

THE UNIVERSITY OF WARWICK



DEPARTMENT OF STATISTICS

Master of Science in
Statistics

Academic Year 2014/2015

Dissertation

**Controls for Multiplicity in Resting State
Functional Connectivity Analysis**

Candidate:

Joseph Patrick Meagher

Supervisor:

Dr Mark Fiecas

“Constant repetition carries conviction”
- Conor McGregor, UFC Interim Featherweight Champion

Acknowledgements

A massive thanks is due to my supervisor on this project, Dr. Mark Fiecas, who put me back on the right path more than once, and always responded to the sillier of my questions with enthusiasm and grace. Also, my parents, Liam and Lucia, who put me in a position to have the freedom to pursue an MSc. in Statistics. And finally, to Clarence Birdseye, for inventing frozen vegetables and making sure that I always got my five a day.

Abstract

When applied to massive multiple testing problems, where the number of tests being performed simultaneously can be in the thousands, classical hypothesis testing methods produce too many false positives to allow meaningful inferences to be made.

This paper presents a review of methods for the control of the rate of false positives in multiple testing problems. These are then applied in the context of a Functional Connectivity study of data produced by a fMRI scan of an individuals brain in a resting state.

As part of this analysis a theorem for the asymptotic distribution of sample cross correlations, proved by Roch Roy, will be implemented to correct for temporal correlation.

This analysis establishes that there are strong positive dependency networks within the brain which are observed even on an individual basis. These networks are also still observable even after correcting for multiplicity and temporal correlation.

Contents

1	Introduction	4
1.1	Multiple Testing	4
1.2	fMRI and the Resting State Brain	5
1.3	Thesis Outline	6
2	Materials and Methods	7
2.1	Participant	7
2.2	Data Acquisition	7
2.3	Preprocessing	7
2.4	Quantifying Functional Connectivity	8
2.5	The Distribution of the Functional Connectivity Metric	8
2.6	Calculation of Test Statistics	9
2.7	Estimation of the Asymptotic Variance	11
2.8	Hypothesis Testing	11
2.8.1	Naïve Hypothesis Testing	12
2.9	Controls for Multiplicity	13
2.9.1	Familywise Error Rate Control	13
2.9.2	False Discovery Rate Control	18
2.10	Simulation Study	19
2.10.1	Bivariate Time Series	19
2.10.2	Mean Squared Error	20
2.11	Simulation Algorithms	20
3	Simulation Study	22
3.1	Experiment I	22
3.2	Experiment II	23
3.3	Experiment III	25
3.4	Simulation Study Conclusions	25
4	Results of Functional Connectivity Analysis of fMRI Data	28
4.1	Analysis of Test Statistics	28
4.2	Naïve Hypothesis Testing	30
4.3	Familywise Error Rate Controls	32
4.4	False Discovery Rate Control	32
5	Discussion and Conclusions	38
5.1	Functional Connectivity in the “Resting State Brain”	38
5.2	Controls for Multiplicity	38

5.3	Roy's Asymptotic Variance Estimate	40
5.4	Limitations	40
	Appendices	45
A	R Code Implementing Roy's Theorem	46
B	Test statistic Graphics	48

Chapter 1

Introduction

1.1 Multiple Testing

Researchers regularly find themselves having to test a family of N hypotheses simultaneously. A family of hypotheses is defined as “any collection of inferences for which it is meaningful to consider some combined measure of error[13]” and whether $N = 10$ or $N = 10,000$ drawing reasonable inference requires more nuanced techniques than the repeated application of classical methods[7]. The fact is that, as the number of hypotheses tested simultaneously increases, the likelihood of at least one of these hypotheses presenting as non-null purely by chance also increases.

Consider an experiment investigating whether a single coin is fair or more likely to land on heads. The null hypothesis (H_0) and alternative hypothesis (H_1) can be characterised as

$$H_0 : p = 0.5, \qquad H_1 : p > 0.5.$$

When a fair coin is flipped 10 times, the probability that it lands on heads 9 or more times is approximately 0.0107. Should this outcome occur in the experiment it is reasonable to reject H_0 , concluding that the coin is not fair. Now consider an experiment investigating whether any of 100 coins are fair or biased in favour of heads. In this case a family of 100 hypotheses, each of which, for $i \in \{1, \dots, 100\}$, can be expressed as

$$H_{0,i} : p = 0.5, \qquad H_{1,i} : p > 0.5.$$

When 100 fair coins are flipped 10 times each, the probability that at least one of them will show heads 9 or more times is approximately 0.76. So finding that the i^{th} coin came up heads 9 or more times does not necessarily imply that it is biased in favour of heads, although it may be of interest for future experiments.

This toy example illustrates the pitfalls of testing multiple hypotheses simultaneously. In the era of “Big Data”, where data mining and biological datasets such as microarrays mean that testing a family of $N = 10,000$ hypotheses simultaneously is not uncommon, an appreciation of the challenges associated with inference on this scale becomes ever more important to analysts.

When presented with a problem involving the simultaneous testing of multiple hypotheses, and analyst can employ a Simultaneous Testing Procedure (STP) to identify which of

the null hypotheses in the family should be rejected. STP's fall into two main categories, those controlling the Familywise Error Rate (FWER), and those which control the False Discovery Rate (FDR). FWER represents the probability that at least one true $H_{0,i}$ is rejected by the testing procedure, while the FDR is the proportion of rejected hypotheses (R) which are in fact null[7].

In this thesis the problem of multiple testing and how to adjust inferences accordingly will be examined, taking a family of hypotheses investigating Functional Connectivity in an individuals brain as an example. This dataset, recorded in a functional Magnetic Resonance Imaging (fMRI) scan, represents the activity of an individuals brain in a resting state over 394 seconds (6 min 34 seconds).

1.2 fMRI and the Resting State Brain

fMRI offers a non-invasive, indirect method for researchers to observe changes in neuronal activity within the brain. It does this by producing a realisation for the Blood Oxygen Level Dependent (BOLD) signal of the brain[9]. This can be observed while the brain is in a “Resting State” The “Resting State”, also referred to as “Spontaneous Activity”[9], or the “Default Mode” [11], describes the behaviour observed in the brain in the absence of specific inputs or outputs. During the data acquisition stage of a “Resting State” study, participants lie awake and still in a scanner, either with their eyes closed or fixated on a crosshair, without thinking about anything in particular[9]. With this data, the Functional Connectivity (FC) of the brain is examined.

FC can be describes as the statistical association, or dependency, between anatomically distinct time series[23] produced by a fMRI scan, or the relationship between the activity of different anatomical regions of the brain. Motivation for examining FC in the “Resting State” stems from the fact that while it accounts for only 2% of the body’s mass, the resting human brain is responsible for 20% of the body’s energy consumption. Only very small changes in the brain’s energy requirements are attributable to task performance, thus it follows that the “Resting State” is a major aspect of normal brain function[9]. Studies have linked disturbances in the FC of the “Resting State” to schizophrenia[15] and normal ageing[6], among other things. This implies that the “Resting State” of the brain may provide important diagnostic and prognostic information.

It has been hypothesised that the brain follows a structure of dynamic, anti-correlated networks which represent a task-related dichotomy[10]. There is strong evidence to support the existence of a “Resting State”, or “Default Mode”, network in the resting brain[11] and regions of the brain which supposedly have opposing functionality have been found to exhibit negative correlation in the “Resting State” [9]. This network, mapped out into Intrinsic Connectivity Networks (ICN), demonstrate both spatial and test-retest consistency across a number of studies and individuals[24].

There is no agreed upon standard metric for quantifying FC. Although this analysis will focus solely on the sample cross-correlation, it is advisable to consider a variety of measures as this offers a more complete overview of the underlying network structure. Popular methods of FC quantification include cross-correlation, partial cross-correlation, cross-coherence, and partial cross-coherence[8]. Correlation measures apply to the time

domain, while coherence measures apply to the frequency domain.

When quantifying FC, isolating the signals of interest can improve estimates. It has been repeatedly shown that these signals exist at low frequency (< 0.1 Hz) bands[23]. By filtering the dataset nuisance signals can be removed. The global BOLD signal, the signal obtained by averaging over the whole brain, is also considered to be a nuisance signal by many researchers, and is often removed before quantifying FC. There are however some question marks surrounding this preprocessing step. It introduces negative correlations into the dataset which may not necessarily exist, as removing this average forces the sum of the cross-correlation values between any particular series and every other series to be negative[18]. However, recent results investigating whether or not to remove the global signal are inconclusive[8]. Thus, as it is not the focus of this paper, the global signal has been removed in this analysis.

It has also been suggested that failing to correct for autocorrelation in fMRI time series is a source of “spurious” estimates of FC. This is based on results from Economics. It was found that the effect of autocorrelation on hypothesis tests of group differences between healthy controls and schizophrenia patients was weak[1]. Despite this, the impact of accounting for the temporal autocorrelation and cross-correlation (temporal correlation) of series when computing test statistics for an individuals “Resting State” FC may be worth further examination. To this effect, implementing a formula for the asymptotic variance of sample cross-correlations[22][17][16] may result in improved test statistics. This is investigated as part of this analysis.

1.3 Thesis Outline

This thesis presents a FC study of an individual brain in a “Resting State”. The analysis will attempt to reproduce well established results identifying a strong positive dependency between the Posterior Cingulate cortex (PCC) and the Medial Prefrontal Cortex (MPF)[10]. Whether these results hold after accounting for temporal correlation and implementing controls for multiplicity will be assessed.

Chapter 2 contains a summary of the materials used and methods employed in completing this piece of research. This covers the collection of the dataset, preprocessing, calculation of test statistics, and implementation of controls for multiplicity.

A simulation study, presented in chapter 3, provides insight into the possible impact of accounting for temporal correlation in the calculation of test statistics.

Chapter 4 presents the results of the analysis.

The thesis concludes in Chapter 5 with a discussion of the results.

Chapter 2

Materials and Methods

2.1 Participant

A resting state fMRI dataset for a single scan of a single individual which is publicly available at NITRC (<http://www.nitrc.org/projects/trt>) was used for this analysis. The individual was chosen at random from a full dataset of 25 participants (mean age 29.44 ± 8.64 , 10 males) which has previously been analysed in a test-retest evaluation of functional connectivity[8], the stability of cross-correlations[24], ICA and dual regression[30], low-frequency oscillations[29], and graph-theoretic network properties[26].

2.2 Data Acquisition

A Siemens Allegra 3.0-Tesla scanner was used to obtain a resting-state scan for the participant. The scan consisted of $T=197$ contiguous EPI functional volumes with a time repetition (TR) = $2000ms$; time echo (TE) = $25ms$; flip angle (FA) = 90° ; number of slices = 39, matrix = 64×64 ; field of view (FOV) = $192mm$; voxel size = $3 \times 3 \times 3mm^3$. During the scan, the participant was asked to relax and remain still with eyes open during the scan. For spatial normalisation and localisation, a highresolution T1-weighted magnetization prepared gradient echo sequence was obtained (MPRAGE, TR = $2500ms$, TE = $4.35ms$; inversion time = $900ms$; FA = 8, number of slices = 176; FOV = $256mm$).

2.3 Preprocessing

A preprocessed dataset in its “native space” was provided for this analysis. This preprocessing was performed using FSL (version 4.1, <http://www.fmrib.ox.ac.uk>). The dataset was motion corrected using FSL’s mcflirt (rigid body transform; cost function normalised correlation; reference volume the middle volume). Using FSL’s fslmaths, the volumes were spatially smoothed using a Gaussian kernel with FWHM = 6mm. This was the dataset, along with a masking array, provided to the analyst in the form of a Nifti file. The purpose of the masking array is to separate the voxels relating to the individuals brain from the rest.

Further preprocessing was performed in R (version 3.1, <http://www.R-project.org/>)[21] requiring the installation of the `oro.nifti`[28] and `signal`[25] packages. The mask was applied to the dataset and the global signal then removed. The dataset was then filtered through a 4-th order Butterworth filter with passband [0.01,0.10] Hertz. Finally the dataset was standardised to mean 0 and variance 1.

2.4 Quantifying Functional Connectivity

The sample cross-correlation at lag 0 is used to quantify FC. It is a popular metric for FC analysis as its statistical properties are well studied and it is straightforward to compute. The sample cross-correlation provides a measure of the linear dependence between series X_i and X_j , where X_1, \dots, X_T is a realisation of a multivariate series of length T . The sample covariance at lag k ($c_{ij}(k)$) is defined as

$$c_{ij}(k) = T^{-1} \sum_{t=1}^{T-k} (x_{i,t} - \bar{x}_i)(x_{j,t} - \bar{x}_j), \quad (2.1)$$

where $x_{i,t}$ is the realisation of X_i at time t and $\bar{x}_i = T^{-1} \sum_{t=1}^T x_{i,t}$.

The sample cross-correlation at lag k is then defined as

$$r_{ij}(k) = c_{ij}(k) \{c_{ii}(0)c_{jj}(0)\}^{-\frac{1}{2}}. \quad (2.2)$$

As this study has set out to reproduce the dependency between the Posterior Cingulate Cortex (PCC) and the Medial Prefrontal Cortex (MPF), a region of interest approximating the PCC is chosen as a seed (s). The seed for this analysis is the $3 \times 3 \times 3$ voxel cube centred on the (33,19,23) co-ordinate of the Nifti file representing the brain in it's "native space". The sample cross-correlation between the seed and each voxel (v) in the brain is computed to quantify FC for the individuals brain.

2.5 The Distribution of the Functional Connectivity Metric

An accurate expression for the distribution of the cross-correlation between the seed and each voxel underpins the calculation of a family of test statistics.

Roy[22] presents a proof of the asymptotic joint normality of any finite set of sample correlations and derives a formula for the asymptotic covariance between two serial correlations. These are stated here without proof.

Note that A' represents the transpose of the matrix A . Let $\{X_t : t \in \mathbb{Z}\}$ be a multivariate weakly stationary stochastic process of dimension $r : X'_t = (X_{1t}, \dots, X_{rt})$. Without loss of generality, assume that $\mathbb{E}[X_t] = 0$ and let $\mathbb{E}[X_t X'_{t+k}] = \Gamma(k) = (\gamma_{ij}(k))$ be the covariance matrix at lag $k, k \in \mathbb{Z}$. The correlation matrix at lag k is denoted by $\rho(k) = (\rho_{ij}(k))$ where $\rho_{ij}(k) = \gamma_{ij}(k) \{\gamma_{ii}(0)\gamma_{jj}(0)\}^{-\frac{1}{2}}$.

