2018

New Tools For Intermodal Analysis And Association Testing In Neuroimaging

Simon Vandekar
University of Pennsylvania, simonv@pennmedicine.upenn.edu

New Tools For Intermodal Analysis And Association Testing In Neuroimaging

Abstract

In the field of neuroimage analysis two key goals are to understand the association of a high- dimensional imaging variable with a phenotype, and to understand relationships between several high-dimensional imaging variables. Several recent studies have shown that the standard “mass- univariate” methods to test an association of an image with a phenotype have inflated type 1 error rates due to invalid assumptions. Here, we propose two new methods to perform association testing in neuroimaging and illustrate the method in two stages of the lifespan. The first is a para- metric bootstrap testing procedure that estimates the joint distribution of test statistical parametric map in order to control the voxel-wise family-wise error rate (FWER). We illustrate the method by identifying sex differences in nonlinear developmental trajectories of cerebral blood flow through adolescence using the Philadelphia Neurodevelopmental Cohort. The second testing procedure is a generalization of Rao’s score test based on projecting the score statistic onto a linear sub-space of a high-dimensional parameter space. The approach provides a way to localize signal in the high-dimensional space by projecting the scores to the subspace where the score test was performed. This allows for inference in the high-dimensional space to be performed on the same degrees of freedom as the score test, effectively reducing the number of comparisons. We illus- trate the method by analyzing a subset of the Alzheimer’s Disease Neuroimaging Initiative dataset. Finally, we propose a new tool to study relationships between neuroimaging modalities that we to describe how the spatial association between cortical thickness and sulcal depth changes in adolescent development.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Epidemiology & Biostatistics

First Advisor
Russell T. Shinohara

Second Advisor
Hongzhe Li

Keywords
Association Test, hypothesis testing, Intermodal analysis, score test

Subject Categories
Biostatistics

NEW TOOLS FOR INTERMODAL ANALYSIS AND ASSOCIATION TESTING IN NEUROIMAGING

Simon N. Vandekar

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

Russell T. Shinohara, Assistant Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Hongzhe Li, Professor of Biostatistics

Theodore D. Satterthwaite, Assistant Professor of Psychiatry

Haochang Shou, Assistant Professor of Biostatistics

NEW TOOLS FOR INTERMODAL ANALYSIS AND ASSOCIATION TESTING IN NEUROIMAGING

ACKNOWLEDGEMENT

I would like to thank my collaborators on this work Theodore D. Satterthwaite, Philip T. Reiss, Armin Raznahan, Adon Rosen, Rastko Ciric, Ryan Hopson, David R. Roalf, Kosha Ruparel, Ruben Gur, Raquel Gur, and my advisor Russell T. Shinohara.

I am most grateful for my wife, Lillie; your emotional support helped me to get up each time I fell. I would also like to thank my family who supported my ambition to obtain a PhD.

ABSTRACT

NEW TOOLS FOR INTERMODAL ANALYSIS AND ASSOCIATION TESTING IN NEUROIMAGING

Simon N. Vandekar

Russell T. Shinohara

In the field of neuroimage analysis two key goals are to understand the association of a high-dimensional imaging variable with a phenotype, and to understand relationships between several high-dimensional imaging variables. Several recent studies have shown that the standard "mass-univariate" methods to test an association of an image with a phenotype have inflated type 1 error rates due to invalid assumptions. Here, we propose two new methods to perform association testing in neuroimaging and illustrate the method in two stages of the lifespan. The first is a parametric bootstrap testing procedure that estimates the joint distribution of test statistical parametric map in order to control the voxel-wise family-wise error rate (FWER). We illustrate the method by identifying sex differences in nonlinear developmental trajectories of cerebral blood flow through adolescence using the Philadelphia Neurodevelopmental Cohort. The second testing procedure is a generalization of Rao's score test based on projecting the score statistic onto a linear subspace of a high-dimensional parameter space. The approach provides a way to localize signal in the high-dimensional space by projecting the scores to the subspace where the score test was performed. This allows for inference in the high-dimensional space to be performed on the same degrees of freedom as the score test, effectively reducing the number of comparisons. We illustrate the method by analyzing a subset of the Alzheimer's Disease Neuroimaging Initiative dataset. Finally, we propose a new tool to study relationships between neuroimaging modalities that we to describe how the spatial association between cortical thickness and sulcal depth changes in adolescent development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

To understand complex associations that arise in neuroscience it is necessary to understand brain-phenotype associations as well as associations between neurological measures within the brain. With neuroimaging data these associations can be understood as image-phenotype associations and image-image associations. An example of an image-phenotype association is in Alzhiemer's disease (AD) where it is critical to understand how an imaging biomarker is associated with increased risk of developing AD. Image-image associations arise in questions from early histological studies of the brain. For example (Economo, 1925) first noted that the cortical sheet is thinner in the folds of a sulcus than on gyri. Image-phenotype and image-image associations can provide important information about how the brain develops and how it is affected by disease.

The most common approach to study image-phenotype associations is call mass-univariate hypothesis testing. This is where a hypothesis test is performed at each location in the image. Recently, several studies have demonstrated that commonly used family-wise error rate (FWER) controlling procedures yield incorrect false positive rates (Eklund, Nichols, and Knutsson, 2016; Eklund et al., 2012; Silver, Montana, and Nichols, 2011). Cluster-based spatial inference procedures (Friston et al., 1994b) that rely on Gaussian random field (GRF) theory can have hugely inflated false positive rates, while voxel-wise GRF MTPs (Friston et al., 1994a) tend to have exceedingly small FWERs that are far below the nominal level. The failure of GRF procedures is due to the fact that the spatial assumptions of Gaussian random field approaches are often violated in neuroimaging data sets (Eklund, Nichols, and Knutsson, 2016). The small type 1 error rate of voxel-wise procedures is due to the reliance on classical FWER procedures (such as the Bonferroni procedure) that do not account for the strong dependence between hypothesis tests in voxel-wise and region-wise analyses. This small type 1 error rate leads to an inflated type 2 error rate and loss of power. These recent studies demonstrate a dire need for robust and powerful inference procedures.

Here we propose two new methods to perform image-phenotype associations. The first is an alternative to mass-univariate testing that employs a dimension reducing projection to conserve power when testing hypotheses about images. The approach, discussed in Chapter 2, is a modification of the score test for high-dimensional data called the projected score test (PST). After performing the test for the image-phenotype associations subsequent regional inference can be performed on the projected scores, which have lower variance than the original score vector. The second method, discussed in Chapter 3, weakens the assumptions of the mass-univariate GRF methods and leverages the joint distribution of the test statistics using a parametric bootstrap joint (PBJ) testing procedure.

