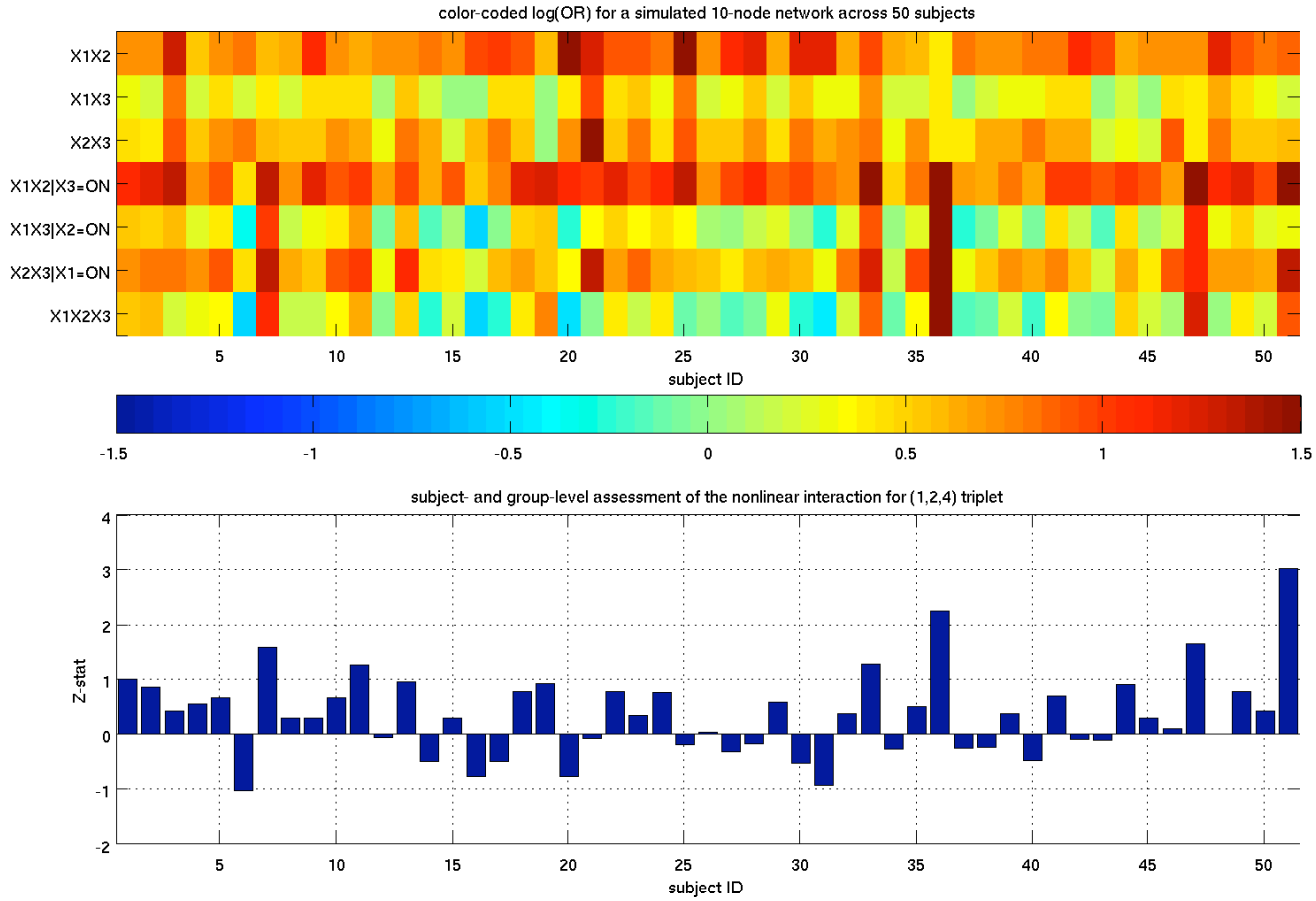


Figure 5.3: Subject-level and pooled $\log_{10}$(OR) and Z-stat for a triplet of nodes with non-additive interaction (X_1, X_2 and X_3 correspond to 1, 2 and 4 in Figure 5.1, respectively. The top row shows $\log_{10}$(OR) for two-way and three-way interactions (see Section 5.2.5 for details on how to interpret OR), while the bottom row shows Z-stats corresponding to three-way interactions. In both these rows, the x-axis is the subject index (1 to 50) with the last value on the x-axis (i.e., 51) being the pooled $\log_{10}$(OR) and Z-stat. These two plots imply that three-way interaction exists in this triplet, and the extent of this non-additive interaction varies across subjects. Also, $\log_{10}$(OR) for $X_1X_2|X_3$=ON shows that X_3 being ON increases the interaction between X_1 and X_2, which is consistent with the simulation ground truth.

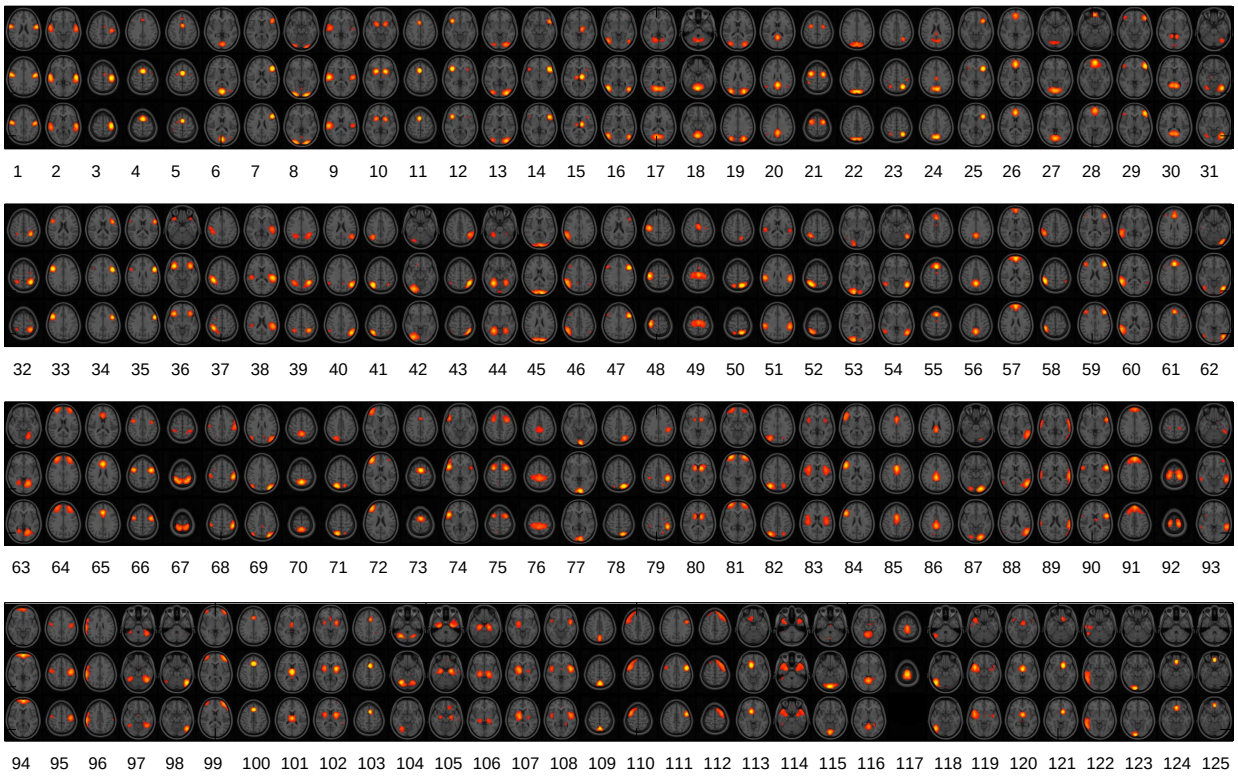


Figure 5.4: The spatial maps of the ICA components. Each column in each row illustrates three important (i.e., those closest to the statistic-maps' centre of gravity) axial slices of the spatial map corresponding to one of the functional nodes.

Figure 5.5 illustrates a sample triplet of nodes with each of its nodes' BrainMap and rFMRI time-series. In this triplet, X_1, X_2 and X_3 match Broca's area, a premotor region and the temporal pole, respectively. For each of these nodes, two time-series are plotted together with their raster plot (group of lines that represent the state in a binary (i.e., ON/OFF) series of events) when each time-series is binarised at the 90th percentile. The information in these raster plots is then used for forming the three-way contingency table.

Figure 5.5 shows just the first three subjects' rFMRI data and a subset of BrainMap studies with the same length (i.e., 528 samples). However, when feeding the data into the LLGM, the binary events in the whole length of rFMRI and BrainMap time-series are used for calculating each triplet's $\{z_{\text{BrainMap}}, z_{\text{rFMRI}}\}$ pair. For example, the triplet in Figure 5.5 results in $\{z_{\text{BrainMap}}, z_{\text{rFMRI}}\} = \{3.46, 4.92\}$, which implies the existence of a significant three-way interaction among its three nodes in both activation and rest. Applying a threshold to such pairs of Z-stats from every triplet results in a list of triplets that show a significant multi-way interaction in activation, rest or both, and hence can provide a measure of correspondence between activation and rest in terms of their functional-regions' multi-way interactions. Figure 5.6 displays counts for this overlap at a Z-stat threshold of 2.5 in its first column (as a function of the binarization threshold applied to rFMRI and BrainMap time-series).

Results in the second, third and fourth columns of Figure 5.6 show the $N_{\text{observed}}/N_{\text{expected}}$ ratio at various rFMRI and BrainMap binarization thresholds, with interaction significance thresholded at different Z-stat values (using half of the data - a "training" subset, with respect to the choice of thresholds). The visual investigation of the top plots implies that the use of 90th percentile for both rFMRI and BrainMap results in a decent correspondence between rFMRI and BrainMap over a range of Z-stat thresholds. Therefore, we advocate this threshold in this chapter for forming the contingency tables before carrying out the LLGM analysis on the other half of the data.