To simplify notation define, for $k \in \mathbb{Z}$,

$$\Delta_k(i, j, l, m) = \sum_{u=-\infty}^{\infty} \rho_{ij}(u)\rho_{lm}(u+k). \quad (2.3)$$

Then write

$$s_{ij}(k) = T^{\frac{1}{2}}(r_{ij}(k) - \rho_{ij}(k)), \quad (2.4)$$

it can be shown that

$$\begin{aligned} \lim_{T \rightarrow \infty} \text{cov}\{s_{ab}(k), s_{de}(h)\} &= \frac{1}{2} \rho_{ab}(k)\rho_{de}(h) \{ \Delta_0(a, d, a, d) + \Delta_0(a, e, a, e) \\ &\quad + \Delta_0(b, d, b, d) + \Delta_0(b, e, b, e) \} \\ &\quad - \rho_{ab}(k) \{ \Delta_h(a, d, a, e) + \Delta_h(b, d, b, e) \} \\ &\quad - \rho_{de}(h) \{ \Delta_k(b, d, a, d) + \Delta_k(b, e, a, e) \} \\ &\quad + \Delta_{h-k}(a, d, b, e) + \Delta_{h+k}(b, d, a, e) \end{aligned} \quad (2.5)$$

The results of the paper are summarised as

Theorem 2.5.1 *Let $\{X_t\}$ be a multivariate second-order stationary process satisfying the assumptions of the central limit theorem of Hannan[12]. Further, suppose that all cumulants of fourth order are zero and that the spectral density of each component of $\{X_t\}$ is square integrable. Let $s_{ij}(k)$ be defined by (2.4), then any finite set of the $s_{ij}(k)$'s are jointly asymptotically normal with mean zero and covariance structure given by (2.5).*

The assumptions made by Hannan[12] are that X_t is ergodic, is a causal linear process, generated according to

$$\begin{aligned} X_t &= \sum_{j=0}^{\infty} \psi_j Z_{t-j} \\ \sum_{j=0}^{\infty} \|\psi_j\|^2 &< \infty \\ \mathbb{E}[Z_j Z_j'] &= \delta_{ij} \Sigma \\ \mathbb{E}[Z_t] &= 0 \end{aligned}$$

where δ_{ij} is Kroneker's Delta, Σ is an $r \times r$ matrix, $\|\cdot\|$ is a norm for $r \times r$ matrices, and Z_t are the prediction errors. Hannan also assumes that the first four moments of Z_t are constant.

2.6 Calculation of Test Statistics

In order to test a family of N hypotheses a family of N test statistics t must first be constructed. Theorem 2.5.1, Roy's Theorem, provides an expression for the asymptotic covariance of the sample cross-correlations. Note that

$$\Gamma(k) = \Gamma(-k), \quad \rho(k) = \rho(-k)'. \quad (2.6)$$

It is now possible to construct two families of test statistics. The first reflects the current standard practice in the FC literature [8][24] and requires the additional assumption that realisations of X_t are independent and identically distributed (IID), that is that each X_t is not temporally correlated. For the IID case (2.5) reduces to

$$\lim_{T \rightarrow \infty} \text{Var}(s_{ij}(0)) = (1 - \rho_{ij}^2(0))^2.$$

Thus by Roy's Theorem

$$T^{\frac{1}{2}}(r_{ij}(0) - \rho_{ij}(0)) \sim \mathcal{N}(0, (1 - \rho_{ij}^2(0))^2)$$

where (\sim) indicates an asymptotic distribution. A Fisher's z transform[4] is applied to the data. This stabilises the asymptotic variance of the cross-correlation meaning that it is not dependent on the population cross-correlation ($\rho_{sv}(0)$). Fisher's Z transform applies the function $g(\cdot)$ to the data where

$$g(x) = \tanh^{-1}(x).$$

By this transformation

$$T^{\frac{1}{2}}g(r_{sv}(0) - \rho_{sv}(0)) \sim \mathcal{N}(0, 1). \quad (2.7)$$

Thus the first family of test statistics t^f can be computed where for each voxel $v \in \{1, \dots, 49441\}$

$$t_{sv}^f = T^{\frac{1}{2}}g(r_{sv}(0) - \hat{\rho}_{sv,0}(0)), \quad (2.8)$$

where $\hat{\rho}_{sv,0}(0)$ is $\rho_{sv}(0)$ under the null hypothesis. Where the null hypothesis is true t_{sv}^f follows an asymptotically $\mathcal{N}(0, 1)$ distribution. This is the standard test statistic used in FC literature.

The second test statistic calculated does not require the additional assumption of independent and identically distributed realisations of X_t . Again, note (2.6), reducing (2.5) to

$$\begin{aligned} \lim_{T \rightarrow \infty} \text{Var}(s_{ij}(0)) = & \frac{1}{2}\rho_{sv}^2(0)\{\Delta_0(i, i, i, i) + 2\Delta_0(i, j, i, j) + \Delta_0(j, j, j, j)\} \\ & - 2\rho_{ij}(0)\{\Delta_0(i, i, i, j) + \Delta_0(j, i, j, j)\} \\ & + \Delta_0(i, i, j, j) + \Delta_0(j, i, i, j). \end{aligned} \quad (2.9)$$

Denote $\lim_{T \rightarrow \infty} \text{Var}(s_{ij}(0))$ as σ_{ij}^2 . By Roy's Theorem

$$T^{\frac{1}{2}}(r_{ij}(0) - \rho_{ij}(0)) \sim \mathcal{N}(0, \sigma_{ij}^2). \quad (2.10)$$

Suppose now that there is an estimate for σ_{ij}^2 , denoted $\hat{\sigma}_{ij}^2$. The second family of test statistics t^r can be calculated for $v \in \{1, \dots, 49441\}$ as

$$t_{sv}^r = \frac{T^{\frac{1}{2}}(r_{sv}(0) - \hat{\rho}_{sv,0}(0))}{\hat{\sigma}_{sv}}. \quad (2.11)$$

with $\hat{\rho}_{sv,0}(0)$ is $\rho_{sv}(0)$ under the null hypothesis and t_{sv}^r is asymptotically $\mathcal{N}(0, 1)$ when the null hypothesis is true.

2.7 Estimation of the Asymptotic Variance

In order to calculate the family of test statistics t^r (2.9) an estimate for σ_{ij}^2 (2.11) must be computed. Mélard & Roy[17] set out and justify an algorithm for computing confidence intervals and test statistics for the autocorrelations of a univariate time series. This algorithm has been adapted here for the multivariate case. The algorithm relies on the use of a window function to ensure that the estimator $\hat{\sigma}_{ij}^2$ is non-negative definite and converges to σ_{ij}^2 . The window function ensures that estimates of $\rho_{ij}(k)$ for large k , which are based on small samples, do not have a disproportionate effect on the estimator $\hat{\sigma}_{ij}^2$.

Recall the definitions for the sample covariance (2.1) and sample correlation (2.2) and choose a window $w_u = w(u/b_{T,H})$, subject to $|u| \leq T-1$, where $b_{T,H} = H\sqrt{T}$. A Modified Bartlett window setting $H = 1$ throughout is used in this analysis, defined as

$$w(x) = \begin{cases} 1 - |x|, & \text{if } |x| \leq 1, \\ 0, & \text{if } |x| > 1. \end{cases}$$

An estimate for $\Delta_k(i, j, l, m)$ (2.3) can now be computed

$$\hat{\Delta}_k(i, j, l, m) = \sum_{u=-T+1}^{T-i-1} w_u w_{u+k} r_{ij}(u) r_{lm}(u+k).$$

An estimate for σ_{ij}^2 (2.9) is calculated as

$$\begin{aligned} \hat{\sigma}_{sv}^2 = & \frac{1}{2} r_{sv}^2(0) \{ \hat{\Delta}_0(s, s, s, s) + 2\hat{\Delta}_0(s, v, s, v) + \hat{\Delta}_0(v, v, v, v) \} \\ & - 2r_{sv}(0) \{ \hat{\Delta}_0(s, s, s, v) + \hat{\Delta}_0(v, s, v, v) \} \\ & + \hat{\Delta}_0(s, s, v, v) + \hat{\Delta}_0(v, s, s, v). \end{aligned} \quad (2.12)$$

It can be shown that $\hat{\sigma}_{ij}^2$ converges to σ_{ij}^2 in probability[17].

2.8 Hypothesis Testing

Inference in neuroimaging is typically focused on hypothesis testing rather than developing point estimates of actual correlation value[1], and this analysis retains this focus.

Borrowing heavily from Casella & Berger[5], a hypothesis is defined as a statement about the underlying parameter. The two complementary hypotheses in the hypothesis testing problem are the null hypothesis H_0 , and the alternative hypothesis H_1 . In this context, a family of hypotheses is being tested where, for each $v \in \{1, \dots, 49441\}$, the hypotheses can be expressed as

$$H_{0,v} : \rho_{sv}(0) = 0, \quad H_{1,v} : \rho_{sv}(0) \neq 0. \quad (2.13)$$

The hypothesis test is a decision rule \mathcal{D} which specifies for which sample values the decision is made to accept H_0 as the truth, and for which sample values H_0 is rejected as false in favour of H_1 . The critical region \mathcal{R} is the subset of the sample space for which \mathcal{D}

rejects H_0 . \mathcal{D} depends on a test statistic t^1 . Knowledge of the distribution of t under H_0 (when H_0 is true), allows for the definition of \mathcal{R} . The families of test statistics, t^f (2.8) and t^r (2.11), have been defined. Both are asymptotically distributed as $\mathcal{N}(0, 1)$ and so \mathcal{R} can be defined for each t_{sv} .

When testing a single hypothesis, a test may make one of two types of error, either Type I or Type II error.

Decision/Truth	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Decision
Fail to Reject H_0	Correct Decision	Type II Error

Table 2.1: Possible outcomes of a hypothesis testing procedure for a single hypothesis

Type I error occurs when t is within \mathcal{R} and H_0 is true. Type II errors occur when t is outside \mathcal{R} and H_1 is true. The power function $\beta(t)$ of a hypothesis test can be defined as

$$\beta(t) = \mathbb{P}[t \in R_{crit}]$$

where

$$\mathbb{P}[t \in R_{crit}] = \begin{cases} \mathbb{P}[\text{Type I Error} | H_0 \text{ True}] \\ \mathbb{P}[\text{Type II Error} | H_1 \text{ True}] \end{cases} \quad (2.14)$$

2.8.1 Naïve Hypothesis Testing

A first attempt at inference for the dataset is provided by naively performing a classical hypothesis test controlling the probability of a Type I error for each member of the family of hypotheses.

A critical value α is chosen. This represents the acceptable probability of a Type I error. Thus

$$\alpha = \mathbb{P}[\text{Reject } H_0 | H_0 \text{ True}]$$

The critical region for the test \mathcal{R}_α is constructed, such that

$$\mathbb{P}[t \in \mathcal{R}_\alpha | H_0 \text{ True}] = \alpha$$

Thus, by rejecting $H_{0,v}$ when $t \in \mathcal{R}_\alpha$ for $v \in \{1, \dots, 49441\}$, the probability of a Type I error occurring at test v is set to α .

The results of the hypothesis tests can be reported simply by stating the critical value and whether or not each null hypothesis has been rejected. Alternatively the p-value for the test can be reported.

Definition 2.8.1 *A p-value $p(t)$ is a test statistic satisfying $0 \leq p(t) \leq 1$ for every possible value of t . Small values of p_v provide evidence that H_0 should be rejected. A p-value is valid if*

$$\mathbb{P}[p(t) \leq \alpha | H_0 \text{ True}] \leq \alpha$$

¹ t is used interchangeably to represent either a general test statistic or situations that apply to both families of test statistics t^f and t^r

H_0 is rejected when for some critical p-value, $p(t) \leq p_{crit}$. In the case of classical hypothesis testing $p_{crit} = \alpha$, therefore when the p-value is less than the critical value, the null hypothesis is rejected.

The family of hypothesis tests described in this case (2.13) is a union-intersection test, or two-tailed hypothesis test. Thus, the critical region is the union of two sets. By (2.7) and (2.10)

$$(t_{sv}|H_{0,v} \text{ True}) \sim \mathcal{N}(0, 1)$$

and so asymptotically \mathcal{R}_α in each case is

$$\mathcal{R}_\alpha = (-\infty, z_{\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty)$$

when $\mathbb{P}[Z > z_\alpha] = \alpha$ with $Z \sim \mathcal{N}(0, 1)$. Z is also referred to as a Gaussian distribution. Therefore, reject $H_{0,v}$ when

$$|t_{sv}| > z_{1-\frac{\alpha}{2}}.$$

The p-value for each hypothesis is also calculated and reported according to Definition 2.8.1, where

$$p_v = 2 \cdot \mathbb{P}(Z > |t_{F,v}|, |t_{R,v}|). \quad (2.15)$$

The null hypothesis is rejected when

$$p_v < \alpha.$$

In this analysis reported p-values are transformed by the $-\log_{10}(\cdot)$ function. This is simply to aid the interpretation of very small values of p_v .

Hypothesis testing in this manner is referred to as naïve as this testing of cross-correlations between 49,441 individual voxels and a seed simultaneously, ensuring a large number of “false positive” results. Suppose $H_{0,v}$ is in fact true for every voxel and all the assumptions underlying the calculation of a test statistic hold. Under these conditions and setting $\alpha = 0.05$, the expected number of rejected null hypotheses in the family is 2,472. These are “false positives” results make it more likely that spurious relationships will be found. Some control for multiplicity is required before meaningful inferences can be drawn from the data.

2.9 Controls for Multiplicity

Controls for multiplicity are methods for controlling the rate at which false positives occur when conducting multiple hypotheses tests simultaneously. These are called Simultaneous Testing Procedures (STP) and involve adjusting either the relevant test statistics or critical value. Two classes of STP are considered here, those based on Familywise Error Rate (FWER) control and those which control the False Discovery Rate (FDR).