Interest in image-image associations is increasing, however, there are important methodological considerations in trying to relate the properties of two cortical surfaces or volumetric brain images: cortical surfaces and brain images are highly autocorrelated, non-independent measures that can be re-sampled to an arbitrary number of observations. Thus, standard parametric statistics are not applicable. Prior work used canonical correlation analysis (Avants et al., 2010; Ouyang et al., 2015) to describe the relationship between two measures in volumetric space. In order to assess the significance of the observed correlation, (Avants et al., 2010) relied on permutation tests that relabeled subjects. This approach seeks to understand the relationship between the two measures across subjects. In our recent work (Vandekar et al., 2015), we introduced a novel spatial permutation testing procedure, adapted from the field of microscopy research, that evaluates the relationship between cortical measures in a statistically rigorous framework. Our approach differed from Avants et al. (2010) in that the interest was in assessing the spatial relationship between two variables within two averaged cortical surfaces. While this approach successfully delineated robust spatial effects associated with development in youth, it was limited in that analyses studied only global effects across subjects. For studies of more subtle individual and group differences in spatial relationships such an approach is not ideal, as it does not allow exploration of local regional effects. In Chapter 4, we propose a new image-image associative measure called coupling that is based on locally weighted regression. We use the new measure to study the association between cortical thickness and sulcal depth and to study how this relationship changes in adolescent development.

CHAPTER 2

INTERPRETABLE HIGH-DIMENSIONAL INFERENCE VIA SCORE PROJECTION WITH AN APPLICATION IN NEUROIMAGING

2.1. Introduction

In scientific fields in which high-dimensional data are prominent, it is often of interest to test the association of a single continuous or categorical outcome with a large number of predictors. A common approach in neuroimaging is to reduce the number of hypothesis tests by testing sequentially. For example, an investigator might first perform a test for the association of a phenotype with an imaging variable averaged across the entire brain. If the test rejects the null hypothesis of no association between brain and phenotype, then subsequent tests are conducted on regional averages of the data or on every voxel in the image using multiplicity correction to address the number of tests performed. Often, location-specific results yield few or no significant findings due to reduced signal and the necessary adjustment for the large number of tests, even though the whole brain average data show a significant association.

In this paper, we propose a unified approach to test the association of an imaging or other high-dimensional predictor with an outcome and perform *post hoc* inference to localize signal. The framework is a modification of Rao's score test for models with a high- or infinite-dimensional parameter defined on a compact space such as the brain. Though the approach is designed for hypothesis testing in neuroimaging, it is applicable to a wide range of scientific domains.

Rao's score test assumes a model where $\mathrm{Y}_i \in \mathbb{R}$ are independent and identically distributed observations from density $f(y;\theta)$ and that the parameter $\theta = (\alpha, \beta) \in \Theta \subset \mathbb{R}^{m+p}$ where $\alpha \in \mathbb{R}^m$ is a nuisance parameter and $\beta \in \mathbb{R}^p$ is the parameter of interest. We seek to test the hypothesis $H_0 : \beta = \beta_0$ for some $\beta_0 \in \mathbb{R}^p$. Define the score function $\mathrm{U} = \mathrm{U}(\theta) = n^{-1}\sum_{i=1}^{n} \frac{\partial \log f(\mathrm{Y}_i|\theta)}{\partial \beta}(\theta)$ and let $\theta_0 = (\alpha, \beta_0)$ be the true value of the parameter under H_0. Let $\mathrm{S} = \mathrm{U}(\hat{\alpha}, \beta_0)$ be the score function evaluated at the maximum likelihood estimate of α under the null hypothesis H_0. Under the null and the conditions described in Appendix A.2, the covariance of S can be obtained from the Fisher information evaluated at the null parameter value,

$$\Omega(\theta_0) = \mathbb{E}\left\{[(\partial/\partial\theta)\log f(\mathrm{Y}_1 \mid \theta)]^T[(\partial/\partial\theta)\log f(\mathrm{Y}_1 \mid \theta)]|_{\theta_0}\right\}.$$

The sum of scores (Sum) test originally discussed by Rao (1948) has been used in genetics and neuroimaging (Kim et al., 2014; Madsen and Browning, 2009; Pan, 2009). The Sum test is based on the statistic

$$n\frac{(\mathrm{S}^T\zeta)^2}{\zeta^T\hat{\Omega}\zeta}, \tag{2.1}$$

where $\zeta \in \mathbb{R}^p$ is a given vector of weights. The denominator is an estimate of the variance of $\mathrm{S}^T\zeta$, so that the statistic is asymptotically χ^2_1 (Rao, 1948). This test has low power when a large number of variables are not associated with the outcome (Pan et al., 2014).

In the case of unknown weights, when $p < n$, Rao (1948) proposed maximizing the Sum test statistic with respect to the weights,

$$\max_{\zeta \neq 0} n\frac{(\mathrm{S}^T\zeta)^2}{\zeta^T\hat{\Omega}\zeta} = n\mathrm{S}^T\hat{\Omega}^{-1}\mathrm{S}. \tag{2.2}$$

When $n > p$ this statistic is approximately distributed as χ^2_p under the null, however when $n < p$, it is not defined because $\hat{\Omega}$ is singular.

For finite dimensional parameters, our proposed test can be thought of as a generalization of Rao's test in the case where the estimate of the information matrix is singular. When $p > n$, the test maximizes the statistic (2.1) with respect to the vector ζ over a subspace, $\mathbb{L}$, of $\mathbb{R}^p$. Maximization of the Sum test in the subspace $\mathbb{L}$ is equivalent to projecting the scores for the original model to a lower dimensional space within which the information matrix is invertible. For this reason, we call the test a projected score test (PST). The procedure does not assume sparsity, but attempts to conserve power by reducing the dimension of the data and performing inference in the lower dimensional space.

In many cases, if a score test rejects H_0, then it is of primary interest to perform *post hoc* inference to identify nonzero parameters. In neuroimaging, this amounts to a high-dimensional testing problem where the association is tested at each location in the image. The standard approach is to perform a hypothesis test at each parameter location and use a multiplicity correction procedure. Such methods in neuroimaging that control the family-wise error rate (FWER) have relied on Gaussian random field theory (Friston et al., 1994a), but have recently been shown to have type 1 error rates far from the nominal level in real data due to unmet assumption (Eklund, Nichols, and Knutsson, 2016; Silver, Montana, and Nichols, 2011). Recently, considerable research activity has focused on leveraging the dependence of the tests to control the false discovery rate (FDR) in high-dimensional settings (Efron, 2007). Sun et al. (2015) develop a procedure to control the FDR for spatial data as well as an approach for controlling the expected proportion of false clusters. Fan, Han, and Gu (2012) discuss estimation of the false discovery proportion (FDP) under dependence for normally distributed test statistics based on a factor approximation. In contrast, the PST *post hoc* inference procedure is performed by projecting the scores onto $\mathbb{L}$, and controlling the FWER of the projected scores.