LLGM results in a group of triplets that measure the extent of three-way interactions when considering BrainMap, rFMRI or both. This extent can vary when going from one

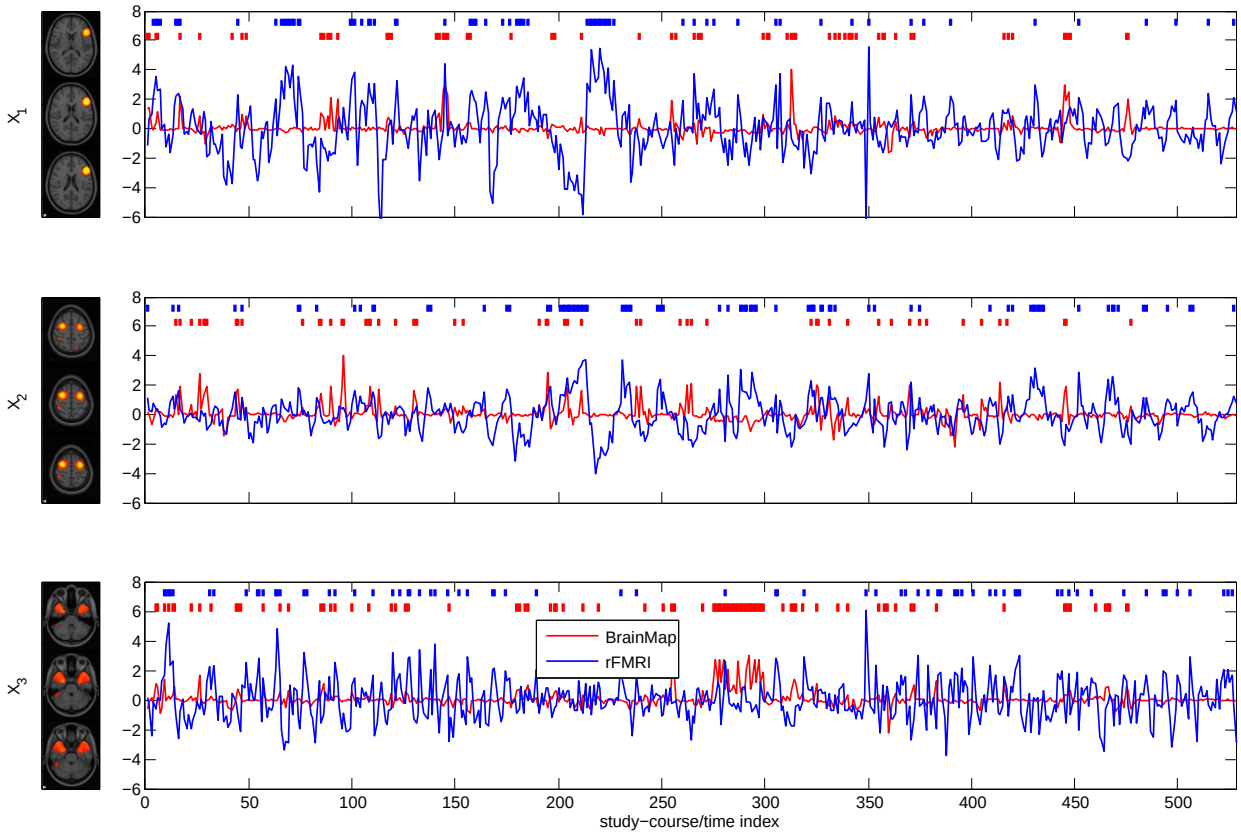


Figure 5.5: A sample triplet with (a subset of) its associated time-series. Each row corresponds to one functional node, which has its BrainMap and rFMRI time-series plotted in red and blue, respectively. Having these time-series binarised at their 90th percentile results in the shown raster plot (representing the state "ON" in a binary (i.e., ON/OFF) series of events). Using these binary time-series, contingency tables are formed whose cell counts are then fed into the LLGM analysis for assessing the H0: "no three-way interaction". For clarity, this figure only shows the first three subjects' rFMRI, i.e., 528 samples, and the first 528 studies from BrainMap.

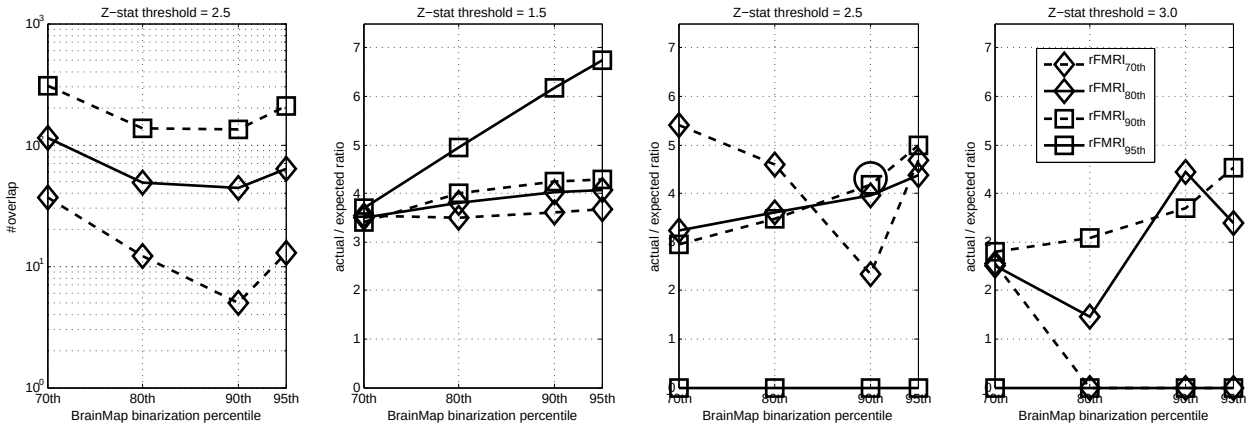


Figure 5.6: The extent of correspondence between the three-way interactions found in BrainMap and rFMRI over a range of binarization thresholds. This shows the correspondence in the first (i.e., training subset) half of the data. The first column (left) of the image displays the number of triplets surviving a Z-stat thresholding of 2.5 in both rFMRI and BrainMap ($N_{observed}$) at various binarization thresholds, with the x-axis representing the BrainMap binarization threshold and each line representing a different rFMRI threshold. The remaining columns display the overlap-to-expected ratio (i.e., $N_{observed}/N_{expected}$) of the number of survivors at Z-stat thresholds of 1.5, 2.5 and 3 (in second, third and fourth columns, respectively). Given that the 90th percentile results in a more reliable correspondence at various Z-stat thresholds, we chose and hence advocate this value as the binarization threshold for both BrainMap and rFMRI time-series. The extent of this correspondence at the chosen thresholds in the second half of the data (i.e., testing subset) is shown as a circle in the third column.

behavioural domain to another (in BrainMap) or from one subject to another (in rFMRI). Therefore, in order to assess such variations and behavioural-domain-specific interaction mechanisms, we analyzed the BrainMap data by dividing it into various behavioural domain groups (e.g., action-execution, working-memory, etc) and applying LLGM to each of them separately. A similar sub-data analysis is carried out for rFMRI where each subset corresponds to a subject; hence results are expected to show the inter-subject variability in the interactions. Results from these analyses are shown in Figures 5.7 and 5.8.

The results thus far show a striking correspondence found between the set of significant rFMRI and BrainMap triplets, extending the single-component matching results in Smith et al. (2009). From these plots, it can be concluded that corresponding non-additive/modulatory interactions exist (on average) in rest and activation, although the extent of which can vary under different interventions (i.e., different behavioural domains) and for different individuals.

In order to find the functional nodes that appear to be involved in the largest numbers of "matched" triplets (i.e., those with both Z_{rFMRI} and Z_{BrainMap} bigger than 2.5), we counted the number of matched triplets each node appears in, with results shown in Figure 5.9. This plot is interpreted as the areas with high modulatory effect between two other regions' functional connection. The descriptions assigned to each region are taken from standard atlases distributed with FSL, such as the Juelich cyto- and myelo-architectonic atlas, the Harvard-Oxford structural brain atlas and the MNI structural atlas. These often-seen modulatory areas are also mapped onto the behavioural domains (experimental paradigm classifications) in the BrainMap database. Given that the spatial maps have their associated time series that quantify their relevance to each of the original 7,342 BrainMap activation images, by multiplying the value at each time point by the corresponding behavioural domain(s) and averaging over all time points, we can derive a measure of how strongly each network relates to each behavioural domain. Each column is normalized to have a mean count of 1, to balance for different domains being represented