2.9.1 Familywise Error Rate Control

FWER is the probability of rejecting at least one true H_0 in a family of hypothesis testing problems[7]. In this case

$$FWER = \mathbb{P}[\text{Reject any true } H_{0,v}]. \quad (2.16)$$

The standard methodology for a FWER control STP takes a family of p-values and indicates which hypotheses are to be rejected while controlling the FWER. This can be done by computing adjusted p-values \tilde{p}_i for which, by rejecting $H_{0,i}$ only when $\tilde{p}_i \leq \alpha$, the analyst ensures

$$FWER \leq \alpha.$$

A STP can control FWER in either a weak sense or a strong sense. Weak FWER control relies on the assumption that all the null hypotheses are in fact true. This is known as the complete null hypothesis. Strong FWER control however, holds even when only some subcollection of the null hypotheses are true. This is the partial null hypothesis[27].

Two methods of FWER control are considered in this analysis. The first is a Holm's step-down STP[14]. The second is Westfall & Young's resampling STP[27].

Holm's step-down STP

In order to implement a step-down STP first arrange family of p-values in order, such that,

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(N)}, \quad (2.17)$$

where N is the number of hypotheses being tested simultaneously. A testing algorithm is then applied to the ordered p-values such that if $H_{0,(i)}$ is rejected then all $H_{0,(j)}$ for $j \in 1, \dots, i - 1$ must also be rejected.

In order to define Holm's step-down STP first consider the STP based on Bonferroni's bound. Bonferroni's STP constitutes a classic and conservative FWER control method. Bonferroni's STP is a single step STP, where the STP performs the equivalent multiplicity control for all tests, regardless of ordering. When implementing Bonferroni's STP reject all $H_{0,i}$ for which

$$p_i \leq \frac{\alpha}{N}$$

in terms of adjusted p-values this can be expressed

$$\tilde{p}_i = \min\{Np_i, 1\}$$

and $H_{0,i}$ is rejected when $\tilde{p}_i \leq \alpha$

To show that implementing these adjusted p-values controls FWER recall Booles inequality

$$\mathbb{P}[\cup A_i] \leq \sum \mathbb{P}[A_i]$$

Now let I_0 index the true null hypotheses, having N_0 members. This implies

$$\begin{aligned} FWER &= \mathbb{P}\left[\bigcup_{I_0} \left(p_i \leq \frac{\alpha}{N}\right)\right] \leq \sum_{I_0} \mathbb{P}\left[p_i \leq \frac{\alpha}{N}\right] \\ &= N_0 \frac{\alpha}{N} \leq \alpha \end{aligned}$$

Thus Bonferroni's bound STP controls the FWER in the strong sense, regardless of the dependence structure of the data.

Holm's step-down STP involves ordering p-values as in (2.17) and rejecting $H_{0,(i)}$, the hypothesis with p-value $p_{(i)}$ if

$$p_{(j)} \leq \frac{\alpha}{N - j + 1}, \text{ for } j = 1, 2, \dots, i.$$

Holm[14] proved that this testing algorithm provided strong FWER control and was a general improvement on Bonferroni’s STP.

In order to implement this STP the adjusted p-values expressed as

$$\tilde{p}_{(i)} = \min \left\{ \max_{j \leq i} \{(N - j + 1)p_{(v)}\}, 1 \right\} \quad (2.18)$$

and $H_{0,(v)}$ rejected when $\tilde{p}_v \leq \alpha$. This exercises a strong control on FWER such that $FWER \leq \alpha$.

Holm’s step-down STP benefits from being computationally straightforward, however, as it relies on probability inequalities, avoiding dependency structures within the data, it can be excessively conservative. This is especially the case when N is large and the hypotheses being tested are correlated, such as in fMRI data. Other methods of FWER control are offered by random field theory and resampling from the data. These methods take into account the dependency structure of the data and can offer a less conservative alternative to methods based on Bonferroni’s bound. Of these approaches, this analysis considers only resampling methods.

Westfall & Young’s Resampling STP

Westfall & Young[27] presented a method for computing adjusted p-values for multiple testing problems by resampling under the null distribution, and outlined the conditions under which this method provided a strong control of the FWER.

There are two general approaches to resampling, bootstrap and permutation methods, which are performed under the null hypothesis. A bootstrap involves repeated resampling from the original dataset with replacement, while permutations resample from the dataset without replacement[19]. These methods construct a null dataset, against which the sample is compared.

Resampling allows the analyst to construct an approximation \hat{G} for the unknown distribution G which generated the data X . The major benefit of resampling over parametric methods is that the distribution \hat{G} implicitly captures the dependency structures and distributional characteristics of G . Accounting for these dependency structures usually reduces the adjusted p-values, as compared to Holm’s STP. This improves the power of the test.

A resampling STP requires methods for testing individual hypotheses, a predefined collection of hypotheses to be tested, and a probabilistic model for the data generating process. With this in place, resampling should be done in a way which reflects H_0 . Provided the condition of subset pivotality holds for the statistic t being resampled, this STP controls FWER in the strong sense.

Definition 2.9.1 *The distribution of t meets the subset pivotality condition if the joint distribution of the subvector $\{t_i; i \in K\}$ is identical under the restrictions $\cap_{i \in K} H_{0,i}$ and H_0^C , for all subsets $K = \{i_1, \dots, i_j\}$ of true null hypotheses.[27]*

Here H_0^C is the complementary set to the true null hypotheses.

Given these conditions hold, a single step resampling method for computing adjusted p-values can be implemented.

To gain an intuition for how such a STP controls the FWER, consider the data X generated according to the distribution G . Some family of test statistics t is then computed for

X . Given an appropriate resampling scheme, X^* is resampled from X under some null hypothesis, such that an estimate for G, \hat{G} is computed. A family of test statistics t^* is also calculated for X^* . Each family of t^* will take on a set of values, which represents a set of values which can occur purely by chance when the null hypothesis is true. Thus any t which fall outside the range of t^* are unlikely to come from the null distribution. By counting how often this occurs, adjusted p-values can be computed. Developing this intuition allows it to be shown that a resampling STP subject to some conditions controls the FWER in a strong sense.

In the case of a FC study employing the metric cross-correlation, subset pivotality does not hold. This is illustrated by Example 2.2 of Westfall & Young. Despite this a resampling STP is implemented for comparison with Holm's STP.

Computation of adjusted p-values requires an appropriate resampling method. The Stationary Bootstrap[20] was employed in this analysis in order to resample from the fMRI time series. This is done by sampling blocks of a geometrically distributed size, of expected length $1/p$, with replacement. Resampling in this manner creates a pseudo time series of length T retains the stationary properties of the original series, in contrast with block resampling.

Algorithm 1 The Stationary Bootstrap

```

procedure RESAMPLING STATIONARY PSEUDO TIME SERIES  $X^*$ 
  Set  $T$ , length of pseudo time series required
  Select  $X_1^* = X_{t_1}$  uniformly at random from  $X_1, \dots, X_T$ 
  for  $i$  in 2 to  $T$  do
    Generate  $u$  from  $U \sim \mathcal{U}[0, 1]$ 
    if  $u < p$  then
      Select  $X_i^* = X_{t_i}$  uniformly at random from  $X_1, \dots, X_T$ 
    else if  $u > p$  then
      Set  $X_i^* = X_{(t_i \bmod T)+1}$ 
    end if
  end for
end procedure

```

Having established the stationary bootstrap resampling scheme a version of the Westfall & Young max T STP is implemented for the construction of adjusted p-values for a one tailed hypothesis test. Algorithm 2 outlines the implementation of this STP

The adjusted p-values here are calculated under the null hypothesis (2.13), that is, the series for each voxel has a correlation of 0 with the seed series. Thus the family of test statistics t^* is calculated for the original seed s and X^* . In order to obtain adjusted p-values for a two tailed test Algorithm 2 is run for the data. The adjusted p-values are then

$$\tilde{p}_v = \min \left\{ 2 \cdot \min \{ \tilde{p}_{upper,v}, \tilde{p}_{lower,v} \}, 1 \right\} \quad (2.19)$$

Westfall & Young also outlined a resampling method for evaluating the a resampling STP.

This algorithm estimates the true critical value of the test α_0 . If $\hat{\alpha}_i >> \alpha$ this indicates that $H_{0,v}$ is rejected too often, while $\hat{\alpha}_i << \alpha$ indicates that $H_{0,v}$ is rejected too rarely.

Algorithm 2 Two-tailed Single Step Resampling Method

```
1: procedure CALCULATING ADJUSTED P-VALUES
2:   Set  $N_{Boot}$ , the number of samples to be taken
3:   Initialise Count Variables, ( $COUNTU$ ,  $COUNTL$ )
4:   for 1 to  $N_{Boot}$  do
5:     Resample data  $X^*$  from  $X$  using the Stationary Bootstrap (Algorithm 1)
6:     Calculate test statistics,  $t^*$  for  $X^*$ 
7:     Find  $t_{max}^* = \max\{t_{sv}^*\}$ 
8:     Find  $t_{min}^* = \min\{t_{sv}^*\}$ 
9:     if  $t_{max}^* > t_{sv}$  then
10:       $COUNTU \leftarrow COUNTU + 1$ 
11:     end if
12:     if  $t_{min}^* < t_{sv}$  then
13:       $COUNTL \leftarrow COUNTL + 1$ 
14:     end if
15:   end for
16:    $\tilde{p}_{upper} = \frac{COUNTU}{N_{Boot}}$ ,  $\tilde{p}_{lower} = \frac{COUNTL}{N_{Boot}}$ 
17:    $\tilde{p} = \min\left\{2 \cdot \min\{\tilde{p}_{upper}, \tilde{p}_{lower}\}, 1\right\}$ 
18:
19: end procedure
```

Algorithm 3 Evaluation of Resampling-Based p-values

```
1: procedure APPROXIMATE  $\alpha_0$ 
2:   Set  $N_{BOOT}$ , the number of bootstrap samples
3:   Set  $NSIM$ , the number of simulations
4:   Initialise simulation counter:  $COUNTS$ 
5:   Initialise bootstrap counter:  $COUNTB$ 
6:   for 1 to  $NSim$  do
7:     Run Algorithm for adjusting p-values (Algorithm 2), setting  $N_{Boot}$  to  $N_{BOOT}$ 
8:     if  $p^* \leq \alpha$  then
9:        $COUNTS \leftarrow COUNTS + 1$ 
10:    end if
11:  end for
12:   $\hat{\alpha} = \frac{COUNTS}{NSIM}$ 
13: end procedure
```

Both cases indicate that the family of null hypotheses and test statistics do not describe the data. Algorithm 3 outlines the procedure required to perform this evaluation.

Algorithm 3 is very computationally demanding and as such was not included in the final analysis, despite the insight it would have provided.

2.9.2 False Discovery Rate Control

FDR control represents a shift in philosophy away from FWER control, offering an alternative, less stringent, STP for assessing multiple testing problems[7].

Consider the situation where N hypotheses are being tested simultaneously and the decision rule \mathcal{D} either rejects, or fails to reject H_0 . For N_0 of the N hypotheses H_0 is true, while for N_1 , H_1 is in fact true. The hypothetical performance of \mathcal{D} is outlined in Table 2.2.

Decision/Truth	H_0 True	H_0 False	Total
Reject H_0	$a_{\mathcal{D}}$	$b_{\mathcal{D}}$	$R_{\mathcal{D}}$
Fail to Reject H_0	$N_0 - a_{\mathcal{D}}$	$N_1 - b_{\mathcal{D}}$	$N - R_{\mathcal{D}}$
Total	N_0	N_1	N

Table 2.2: \mathcal{D} rejects $R_{\mathcal{D}}$ out of N null hypothesis. $a_{\mathcal{D}}$ of the hypotheses rejected were incorrect decisions or “false discoveries”, whereas $b_{\mathcal{D}}$ represents the number of “true discoveries.” The false discovery proportion (Fdp) is $a/R_{\mathcal{D}}$.

Recalling the definition of FWER, thus according to Table 2.2, $FWER = \mathbb{P}[a > 0]$. FDR control STP as proposed by Benjamini & Hochberg[2] (BHSTP) controls $\mathbb{E}[a/R]$.

The BHSTP first orders the p-values as in (2.17). Then, for a fixed value of q in $(0,1)$, let i_{max} be the largest index for which

$$p_{(i)} \leq \frac{i}{N}q \quad (2.20)$$

and reject $H_{0,(i)}$ if

$$i \leq i_{max}.$$

The BHSTP returns the adjusted p-values

$$\tilde{p}_{(v)} = \min \left\{ \frac{N}{v}p_{(v)}, 1 \right\} \quad (2.21)$$

where $H_{0,(v)}$ is rejected when $\tilde{p}_{(v)} \geq q$.

The theorem proved by Benjamini & Hochberg[2] states

Theorem 2.9.1 *For independent test statistics and under a partial null hypothesis, the BHSTP controls the expected false discovery proportion (Fdp) at q .*

$$\mathbb{E}[Fdp_{BHSTP(q)}] = \frac{N_0}{N}q \leq q$$

Note that the theorem assumes that the test statistics are independent under the null hypothesis. Although this assumption can be weakened to *positive regression dependence on each one from a subset* I_0^2 , neither assumption holds in the case of a FC study of fMRI data. For cases in which the test statistics have a general dependence structure Benjamini & Yekuteili[3] proved

Theorem 2.9.2 *When the BHSTP is performed with $q/(\sum_{i=1}^N \frac{1}{i})$ replacing q in 2.20, the procedure always controls the expected Fdp at level less than or equal to $\frac{N_0}{N}q$*

Implementing this theorem returns the adjusted p-values

$$\tilde{p}_{(v)} = \min \left\{ \left(\sum_{j=1}^N \frac{1}{j} \right) \frac{N}{v} p_{(v)}, 1 \right\} \quad (2.22)$$

This represents a severe penalty but does preserve FDR control for families of hypotheses with general dependency structures, as is the case for this analysis. Thus, this is the FDR control STP implemented for this analysis.