Several recent studies have considered hypothesis tests for functional data, which is conceptually similar to our approach for an infinite-dimensional parameter. Reiss and Ogden (2010) propose inverting simultaneous confidence bands for the parameter of a functional predictor to test which locations of the image are associated with the outcome. Smith and Fahrmeir (2007) use a binary Markov random field model to compute the joint probability that the marginal parameter estimates are equal to zero. Our *post hoc* inference is most similar to Smith and Fahrmeir (2007) as the interpretation of the contribution of the scores retains a marginal interpretation.

We derive the asymptotic null distribution of the PST statistic under some standard regularity conditions. For a normal linear model, we show how the finite sample distribution of our statistic can be calculated exactly for fixed n and p. For data that are measured on a compact space, such as brain images, we discuss sufficient theoretical assumptions for characterizing test behaviors as both n and p approach infinity.

Our approach to asymptotics in p studies the growth of dimension of the grid at which the underlying stochastic process is observed. That is, as $p \to \infty$ we assume that the data are observed at increasing resolution. The rate that p increases is thus not dependent on n. In contrast, high-dimensional tests that do not make this assumption often have restrictions on the rates of growth of n and p. For example, Xu et al. (2016) bound the rate at which p increases by a function of n, and Cai, Liu, and Xia (2014) require that the maximum expected value of the false null statistics is larger than a given function of n and p.

To demonstrate how the test can be used in neuroimaging, we investigate the association of cortical thickness with mild cognitive impairment (MCI) in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, a data set with p =18,715 and $n = 628$. The outer surface of the brain (cortex) represents a highly folded sheet in 3-dimensional space. The thickness of the cortex is known to be affected in individuals with psychopathology and neurological illness. MCI is a subtle pre-Alzheimer's disease decline in cognitive functioning. There is significant clinical interest in finding biological markers of MCI in order to identify those at risk for developing Alzheimer's disease, as prevention strategies and therapies for early disease are increasingly common. In this data set, we seek to localize regions of the brain where cortical thinning provides additional information with regard to the diagnosis of MCI beyond what can be ascertained by neurocognitive scales alone.

For the remainder of the paper, we denote matrices by uppercase italic letters (X), vectors by lowercase (x), and random vectors by uppercase Roman letters (X). Hilbert spaces are denoted with black-board letters $(\mathbb{X})$ and Greek letters denote model parameters. For the singular value decomposition (SVD) of any matrix we will assume that the smallest dimensions of the matrices obtained are equal to the rank of the matrix X. $\xrightarrow{L}$ denotes convergence in law and $\xrightarrow{P}$ denotes convergence in probability.

2.2. The Projected Score Test

In Section 2.2.1, we define the finite parameter PST statistic and give its asymptotic distribution for fixed p. In Section 2.2.2, we describe conditions sufficient for studying asymptotics in p. We discuss the PST for normal linear models in Section 2.2.3.

2.2.1. PST for finite-dimensional parameters

We assume the observed data are finite-dimensional representations that are generated from an underlying stochastic process. Here, we informally define the finite-dimensional likelihood and refer the interested reader to Appendix A.1 for further details on deriving the finite-dimensional likelihood from the infinite-dimensional likelihood.

Let $\mathbb{V}$ be a nonempty compact subset of $\mathbb{R}^3$ and $\mathcal{L}^2(\mathbb{V})$ be the space of square integrable functions from $\mathbb{V}$ to $\mathbb{R}$. $\mathbb{V}$ represents the space on which data can be observed; in neuroimaging this space is the volume of the brain. The underlying images G_i take values in $\mathcal{L}^2(\mathbb{V})$; however, the observed data are p-dimensional discretizations of this image. Throughout this section, we model an outcome $\mathrm{Y}_i \in \mathbb{R}$ as a function of observed image data $G_{ip} \in \mathbb{R}^p$ and a set of covariates $X_i \in \mathbb{R}^k$. Thus, the observed data can be described as independent and identically distributed observations $(\mathrm{Y}_i, X_i, G_{ip})$. We denote the collection of data by $\mathrm{Y} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_n)$ and similarly define $X = (X_1, \ldots, X_n)$ and $G_p = (G_{1p}, \ldots, G_{np})$. We define a parameter $\theta_p = (\alpha, \beta_p) \in \Theta \subset \mathbb{R}^{m+p}$, where $\alpha \in \mathbb{R}^m$ is a nuisance parameter and $\beta_p \in R^p$ is the parameter of interest. Together these parameters describe the conditional distribution of Y given the imaging data and covariates. We allow $m \geq k$ in order to flexibly model the relationship of the covariates and outcome, for example with unpenalized splines.

Denote the finite-dimensional likelihood by $\ell(\theta_p; \mathrm{Y}) = n^{-1} \sum_{i=1}^n \partial \log f(\mathrm{Y}_i \mid \theta_p, X, G_p)/\partial\beta_p$. Define the score function $\mathrm{U}_{np} = \frac{\partial \ell}{\partial \beta_p}\{\theta_p; \mathrm{Y}\}$ and let

$$\mathrm{S}_{np} = \mathrm{U}_{np}(\hat{\alpha}, \beta_{p0}) \in \mathbb{R}^p \tag{2.3}$$

denote the score function evaluated at the maximum likelihood estimate (MLE) under the null hypothesis

$$H_0 : \beta_p = \beta_{p0}. \tag{2.4}$$

Let the Fisher information for the full model be

$$\Omega_F(\theta_{p0}) = \mathbb{E}_{\theta_{p0}} \left(\{ \frac{\partial}{\partial \theta_p} \log f(\mathrm{Y}_1 \mid \theta_p, X, G_p)\} \{ \frac{\partial}{\partial \theta_p} \log f(\mathrm{Y}_1 \mid \theta_p, X, G_p)\}^T \Big|_{\theta_{p0}} \right) = \begin{bmatrix} \Omega_\alpha & \Omega_{\alpha\beta} \\ \Omega_{\beta\alpha} & \Omega_\beta \end{bmatrix}, \tag{2.5}$$

where $\theta_{p0} = (\alpha, \beta_{p0})$. Then the asymptotic variance of $\sqrt{n}\mathrm{S}_{np}$ under H_0 is (see e.g. Vaart (2000)

$$\Omega(\theta_{p0}) = \Omega_\beta - \Omega_{\beta\alpha}\Omega_\alpha^{-1}\Omega_{\alpha\beta}. \tag{2.6}$$

With the finite parameter scores defined, we can define the PST.