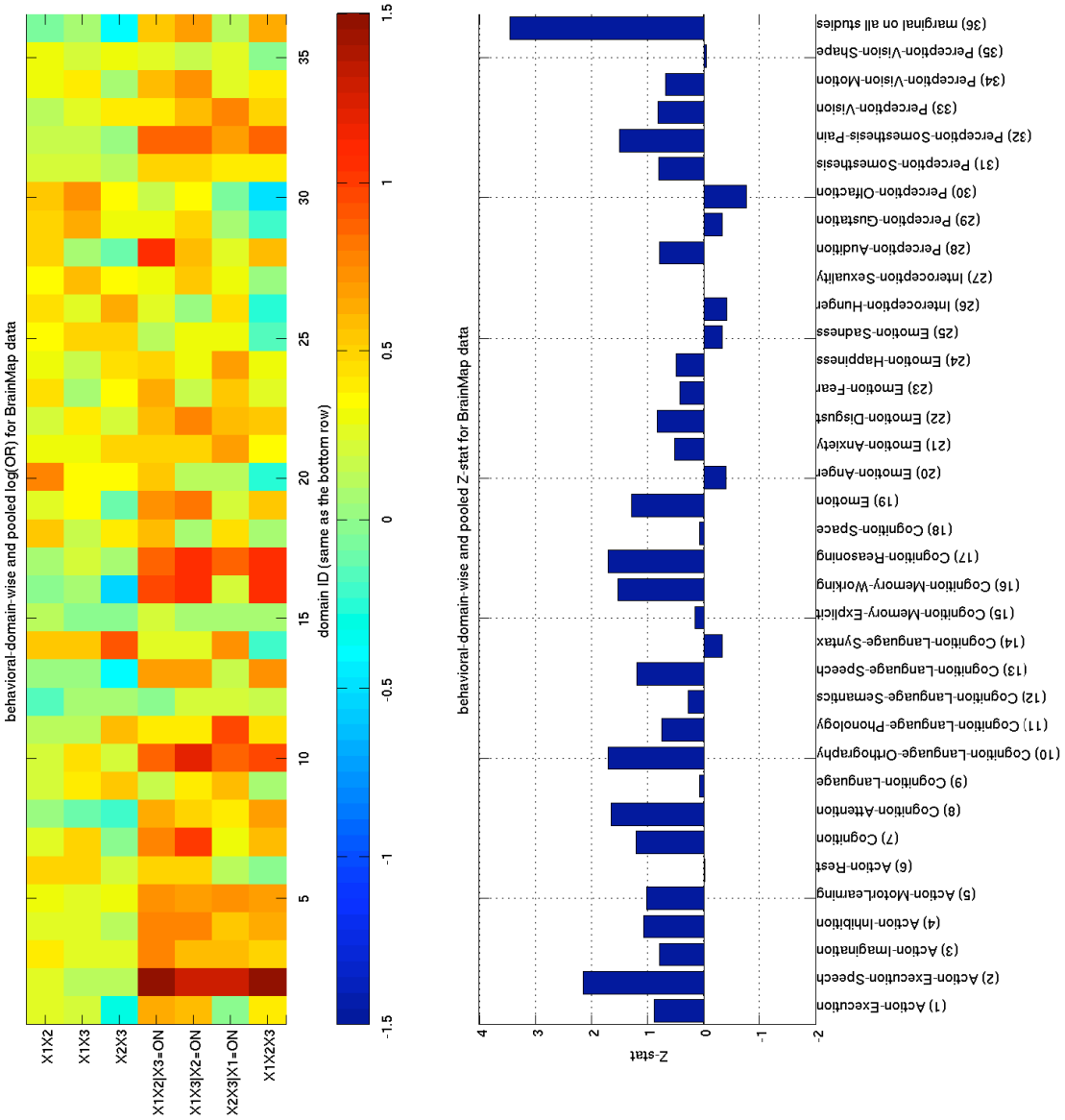


Figure 5.7: Assessing the inter-domain variability of three-way interactions in BrainMap data for the triplet shown in Figure 5.5. The bottom row displays the non-additive interaction's Z-stat when using the BrainMap 'time-series' from all or a subset of studies; the $\log_{10}(\text{OR})$ for the same data that generated the Z-stats can be found in the top row (see Section 5.2.5 for details on how to interpret OR). The $\log_{10}(\text{OR})$ is used for assessing the effect of one region on the other two regions' apparent causal connection. The strength of the three-way interaction varies across different behavioural domains, e.g., is high in action-execution-speech and action-motor-learning, and low in cognition-attention. When assessed on average ("marginal") across all behavioural domains (the bottom row) it still shows a large three-way interaction.

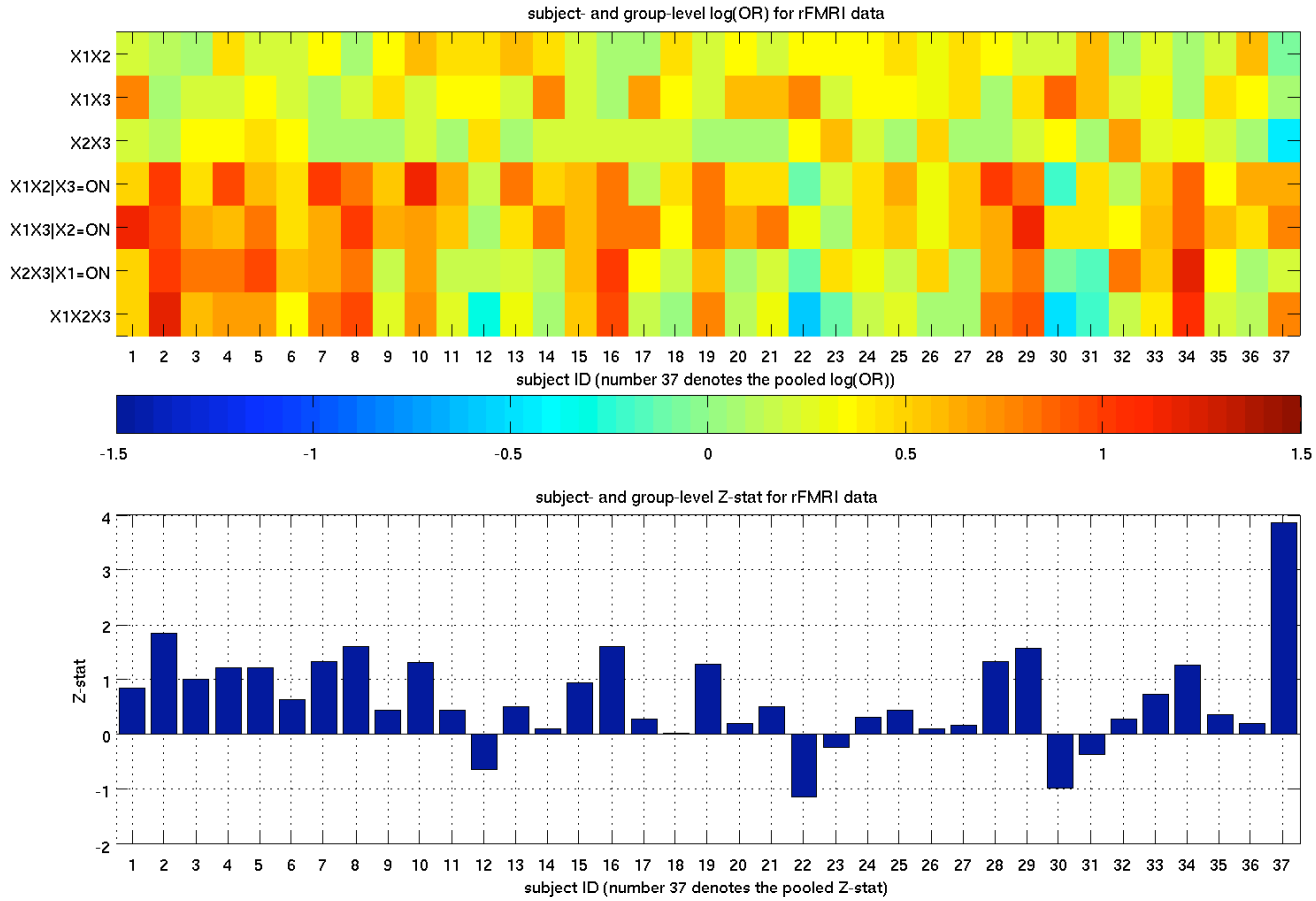


Figure 5.8: Assessing the inter-subject variability of three-way interactions in rFMRI data for the triplet shown in Figure 5.5. The bottom row displays the non-additive interaction's Z-stat when using the rFMRI time-series of a given subject; the $\log_{10}(\mathrm{OR})$ for the same data that generated the Z-stats can be found in the top row. The $\log_{10}(\mathrm{OR})$ is used for assessing the effect of one region on the other two regions' interaction. The extent of this three-way interaction varies across subjects (and/or sessions), e.g., being low in subject 14 and high in subjects 2 and subject 8. However, when all subjects' Z-stats are pooled (i.e., using a T-test) the result shows a significant three-way interaction (#37).

different numbers of times in the database.

5.4 Conclusions and Discussion

We have applied a multivariate exploratory method (ICA) *jointly* to meta-analytic task FMRI pseudo-images and rFMRI data in order to find functionally-distinct brain regions (i.e., nodes) and their interconnections. With over seven thousand functional contrasts and rFMRI from thirty-six subjects involved, we hoped for statistically powerful (i.e., small likelihood of type II error) and hence generalisable results. Modulatory effects between the nodes are found by using a log-linear graphical model, identifying just small-scale interactions, as we have found that network modelling using a high number of nodes is challenging in these datasets. In this approach, when one node modulates the functional connection between two others (i.e., a three-way interaction), LLGM assigns this triplet of nodes a significant Z-stat. The interactions are further analyzed by employing an odds-ratio analysis, which pinpoints the directionality of such interactions.

Our results confirm the existence of strong non-additive interactions among functionally-distinct brain regions in both activation and rest, with a striking correspondence. The extent of this correspondence, however, varies across different behavioural domains and in different individuals. The pattern of some such variations matches expectations, e.g., the premotor region has a stronger involvement in non-additive interactions in motor-domain (e.g., action-execution) studies.

The proposed method is capable of capturing the local graphical structure of multi-node interactions, i.e., the method can also be extended from a 3-way to an N-way interaction simply by using N variables to form the contingency table and LLGM. Given the interest in the application of graphical models to neuroimaging data, we expect LLGM to be capable of offering new solutions to this problem. This is in essence very similar to factor graphs (Kschischang et al., 2001) that are based on the idea that dealing with complicated global functions of many variables often exploit the manner in which the given functions factor as a product of "local" functions, each of which depends on a

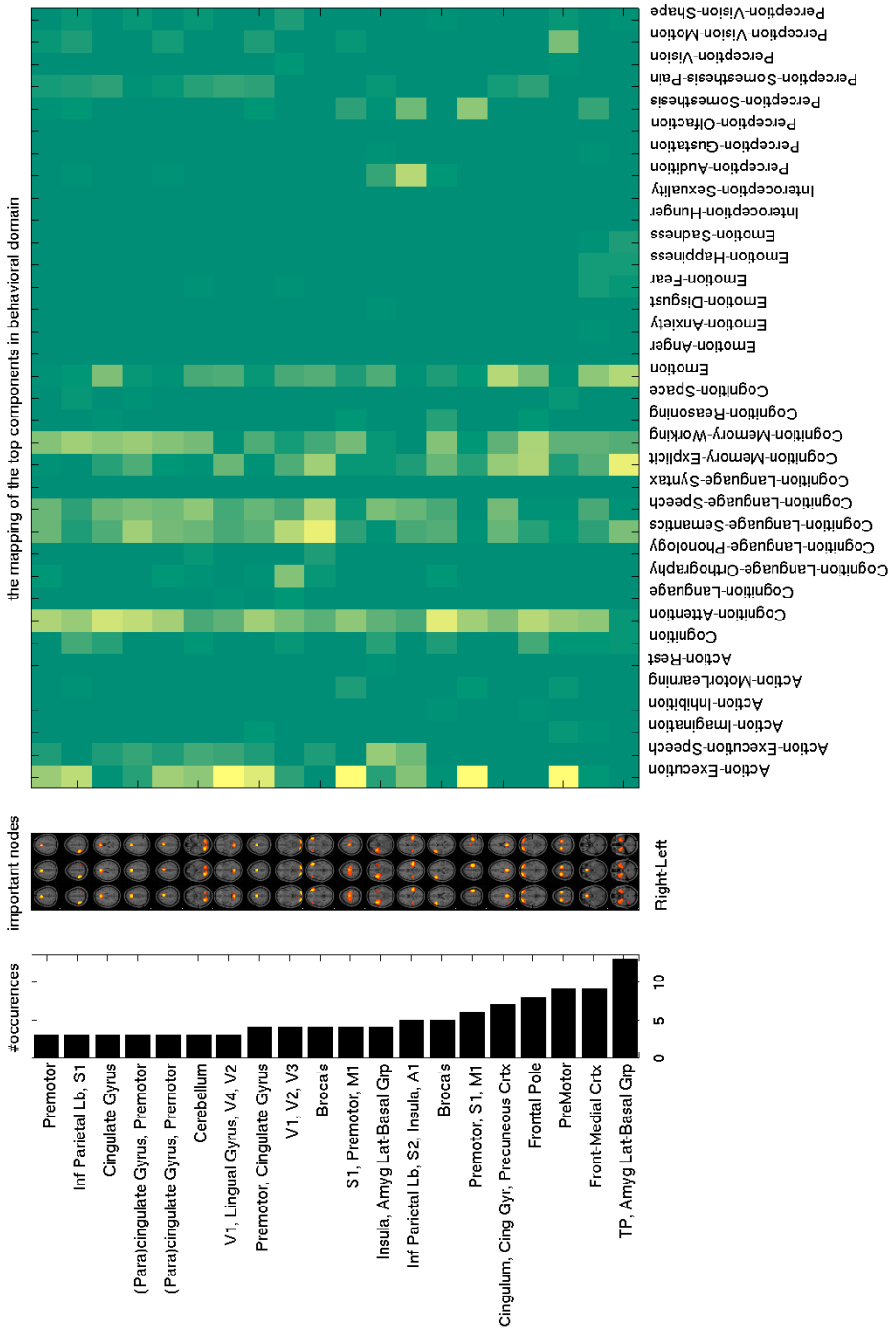


Figure 5.9: Regions with high frequency of appearance in three-way-interactive triplets. Left shows the number of matching triplets that a particular node appeared in, with each bar representing one node (the important slices for each of these nodes are shown in the middle panel). The effect of the behavioural domain on the multi-way interactions of each of these nodes is shown on the right, with its x-axis being the behavioural domain. Given that the spatial maps have their associated time series that quantify their relevance to each of the original 7,342 BrainMap activation images, by multiplying the value at each time point by the corresponding behavioural domain(s) and averaging over all 'time points', we can derive a measure of how strongly each network relates to each behavioural domain.