2.10 Simulation Study

A simulation study is conducted, investigating the impact of considering temporal correlation when performing hypothesis tests. Simulations of moving average (MA) and vector autoregressive (VAR) bivariate time series are tested and the performance of the test statistics t_{ij}^f (2.8) and t_{ij}^r (2.11) assessed.

2.10.1 Bivariate Time Series

First, define the bivariate white noise process Z_t where

$$Z_t \sim \mathcal{N}(0, \Sigma). \quad (2.23)$$

Note that Σ is symmetric, that is

$$\Sigma = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{pmatrix}$$

where $\sigma_{1,2} = \sigma_{2,1}$.

Now consider the bivariate time series

$$X_t = \begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix}.$$

X_t is defined as a MA(1) process when

$$X_t = \Theta Z_{t-1} + \mathbb{I}Z_t$$

²For any increasing set D and for each $i \in I_0$, $\mathbb{P}(\mathbf{X} \in D | X_i = x)$ is non-decreasing in x [3]

where

$$\Theta = \begin{pmatrix} \theta_{1,1} & \theta_{1,2} \\ \theta_{2,1} & \theta_{2,2} \end{pmatrix}, \mathbb{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Similarly, X_t is defined as a VAR(1) process when

$$X_t = \Phi X_{t-1} + \mathbb{I}Z_t, \quad (2.24)$$

where

$$\Phi = \begin{pmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{2,1} & \phi_{2,2} \end{pmatrix}$$

Note that in both instances X_t is a linear process, and can be represented in the form

$$X_t = \sum_{j \in \mathbb{Z}} C_j Z_{t-j}$$

which implies that the covariance matrix function of X_t is

$$\Gamma(h) = \sum_{j \in \mathbb{Z}} C_{j+h} \Sigma C_j'. \quad (2.25)$$

2.10.2 Mean Squared Error

The Mean Squared Error (MSE) is a useful statistic for comparing estimators in simulation studies. Casella & Berger [5] define MSE as

$$\mathbb{E}[W - \lambda]^2 \quad (2.26)$$

where W is an estimator for the parameter λ . It can be shown that

$$\mathbb{E}[W - \lambda]^2 = \text{Var}(W) + (\text{Bias}(W))^2 \quad (2.27)$$

where $\text{Bias}(W) = \mathbb{E}|W - \lambda|$.

2.11 Simulation Algorithms

Having outlined the theory required for the simulation study, two simulation algorithms are specified. The first calculates $\hat{\sigma}_{X_1 X_2}$ 2.12 and $r_{X_1 X_2}(0)$ 2.2 for N simulated bivariate time series of length T . The second algorithm simulates N bivariate time series of length T and for each series computes t_F , and t_R .

With these algorithms in place, a simulation study investigating the impact of accounting for temporal correlation on the power of a hypothesis test can be performed.

Algorithm 4 Calculation of $\hat{\sigma}_{X_1X_2}$ & $r_{X_1X_2}(0)$ for series X_t

```
1: procedure SIMULATION PROCEDURE
2:   Specify the number of series to be generated -  $N$ 
3:   Specify the length of each series to be generated -  $T$ 
4:   for 1 to  $N$  do
5:     Simulate specified time series of length  $T$ .
6:     Calculate estimate  $\hat{\sigma}_{X_1X_2}$ 
7:     Calculate  $r_{X_1X_2}(0)$ 
8:   end for
9: end procedure
```

Algorithm 5 Calculation of t_F and t_R for series X_t

```
1: procedure SIMULATION PROCEDURE
2:   Specify the number of series to be generated -  $N$ 
3:   Specify the length of each series to be generated -  $T$ 
4:   for 1 to  $N$  do
5:     Simulate specified time series of length  $T$ .
6:     Calculate  $t_F$  for  $X_t$ .
7:     Calculate  $t_R$  for  $X_t$ .
8:   end for
9: end procedure
```

Chapter 3

Simulation Study

This simulation study comprises of three experiments, each investigating a different aspect of the implementation of Roy's Theorem. The first experiment assesses the accuracy of the estimate $\hat{\sigma}_{ij}$ (2.12) of σ_{ij} (2.9) and investigates the estimate asymptotic properties. The second and third experiments compare the power functions (2.14) of hypothesis tests implementing t_F (2.8) and t_R (2.11), under various conditions, where the hypothesis being tested is $H_0 : \rho_{X_1X_2}(0) = 0$ and $H_1 : \rho_{X_1X_2}(0) \neq 0$

3.1 Experiment I

The first experiment simulates $N = 1000$ bivariate MA(1) time series for series lengths $T = 200, T = 400, T = 800$. The MA(1) series (2.24) is specified by

$$\Theta = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.9 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \quad (3.1)$$

From the expression for the covariance function of a linear process (2.26), it is found analytically that in this case

$$\sigma_{X_1X_2} = 0.05155$$

Simulations were performed according to Algorithm 4, the results of which are presented in Table 3.1 and Figure 3.1. Mean Squared Error of $\hat{\sigma}_{X_1X_2}$ and an estimate of the variance if IID realisations of X_t ($(1 - r_{X_1X_2}^2(0))^2$) could be compared.

T	Mean	Variance	Min	Median	Max	MSE($\hat{\sigma}_{X_1X_2}$)	MSE($(1 - r^2)^2$)
200	0.0519	2.66×10^{-4}	0.0176	0.0504	0.1263	2.66×10^{-4}	3.43×10^{-4}
400	0.0527	1.33×10^{-4}	0.0242	0.0515	0.1047	1.34×10^{-4}	2.82×10^{-4}
800	0.0527	6.79×10^{-5}	0.0326	0.0519	0.0885	6.93×10^{-5}	2.65×10^{-4}

Table 3.1: Summary statistics of $\hat{\sigma}_{X_1X_2}$ for $N = 1000$ simulations of a bivariate MA(1) time series specified by (3.1) of length T and comparison of MSE under IID assumption and under temporal correlation.

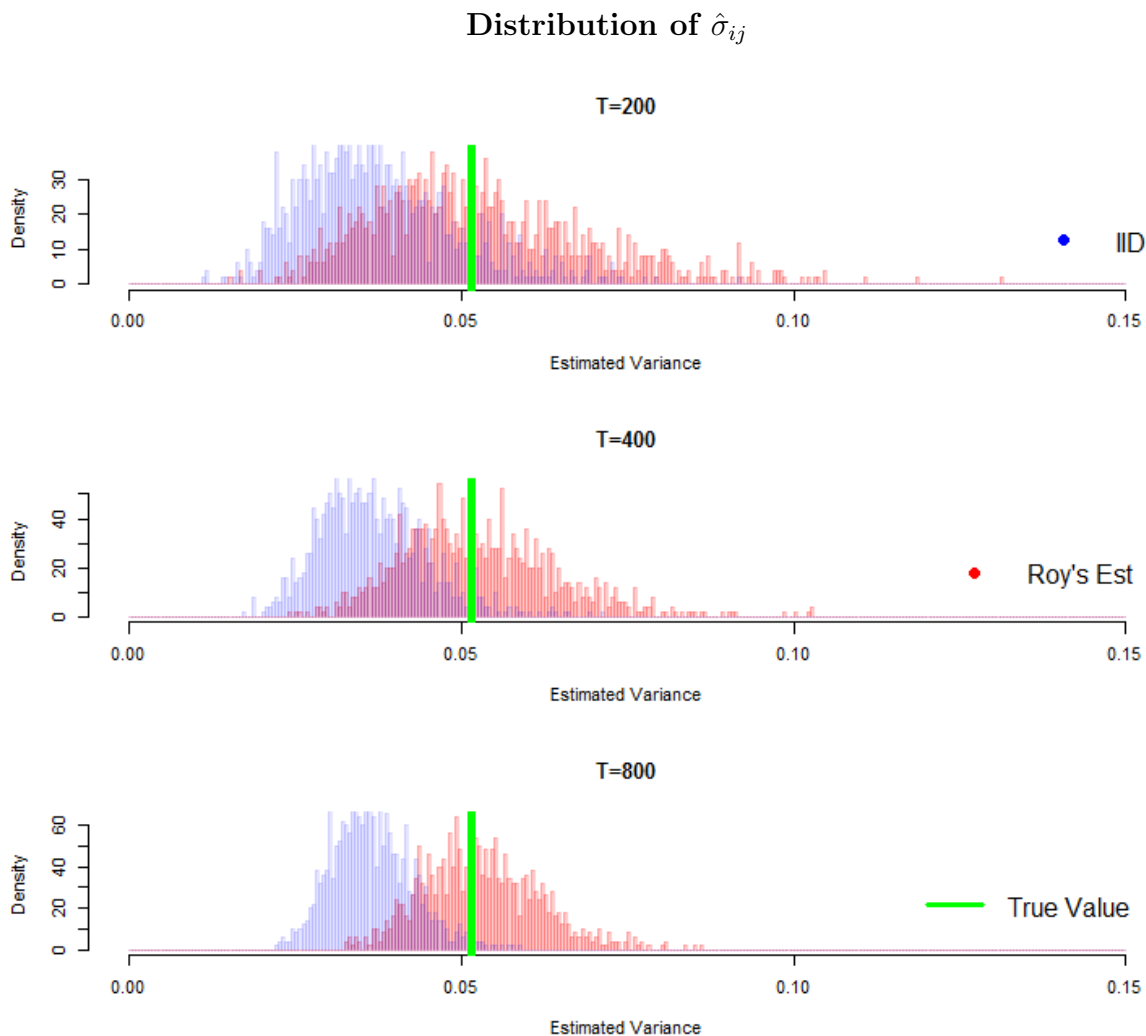


Figure 3.1: Histograms of $\hat{\sigma}_{X_1 X_2}$ and $(1 - r^2(0))^2$ for $N = 1000$ simulations of a bivariate MA(1) time series specified by (3.1) of length T . The green line represents the analytic solution given by Roy's Theorem.

This experiment shows that, in the case of a bivariate MA(1) series at least, $\hat{\sigma}_{ij}$ is a good estimator for σ_{ij} , correcting the bias introduced by the IID assumption and showing a lower MSE, as predicted by Roy's Theorem. Focussing on $\text{MSE}(\hat{\sigma}_{ij})$, it is close to the sample variance of $\hat{\sigma}_{ij}$ for each T , suggesting that it is not a biased estimator. It is also decreasing as T increases, suggesting that $\hat{\sigma}_{ij}$ does in fact converge to σ_{ij} as $T \rightarrow \infty$.

3.2 Experiment II

The second experiment is designed to compare the power function (2.14) of tests using t_{ij}^f and t_{ij}^r , when H_0 is in fact true. This is the Type I error rate and ideally should be approximately equal to the critical value α .

The analysis will be performed on $N = 1000$ series of length $T = 500$. The series to be

analysed are MA(1) (2.24) specified by

$$\Theta = \begin{pmatrix} \theta & 0 \\ 0 & \theta \end{pmatrix}, \text{ for } \theta \in \{0.1, 0.25, 0.5, 0.75, 0.9\}, \quad (3.2)$$

and VAR(1) (2.25) specified by

$$\Phi = \begin{pmatrix} \phi & 0 \\ 0 & \phi \end{pmatrix}, \text{ for } \phi \in \{0.1, 0.25, 0.5, 0.75, 0.9\}. \quad (3.3)$$

For all series

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

thus $\rho_{X_1 X_2}(0) = 0$ and H_0 is true.

In each case simulations are run according to Algorithm 5. The critical region \mathcal{R}_α for each test is constructed by setting $\alpha = 0.05$

Power Function when H_0 True

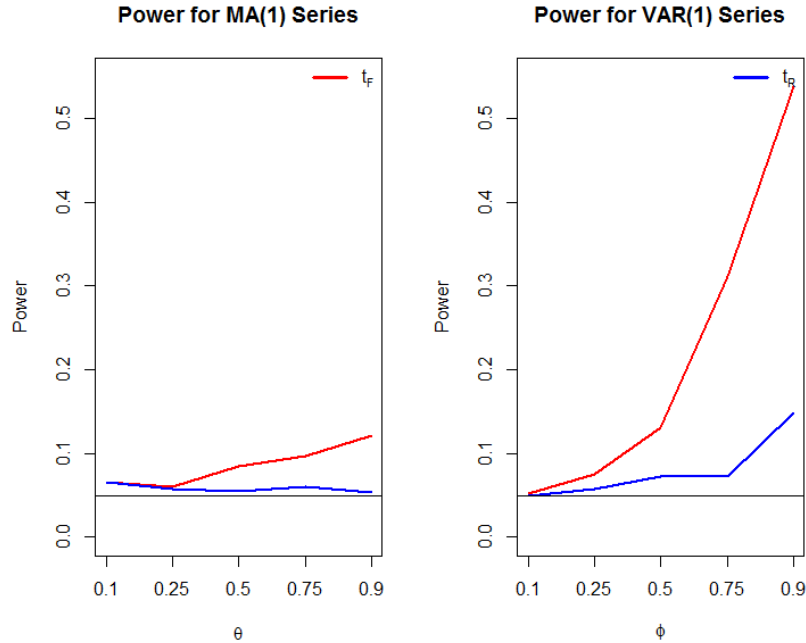


Figure 3.2: Plots of the power of hypothesis tests when H_0 true (Type I error rate) for values of θ in bivariate MA(1) series and for values of ϕ in bivariate VAR(1) series. For construction of \mathcal{R}_α $\alpha = 0.05$, thus a test is performing well when it's power is approximately 0.05. The solid black line represents this $\alpha = 0.05$

Figure 3.2 demonstrates that testing hypothesis using $t_{X_1 X_2}^r$ results in a reduced Type I error rate in all instances. This is to be expected as temporal correlation is now being accounted for in the test statistics. It is important to note that while both test statistics have elevated Type I error rates when ϕ is large for a VAR(1) series, $t_{X_1 X_2}^r$ far outperforms $t_{X_1 X_2}^f$, which offers very little insight into the underlying correlation between series. It is

also worth noting that in situations where temporal correlation over many time lags is suspected to be a feature of the data the window function in the $\hat{\sigma}_{ij}$ can be adjusted to account for this.

3.3 Experiment III

The third and final simulation experiment investigates the power function (2.14) of the test statistics when H_1 is true, and $\rho_{X_1X_2}(0) \neq 0$. In this case the power is (1-Type II Error Rate) and ideally should be close to 1.