Definition 2.2.1. *Let $P_{\mathbb{L}}$ be the orthogonal projection matrix onto a linear space $\mathbb{L} \subset \mathbb{R}^p$ with $r =$ dim$(\mathbb{L}) < n - m$. Let S_{np} be as defined in* (2.3) *and $\hat{\Omega}$ be the plug-in estimator of the covariance*

(2.6) *obtained from*

$$\hat{\Omega}_F = n^{-1} \sum_{i=1}^{n} \left(\frac{\partial \log f(\mathrm{Y}_i \mid \theta_p, X, G_p)}{\partial \theta_p} \right) \left(\frac{\partial \log f(\mathrm{Y}_i \mid \theta_p, X, G_p)}{\partial \theta_p} \right)^T \Big|_{\hat{\theta}_{p0}}, \tag{2.7}$$

where $\hat{\theta}_{p0} = (\hat{\alpha}, \beta_{p0})$ *denotes the maximum likelihood estimate of the parameter vector under the null hypothesis* (2.4)*. Then the PST statistic with respect to* $\mathbb{L}$ *is defined as*

$$\mathrm{R}^{\mathbb{L}} = \max_{\zeta \in \mathbb{L} \setminus \{0\}} n \frac{(\zeta^T \mathrm{S}_{np})^2}{\zeta^T \hat{\Omega}(\theta_{p0}) \zeta} = \max_{\gamma \in \mathbb{R}^p \setminus \{0\}} n \frac{(\gamma^T P_{\mathbb{L}} \mathrm{S}_{np})^2}{\gamma^T P_{\mathbb{L}} \hat{\Omega}(\theta_{p0}) P_{\mathbb{L}} \gamma}. \tag{2.8}$$

The following theorem gives the asymptotic distribution (with respect to n) of the PST statistic provided the same regularity conditions required for the convergence of the scores to a multivariate normal random variable.

Theorem 2.2.2. *Assume all objects are as described in Definition 2.2.1. Let* $P_{\mathbb{L}} = QQ^T$ *where the columns of the* $r \times p$ *matrix* Q *are any orthonormal basis for* $\mathbb{L}$*. Define*

$$V_p = V(\theta_{p0}) = Q^T \Omega(\theta_{p0}) Q,$$

and assume the estimate $\hat{\mathrm{V}}_{np} = Q^T \hat{\Omega}(\theta_{p0}) Q$ *is invertible, and that the conditions given in Appendix A.2 are satisfied. Then, under the null* (2.4)*, the rotated scores* $\mathrm{S}_{np}^Q \equiv Q^T \mathrm{S}_{np}$ *satisfy*

$$n^{1/2} \mathrm{S}_{np}^Q \xrightarrow{P} \mathrm{S}_p^Q \sim N_r(0, V), \tag{2.9}$$

the PST statistic (2.8) *is*

$$\mathrm{R}^{\mathbb{L}} = n (\mathrm{S}_{np}^Q)^T \hat{\mathrm{V}}_{np}^{-1} \mathrm{S}_{np}^Q, \tag{2.10}$$

and $\mathrm{R}^{\mathbb{L}} \xrightarrow{L} \chi_r^2$ *as* $n \to \infty$*.*

Theorem 2.2.2 requires that $\hat{\mathrm{V}}_{np}$ is nonsingular; however, in practice it is possible to ensure that Q is in the column space of $\hat{\Omega}(\theta_{p0})$, so that $\hat{\mathrm{V}}_{np}^{-1}$ exists. The proof of Theorem 2.2.2 is given in Appendix A.3. We also demonstrate there that the result of Theorem 2.2.2 does not depend on the choice of Q. We show how $\mathbb{L}$ can be chosen for GLMs in Sections 2.3.1 and 2.3.2, and for imaging data in the analysis of the ADNI dataset in Section 2.5.

2.2.2. The PST as $p \to \infty$

We will show that as $p \to \infty$ the PST statistic converges to an integral over a stochastic process. The rate that p approaches infinity does not depend on the sample size. Here, we assume the data take values on the space $\mathbb{Y} = \mathbb{R} \times \mathbb{R}^k \times \mathcal{L}^2(\mathbb{V})$, where $\mathbb{V}$ is a nonempty compact subset of $\mathbb{R}^3$ and $\mathcal{L}^2(\mathbb{V})$ is the space of square integrable functions, with respect to the Lebesgue measure, from $\mathbb{V}$ to $\mathbb{R}$. Let $\mathrm{O}_i = (\mathrm{Y}_i, X_i, G_i)$, for $i = 1, \ldots, n$, be independent and identically distributed with $\mathrm{Y}_i \in \mathbb{R}$, $X_i \in \mathbb{R}^k$, and $G_i \in \mathcal{L}^2(\mathbb{V})$. Realizations of O_i are the outcome variable, a vector of k covariates and a function G_i. The infinite-dimensional score function, U_n, is defined in Appendix A.1 as the Fréchet derivative of the log likelihood with respect to the parameter $\beta \in \mathcal{L}^2(\mathbb{V})$, $\mathrm{U}_n = \mathrm{U}_n(v) = \frac{\partial \ell}{\partial \beta}\{(\alpha, \beta(v)); \mathrm{O}(v)\}$. The score is defined for fixed β_0 as the stochastic process

$$\mathrm{S}_n = \mathrm{U}_n\{\cdot; (\hat{\alpha}, \beta_0)\} \in \mathcal{L}^2(\mathbb{V}). \tag{2.11}$$

Throughout Section 2.2.2, we assume that the infinite-dimensional scores converge in probability to a mean zero Gaussian process, S, i.e.

$$n^{1/2}\mathrm{S}_n \to_P \mathrm{S}, \tag{2.12}$$

and that the dimension of the finite parameter α is fixed. Theorem A.4.1 in Appendix A.4 gives conditions under which this convergence holds (Vaart, 2000). The following definition of the PST statistic extends formula (2.10) to infinite-dimensional parameters.