subset of the variables; visualizing such a factorization with a bipartite graph is called a factor graph.

When extracting a larger number of functional components from joint rFMRI-BrainMap data, we probe a different level in the hierarchy of functional networks and their subnetworks, which according to Smith et al. (2009) are in correspondence between activation networks and resting networks at "low" and "medium" levels of decomposition. Previously reported evidence on the existence of correspondence between such networks across different individuals, provides convincing evidence that, although the quality of our results (e.g., in terms of spatial detail and signal-to-noise ratio) is aided by having multiple subjects' resting datasets combined, the close matching of the rFMRI onto BrainMap in their modulatory structure is not an artifact of combining many subjects together (Smith et al., 2009).

An important issue to note when pooling sub-populations of data in order to achieve a population-level conclusion is the possibility of encountering Simpson's paradox (Blyth, 1972), in which a trend present in different groups is reversed when the groups are combined. This occurs when the sizes of the groups, which are combined in the presence of an ignored hidden variable (or confounding variable that is an extraneous variable in a statistical model that correlates with both the dependent and independent variables) are very different. Thus, like other paradoxes, it only appears to be a paradox because of incorrect assumptions, incomplete or misguided information, or a lack of understanding a particular concept. In this study, however, in spite of the existence of such a heterogeneous sample size (see Figure 5.10), using the appropriate models for pooling the sub-populations' Z-stats, we hope to have minimized the risk of this effect.

The quality of the described correspondence between the interactive-networks in BrainMap and rFMRI is particularly compelling given the fundamentally different nature of the two data-types jointly fed into the ICA analysis. For the resting FMRI analyses, we have data from just 36 subjects (compared with nearly 30,000 subjects in BrainMap), comprising just a few minutes' rFMRI data from each. For the BrainMap analyses, the

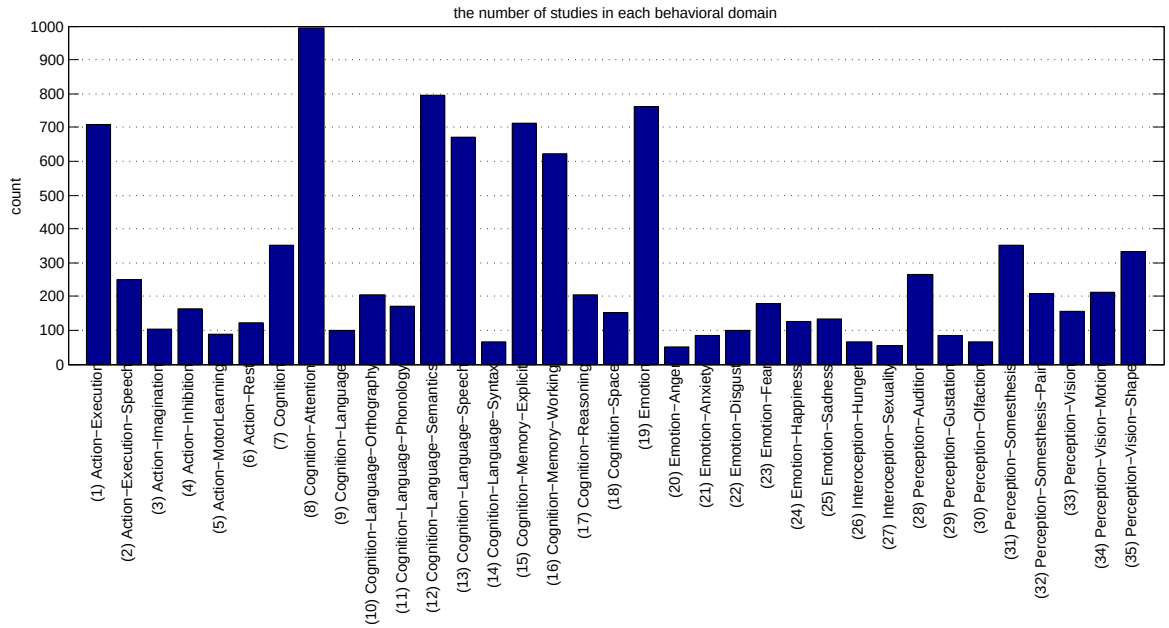


Figure 5.10: The number of studies in each sub-population of the BrainMap database. This plot implies the importance of taking the sample size (as a potential confound variable) into account when pooling the sub-populations for a more global inference.

pseudo-activation images have to be created purely from the coordinates of the peaks of just a few locations in the brain for each activation condition. In addition to not having the full richness of the original activation maps, BrainMap comprises results from across a broad range of imaging hardware types, data qualities, analysis software implementations, and task paradigm specifics. The fact that the major functional networks found from the resting and task domains match so closely indicates that BrainMap has been largely successful in its goal of objectively representing the brain's activation dynamics and hence can be a valuable resource providing neuroimaging researchers with useful data.

The LLGM employed in this study assumes that the samples (i.e., time-/study-points) are independent realizations from Poisson distributions. While this assumption may perfectly hold in the BrainMap case, for RSN data, however, we expect a temporal autocorrelation. Given the short-scale nature of such autocorrelations and that we binarise the time-series, we expect this issue to be less serious. Please note that the current results require further analysis and consideration for neuroscientific interpretations, which is our next step (i.e., future works).

Chapter 6

Conclusions and Future Works

This thesis was initiated by "input", "output", "model", "nonstationarity" and "approach" problems that existed in neuroimaging meta-analysis, and concluded by presenting several novel approaches for both coordinate- and image-based meta-analysis. Given the growing interest in such meta-analyses in the field of neuroimaging, any new development that can result in more accurate inference is of great value. In addition to the described improvements, evaluation of the techniques that are reported in this thesis (e.g., GPR, ALE and KDA) sheds light on the strengths and weaknesses of each of them. Such information can provide meta-analysis researchers with awareness about the issues that should be taken care of prior to reporting the results or/and even carrying out the analysis. In the rest of this chapter, we briefly present a list of our contributions, introduce a series of issues where neuroimaging meta-analysis still requires further developments and conclude the thesis with a brief overall summary.

6.1 Summary of Contributions

In the introduction chapter, we introduced a list of problems that neuroimaging meta-analysis was to address. The focus of the rest of this thesis was to present solutions for overcoming these problems in order to enable neuroimaging meta-analysis to result in more accurate inference. Below, an itemized list of these solutions/contributions is given.

Image-based meta-analysis: As the first step in our approach toward meta-analysis, we attempted to encourage the field of neuroimaging to start sharing the full-image data (i.e., volumes of effect-size estimates and their uncertainties) for future meta-analyses. In order to make such a data sharing as simple as possible, we introduced a hierarchical Bayesian model that *only* requires the voxel-wise effect-size estimates and their uncertainties. It was previously shown that passing such summary statistics (i.e., sufficient statistics) from lower levels of this hierarchy to their next levels is equal to fitting a large single-layer model to the lower-level time series. This approach provides the meta-analysis with uncensored data and offers the richness of the possible models that can be employed for the meta-analytic pooling and inference. The results of such IBMA were shown to be more accurate than those from incorporating coordinate summaries (i.e., CBMA).

Adjusting the image-based inference for nonstationarity: In order to overcome the nonstationarity problem in cluster-related inference, we next introduced a new empirical approach for adjusting two cluster-related inferences: cluster-size and threshold-free cluster-enhancement (TFCE) and provided an extensive review and comparison of all possible existing approaches in the field. Our results on both simulated and real data show that our empirical approach (when applied to both cluster-size and TFCE statistics) is better than or at least as good as existing solutions that are based on random-field theory.

Coordinate-based meta-analysis using Gaussian-process regression: Given that the input to neuroimaging meta-analysis is limited to only the (x,y,z) coordinates of activation peaks, and how the existing CBMA approaches suffer from a series of weaknesses, we next introduced an appropriate nonparametric CBMA model. Assuming the statistical landscape being a Gaussian process with (de)activation foci as random samples drawn from it, our approach is simply a Gaussian-process regression that starts by learning its hyperparameters from the observations and then attempts to predict the

effect-size values in the rest of the image. Our results on both simulated and real data show that GPR outperforms the existing CBMA techniques and is capable of addressing the problems that existing CBMA techniques suffer from.

Joint analysis of FMRI time-series and BrainMap data: In addition to the new ways of carrying out a coordinate- or image-based meta-analysis, we introduced a new approach possibility: joint analysis of both BrainMap and FMRI data in order to pinpoint the functional networks of the brain. Given the striking correspondence found between the covarying components in activation and rest, such analyses can provide an answer to the next question: is there a correspondence between the resting and active brain in terms of their network structure. Note that, generalizing such concatenations to slightly different problems requires careful considerations (e.g., compatibility of the datasets and preprocessing applied to each one).

Network modelling: Network modelling using BrainMap falls at the intersection of two big trends in the field of neuroimaging: (1) meta-analysis and (2) functional-connectivity analysis. However, extracting the accurate structure of such networks becomes more difficult when the number of nodes becomes larger. Therefore, we approached this problem in a slightly different way by using a different class of such models (i.e., log-linear graphical models) for a reformulated problem (i.e., finding triplets of functional nodes which appear to interact in such a way that one node modulates the functional connection between the other two).