As in Experiment II, the analysis simulated $N = 1000$ series of length $T = 500$ for all MA(1) and VAR(1) series specified by (3.2), (3.3), but in this case

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \text{ for } \rho \in \{0.5, 0.075, 0.1, 0.125, 0.15\}$$

thus $\rho_{X_1X_2}(0) \neq 0$ and H_1 is true in each case. Figure 3.3 illustrates the results of the experiment run fo MA(1) series with Figure 3.4 illustrates the same for VAR(1) series.

In most cases hypothesis testing with $t_{X_1X_2}^f$ is slightly more powerful than using $t_{X_1X_2}^r$, indicating that it rejects more false null hypotheses. Major discrepancies occur in VAR(1) series with large values for ϕ , where $t_{X_1X_2}^f$ is far more powerful than $t_{X_1X_2}^r$. However it is worth noting that the slope of the power function for these values ($\phi \in \{0.75, 0.9\}$) is relatively low, indicating that it is not very responsive to changes in the underlying value of ρ .

3.4 Simulation Study Conclusions

The simulation study found that hypothesis tests using $t_{X_1X_2}^f$ to approximate the null distribution were always more powerful than those using $t_{X_1X_2}^r$, however this does not necessarily indicate that $t_{X_1X_2}^f$ is the most appropriate test statistic. In fact, the situations where there was a major, desirable difference in power, arose only because $t_{X_1X_2}^f$ utterly failed to describe the null distribution, as illustrated by Figure 3.5.

Figure 3.5 demonstrates the intuition that the variance of a series with high serial autocorrelation will be severely underestimated when the assumption of IID realisations of the series is made. Thus, it is to be expected that the family of test statistics t^r provide a far more accurate representation of the truth. In cases such as this increasing the width of the window function, i.e. adjusting H so the window takes more lags into account, will improve the representation of the truth further.

It is fair to conclude that $t_{X_1X_2}^r$ offers a better alternative to $t_{X_1X_2}^f$ for hypothesis testing, especially in the context of strongly autocorrelated series. Precisely what this implies for FC studies of fMRI data, remains to be seen.

Power Function when H_1 True for MA(1) Series

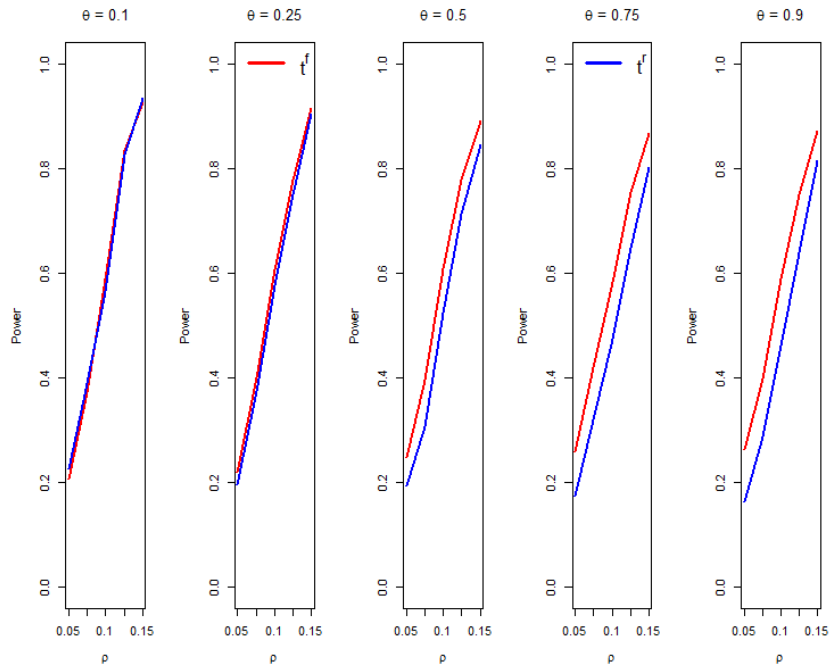


Figure 3.3: Plots of the power of hypothesis tests when H_1 true (1 - Type II error rate) for values of θ & ρ in bivariate MA(1) series. For construction of \mathcal{R}_α , $\alpha = 0.05$. The test is performing well when the power is close to 1

Power Function when H_1 True for VAR(1) Series

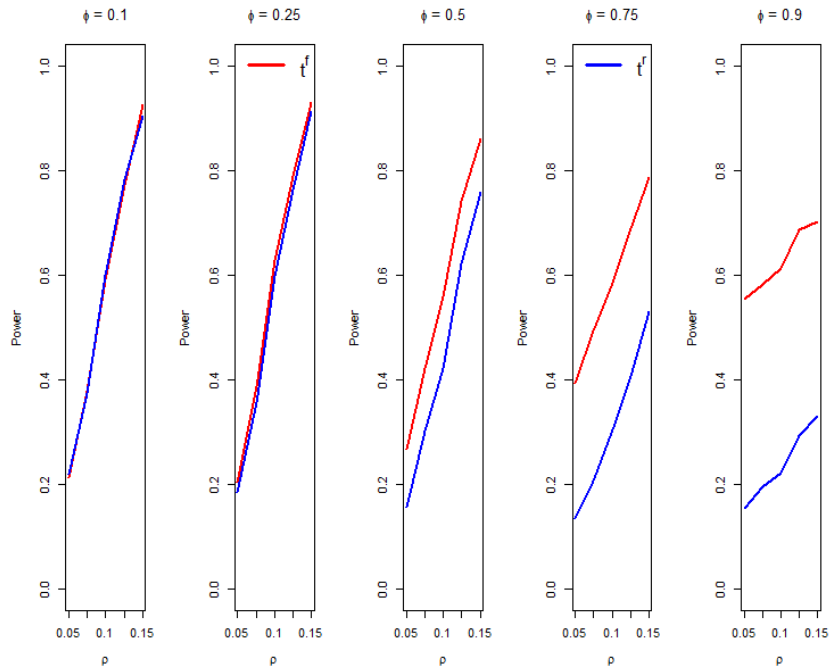


Figure 3.4: Plots of the power of hypothesis tests when H_1 true (1 - Type II error rate) for values of ϕ & ρ in bivariate VAR(1) series. For construction of \mathcal{R} , $\alpha = 0.05$. The test is performing well when the power is close to 1

Distributions of t^r & t^f when $\phi = 0.9$ and H_0 True

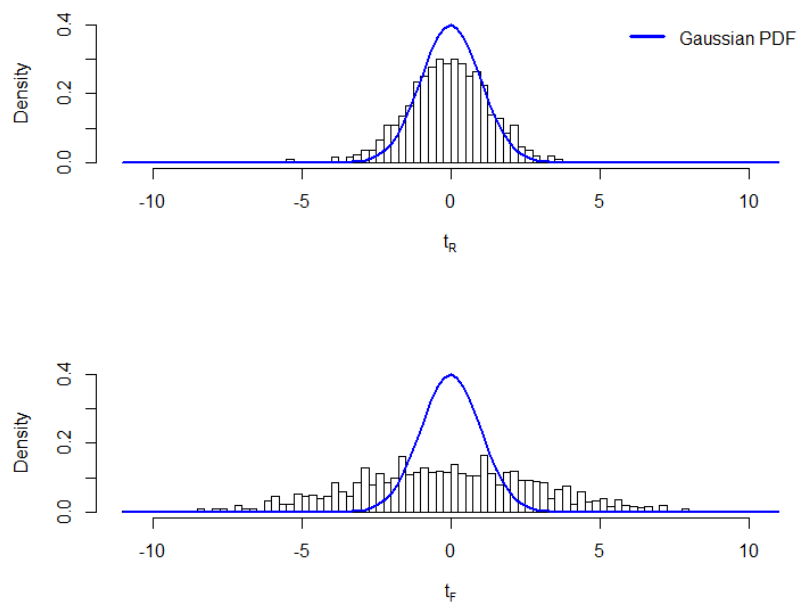


Figure 3.5: Histograms comparing the distributions of test statistics t^f & t^r to the theoretic asymptotic null distribution when H_0 is in fact true.

Chapter 4

Results of Functional Connectivity Analysis of fMRI Data

The methods detailed in Chapter 2 were applied to the analysis of a fMRI dataset of an individual's brain. Functional Connectivity, as quantified by the sample cross-correlation between the seed cube centred on voxel (33,19,23) of the native space and every other voxel in the brain is examined. Presented in this chapter is nothing more than the results of implementing the methods for the dataset.

The first look at the data involves the calculation and comparison of the families of test statistics t^f (2.8) and t^r (2.11). Having done this the families of hypotheses can be tested. Firstly, the naive hypothesis testing outlined in section 2.8.1 is performed. Secondly, controls of the Familywise Error Rate (FWER) are implemented, as per section 2.9.1. Finally, the False Discovery Rate (FDR) control detailed in section 2.9.2 is applied to the test statistics.

4.1 Analysis of Test Statistics

Two families of test statistics, as presented in sections 2.4-2.7, are calculated for the fMRI dataset. These are the metrics for analysing FC.

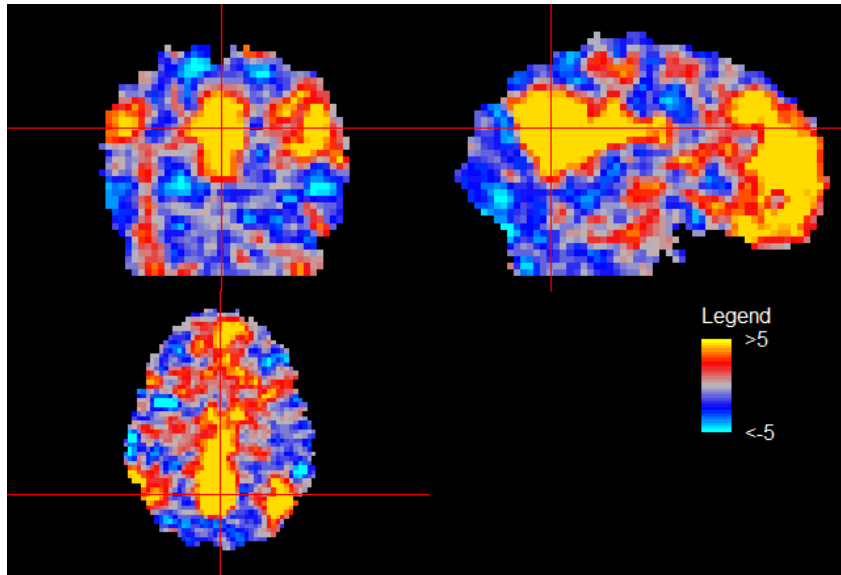
A first look at the dataset is provided by mapping the test statistic for each voxel onto an orthographic presentation of the brain, centred on the seed. This provides an insight into the distribution of FC networks throughout the brain.

In these maps, test statistics were limited to the interval $[-5,5]$. In cases where $t_{sv} < -5$, t_{sv} is set to -5, and in cases where $t_{sv} > 5$, t_{sv} is set to 5. The maps are presented in Figure 4.1.

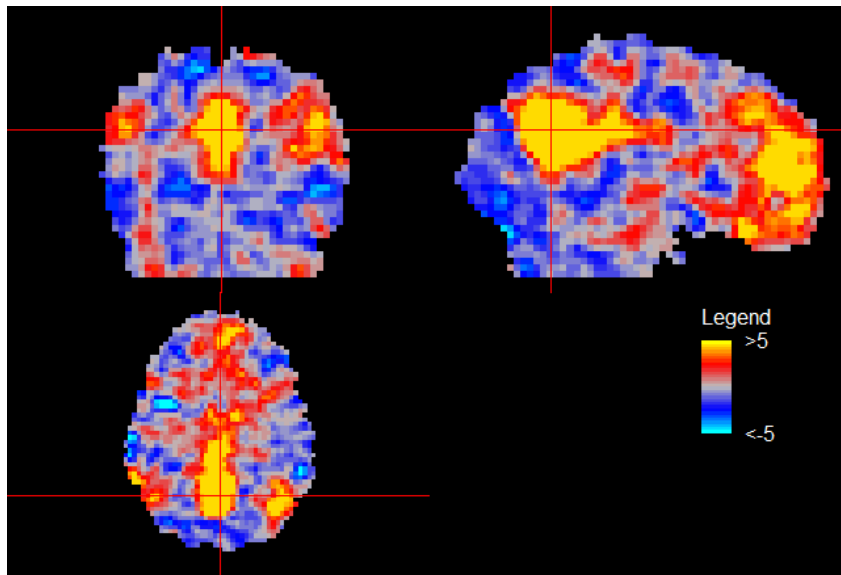
Both test statistics follow a similar spatial distribution within the brain. Both show strong positive dependencies between the PCC and MPF, reproducing the work of Fox [10]. However t_{sv}^r appears to have more values close to 0 and fewer extreme values. This warrants a closer investigation of the distribution of the test statistics themselves.