Definition 2.2.3. *Let $(q_1(v), \ldots, q_r(v))$ be an orthonormal basis for the linear subspace $\mathbb{L} \subset \mathcal{L}^2(\mathbb{V})$ where q_j are continuous almost everywhere, and $r = dim(\mathbb{L}) < n - m$. Also, assume that S_n and $\frac{\partial}{\partial \beta} \log f\{\mathrm{Y}_i(v) \mid \hat{\alpha}, \beta_0(v)\}$, have continuous sample paths with respect to v, where $\frac{\partial}{\partial \beta}$ denotes the Fréchet derivative. Define the column vector $\mathrm{S}_n^Q \in \mathbb{R}^r$, with jth element*

$$(\mathrm{S}_n^Q)_j = \int_{\mathbb{V}} q_j(v)\mathrm{S}_n(v)dv, \tag{2.13}$$

and let $\hat{\mathrm{V}}_n$ be the $r \times r$ matrix with (j, k)th element

$$\begin{aligned}\hat{\mathrm{V}}_n^{j,k} =& n^{-1} \sum_{i=1}^{n} \left(\int_{\mathbb{V}} q_j(v) \left[\frac{\partial}{\partial \beta} \log f\{\mathrm{Y}_i \mid \hat{\alpha}, \beta_0(v)\} \right] dv \right) \\ & \times \left(\int_{\mathbb{V}} q_k(v) \left[\frac{\partial}{\partial \beta} \log f\{\mathrm{Y}_i \mid \hat{\alpha}, \beta_0(v)\} \right] dv \right),\end{aligned}$$

which is readily shown to be the covariance matrix of S_n^Q. Assume that $\hat{\mathrm{V}}_n$ is invertible. Then the PST statistic with respect to $\mathbb{L}$ is defined as

$$\mathrm{R}^{\mathbb{L}} = n(\mathrm{S}_n^Q)^T \hat{\mathrm{V}}_n^{-1} \mathrm{S}_n^Q. \tag{2.14}$$

While we have given a definition of the PST statistic in infinite dimensions, in practice this statistic is not estimable because it depends on functions which are only observed on a finite grid. The following theorem states that as the resolution of the grid is increased then the finite parameter PST statistic (2.10) converges to the infinite-parameter PST statistic (2.14). Moreover, as the sample size increases the statistic converges to a function of the Gaussian process S in (2.12).

Theorem 2.2.4. *Let S_{np} be as defined in* (2.3)*. Let Q_p be the $p \times r$ matrix with jth column $q_{jp} = (q_j(v_{1p})\nu(\mathbb{V}_{1p}), \ldots, q_j(v_{pp})\nu(\mathbb{V}_{pp}))^T$, where v_{jp} and $\mathbb{V}_{jp}$ are defined in Appendix A.1, $\mathcal{V}_p = \{\mathbb{V}_{1p}, \ldots, \mathbb{V}_{pp}\}$, and ν denotes Lebesgue measure. Denote $\mathrm{S}_{np}^Q = Q_p^T \mathrm{S}_{np}$. Define $\mathrm{S}^Q \in \mathbb{R}^r$ as the vector with jth element*

$$(\mathrm{S}^Q)_j = \int_{\mathbb{V}} q_j(v) \mathrm{S}(v) dv.$$

Assume the conditions for Theorems 2.2.2 and A.4.1, and that S_n, $\frac{\partial}{\partial \beta} \log f\{\mathrm{Y}_i(v) \mid \hat{\alpha}, \beta_0(v)\}$, and S have continuous sample paths with respect to v. Let $V = \mathbb{E}\hat{\mathrm{V}}_n$. For $p_1 > p_2$, let $\mathcal{V}_{p_1}$ be a refinement of $\mathcal{V}_{p_2}$, and assume that

$$\lim_{p \to \infty} \sup_k \nu(\mathbb{V}_{kp}) = 0. \tag{2.15}$$

Then as $n, p \to \infty$,

$$n(\mathrm{S}_{np}^Q)^T \hat{\mathrm{V}}_{np}^{-1} \mathrm{S}_{np}^Q \xrightarrow{P} (\mathrm{S}^Q)^T V^{-1} \mathrm{S}^Q. \tag{2.16}$$

The proof is given in Appendix A.4.

2.2.3. The PST in Normal Linear Models

The finite-sample distribution of the PST statistic for a normal linear model can be found exactly. Define $X = [X_1, \ldots, X_n]^T$ to be an $n \times m$ full rank matrix of covariates for each observation, $G = [G_1, \ldots, G_n]^T$ an $n \times p$ full-rank matrix of predictor variables of interest with $p > n$, and $\mathrm{Y} = [\mathrm{Y}_1, \ldots, \mathrm{Y}_n]^T$ $n \times 1$ normal random vector with independent elements conditional on X and G. The score test with normal error is based on the model

$$\mathrm{Y}_i = \alpha^T X_i + \beta^T G_i + \mathrm{E}_i, \tag{2.17}$$

where $\mathrm{E}_i \sim N(0, \sigma^2)$ are independent.

Theorem 2.2.5. *Under model* (2.17) *and the null $H_0 : \beta = 0$,*

$$\mathrm{R}^{\mathbb{L}} =_L \frac{r(n-m)}{r + (n-m-r)\mathrm{F}_{(n-m-r),r}}, \tag{2.18}$$

where $\mathrm{F}_{(n-m-r),r}$ is F-distributed with $(n-m-r)$ and r degrees of freedom.

The proof can be found in Appendix A.5. The finite-sample distribution of $\mathrm{R}^{\mathbb{L}}$ depends only on the sample size and the dimension of the basis, but not on the particular choice of $\mathbb{L}$.

2.3. Specifying the linear subspace $\mathbb{L}$

2.3.1. Specifying $\mathbb{L}$ in generalized linear models

Here, we discuss choices for the selection of $\mathbb{L}$ in the context of GLMs with the canonical link function. We restrict attention to finite dimensional parameters. Let X and G be as defined in Section 2.2.3. Assume the outcome $\mathrm{Y} = [\mathrm{Y}_1, \ldots, \mathrm{Y}_n]^T$ is from an exponential family where the expectation can be written

$$h(\mathbb{E}\mathrm{Y}_i) = \alpha^T X_i + \beta^T G_i,$$

where h is the canonical link function. For the GLM with canonical link, the scores are (McCullagh and Nelder, 1989)

$$\mathrm{S}_{np} = n^{-1} G^T(\mathrm{Y} - \hat{\mathrm{Y}}),$$

where

$$\hat{\mathrm{Y}} = [\hat{\mathrm{Y}}_1, \dots \hat{\mathrm{Y}}_n]^T$$

and $\hat{\mathrm{Y}}_i = h^{-1}(x_i^T \hat{\alpha})$ is the ith fitted value under the null. Let Γ be the $n \times n$ diagonal matrix with ith diagonal element $\Gamma_{ii} = (\mathrm{Y}_i - \hat{\mathrm{Y}}_i)^2$. Then the estimate of the covariance (2.6) obtained using (2.7) is

$$\hat{\Omega} = n^{-1}\{G^T \Gamma G - G^T \Gamma X (X^T \Gamma X)^{-1} X^T \Gamma G\} \tag{2.19}$$

The score statistic is obtained from the scores and the estimated information as in expression (2.2).