Neuroscience application: The aim of this research is to develop statistical models and tools for modelling and analysis of the neuroimaging data. Therefore, the proposed methods were tested on various neuroscientific applications such as pain perception (image- and coordinate-based meta-analysis, and network modelling), neuro-degenerative diseases (e.g., voxel-based morphometry that is adjusted for nonstationarity), and resting brain (network modelling and source extraction). In all these applications, we introduced

methods that show a better performance than (or are equally as good as) the ones being currently used by neuroimaging researchers.

6.2 Problems to Overcome and Future Directions

In spite of the improvements that this research has provided for meta-analysis research in neuroimaging, there are still things that require further improvement. The traditional CBMA methods were only capable of incorporating the (x,y,z) coordinates of the peaks. We took CBMA one step forward by enabling it to use more information such as the effect sizes. However, there are still other pieces of information in the literature than can be incorporated in the CBMA models. For example, given the cluster size in a cluster-based inference along with the cluster-forming threshold and local smoothness estimate, random-field theory can provide an estimate for the shape of the landscape around a cluster's maximum. Such information is provided in papers that carry out cluster-based inference.

In the meta-analysis literature, it is strongly recommended to carry out a bias and heterogeneity assessment prior to the actual pooling. Bias and heterogeneity can be caused by differences between the constituent studies and as a result of the impact of omitted studies. Even the best-designed meta-analyses will have differences in subject population, interventions, outcome definition, and study design. There are various graphical tools and statistical tests to assess the issues of heterogeneity and bias in neuroimaging meta-analyses (e.g., funnel plot, Galbraith plot and forest plots). One of the reasons why such tools were not used prior to coordinate-based analyses is not having access to study-level effect sizes and their standard deviations, which are required for plotting such graphs. With our proposed IBMA approach and GPR CBMA, however, such tools can be employed for a more appropriate meta-analysis, and hence an assessment of their performance needs to be reported by future research.

In all CBMA methods, including our GPR, the reported coordinates are taken for granted and assumed to carry zero uncertainty. Our GPR model does not explicitly model

the uncertainty in the location of the foci. However, such uncertainty is an implicit feature of the GPR model; as it assumes that the nearby voxels have similar values, modelling the uncertainty of the target values indirectly has a location uncertainty implication in it. In order to have such uncertainties explicitly modeled, for example, a new layer can be added to the meta-analytic models such as GPR, which models the foci as samples from a spatial point process. After this development, both coordinates and effect-sizes, and the uncertainty associated with them are explicitly modeled and hence can make detailed interpretation of the result possible.

Our graphical modelling approach overcomes some of the problems encountered in modelling large brain networks. However, there are still many issues that we believe necessitate customized network modelling. For instance, our results indicate that there are modulatory structures in the brain where one node changes the strength of the other nodes' interaction; this is not reflected in the existing models. One example solution would be to use factor graphs, while there are variety of other possible more customized models that can be developed for this particular application.

Another important point that is missing in graphical modelling of the brain's functional networks, is the fact that there is no single network that can explain the brain's function. That is, most of the graphical models that are employed for the analysis of functional connectivity are not "dynamic"; even those models that are known as dynamic (e.g., dynamic Bayesian networks) are no more than two static networks. One solution for describing the dynamic behavior of the brain's functional time-series is the use of "mixture of stochastic processes", e.g., mixture of Gaussian processes with temporal covariance structures. Under these models, each time-series at a time can belong to a GP group with a mean characteristic time-series. Also, instead of a group of static network structure, one can employ a temporal covariance for the time-series that captures the dynamical behavior.

6.3 Final Conclusions

Meta-analysis is important for pooling data from under-powered studies such as many of the ones in the field of neuroimaging, which justifies its increasing popularity in neuroimaging. However, due to the spatiotemporal characteristic of the neuroimaging data, traditional meta-analyses cannot be employed without customization. We addressed a series of such customizations by introducing a flexible hierarchical mixed-effects model which can take advantage of the sensitivity that cluster-related inferences provide. Additionally, we provide a series of adjustment procedures that can overcome the problem of spatial heterogeneity of smoothness (i.e., nonstationarity) when carrying out a cluster-related inference. Due to the nature of the neuroimaging data and the sparseness of its corresponding shared information, however, we need to take the traditional univariate meta-analysis to a different level by adding the spatial covariance to it. We tackled this problem by introducing Gaussian-process regression, which lies under the same hierarchical principles as the initial model we started with (i.e., IBMA) and hence can provide very similar outputs (i.e., voxel-wise estimates for effect-size mean and standard deviation). Additionally, we provided a neuroimaging meta-analysis research with a new approach toward meta-analysis: meta-analytic functional connectivity. In summary, we expect this research to be of neuroimaging researchers' great interest and provide them with some useful solutions.

Appendix A

Image-based meta-analysis

A.1 Pseudo-code for ALE Method

Activation Likelihood Estimation (ALE) Pseudo-code
begin
  for *all foci in the list*
  begin
    for *all voxels in the brain*
      calculate the activation likelihood using the Gaussian kernel
  end
  for *a necessary number of times (i.e., 10000 times)*
  begin
    generate a group of randomly located foci (uniformly distributed over whole gray matter)
    for *all these fake foci*
    begin
      for *all voxels in the brain*
        calculate the activation likelihood using the Gaussian kernel
    end
  end
  test the real AL map wrt AL maps from Monte Carlo approach
end

A.2 Pseudo-code for KDA Method

Kernel Density Approximation (KDA) Pseudo-code
begin
  for *all foci in the list*
  begin
    for *all voxels in the brain closer than a radius to each foci*
      increment voxel's intensity(with step of one)
  end
  for *a necessary number of times (i.e., 10000 times)*
  begin
    generate a group of randomly located foci (uniformly distributed over whole gray matter)

```
    for all these fake foci
    begin
      for all voxels in the brain
        increment voxel's intensity(with step of one)
    end
  end
  test the real density map wrt density maps from Monte Carlo approach
end
```

A.3 Pseudo-code for KDA Method

```
Multi-level Kernel Density Analysis (KDA) Pseudo-code
begin
  for all studies
  begin
    generate the KDA map
    make comparison indicator maps (CIM) by binarising the map wrt 0.5
  end
  weight and average CIMs to make the proportion of study comparison maps (PSCM)
  for a necessary number of times (i.e., 10000 times)
  begin
    for all CIMs
    begin
      extract the blobs
      move the centers of blobs randomly (Monte Carlo part)
    end
    make the PSCM maps
  end
  test the real PSCM map map wrt distribution of Monte Carlo-generated PSCM maps
end
```

Appendix B

Nonstationarity

B.1 Estimation of Smoothness/Roughness

There are two methods for estimating the smoothness of the component fields available. One is directly motivated by the definition of Λ and is based on the variance of the partial derivatives of the residuals. The other uses the assumption that the spatial autocorrelation function has the form of a Gaussian density, which leads to an estimate based on sample correlations of adjacent voxels. We demonstrate that both methods can be seen to have additive contributions from each residual image, and so allow smoothness estimation to be based on only a subset of the images.

B.1.1 Kiebel's Method

Using random field theory (RFT) concepts, Kiebel et al. (1999) propose an unbiased estimator for the covariance of the partial derivatives at voxel i in a D-dimensional Gaussian random field as

$$\lambda_{i,d} = \frac{\nu-2}{\nu-1} \cdot \frac{1}{M} \sum_{t=1}^{M} \left(\frac{\partial \hat{S}_{i,t}}{\partial x_d} \right)^2 \tag{B.1}$$

where $\hat{S}$ is the standardized estimated residual, ν is the number of degrees of freedom, and M is the number of observations (time points or subjects). In Eq. B.1, x_d represents three main directions of the image (x, y, and z) at which the partial derivative is calculated via the gradient operator ($\nabla \hat{S}_i = \frac{\hat{S}_{i+} - \hat{S}_i}{\delta d}$) and used to estimate the voxel-wise roughness

measures as

$$\begin{aligned} RPV_{i,d} &= (8 \cdot \ln(2))^{-1/2} \cdot (2\lambda_{i,d})^{1/2} \\ RPV_i &= \prod_{d=1}^{D} RPV_{i,d}. \end{aligned} \tag{B.2}$$

where $\sigma_i = (8 \cdot \ln(2))^{-1/2} RPV_i^{-1}$ and $FWHM_i = (8 \cdot \ln(2))^{1/2} \sigma_i$.

As an alternative, a robust two-step estimate can be made by using just the "control group", or using all observations after excluding the outlier observations. For the latter estimate, using

$$\lambda_{i,d,t} = \frac{\nu - 2}{\nu - 1} \cdot \left(\frac{\partial \hat{S}_{i,t}}{\partial x_d} \right)^2 \tag{B.3}$$

in the first step (instead of Eq. B.1) results in a voxel-wise array of FWHM estimates (one per observation). Truncating both tails of the FWHM histogram at each voxel and keeping the remaining observations results in a voxel-wise list (l) of L observations, which changes Eq. B.1 to

$$\lambda_{i,d} = \frac{\nu - 2}{\nu - 1} \cdot \frac{1}{L} \sum_{t \in l} \left(\frac{\partial \hat{S}_{i,t}}{\partial x_d} \right)^2 \tag{B.4}$$

in the second step of the calculations, which is to be followed by Eq. B.2 for the robust estimation.

B.1.2 Jenkinson's Method

As the smoothing filter's width decreases, the gradient operator and hence Kiebel's estimator increasingly become inaccurate. An alternative estimate can be established based on the correlation of neighbouring voxels (together with a Gaussian autocorrelation function assumption) for a more robust estimate (Flitney and Jenkinson, 2000; Nichols, 2008). Using

$$\begin{aligned} \hat{S}_i^2 &= \frac{1}{M} \sum_{t=1}^{M} \hat{S}_{i,t}^2 \\ \hat{SS}_{i,d} &= \frac{1}{M} \sum_{t=1}^{M} \left(\hat{S}_{i+_d,t} \cdot \hat{S}_{i,t} \right) \end{aligned} \tag{B.5}$$

results in an estimation for voxel i's autocorrelation ($\hat{S}_i^2$) as well as its cross-correlation with its next voxel in the d direction ($\hat{SS}_{i,d}$). Using these estimations in

$$\sigma_{i,d}^2 = \left(4 \cdot \ln\left(\frac{\hat{S}_i^2}{\hat{SS}_{i,d}}\right)\right)^{-1} \tag{B.6}$$

followed by Eq. B.2 results in an estimation for smoothness. Similar to Kiebel's method, using only the control group or excluding the outlier observations is expected to result in a more robust smoothness estimation. In order to find the outlier observations,

$$\begin{aligned} \hat{SS}_{i,t}^2 &= \hat{S}_{i,t}^2 \\ \hat{SS}_{i,d,t} &= \hat{S}_{i+_d,t} \cdot \hat{S}_{i,t} \end{aligned} \tag{B.7}$$

replaces the Eq. B.5 in the first step of the calculations in order to result in a voxel-wise array of FWHM estimates. Truncating both tails of the FWHM histogram at each voxel results in a list (l) of L observations, which changes Eq. B.5 to

$$\begin{aligned} \hat{S}_i^2 &= \frac{1}{L}\sum_{t\in l} \hat{S}_{i,t}^2 \\ \hat{SS}_{i,d} &= \frac{1}{L}\sum_{t\in l} \left(\hat{S}_{i+_d,t} \cdot \hat{S}_{i,t}\right) \end{aligned} \tag{B.8}$$

in the second step of the calculations, which is to be followed by Eq. B.2 for the robust estimation. In this study, 15% of the observations on each tail of the histogram are labelled as outliers.