Figure 4.2 presents the distribution of each family of FC test statistics. Some members of each family were greater than 10, however these were not presented as most of these

Orthographic Map of Test Statistics



(a) Orthographic Map of t^F



(b) Orthographic Map of t^R

Figure 4.1: Map of test statistics limited to $[-5, 5]$, centred on the central seed voxel (33,19,23)

Histogram of Test Statistics

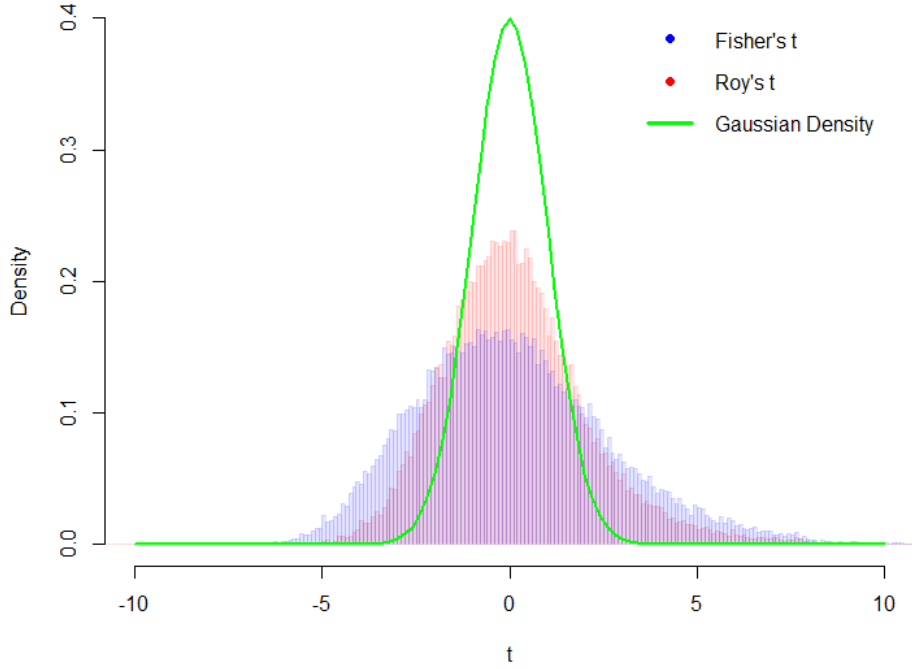


Figure 4.2: Histogram illustrating the distributions t^f and t^r over the interval $[-10,10]$

values related to the voxels making up the seed cube and skewed the distribution. The histogram of test statistics for the interval $[10,50]$ and the orthographic map presenting the spatial distribution of the differences ($t^r - t^f$) are included as appendix B. Figure 4.2 demonstrates clearly that neither family of test statistics are distributed according to the null hypothesis. However, the histogram presents an effect of correcting for temporal correlation similar to that seen in Figure 3.5 of the simulation study.

4.2 Naïve Hypothesis Testing

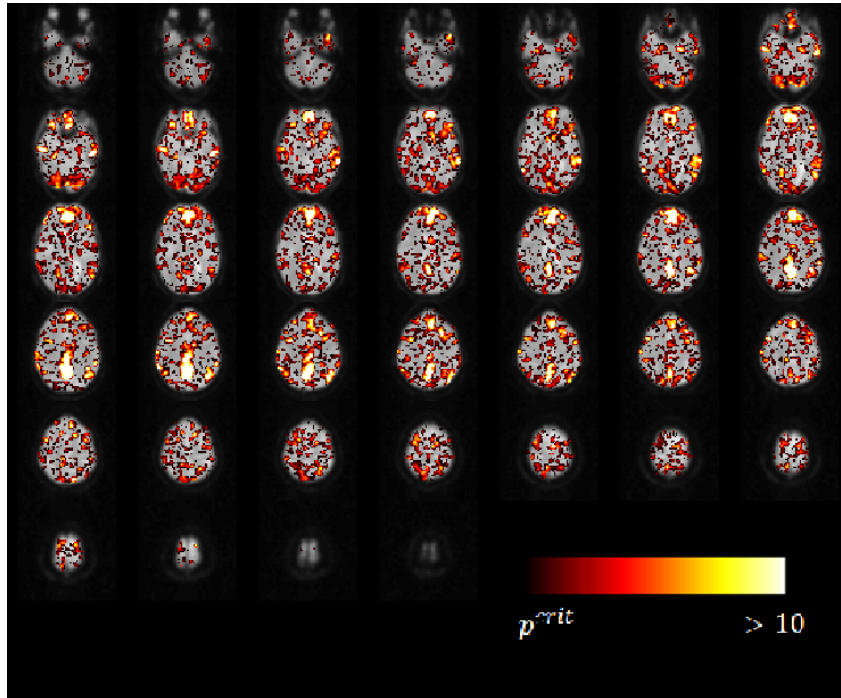
The naïve hypothesis testing procedure was outlined in section 2.8.1. This procedure is now applied to the fMRI dataset through the families of test statistics t^f and t^r testing the hypotheses 2.13 for the critical value $\alpha = 0.05$.

	H_0 rejected with $t < 0$	Failed to reject H_0	H_0 rejected with $t > 0$
t^f	10,914	27,744	10,783
t^r	6,196	36,088	7,157

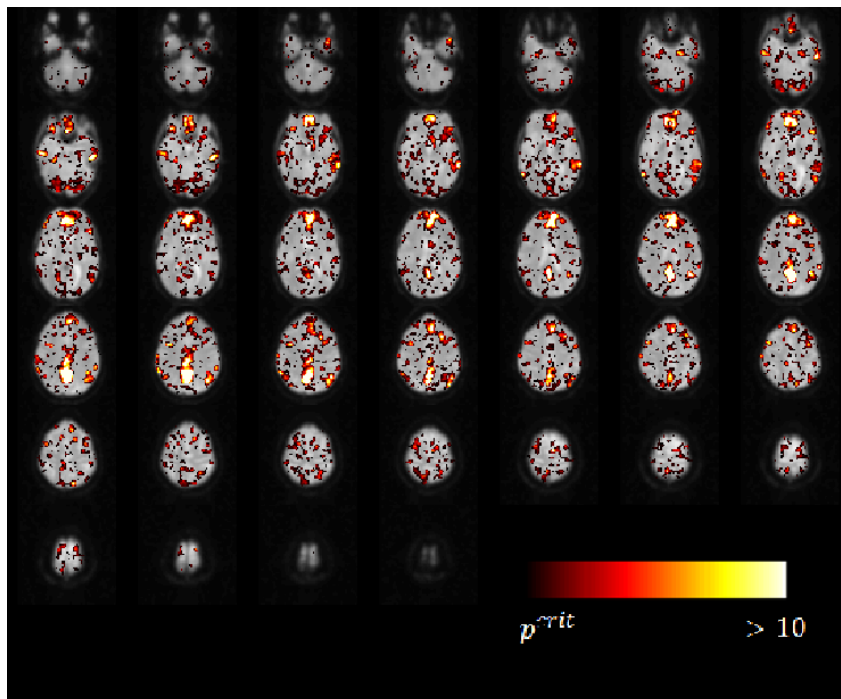
Table 4.1: Summary of results for naïve hypothesis testing on each family of test statistics.

The family of test statistics t^r rejects far fewer hypotheses than t^f . The spatial distribution of the family of p-values on the $-\log_{10}(\cdot)$ is illustrated in Figure 4.3.

Naive Hypothesis Testing Results



(a) Voxels exhibiting non-zero cross-correlation according to t^f



(b) Voxels exhibiting non-zero cross-correlation according to t^r

Figure 4.3: Voxels exhibiting non-zero cross-correlation with the seed region when naively testing each hypothesis for a critical value $\alpha = 0.05$. All p -values plotted on a $-\log_{10}(\cdot)$ scale

The least likely null hypotheses occur in the PCC and MPF regions of the brain.

4.3 Familywise Error Rate Controls

Two FWER control STP's were implemented for the calculation of adjusted p-values, Holm's STP and a Westfall-Young single-step resampling STP, as discussed in section 2.9.1. Each STP was constructed to ensure that $FWER \leq \alpha$ for $\alpha = 0.5$, rejecting $H_{0,v}$ when $\tilde{p}_v \leq \alpha$. Thus $p_{crit} = \max\{p_i \text{ such that } \tilde{p}_i \leq \alpha\}$.

The first of these implemented is Holm's procedure, the results of which are summarised in Table 4.2 with the spatial distribution of adjusted p-values for rejected hypotheses presented in Figure 4.4.

	H_0 rejected with $t < 0$	Failed to reject H_0	H_0 rejected with $t > 0$
t^f	555	46,523	2,363
t^r	69	48,184	1,188

Table 4.2: Summary of results for implementation of Holm's STP on each family of test statistics.

As expected, most of those hypotheses which were naively rejected are no longer deemed to be of interest, although positive test statistics are more robust.

The second method of controlling FWER employed a Westfall-Young single-step Resampling STP, using a stationary bootstrap resampling scheme for 500 bootstrap samples. Only 500 bootstrap samples were obtained due to the computational expense of the Algorithm. The results of this are summarised in Table 4.3 with the spatial distribution of adjusted p-values for rejected hypotheses presented in Figure 4.6.

	H_0 rejected with $t < 0$	Failed to reject H_0	H_0 rejected with $t > 0$
t^f	0	49,210	231
t^r	1	48998	442

Table 4.3: Summary of results for implementation of Westfall & Young's STP on each family of test statistics.

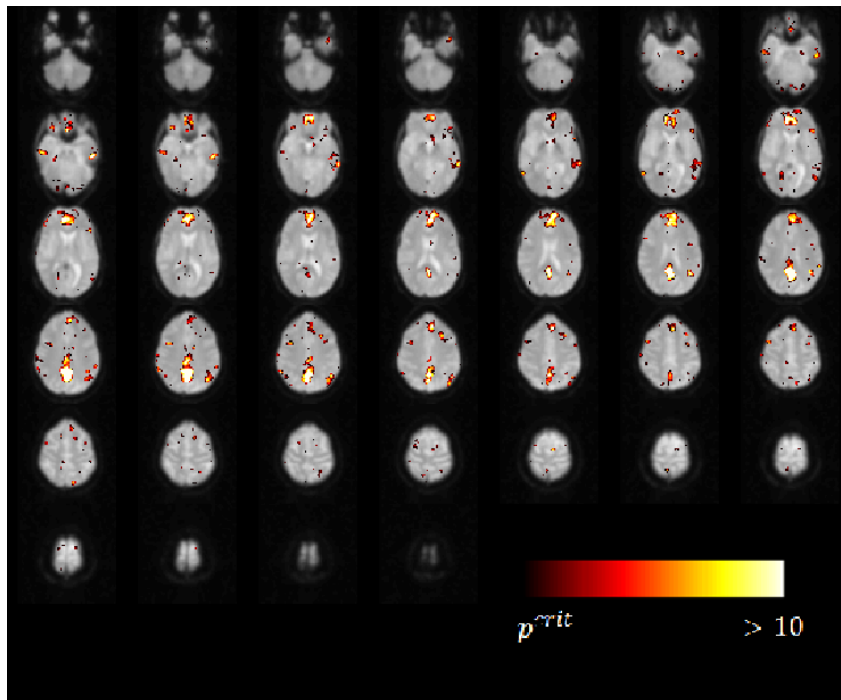
Somewhat surprisingly the Westfall & Young resampling STP is less powerful than Holm's STP and the test t^r is more powerful than t^f . Despite this dependency between the PCC and MPF are still observed to a degree.

It is worth including a histogram examining the test statistics for one resampled pseudo series illustrated in Figure 4.5. Note that neither family of test statistics follow the theoretical null distribution, both having a lower peak and fatter tails.

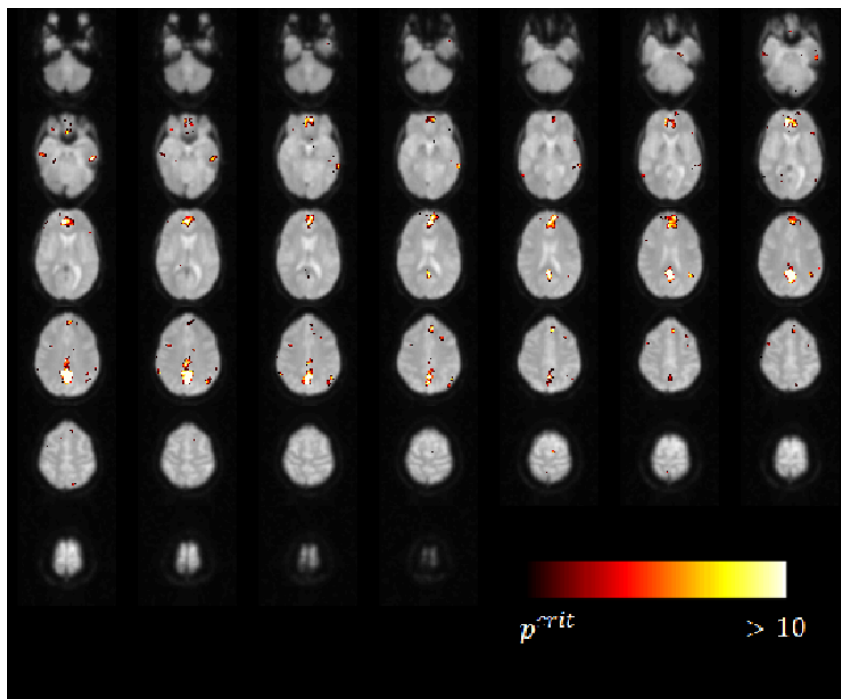
4.4 False Discovery Rate Control

FDR control was carried out for each family of test statistics. As detailed in section 2.9.2 the families of p-values did not satisfy the assumptions underpinning Benjamini &

Holm's STP Results



(a) Voxels exhibiting non-zero cross-correlation according to t^f



(b) Voxels exhibiting non-zero cross-correlation according to t^r

Figure 4.4: Voxels exhibiting non-zero cross-correlation with the seed region when controlling the FWER $\leq \alpha$ using Holms Procedure. All adjusted p-values plotted on a $-\log_{10}(p)$ scale