In this setup, the basis for $\mathbb{L}$ can be constructed from the principal component analysis (PCA) of G. We write the PCA of G in terms of the SVD $G = T_* D Q^T$, where the principal scores are $T = T_* D = GQ$.

With this basis, the PST is equivalent to performing Rao's score test in a principal components regression model. To see this, first note that principal component regression is defined by

$$h(\mathbb{E}\mathrm{Y}) = X\alpha + T\beta_T.$$

The scores for β_T are

$$\mathrm{S}_{npT} = n^{-1} Q^T G^T (\mathrm{Y} - \hat{\mathrm{Y}}) = Q^T \mathrm{S}_{np},$$

which are the same as the rotated scores in (2.9). The information estimate is also equivalent. Thus the score test statistic, $n\mathrm{S}_{npT}^T \hat{\Omega}_T^{-1} \mathrm{S}_{npT}$, in principal component regression is equivalent to the PST statistic (2.10).

Another useful basis for $\mathbb{L}$ can be constructed from vectors that are indicators of variables that are expected to have a similar relationship with the outcome. The anatomical basis we use in Section 2.5 is an example. To define the basis vectors q_j, $j = 1, \dots, r$, we let $\mathcal{Q}_j \subset \{1, \dots, p\}$ such that $\mathcal{Q}_j \cap \mathcal{Q}_{j'} = \varnothing$ for $j \neq j'$, and then set the kth element of the jth basis vector to be $q_{kj} = \mathbb{1}(k \in \mathcal{Q}_j)$. These define orthogonal basis vectors since the sets $\mathcal{Q}_j$ are disjoint. This basis is equivalent to averaging r subsets of the p predictor variables and performing a hypothesis test of the regression onto the r averaged variables.

The choice of the basis is a critical decision as it affects the power and interpretation of the *post* *hoc* inference. To clarify, under the alternative the scores have nonzero mean

$$\mathbb{E}\mathrm{S}_{np} = \mu \in \mathbb{R}^p. \tag{2.20}$$

If the projection is orthogonal to μ then the test will have power equal to the type 1 error rate. The PCA basis assumes that μ has a spatial pattern similar to the covariance structure of the predictor variables. The anatomical basis assumes that all locations within a region have similar parameter values. We discuss the effect of the basis on the interpretation of the *post hoc* inference in Section 2.4.

2.3.2. Choosing a dimension for the PCA basis

In order to choose a dimension for the PCA basis, we propose an automatic procedure that sequentially tests bases of increasing dimension while controlling the type 1 error rate. To do this we first condition on the parameter estimate $\hat{\alpha}$ for the reduced model and perform the SVD $(\Gamma - \Gamma X(X^T\Gamma X)^{-1}X^T\Gamma)^{1/2}G = TDQ^T$. We use subsets of the columns of Q as the basis $\mathbb{L}$. For $j \neq k$

$$\text{Cov}(q_j^T\mathrm{S}_{np}, q_k^T\mathrm{S}_{np} \mid \hat{\alpha}) = q_j^T G^T(\Gamma - \Gamma X(X^T\Gamma X)^{-1}X^T\Gamma)Gq_k = q_j^T QD^2Q^T q_k = 0.$$

Thus, each rotated score $n^{1/2}q_j^T\mathrm{S}_{np}$ is asymptotically independent, conditional on $\hat{\alpha}$, and can be tested by a separate chi-square test at level α^*. If this is done sequentially for $r = 1, \ldots, n-m$, then, due to asymptotic independence, the probability of a type 1 error under the global null is

$$\begin{aligned}\sum_{r=1}^{n-m} \mathbb{P}\left((n^{1/2}q_j^T\mathrm{S}_{np})^2 > \chi_1^2(\alpha^*) \text{ for all } j \leq r\right) &\approx \sum_{r=1}^{n-m}(\alpha^*)^r \\ &\leq \sum_{r=1}^{\infty}(\alpha^*)^r = \frac{\alpha^*}{1-\alpha^*},\end{aligned}$$

where $\chi_1^2(\alpha^*)$ denotes the $1-\alpha^*$ quantile of the chi-squared distribution with one degree of freedom. The approximate equality is due to the asymptotic approximation. In order to control the type 1 error at level α, we choose $\alpha^* = \alpha/(1+\alpha)$, then we sequentially test $r = 1, \ldots, (n-m)$ until we fail to reject a test at level α^*. Note that the power depends critically on the first test in the sequence; subsequent tests serve only to increase the dimension of the basis. If the first component is orthogonal to μ in

(2.20), the probability of reaching other components that are not orthogonal to μ is less than α^*.

A potentially more robust procedure is to test chunks of PCs by choosinging $r_1 = 0, r_2, r_3, \ldots, r_k = n - m$ and for the jth test perform a chi-square test of PCs $(r_j + 1), \ldots, r_{j+1}$ on $r_{j+1} - r_j$ degrees of freedom. So long as the tests are independent, which is true under the global null, the rejection threshold α^* will control the type 1 error rate at level α. This automatic PCA (aPCA) procedure is implemented below by testing the first 5 components together and sequentially testing components $6, \ldots, n - m$ one-at-a-time. We demonstrate the procedure in the ADNI data analysis below and type 1 error rates are assessed in Section 2.6.

2.4. *Post hoc* Inference for Localizing Signal

After performing the test of association using the PST, it is of primary interest to investigate the contribution of the scores to the statistic in order to identify which locations in the image are associated with the outcome and the direction of the effect. This can be done by projecting the scores onto $\mathbb{L}$ and performing inference that controls the FWER for the projected scores. Because the projected scores are distributed in a linear subspace of $\mathbb{R}^p$, inference is much less conservative than performing inference on the original score vector.