B.2 Validity of Empirical Cluster Size Adjustment

Our "2-pass" nonstationarity-adjustment method for cluster and TFCE inference is a resampling-based estimation of a nuisance variable, followed by a permutation test that incorporates that estimated nuisance. In this appendix we precisely define the procedure (for cluster statistic, which is generalizable for TFCE statistic) and justify its use despite the apparent "double dipping" or double-use of the data.

The cluster-size statistic is affected by both nuisance variation, due to variable smoothness, as well as inflation of cluster size due to true signal. Specifically, under

the null hypothesis we can write the mean size of a cluster occurring at voxel i as

$$E(S_i|S_i > 0, H_0) = \gamma_i \tag{B.9}$$

and when there is a signal present,

$$E(S_i|S_i > 0) = \mu_i\gamma_i, \tag{B.10}$$

where we have conditioned on $S_i > 0$ (since we are only interested in the case when a cluster actually occurs at i). ECSPV_i (Eq. 3.10) is our resampling-based estimator of γ_i; simply, the empirical mean cluster size over null hypothesis samples of the image (with the outlier/skew adjustment parameter E (=2/3 for example)).

First consider adjusting the size of an individual cluster; a randomly selected cluster comprising of voxels $\mathcal{S}$ and size statistics S_i. If γ_i was known, the test statistic S_i/γ_i would be perfectly adjusted for nonstationarity. However, we would no longer have a cluster-size test, as voxel-wise $\{S_i/\gamma_i\}$ comprises an image and doesn't tell us how to make inference on any particular given cluster.

For a cluster comprising of voxels $\mathcal{S}$, there are two possible adjustment strategies: to normalise cluster size S (a cluster-wise approach) or to normalise each voxel in $\mathcal{S}$ (a voxel-wise approach). To adjust cluster size S using the former approach, we could compute $\bar{\gamma}_{\mathcal{S}} = \sum_{i\in\mathcal{S}} \gamma_i/S$, and use test statistic $S^{\text{clus}}_{\text{adj}} = S/\bar{\gamma}_{\mathcal{C}}$. Alternatively, we could adjust each voxel to create a test statistic $S^{\text{vox}}_{\text{adj}} = \sum_{i\in\mathcal{S}} 1/\gamma_i$ (instead of unadjusted "$S = \sum_{i\in\mathcal{S}} 1$" statistic). When a cluster arises in a homogeneous region, both a cluster-wise and voxel-wise approach should give the same adjusted value. When there is heterogeneity within the cluster, the voxel-based adjustment should give better correction as each voxel is considered individually. For example, if a single voxel in a cluster had an arbitrarily large γ_i, the average $\bar{\gamma}_{\mathcal{C}}$ could be dramatically altered while the sum over $1/\gamma_i$ would be negligibly affected.

Replacing γ_i with ECSPV_i, we in fact considered both approaches. However we found that the voxel-wise approach of adjusting individual voxels consistently produced similar or better results, and so we only considered the voxel-wise approach in this work.

Null Hypothesis Validity The estimator ECSPV_i is computed from "first pass" permutation of the data, and then is fixed upon the subsequent "second pass" permutation test. At any one voxel, this "double use" of the data cannot invalidate a local (uncorrected) permutation test, as the ECSPV-based adjustment is fixed over all permutations. Over space, one *could* imagine invalid two pass procedures, such as one suggested by a reviewer: In the first pass, find the location i^* of the maximum statistic, and for the second pass use a statistic that zeros all voxels except i^*. However, that example actually uses the original data, permutation $k = 1$ in our notation. Our procedure only uses resampled data $k > 1$, and it is not evident how to construct such a pathological procedure without using the original data. In as much as our procedure *reduces* the heterogeneity of the statistic from an image of $\{\gamma_i\}$ to a constant value on average, it will not produce any increase in false positive rates.

Alternative Hypothesis Sensitivity Under the alternative $\mu_i > 1$ and for our method to succeed we need to argue that adjustments made with ECSPV do not cancel out the signal. Again, since we do not use the original data, permutation should eliminate any signal present in the data, rendering ECSPV_i an accurate estimate of γ_i. However, when the signal is very strong, it is possible for the effect to "leak" into the permutation distribution. That is, some permutations will consist of nearly the original dataset, plus only a minor perturbation. In this case ECSPV_i could be overestimated and degrade power. We checked our true signal simulations and found that, for very high signal magnitudes, the images of ECSPV_i faintly showed the location of the signal. Using an alternate permutation method (Ter Braak, 1992), which uses permutation of full model (instead of null model) residuals, eliminated this "leakage" problem. Changing from the standard permutation method (Freedman and Lane, 1983) to the alternate method, however, had negligible impact on the real data analysis; this is not surprising, since we only observed the leakage at extremely high signal magnitudes.

B.3 Storing the Empirical Statistics

Since the number of permutations is not extremely large, it is possible to store each and every cluster size across the permutations. However, in order to overcome the computational problem in storing the TFCE scores of all voxels across all permutations (a matrix of size 'number-of-voxels' × 'number-of-permutations'), instead of saving each and every TFCE score, they are stored in a histogram format, i.e., an intensity- down-sampled version of the data. TFCE scores have an arbitrary scale but typically range into the thousands, while RFT-adjusted TFCE scores (TFCE_{RFT}) are quite small. Thus we use a unit bin size for raw TFCE scores and a 0.001 bin size for TFCE_{RFT}.

B.4 Pseudo-codes

In order to describe the methodological issues employed in this paper, they are illustrated in a pseudo-code format as follows. In order to assess each method's performance in terms of their statistical sensitivity and power, Algorithm 2 is used to simulate the nonstationary data for the 'Signal+Noise vs. Noise' analysis, i.e., two groups of subjects. This data is then analyzed by different RFT-based and empirical adjustment techniques whose performance is then quantified by an area under the ROC curve analysis. The nonstationary null data can be thought of as a special case of this simulation where SNR=0, with σ values of 2/3/4, 2/4/3, 3/2/4, 3/4/2, 4/2/3 and 4/3/2. Note that σ_1, σ_2 and σ_3 are the three numbers in σ, i.e., $\sigma_1/\sigma_2/\sigma_3$.

Algorithm 2 Simulation of the Nonstationary Data for 'Signal+Noise vs. Noise' Analysis

for realization=1 to 50 **do**
  noise_img ← randn(40,150,150,150)
  for i=1, 2 and 3 **do**
    $\text{noise\_img}_{\sigma_i}$ ← smooth(noise_img, σ_i)
    $\text{noise\_img}_{\sigma_i}$ ← $\text{noise\_img}_{\sigma_i}$/SD($\text{noise\_img}_{\sigma_i}$)
  end for
  for σ=2/3/4, 2/4/3 and 3/2/4 **do**
    $\text{nonstat\_noise\_img}_{\sigma}$ ← combine($\text{noise\_img}_{\sigma_1}$,$\text{noise\_img}_{\sigma_2}$,$\text{noise\_img}_{\sigma_3}$)
    $\text{nonstat\_noise\_img}_{\sigma}$ ← smooth($\text{nonstat\_noise\_img}_{\sigma}$, 1.5)
    $\text{nonstat\_noise\_img}_{\sigma}$ ← $\text{nonstat\_noise\_img}_{\sigma}$/SD($\text{nonstat\_noise\_img}_{\sigma}$)
    for SNR = 6 and 9 **do**
      for subject=1 to 10 **do**
        $\text{nonstat\_noise\_img}_{\sigma}$(subject) ← $\text{nonstat\_noise\_img}_{\sigma}$(subject) + SNR× signal
      end for
      for subject=11 to 20 **do**
        $\text{nonstat\_noise\_img}_{\sigma}$(subject) ← $\text{nonstat\_noise\_img}_{\sigma}$(subject)
      end for
      $\text{nonstat\_noise\_img}_{\sigma}$ ← exclude_boundary_voxels($\text{nonstat\_noise\_img}_{\sigma}$)
      for all adjustment methods **do**
        output ← randomise($\text{nonstat\_noise\_img}_{\sigma}$, adjustment_method)
        $\text{AUC}_{\text{realization, method, SNR, } \sigma}$ ← ROC_analysis(output, signal)
      end for
    end for
  end for
end for

Appendix C

Triplet Analysis

C.1 Analysis Procedure

Algorithm 3 Summary of the rFMRI-BrainMap LLGM Analysis

rFMRI←*preprocess*(rFMRI)
BrainMap←*preprocess*(BrainMap)
JointData←*ScaleConcatenateAndMakeJoint2D*(BrainMap, rFMRI)
(ComponentsMaps, ComponentsTimeSeries)←*ICAto150Components*(JointData)
(ComponentsMaps, BrainMapTimeSeries, $\text{rFMRITimeSeries}_{\text{subject}=1,\ldots,36}$)
←*DualRegression*(BrainMap, rFMRI, ComponentsMaps, ComponentsTimeSeries)

for triplet = 1 TO 317750 **do**

  (Xo_1,Xo_2,Xo_3)←ExtractTrivariateTimeSeries(rFMRITimeSeries, triplet)
  for subject = 1 TO 36 **do**
    (X_1,X_2,X_3)←GetThisSubjectsChunck(Xo_1,Xo_2,Xo_3, subject)
    (X_1,X_2,X_3)←BinarizeAtPercentileThreshold(X_1,X_2,X_3, percentile)
    $\text{Y}_{\text{CellCounts}}$←FormTheContingencyTableAndGiveTheCounts(X_1,X_2,X_3)
    Z_{rFMRI}(subject)← ApplyLLGM($\text{Y}_{\text{CellCounts}}$, X_{simple}, $\text{X}_{\text{saturated}}$)
  end for
  $\text{Zvector}_{\text{rFMRI}}$(triplet)←OneSampleTtest(Z_{rFMRI})

  (X_1,X_2,X_3)←ExtractTrivariateTimeSeries(BrainMapTimeSeries,triplet)
  (X_1,X_2,X_3)←BinarizeAtPercentileThreshold(X_1,X_2,X_3, percentile)
  $\text{Y}_{\text{CellCounts}}$←FormTheContingencyTableAndGiveTheCounts(X_1,X_2,X_3)
  $\text{Zvector}_{\text{BrainMap}}$(triplet)← ApplyLLGM($\text{Y}_{\text{CellCounts}}$, X_{simple}, $\text{X}_{\text{saturated}}$)

end for

Bibliography

A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, New York, 2002.