Histogram of pseudo Test Statistics

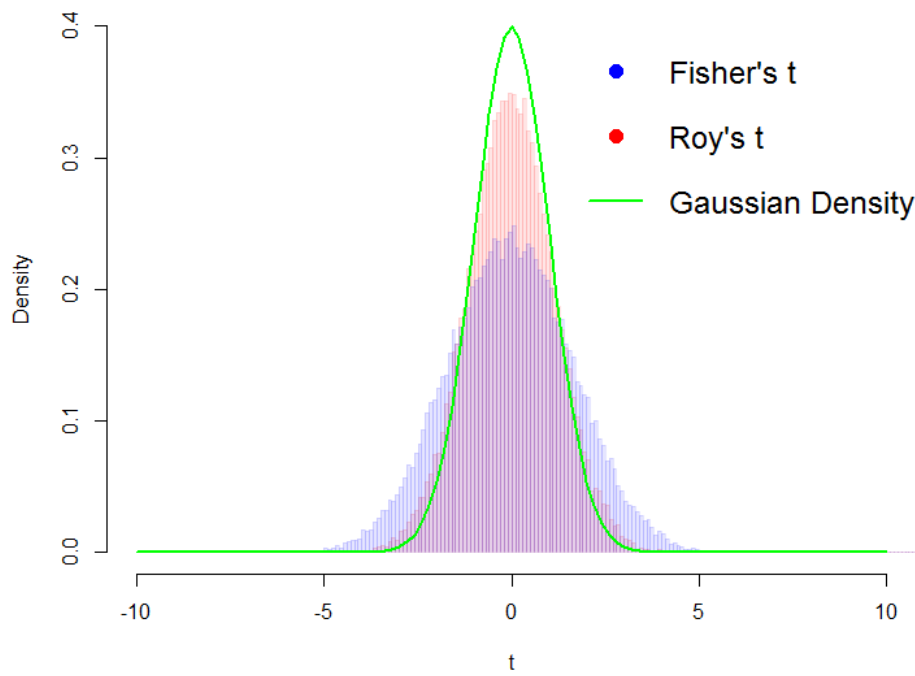
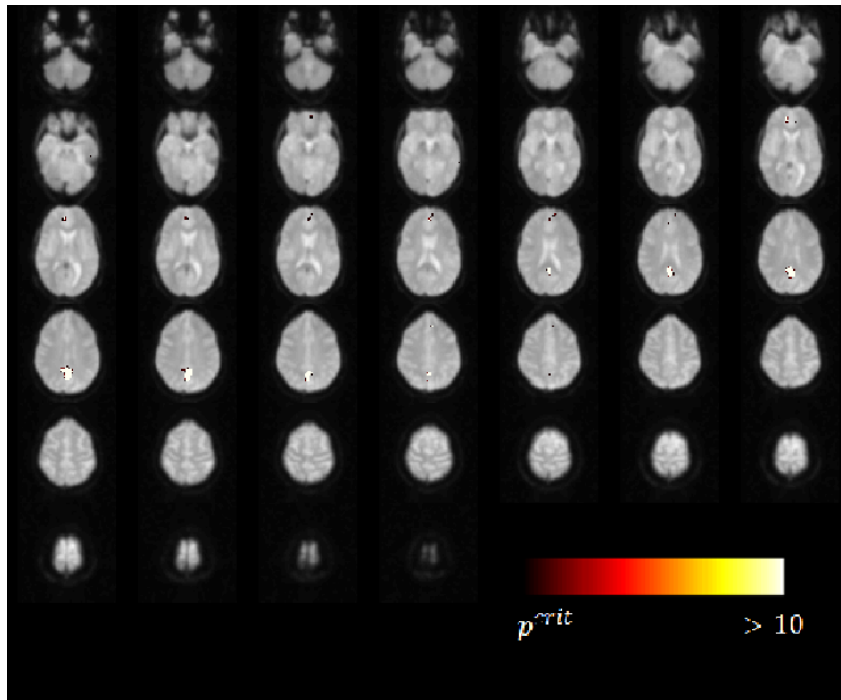
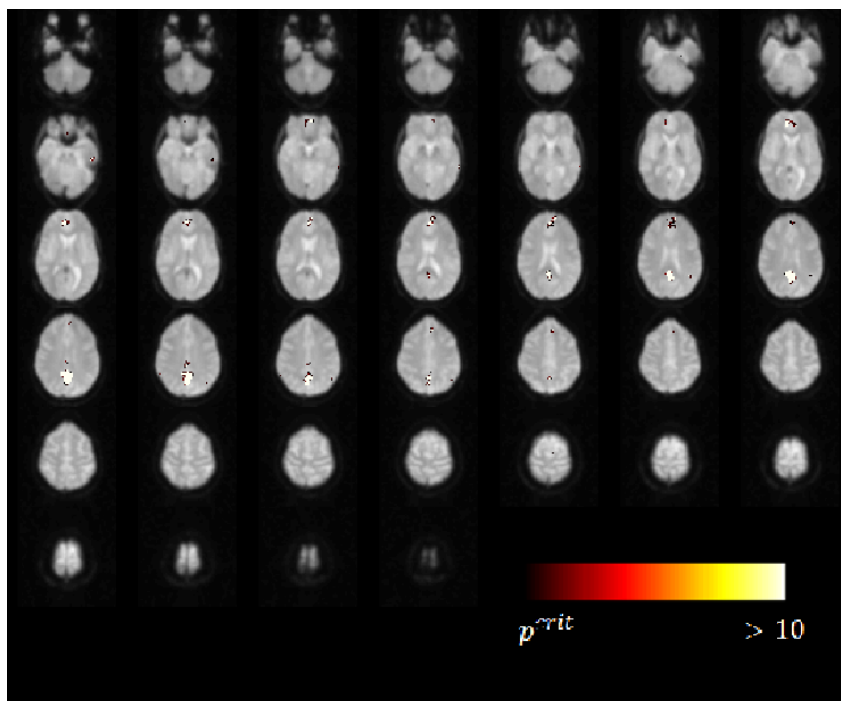


Figure 4.5: Histogram illustrating the distributions t^{f*} and t^{r*} over the interval $[-10,10]$

Westfall & Young's STP Results



(a) Voxels exhibiting non-zero cross-correlation according to t^f



(b) Voxels exhibiting non-zero cross-correlation according to t^r

Figure 4.6: Voxels exhibiting non-zero cross-correlation with the seed region when implementing Westfall & Young's resampling STP. All adjusted p-values plotted on a $-\log_{10}(\cdot)$ scale

Hochberg’s STP and so Benjamini & Yekuteili’s STP was implemented. This controls the $\mathbb{E}[\text{False discovery proportion}] \leq q$, in this case $q = 0.05$, by rejecting $H_{0,v}$ when $\tilde{p}_v \leq q$. Thus $p_{crit} = \max\{p_i \text{ such that } \tilde{p}_i \leq q\}$.

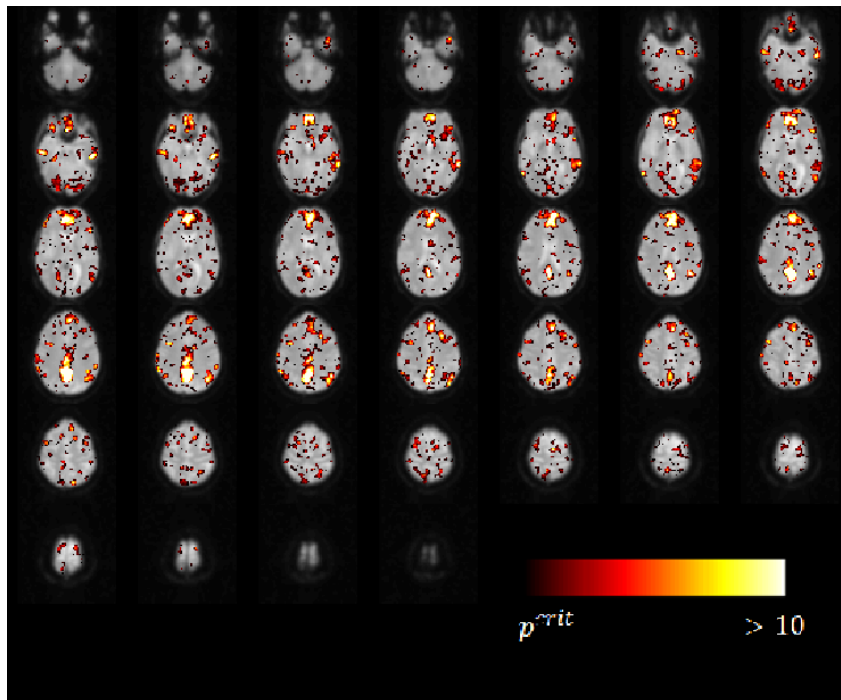
The results of the implementation of Benjamini & Yekuteili’s FDR control STP are summarised in Table 4.4 and the spatial distribution of the p-values for rejected hypotheses presented in Figure 4.7.

	H_0 rejected with $t < 0$	Failed to reject H_0	H_0 rejected with $t > 0$
t^f	3,771	40,285	5,385
t^r	608	46,340	2,487

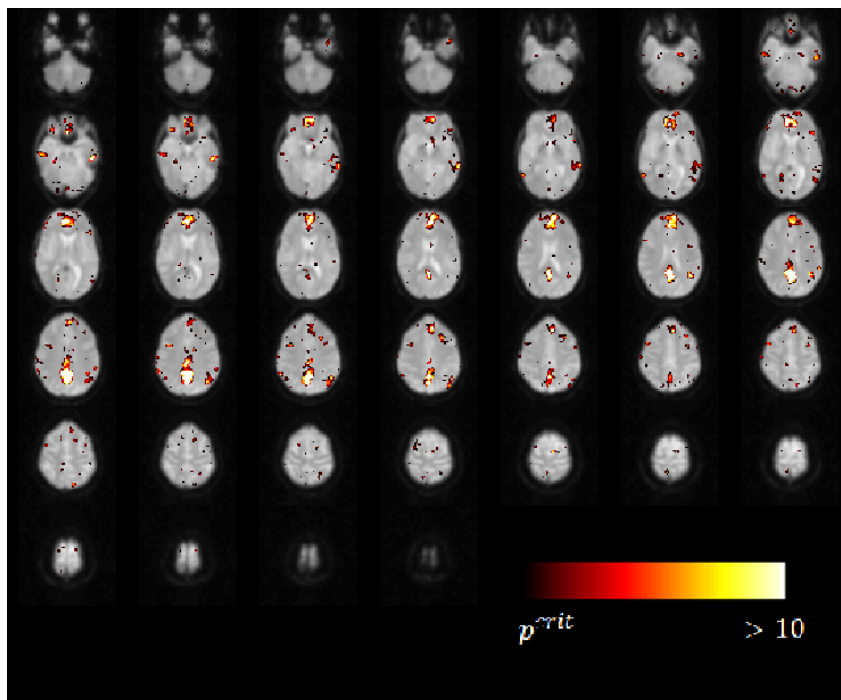
Table 4.4: Summary of results for implementation of Benjamini & Yekuteili’s FDR STP on each family of test statistics.

This STP is far more powerful than FWER controls, though still less powerful than naive methods. In this case the difference between the tests t^f and t^r is striking.

Benjamini & Yekuteili's FDR STP Results



(a) Voxels exhibiting non-zero cross-correlation according to t^f



(b) Voxels exhibiting non-zero cross-correlation according to t^r

Figure 4.7: Voxels exhibiting non-zero cross-correlation with the seed region when implementing Benjamini & Yekuteili's FDR control STP. All adjusted p-values plotted on a $-\log_{10}(\cdot)$ scale

Chapter 5

Discussion and Conclusions

5.1 Functional Connectivity in the “Resting State Brain”

This study of Functional Connectivity in the “Resting State” human brain concludes that there is an Intrinsic Connectivity Network (ICN) in the resting brain. This manifests itself as a high degree of linear dependency, as quantified by the sample cross-correlation, between the Posterior Cingulate Cortex (PCC) and the Medial Prefrontal Cortex (MPF). These results demonstrate that this ICN is observable in an individual brain. This is still the case even after correcting test statistics for temporal correlation and applying very strict controls for multiplicity to these test statistics. This is demonstrated by figures 4.4, 4.6, and 4.7. In each case positive dependencies were more robust, suggesting a greater reliability associated with areas of positive dependency, as per previously established results [24].

5.2 Controls for Multiplicity

This analysis considered two classes of controls for multiplicity. Familywise Error Rate (FWER) control offers a stringent set of criteria for identifying result which deviate from the null hypotheses. False Discovery Rate (FDR) control on the other hand represents a far more liberal alternative. Each Simultaneous Testing Procedure offers it’s own set of benefits and drawbacks.

Considering first each of the STP’s implemented in a general sense. Holm’s STP offers a benchmark Familywise Error Rate (FWER) control STP. It is not computationally demanding, controls the FWER in the strong sense, and requires very few assumptions. It offers a good first look at any multiple testing problem. However, adhering strictly to Holm’s procedure results in a test with very low power. It ignores the dependency structure of the data which can result in an overly conservative test.

The alternative method for FWER control implementes here is offered by Westfall & Young’s resampling STP. This takes into account the dependency structure of the data by estimating the distribution from which the data is generated under the null hypothesis. If

the test statistic then satisfies the subset pivotality condition, this also provides FWER control in the strong sense, often with lower adjusted p-values than Holm’s STP. Thus it can provide an improved testing procedure when compared with Holm’s STP. However this method of FWER control can be extremely computationally demanding, especially if the analyst wishes to assess the true significance level of the testing procedure. Also, it may not be possible for the test statistic to meet subset pivotality, as was the case in this analysis.

FDR control for families of test statistics with a general dependence structure is provided by the Benjamini & Yekuteili STP. Even under this STP, which is far more conservative than Benjamini & Hochberg’s FDR control, FDR control provides a far more powerful, easy to compute alternative to FWER control. However as it only controls the expectation of the false discovery proportion, rather than a probability, any relationships identified in by this STP must be qualified as such. It is often more useful to consider FDR control as a method for identifying an “interesting” sub-collection of the hypotheses in any particular family. This sub-collection may then be ear-marked for further examination.

In the context of this particular FC study, it would appear that FDR control provides the most useful insight into the data. When the number of hypotheses being tested simultaneously is in the thousands, producing some false positives does not necessarily undermine a study. Thus it is difficult to justify a strict adherence to FWER control here, as it may give rise to a case of “throwing the baby out with the bathwater”. This is especially the case when the analysis is simply identifying cases for which the underlying dependency is different from zero.

Holm’s STP is ultra conservative, rejecting only 2.5% of the family of hypotheses from t^r . The performance of Westfall & Young’s resampling STP was particularly disappointing, as it produced a STP with even lower power than Holm’s procedure.

The reason for this is illustrated by Figure 4.5. Both t^{f*} and t^{r*} have fatter tails than a gaussian distribution. Therefore controlling the FWER with these distributions leads to a less powerful test than one assuming test statistics do in fact follow a gaussian distribution under the null hypothesis. Figure 4.5 also explains why t^r is a more powerful family of test statistics only for the resampling STP.

Even when applying the more stringent, general FDR control of Benjamini & Yekuteili, which has been shown in many cases to be unnecessary [3], FDR control allows the identification of far more cases of interest than FWER control. It should be noted however that q and α are not comparable. α represents the probability of any Type I error, while q is the expected false discovery proportion. Decisions on exactly what value is used for each are at the analysts discretion.