Our aim is to construct a rejection region for each element of the projected score vector $(P_{\mathbb{L}}\mathrm{S}_{np})_j$, for $j = 1, \ldots p$. Under the null the projected scores are asymptotically normal,

$$P_{\mathbb{L}}\mathrm{S}_{np} \sim N(0, P_{\mathbb{L}}\Omega P_{\mathbb{L}}).$$

The diagonal elements of $P_{\mathbb{L}}\Omega P_{\mathbb{L}}$ are not equal, so defining a single rejection threshold for all elements favors rejection for elements with larger variances. To resolve this issue we scale by the inverse of the standard deviation of the projected scores. Let Δ be the diagonal matrix with jth diagonal element $\Delta_{jj} = 1/\sqrt{(P_{\mathbb{L}}\Omega P_{\mathbb{L}})_{jj}}$. Then the rejection threshold that controls the FWER for the standardized projected scores is defined by c that satisfies

$$1 - \mathbb{P}(|(\Delta P_{\mathbb{L}}\mathrm{S}_{np})_j| > c \text{ for some } j) = \mathbb{P}(\max_j |(\Delta P_{\mathbb{L}}\mathrm{S}_{np})_j| < c) = 1 - \alpha. \tag{2.21}$$

Thus, the distribution of the infinity norm of $\Delta P_{\mathbb{L}}\mathrm{S}_{np}$ can be used to compute a rejection threshold for the standardized projected scores that controls the FWER for the test of hypotheses about the projected scores

$$H_{0j}^{\mathbb{L}} : \mathbb{E}(\Delta P_{\mathbb{L}} \mathrm{S}_{np})_j = 0. \tag{2.22}$$

We reject the null hypothesis (2.22) at location j if the observed projected score $|(\Delta P_{\mathbb{L}} s_{np})_j| > c$. This threshold corresponds to a single-step "maxT" joint multiple testing procedure (Westfall and Young, 1993) and provides strong control of the FWER (Romano and Wolf, 2005).

By (2.9) we have

$$\Delta P_{\mathbb{L}} \mathrm{S}_{np} \xrightarrow{L} \Delta Q V^{1/2} \mathrm{Z},$$

where $\mathrm{Z} \sim N_r(0, I)$. Thus we can approximate the region in (2.21) by finding c so that

$$\int_{|\Delta Q V^{1/2} z|_\infty \leq c} \phi_r(z) dz = 1 - \alpha,$$

where ϕ_r denotes the PDF of Z. In practice we approximate this integral by plugging in estimates for Δ and $V^{1/2}$.

The integral is difficult to calculate due to the large dimensions of Q, but can be approximated quickly and easily using Monte Carlo simulations. B simulations are used to estimate the CDF of the infinity norm, $\hat{F}_B(\cdot)$, which we use to obtain adjusted p-values for each observed standardized projected score, $(\Delta P_{\mathbb{L}} s_{np})_j$, by evaluating

$$p_j = 1 - \hat{F}_B\left((\Delta P_{\mathbb{L}} s_{np})_j\right), \tag{2.23}$$

or a rejection threshold can be obtained by using

$$c = \hat{F}_B^{-1}(1 - \alpha). \tag{2.24}$$

The p-value (2.23) for a given element of the standardized projected score vector is the probability of observing a jth projected score as large as $(\Delta P_{\mathbb{L}} s_{np})_j$ under the global null $H_0 : \beta = \beta_0$. The standard deviation of the Monte Carlo estimate (2.23) decreases at a $\sqrt{B}$ rate and depends only on the volume of the space being integrated, so the procedure will perform well for computing adjusted p-values with a small error (Press et al., 2007). For example, with 10,000 simulations the standard deviation is on the order of $B^{-1/2} = 0.01$.

Rejection of the null hypothesis $H_0 : \beta = \beta_0$ is not strictly necessary to proceed with the *post hoc* inference procedure; the *post hoc* procedure can be used separately from the PST. In addition, it is important to note that the *post hoc* inference is restricted to the projected scores. When the alternative hypothesis is true, the rejection regions for the projected scores do not necessarily control the type 1 error for the unprojected scores. This is demonstrated in the simulations in Section 2.6.

As mentioned above, the basis affects the interpretation of the inference on the projected scores. For the PCA basis the interpretation is as follows: over repeated experiments, if the data are projected onto $\mathbb{L}$, then the probability of falsely rejecting one or more scores j with $(\Delta P_{\mathbb{L}}\mu)_j = 0$ is at most α, where μ is as defined in (2.20). The anatomical basis assumes all locations in each basis vector have similar parameter values.

2.5. ADNI Data Analysis

We obtained data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The ADNI is a longitudinal observational study designed to investigate the early biomarkers of Alzheimer's disease; detailed MRI methods are given by Jack et al. (2008). Mild cognitive impairment (MCI) represents a subtle pre-Alzheimer's Disease decline in cognitive performance. The goal of our analysis is to identify whether a subset of the neuroimaging data from the ADNI can provide more information regarding diagnosis of MCI than the standardized memory tests obtained as part of the study. Moreover, we are interested in localizing areas of the cortex that differ on average between healthy controls (HC) and individuals with MCI. Three-dimensional T1-weighted structural images for 229 healthy controls and 399 subjects with MCI were obtained as part of the ADNI. This sample consists of subjects who had images and a composite memory score available at baseline. Cortical thickness was estimated using Freesurfer (Dale, Fischl, and Sereno, 1999; Fischl and Dale, 2000). Subjects' thickness data were registered to a standard template for analysis and smoothed at 10mm FWHM to reduce noise related to preprocessing and registration. The template contains 18,715 vertex locations where cortical thickness is measured for each subject. Our goal is to identify whether the 18,715 cortical thickness measurements provide any additional information regarding the diagnosis of the individuals.

We perform the analysis in two ways: First, we proceed with standard analysis methods currently available for neuroimaging data in open access software (Fischl, 2012). Second, we use the PST statistic and the high-dimensional inference procedure described above. For all analyses we include age, sex, and the composite memory score as covariates (Crane et al., 2012).

2.5.1. Standard Neuroimaging Analysis Procedure: Average and Vertex-wise Testing

Because neuroimaging studies typically collect many types of images with many covariates and possible outcomes, it is common to obtain a summary measure of a high-dimensional variable, and then proceed with further analysis if the summary measure appears to be associated with an endpoint of interest. In this analysis we first take the average of all the cortical thickness measurements across the cortical surface for each subject and perform a regression with diagnosis as the outcome using logistic regression. Specifically let C_i denote the average cortical thickness measurement for subject i, and X_i denote a vector with an intercept term, age, an indicator for sex, and the composite memory score for subject i. Then we fit the model

$$\text{logit}\{\mathbb{P}(Y_i = 1 \mid C_i, X_i)\} = X_i^T \alpha + C_i \beta_C.$$

If there is a significant relationship with the average cortical thickness measurements, i.e. if we reject $H_0 : \beta_C = 0$, then we will proceed by performing mass-univariate vertex-wise analyses by running a separate model at each point on the cortical surface.

The analysis using the average cortical thickness variable suggests a highly significant association of cortical thickness with diagnosis, indicating that subjects with thinner cortices are more likely to have MCI (Table 2.1). Based on these results we choose to investigate the relationship at each vertex to localize where in the cortex the association occurs.