E. Amaro and G. Barker. Study design in fmri: Basic principles. *Brain and Cognition*, 60(3):220–32, 2006.

J. Andersson, S. Smith, and M. Jenkinson. Non-linear optimisation. Internal Technical Report TR07JA1, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, Oxford Univ, Oxford, UK, available at www.fmrib.ox.ac.uk/analysis/techrep, 2007.

M. Avriel. *Nonlinear Programming: Analysis and Methods*. Dover Publishing, 2003.

L. Becerra, S. Morris, S. Bazes, R. Gostic, S. Sherman, J. Gostic, G. Pendse, E. Moulton, S. Scrivani, D. Keith, B. Chizh, and D. Borsook. Trigeminal neuropathic pain alters responses in cns circuits to mechanical (brush) and thermal (cold and heat) stimuli. *J Neurosci*, 26(42):10646–57, 2006.

C. Beckmann, C. Mackay, N. Filippini, and S. Smith. Group comparison of resting-state fmri data using multi-subject ica and previous termdual regression. In *15th Annual Meeting of Organization for Human Brain Mapping, poster 441 SU-AM*, 2009.

C. F. Beckmann, M. Jenkinson, and S. M. Smith. General multilevel linear modeling for group analysis in fMRI. *NeuroImage*, 20(2):1052–63, 2003.

C. F. Beckmann, M. DeLuca, J. Devlin, and S. M. Smith. Investigations into resting-state connectivity using independent component analysis. *Philos Trans R Soc Lond B Biol Sci.*, 360(1457):1001–13, 2005.

C. R. Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–66, 1972.

D. Borsook, E. A. Moulton, S. Tully, J. D. Schmahmann, and L. Becerra. Human cerebellar responses to brush and heat stimuli in healthy and neuropathic pain subjects. *Cerebellum*, pages 1–21, April 2008.

E. Bullmore, J. Suckling, S. Overmeyer, S. Rabe-Hesketh, E. Taylor, and M. Brammer. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans Med Imaging*, 18:32–42, 1999.

P. Bunch, J. Hamilton, G. Sanderson, and J. Hamilton. A free response approach to the measurement and characterization of radiographic observer performance. *J. Appl. Photogr. Eng*, 4:166–172., 1978.

J. Burge, T. Lane, S. Link, H. Qiu, and V. Clark. Discrete dynamic bayesian network analysis of fMRI data. *Hum Brain Mapp*, 30:122–37, 2009.

R. Buxton, E. Wong, and L. Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn. Reson. Med.*, 39(6):855–64, 1998.

A. Cameron and P. Trivedi. *Regression analysis of count data*. Cambridge University Press, 1998.

J. Cao and K. Worsley. Applications of random fields in human brain mapping. In M. Moore, editor, *Spatial Statistics: Methodological Aspects and Applications, Springer Lect. Notes Stat.*, volume 159, pages 169–182, New York, 2001. Springer.

C. Chang, M. Thomason, and G. Glover. Mapping and correction of vascular hemodynamic latency in the BOLD signal. *NeuroImage*, 43:90–102, 2008.

J. M. Chien, K. Fissell, S. Jacobs, and J. A. Fiez. Functional heterogeneity within broca's area during verbal working memory. *Physiol Behav*, 77(4-5):635–9, 2002.

D. M. Cole, S. M. Smith, and C. F. Beckmann. Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. *Front Syst Neurosci*, 4:8, 2010. doi: 10.3389/fnsys.2010.00008.

J. Copas and J. Shi. A sensitivity analysis for publication bias in systematic reviews. *Stat Methods Med Res*, 10:251–265, 2001.

S. G. Costafreda, A. S. David, and M. J. Brammer. A parametric approach to voxel-based meta-analysis. *NeuroImage*, 46(1):115–22, May 2009.

J. Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, C. Stam, S. Smith, and C. F. Beckmann. Consistent resting-state networks across healthy subjects. *Proc Natl Acad Sci U S A*, 103(37):13848–53, 2006.

M. D'Esposito, E. Zarahn, and G. K. Aguirre. Event-related functional MRI: Implications for cognitive psychology. *Psychological Bulletin*, 125:155–164, 1999.

L. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

G. Douaud, S. Smith, M. Jenkinson, T. E. Behrens, H. Johansen-Berg, J. Vickers, S. James, N. Voets, K. Watkins, P. Matthews, and A. James. Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain*, 130(9): 2375–86, 2007.

N. Draper and H. Smith. *Applied Regression Analysis*. Series in Probability and Statistics, 1998.

S. B. Eickhoff, A. R. Laird, C. Grefkes, L. E. Wang, K. Zilles, and P. T. Fox. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Hum Brain Mapp*, 2009.

R. Fisher. Combining independent tests of significance. *Am. Stat.*, 2:30, 1948.

D. Flitney and M. Jenkinson. Cluster analysis revisited. Technical Report TR00DF1, FMRIB Centre, University of Oxford, 2000.

S. Forman, J. Cohen, J. Fitzgerald, W. Eddy, M. Mintun, and D. Noll. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn. Reson. Med.*, 33:636–647, 1995.

M. D. Fox and M. E. Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci*, 8(9):700–11, Sep 2007. doi: 10.1038/nrn2201.

P. T. Fox, J. Lancaster, L. Parsons, X. J., and F. Zamarripa. Functional volumes modeling: Theory and preliminary assessment. *Hum Brain Mapp*, 5(4):306–311, 1997.

P. T. Fox, L. Parsons, and J. Lancaster. Beyond the single study: function/location metanalysis in cognitive neuroimaging. *Curr Opin Neurobiol*, 8(2):178–87, 1998.

P. T. Fox, A. Huang, L. Parsons, X. J., L. Rainey, and J. Lancaster. Functional volumes modeling: Scaling for group size in averaged images. *Hum Brain Mapp*, 8:143–150, 1999.

D. Freedman and D. Lane. A Nonstochastic Interpretation of Reported Significance Levels. *Journal of Business & Economic Statistics*, 1(4):292–98, 1983.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–41, 2008.

L. Friedman, G. H. Glover, D. Krenz, and V. Magnotta. Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. *NeuroImage*, 32(4):1656–68, 2006.

L. Friedman, H. Stern, G. G. Brown, D. H. Mathalon, J. Turner, G. H. Glover, R. L. Gollub, J. Lauriello, K. O. Lim, T. Cannon, D. N. Greve, H. J. Bockholt, A. Belger,

B. Mueller, M. J. Doty, J. He, W. Wells, P. Smyth, S. Pieper, S. Kim, M. Kubicki, M. Vangel, and S. G. Potkin. Test-retest and between-site reliability in a multicenter fMRI study. *Hum Brain Mapp*, 2007.

K. Friston, K. Worsley, R. Frackowiak, J. Mazziotta, and A. Evans. Assessing the significance of focal activations using their spatial extent. *Hum. Brain Mapp.*, 1:210–220, 1994.

K. Friston, A. Holmes, J. Poline, C. Price, and C. Frith. Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage*, 4:223–235, 1996.

K. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *NeuroImage*, 19(4): 1273–1302, 2003.

K. J. Friston, E. Zarahn, O. Josephs, R. N. Henson, and A. M. Dale. Stochastic designs in event-related fmri. *NeuroImage*, 10:607–619, 1999.

C. R. Genovese, N. A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4):870–8, 2002.

D. Greve and B. Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72, 2009.

A. Groves, M. Chappell, and M. W. Woolrich. Combined spatial and non-spatial prior for inference on MRI time-series. *NeuroImage*, 45:795–809, 2009.

S. Hayasaka, K. Phan, I. Liberzon, K. Worsley, and T. Nichols. Nonstationary cluster-size inference with random field and permutation methods. *NeuroImage*, 22(2):676–87, 2004.

J. P. T. Higgins, S. G. Thompson, and D. J. Spiegelhalter. A re-evaluation of random-effects meta-analysis. *J. R. Statist. Soc.*, 172(1):137–159, 2009.

A. Holmes, R. Blair, J. Watson, and I. Ford. Nonparametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Metab.*, 16(1):7–22, 1996.

S. Huang, J. Li, L. Sun, J. Ye, A. Fleisher, T. Wu, K. Chen, E. Reiman, and A. D. N. Initiative. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–49, 2010.

G. D. Iannetti, L. Zambreanu, R. G. Wise, T. J. Buchanan, J. P. Huggins, T. S. Smart, W. Vennart, and I. Tracey. Pharmacological modulation of pain-related brain activity during normal and central sensitization states in humans. *Proc Natl Acad Sci U S A*, 102(50):18195–200, 2005.

M. Jenkinson and S. Smith. A global optimisation method for robust affine registration of brain images. *Med Image Anal*, 5(2):143–56, 2001.

M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–41, 2002.

P. Jezzard, P. Matthews, and S. Smith, editors. *Functional MRI: An Introduction to Methods*. Oxford University Press, Oxford, UK, 2003.

G. Karas, P. Scheltens, S. Rombouts, P. Visser, R. van Schijndel, N. Fox, and F. Barkhof. Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, 22(2):708–16, 2004.

S. Kiebel, J. Poline, K. Friston, A. Holmes, and K. Worsley. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *NeuroImage*, 10(6):756–66, 1999.

I. Kirsch, B. J. Deacon, T. Huedo-Medina, A. Scoboria, T. Moore, and B. Johnson. Initial severity and antidepressant benefits: A meta-analysis of data submitted to the food and drug administration. *PLoS Medicine*, 5(2):260–8, 2008.

F. Kschischang, B. Frey, and H. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):598–519, 2001.

A. Laird, P. Fox, C. Price, D. C. Glahn, A. Uecker, J. Lancaster, P. Turkeltaub, P. Kochunov, and P. Fox. ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Hum Brain Mapp*, 25(1):155–64, 2005a.

A. Laird, J. Lancaster, and P. T. Fox. BrainMap: The social evolution of a functional neuroimaging database. *Neuroinformatics*, 3:65–78, 2005b.

N. Lazar, B. Luna, J. A. Sweeney, and W. F. Eddy. Combining brains: a survey of methods for statistical pooling of information. *NeuroImage*, 16(2):538–50, 2002.