In general, when faced with a multiple testing problem an analyst should consider the cost of a false positive result, the acceptable level of computational resources to allocate to a solution, and the dependency structure of the data. This analysis suggests that, at least in cases where N is large, FDR control offers a better approach to controlling for multiplicity than FWER control. Perhaps when N is smaller FWER control is more appropriate, for what size of N this shift may occur however, has not been examined.

One insight provided by Figure 4.5 is that, when implementing a resampling STP, comparing the distribution of one resampled realisation to the theoretical null indicates

how the STP will perform as compared to an inequality based STP such as Holm's STP. For the resampling STP to offer a more powerful alternative, the maximum and minimum values of t^* must be within the tails of the null distribution.

5.3 Roy's Asymptotic Variance Estimate

The simulation study in chapter 3 demonstrated that test statistics based on Roy's theorem for the asymptotic covariance of cross-correlations provided a better insight to the underlying characteristics cross-correlation than those making the assumption of independent and identically distributed realisations of X_t .

Figure 4.5 offers an interesting insight into the behaviour of each test statistic. When resampled under the null hypothesis t^{r*} is much closer to the null distribution, with lighter tails, than t^{f*} . This indicates that tests using t^r will have much lower rate of false positives. While t^{r*} does still have heavier tails than a gaussian distribution, it is worth noting that no efforts have been made in this analysis to optimise the window function used, which may offer some improvement in the estimate $\hat{\sigma}_{ij}$. These observations, along with Figures 4.2 and 3.5, imply that test statistics based implementing a Fisher's z transform underestimate the variance of the autocorrelations and making the tests power less responsive to the truth.

Although the calculation of $\hat{\sigma}_{ij}$ does create additional computational demands, there does not appear to be any reason why this method should not be considered as the standard practise for computing test statistics, at least in the case of an individual level brain study.

Figure 4.2 demonstrates that neither set of test statistics follows the theoretical null distribution. The most likely explanation for this is that the empirical distribution is produced by some voxels which follow the null distribution and others which follow alternative distributions. Another explanation, that there is some other major source of variance, does not seem credible, although this analysis did not include any spatial analysis of variance. Further investigation could include the use of an Expectation-Maximisation algorithm to identify possible alternative distributions.

The most obvious area for further study following from the implementation of Roy's Theorem would involve optimising the window functions used. In this analysis a Bartlett window with $H = 1$ was arbitrarily chosen. How to optimise these windows remains an open question.

Estimating the variance of series by $\hat{\sigma}_{ij}$ could also prove useful in the construction of confidence intervals for point estimates. This may provide an avenue for further study.

5.4 Limitations

This project was subject to the time constraints, as is the nature of any MSc. dissertation. The most time consuming aspects of the analysis were the writing of code to implement Roy's Theorem and the development of the Westfall & Young resampling STP.

While correcting test statistics for temporal correlation proved to be a worthwhile exercise, the time invested in the resampling STP remains a frustration. This is especially the case when Figure 4.5 intuitively demonstrates that the resampling STP was unlikely to offer any improvement in power over Holm's STP, and that the data did not satisfy the condition of subset pivotality in any case. The time involved in these aspects of the analysis meant that other method of FWER control, such as Random Field Theory, were not explored.

When implementing each control for multiplicity the most stringent criteria was applied for identifying the cases for which H_0 was rejected. Further work could be done to identify methods for increasing the power of the tests while still being able to make a definite statement about the control that was applied.

This analysis did not address explicitly any sort of spatial relationship between voxels. Implementation of such a model may offer further insight into sources of variance, further improving the test statistics used.

Finally, the major limitation of this study is that it took place on a single subject for a single scan. In order to reach any meaningful general conclusions the study would have to be expanded to include more participants. Some work on the impact of correcting for temporal correlation in this case may have provided useful insights.

Bibliography

- [1] Mohammad R Arbabshirani, Eswar Damaraju, Ronald Phlypo, Sergey Plis, Elena Allen, Sai Ma, Daniel Mathalon, Adrian Preda, Jatin G Vaidya, Tülay Adali, et al. Impact of autocorrelation on functional connectivity. *Neuroimage*, 102:294–308, 2014.
- [2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.
- [3] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.
- [4] Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2013.
- [5] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.
- [6] JS Damoiseaux, CF Beckmann, EJ Sanz Arigita, F Barkhof, Ph Scheltens, CJ Stam, SM Smith, and SARB Rombouts. Reduced resting-state brain activity in the “default network” in normal aging. *Cerebral cortex*, 18(8):1856–1864, 2008.
- [7] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.
- [8] Mark Fiecas, Hernando Ombao, Dan van Lunen, Richard Baumgartner, Alexandre Coimbra, and Dai Feng. Quantifying temporal correlations: A test–retest evaluation of functional connectivity in resting-state fmri. *Neuroimage*, 65:231–241, 2013.
- [9] Michael D Fox and Marcus E Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8(9):700–711, 2007.
- [10] Michael D Fox, Abraham Z Snyder, Justin L Vincent, Maurizio Corbetta, David C Van Essen, and Marcus E Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9673–9678, 2005.
- [11] Michael D Greicius, Ben Krasnow, Allan L Reiss, and Vinod Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258, 2003.

- [12] Edward J Hannan. The asymptotic distribution of serial covariances. *The Annals of Statistics*, pages 396–399, 1976.
- [13] Yosef Hochberg and Ajit C Tamhane. Multiple comparison procedures. 2009.
- [14] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.
- [15] Mary-Ellen Lynall, Danielle S Bassett, Robert Kerwin, Peter J McKenna, Manfred Kitzbichler, Ulrich Muller, and Ed Bullmore. Functional connectivity and brain networks in schizophrenia. *The Journal of Neuroscience*, 30(28):9477–9487, 2010.
- [16] Guy Melard, Marianne Paesmans, and Roch Roy. Consistent estimation of the asymptotic covariance structure of multivariate serial correlations. *Journal of Time Series Analysis*, 12(4):351–361, 1991.
- [17] Guy Melard and Roch Roy. On confidence intervals and tests for autocorrelations. *Computational Statistics & Data Analysis*, 5(1):31–44, 1987.
- [18] Kevin Murphy, Rasmus M Birn, Daniel A Handwerker, Tyler B Jones, and Peter A Bandettini. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *Neuroimage*, 44(3):893–905, 2009.
- [19] Thomas Nichols and Satoru Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5):419–446, 2003.
- [20] Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.
- [21] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
- [22] Roch Roy. Asymptotic covariance structure of serial correlations in multivariate time series. *Biometrika*, 76(4):824–827, 1989.
- [23] Raymond Salvador, John Suckling, Christian Schwarzbauer, and Ed Bullmore. Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):937–946, 2005.
- [24] Zarrar Shehzad, AM Clare Kelly, Philip T Reiss, Dylan G Gee, Kristin Gotimer, Lucina Q Uddin, Sang Han Lee, Daniel S Margulies, Amy Krain Roy, Bharat B Biswal, et al. The resting brain: unconstrained yet reliable. *Cerebral cortex*, 19(10):2209–2229, 2009.
- [25] signal developers. *signal: Signal processing*, 2014.
- [26] Jin-Hui Wang, Xi-Nian Zuo, Suril Gohel, Michael P Milham, Bharat B Biswal, and Yong He. Graph theoretical analysis of functional brain networks: test-retest evaluation on short-and long-term resting-state functional mri data. *PloS one*, 6(7):e21976, 2011.

- [27] Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.
- [28] Brandon Whitcher, Volker J. Schmid, and Andrew Thornton. Working with the DICOM and NIFTI data standards in R. *Journal of Statistical Software*, 44(6):1–28, 2011.
- [29] Xi-Nian Zuo, Adriana Di Martino, Clare Kelly, Zarrar E Shehzad, Dylan G Gee, Donald F Klein, F Xavier Castellanos, Bharat B Biswal, and Michael P Milham. The oscillating brain: complex and reliable. *Neuroimage*, 49(2):1432–1445, 2010.
- [30] Xi-Nian Zuo, Clare Kelly, Jonathan S Adelstein, Donald F Klein, F Xavier Castellanos, and Michael P Milham. Reliable intrinsic connectivity networks: test–retest evaluation using ica and dual regression approach. *Neuroimage*, 49(3):2163–2177, 2010.

Appendices

Appendix A

R Code Implementing Roy's Theorem

```
#####  
## FUNCTIONS FOR IMPLEMENTATION OF ROY'S THEOREM ##  
#####  
  
#####  
## Functions required to run Roy  
#####  
## Auto and cross correlations  
A <- function(x, l=20){  
  z <- acf(x,lag.max=1,plot=F)  
  z$acf  
}  
## returns an estimate of the autocorrelation for  
## l = 20  
## lags, length 21  
  
C <- function(x,y, l=20){  
  z <- ccf(x,y,lag.max=1,plot = F)  
  z$acf  
}  
## returns an estimate of the crosscorrelation for  
## l = 20  
## lags, length 41  
  
## window functions  
  
## Bartlett - w.b()  
w.b <- function(x){  
  if(abs(x)<=1){  
    1-abs(x)  
  }else{  
    0  
  }  
}  
  
## Adjustment for x  
w <- function(x, N, H){  
  x/(H*sqrt(N))  
}
```

```

#####
## Function to calculate Variance
## according to Roy's Theorem
## using Bartlett window
#####
Roy.Var.b <- function(seed, voxel, H = 1){
N <- length(seed)
# set number of observations

M <- ceiling(sqrt(H*N))
# make loops/correlations as small as possible to
## be efficient
s.a <- A(seed, l=M)
v.a <- A(voxel, l=M)
c.sv <- C(seed, voxel, l=M)
c.vs <- C(voxel, seed, l=M)
# Auto and Cross Correlation vectors

d.s <- 0; d.v <- 0; d.vs <- 0; d.sv <- 0
d.sssv <- 0; d.vsvv <- 0
d.vsss <- 0; d.vvsv <- 0
d.ssvv <- 0; d.vssv <- 0
# set each sum to initial value of 0

for(u in -M:M){
d.s <- d.s + ((w.b(w(u, N, H)))^2)*
              ((s.a[abs(u)+1])^2)
d.v <- d.v + ((w.b(w(u, N, H)))^2)*
              ((v.a[abs(u)+1])^2)
d.sv <- d.sv + (w.b(w(u, N, H))^2)*
              ((c.sv[u+M+1])^2)
d.vs <- d.vs + (w.b(w(u, N, H))^2)*
              ((c.vs[u+M+1])^2)
d.sssv <- d.sssv + ((w.b(w(u, N, H))^2)*
                   ((s.a[abs(u)+1])*(c.sv[u+M+1])))
d.vsvv <- d.vsvv + ((w.b(w(u, N, H))^2)*
                   ((c.vs[u+M+1])*(v.a[abs(u)+1])))
d.vsss <- d.vsss + ((w.b(w(u, N, H))^2)*
                   ((c.vs[u+M+1])*(s.a[abs(u)+1])))
d.vvsv <- d.vvsv + ((w.b(w(u, N, H))^2)*
                   ((v.a[abs(u)+1])*(c.sv[u+M+1])))
d.ssvv <- d.ssvv + ((w.b(w(u, N, H))^2)*
                   ((s.a[abs(u)+1])*(v.a[abs(u)+1])))
d.vssv <- d.vssv + ((w.b(w(u, N, H))^2)*
                   ((c.vs[u+M+1])*((c.sv[u+M+1]))))
}
# Estimate for each infinite sum

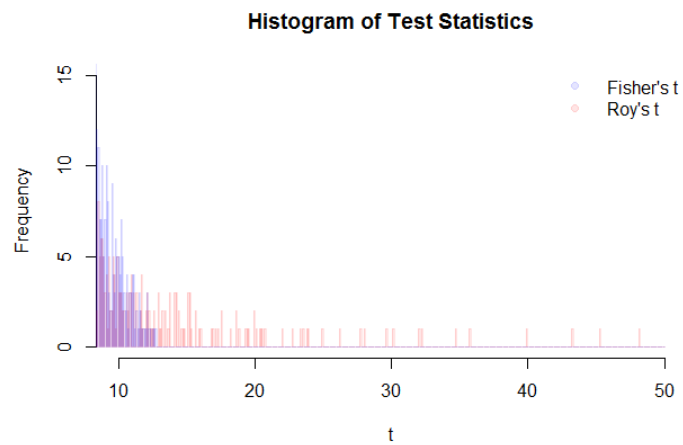
c <- c.sv[M+1]
# cross cor at lag 0

V <- (((1/2)*(c^2)*(d.s+d.sv+d.vs+d.v)) -
c*(d.sssv + d.vsvv) - c*(d.vsss + d.vvsv) +
d.ssvv + d.vssv)
V
}
#####

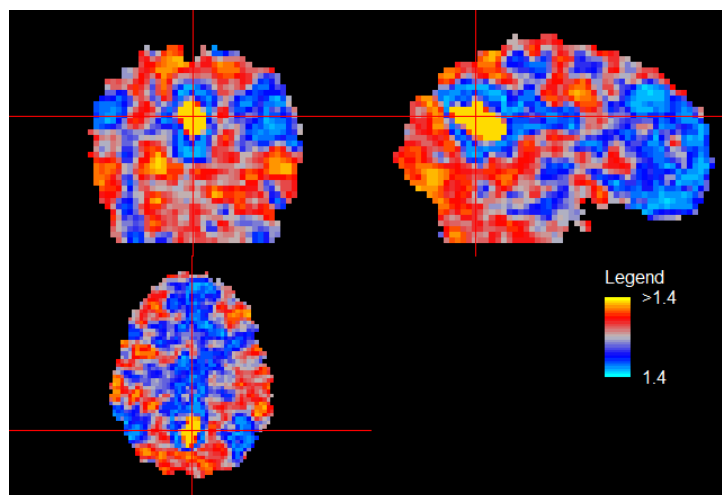
```


Appendix B

Test statistic Graphics



(a) Distribution of t^f & t^r for values greater than 10



(b) Spatial distribution of differences $t^r - t^f$

Figure B.1: Further information on distributions of the families of the statistics t^f & t^r .