A. Ledberg, S. Åkerman, and P. Roland. Estimation of the probability of 3D clusters in functional brain images. *NeuroImage*, 8:113– 128, 1998.

T. T. Liu. Efficiency, power and entropy in event-related fmri with multiple trial types. part ii: design of experiments. *NeuroImage*, 21:400–413, 2004.

T. T. Liu and L. R. Frank. Efficiency, power, and entropy in event-related fmri with multiple trial types: Part i. theory. *NeuroImage*, 21:387–400, 2004.

G. Marrelec, A. Krainik, H. Duffau, M. Pelegrini-Issac, S. Lehericy, J. Doyon, and H. Benali. Partial correlation for functional brain interactivity investigation in functional mri. *NeuroImage*, 32:228–37, 2006.

P. McCullagh and J. Nelder. *Generalized linear models*. Chapman & Hall, CRC, 1999.

A. McIntosh and F. Gonzalez-Lima. Structural equation modeling and its application to network analysis in functional brain imaging. *Hum Brain Mapp*, 2(1-2):2–22, 1994.

K. Miller, W. Luh, T. Liu, A. Martinez, T. Obata, E. Wong, L. Frank, and R. Buxton. Nonlinear temporal dynamics of the cerebral blood flow response. *Hum Brain Mapp*, 13(1):1–12, 2001.

T. Moorhead, D. Job, M. Spencer, H. Whalley, E. Johnstone, and S. Lawrie. Empirical comparison of maximal voxel and non-isotropic adjusted cluster extent results in a voxel-based morphometry study of comorbid learning disability with schizophrenia. *NeuroImage*, 28(3):544–52, 2005.

J. Neumann, G. Lohmann, J. Derrfuss, and D. von Cramon. Meta-analysis of functional imaging data using replicator dynamics. *Hum Brain Mapp*, 25(1):165–73, 2005.

J. Neumann, D. Cramon, and G. Lohmann. Model-based clustering of meta-analytic functional imaging data. *Hum Brain Mapp*, 29(2):177–92, 2008.

J. Neumann, P. T. Fox, R. Turner, and G. Lohmann. Learning partially directed functional networks from meta-analysis imaging data. *NeuroImage*, 49(2):1372–84, 2009.

T. Nichols. Cluster analysis revisited - again: Implementing nonstationary cluster size inference. Technical Report TR08TN1, FMRIB Centre, University of Oxford, 2008.

T. Nichols and A. Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp*, 15(1):1–25, 2002.

T. E. Nichols and S. Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res*, 12(5):419–46, 2003.

F. Nielsen and L. K. Hansen. Finding related functional neuroimaging volumes. *Artif Intell Med*, 30(2):141–51, 2004.

F. A. Nielsen. Visualizing data mining results with the brede tools. *Frontiers in Neuroinformatics*, 3(26), 2009.

F. A. Nielsen and L. K. Hansen. Modeling of activation data in the BrainMap database: detection of outliers. *Hum Brain Mapp*, 15(3):146–56, 2002.

R. Patel, F. Bowman, and J. Rilling. Determining hierarchical functional networks from auditory stimuli fMRI. *Hum Brain Mapp*, 27(5):462–70, 2006.

K. Petersson, T. Nichols, J. Poline, and A. Holmes. Statistical limitations in functional neuroimaging ii. signal detection and statistical inference. *Philos. Trans. R. Soc. (Lond)*, Series B 354:1261–1281, 1999.

K. L. Phan, T. Wager, S. F. Taylor, and I. Liberzon. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage*, 16(2): 331–48, Jun 2002. doi: 10.1006/nimg.2002.1087.

J. Poline and B. Mazoyer. Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise ratio pixel clusters. *J. Cereb. Blood Flow Metab.*, 13:425–437, 1993.

J. Poline, K. Worsley, A. Evans, and K. Friston. Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, 5(2):83–96, 1997.

J. Ramsey, P. Spirtes, and C. Glymour. On meta-analyses of imaging data and the mixture of records. *NeuroImage*, 2010.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

S. Risacher, A. Saykin, J. West, L. Shen, H. Firpi, B. McDonald, and A. D. N. I. (ADNI). Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr Alzheimer Res.*, 6(4):347–61, 2009.

P. Roland, B. Levin, R. Kawashima, and S. Åkerman. Three-dimensional analysis of clustered voxels in 15-o-butanol brain activation images. *Hum. Brain Mapp.*, 1:3–19, 1993.

G. Salimi-Khorshidi, S. Smith, and T. Nichols. Adjusting the neuroimaging statistical inferences for nonstationarity. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, volume 5761/2009 of *Lecture Notes in Computer Science*, pages 992–999. Springer Berlin / Heidelberg, 2009a.

G. Salimi-Khorshidi, S. Smith, and T. E. Nichols. Bias and heterogeneity in neuroimaging meta-analysis. In *15th Annual Meeting of the Organization for Human Brain Mapping Abstracts Online*, volume SA-PM, page 406, 2009b.

G. Salimi-Khorshidi, S. Smith, J. Keltner, T. D. Wager, and T. E. Nichols. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *NeuroImage*, 45(3):810–23, April 2009a.

G. Salimi-Khorshidi, S. Smith, and T. Nichols. Adjusting the Effect of Nonstationarity in Cluster-based and TFCE Inference. *NeuroImage*, 2010.

S. Smith. Fast robust automated brain extraction. *Hum Brain Mapp*, 17(3):143–55, 2002.

S. Smith, P. R. Bannister, C. Beckman, M. Brady, S. Clare, D. Flitney, P. Hansen, M. Jenkinson, D. Leibovici, B. Ripley, M. Woolrich, and J. Zhang. FSL: New tools for functional and structural brain image analysis. *NeuroImage*, 13(6):249, 2001.

S. Smith, M. Jenkinson, M. Woolrich, C. Beckmann, T. Behrens, H. Johansen-Berg, P. Bannister, M. De Luca, I. Drobnjak, D. Flitney, R. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. Brady, and P. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. neuroimage, 23(s1):208-219, 2004. *NeuroImage*, 23(S1):208–219, 2004.

S. Smith, M. Jenkinson, C. Beckmann, K. Miller, and M. Woolrich. Meaningful design and contrast estimability in FMRI. *NeuroImage*, 34(1):127–36, 2007.

S. Smith, G. Douaud, G. Salimi-Khorshidi, M. Webster, C. Mackay, A. Groves, and T. Nichols. Threshold-free cluster enhancement: Practical examples. In 14th Anuual Meeting of the Organization for Human Brain Mapping, Melbourne, Australia, 2008.

S. Smith, P. Fox, K. Miller, D. Glahn, P. Fox, C. Mackay, N. Filippini, K. Watkins, R. Toro, A. Laird, and C. Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proc Natl Acad Sci U S A*, 106(31):13040–5, 2009.

S. M. Smith and T. E. Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1):83–98, 2009.

S. M. Smith, K. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. Nichols, J. Ramsey, and M. W. Woolrich. Network modelling methods for FMRI. *NeuroImage*, 2010.

M. Stein. *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.

A. Sutton, D. Jones, K. Abrams, T. Sheldon, and F. Song. *Methods for Meta-analysis in Medical Research*. John Wiley, London, 2000.

J. Talairach and P. Tournoux. *Co-Planar Stereotaxic Atlas of the Human Brain*. Thieme, New York, 1987.

C. Ter Braak. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In *Bootstrapping and related techniques*, pages 79–85. Springer Verlag, Berlin, 1992.

B. Thirion, P. Pinel, S. Mériaux, A. Roche, S. Dehaene, and J. Poline. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage*, 35(1):105–20, 2007.

A. W. Toga. Neuroimage databases: the good, the bad and the ugly. *Nat Rev Neurosci*, 3(4):302–9, 2002.

R. Toro, P. T. Fox, and T. Paus. Functional coactivation map of the human brain. *Cereb Cortex.*, 18(11):2553–9, 2008.

P. Turkeltaub, G. Eden, K. Jones, and T. A. T. Zeffiro. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *NeuroImage*, 16(3.1): 765–80, 2002.

J. Van Horn, S. T. Grafton, D. Rockmore, and M. S. Gazzaniga. Sharing neuroimaging studies of human cognition. *Nat Neurosci.*, 7(5):473–81, 2004.

I. Veer, C. Beckmann, M. van Tol, L. Ferrarini, J. Milles, D. Veltman, A. Aleman, M. van Buchem, N. van der Wee, and S. Rombouts. Whole brain resting-state analysis reveals decreased functional connectivity in major depression. *Front Syst Neurosci*, 4(41), 2010.

T. D. Wager, J. Jonides, and S. Reading. Neuroimaging studies of shifting attention: a meta-analysis. *NeuroImage*, 22(4):1679–93, 2004.

T. D. Wager, M. Lindquist, and L. Kapla. Meta-analysis of functional neuroimaging data: current and future directions. *Soc Cogn Affect Neurosci*, 2(2):150–158, 2007.

T. D. Wager, M. A. Lindquist, T. E. Nichols, H. Kober, and V. S. J. X. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage*, 2009.

M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith. Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, 14(6):1370–86, 2001.

M. W. Woolrich, T. E. Behrens, C. F. Beckmann, M. Jenkinson, and S. M. Smith. Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21(4):1732–47, 2004.

M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. E. Behrens, C. F. Beckmann, M. Jenkinson, and S. M. Smith. Bayesian analysis of neuroimaging data in fsl. *NeuroImage*, 45(1 Suppl):S173–86, 2009.

K. Worsley, A. Evans, S. Marrett, and P. Neelin. Three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.*, 12:900–918, 1992.

K. Worsley, S. Marrett, P. Neelin, A. Vandal, K. Friston, and A. C. Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp*, 4:58–73, 1996.

K. Worsley, M. Andermann, T. Koulis, D. MacDonald, and A. Evans. Detecting changes in nonisotropic images. *Hum. Brain Mapp.*, 8:98–101, 1999.

K. Worsley, C. Liao, J. Aston, V. Petre, G. Duncan, F. Morales, and A. Evans. A general statistical analysis for fMRI data. *NeuroImage*, 15(1):1–15, 2002.

H. Zhang, S. Wang, B. Liu, Z. Ma, M. Yang, Z. Zhang, and G. Teng. Resting brain connectivity: changes during the progress of Alzheimer disease. *Radiology*, 256(2): 598–606, 2010.

K. H. Zou, D. N. Greve, M. Wang, S. D. Pieper, S. K. Warfield, N. S. White, S. Manandhar, G. G. Brown, M. G. Vangel, R. Kikinis, and W. M. Wells, 3rd. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. *Radiology*, 237(3):781–9, 2005.