For the vertex-wise analyses, we use the software package Freesurfer to perform Benjamini-Hochberg (BH) correction separately across each hemisphere (Figure 2.1 A). The spatial extent of the FDR-corrected results is more limited than what we might expect given the very strong association between diagnosis and average cortical thickness. We conducted uncorrected exploratory analyses to further identify regions related to the whole-brain results (Figure 2.1 B). The most significant re-

	Estimate (SE)	p-value
Age	-0.08 (0.02)	< 0.001
Sex (Male)	-0.37 (0.26)	0.15
Memory score	-3.22 (0.27)	< 0.001
Average cortical thickness	-4.23 (1.02)	< 0.001

Table 2.1: Results for the logistic regression of diagnosis onto covariates and whole-brain average cortical thickness. Results for average cortical thickness indicate a highly significant association between cortical thickness and diagnosis. SE denotes standard error.

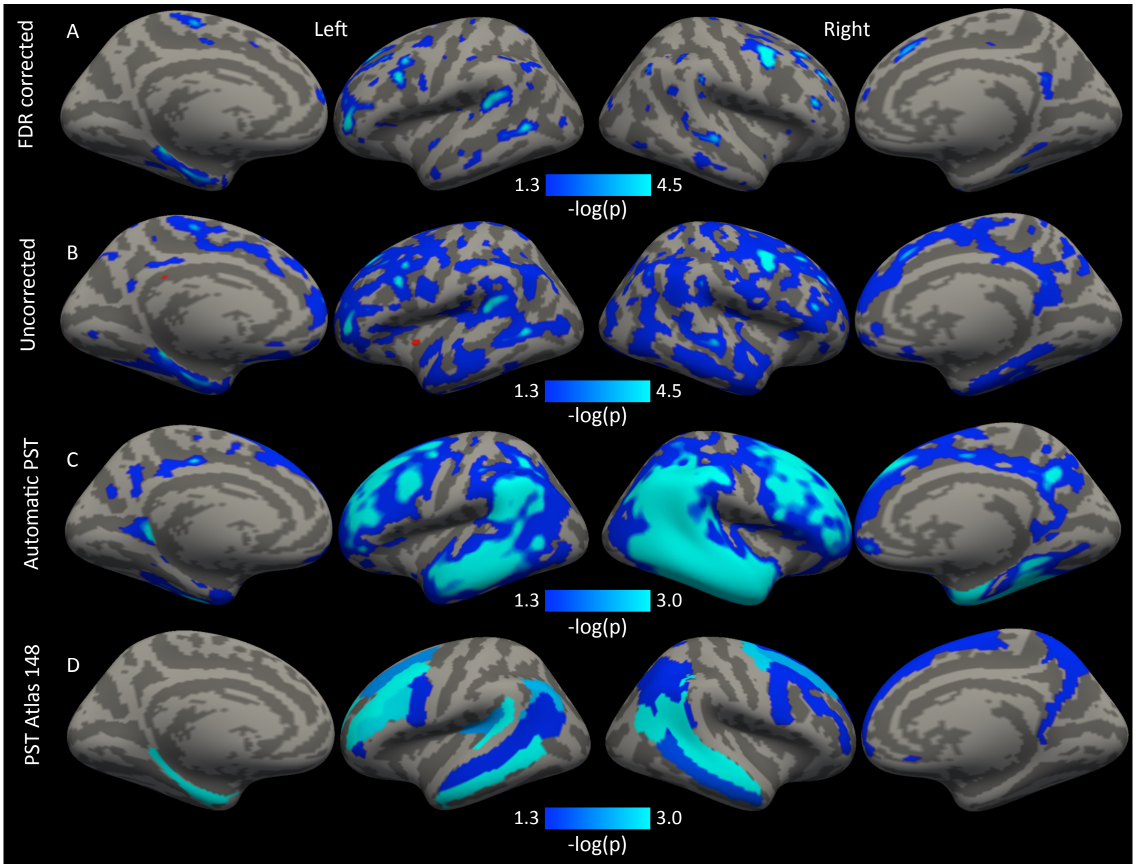

Figure 2.1: A comparison of inference procedures of the association between the imaging data and diagnosis. (A) Benjamini-Hochberg corrected vertexwise results, (B) Uncorrected vertexwise results, and (C & D) results based on PST high-dimensional inference that control the FWER of the projected scores. (C) The dimension ($r = 7$) of the PCA basis for $\mathbb{L}$ was selected using the automatic procedure. (D) The 148 dimensional basis constructed from the Destrieux anatomical atlas. Blue values show significant ($\alpha = 0.05$) negative association with diagnosis indicating that thinner cortex in these regions is associated with an MCI diagnosis.

sults occur in left and right frontal lobes. These analyses suggest that distributed thinning in large portions of the frontal and temporal lobes is associated with increased risk of MCI; however, these results are not found using a method that guarantees control of the FWER or FDR.

2.5.2. PST and High-Dimensional Inference Procedures

To use the PST procedure we perform the following steps:

1. Select a subspace $\mathbb{L}$.

2. Perform the PST for the association between the image and diagnosis.

3. If the test in step 2 rejects, then perform *post hoc* inference as in Section 2.4.

We select a basis for $\mathbb{L}$ in the two ways described in Section 2.3.1. For this analysis we use the aPCA procedure described in Section 2.3.2 to choose the best PCA basis by testing the first 5 components together and sequentially testing components $6, \ldots, n-m$. We also present results for the PCA basis fixed at several other dimensions ($r = 10, 20, 50$) to demonstrate how the basis dimension affects the results of the analysis. In addition we consider a basis constructed from the $r = 148$ regions (74 per hemisphere) of the anatomical atlas of Destrieux et al. (2010). If we were unwilling to condition on the covariance structure of the scores or the anatomical atlas, a basis could be constructed that approximates a predetermined covariance structure (e.g. a spatial AR(1)), or a covariance structure estimated from an independent sample, which can be used to construct the PCA basis. In addition to the PST we perform the sequence kernel association test (SKAT) (Wu et al., 2011), the sum of powered scores (SPU) test using the infinity norm, which corresponds to testing the max across the scores (Pan et al., 2014), and the adaptive sum of powered scores test (aSPU), which has competitive power to many other score tests (Pan et al., 2014). The SKAT is known to be powerful if there is a distributed signal, and the SPU infinity norm will be powerful for a sparse signal. The aSPU test combines multiple tests based on the norms $\|\mathrm{S}\|_\gamma^\gamma$ for γ varying over a finite subset of $\mathbb{N}$ by choosing one with the smallest p-value. Permutation testing is used to assess the significance of these statistics, however, recently, Xu et al. (2016) derived the asymptotic distributions of these tests under mild restrictions on the rate that p grows with respect to n.

The aPCA basis selected $r = 7$ by testing for $r = 5$ and then sequentially testing the next two PCs. With this and all other bases we reject the null hypothesis using the PST (Table 2.2), indicating that there is an association between the image and diagnosis conditioning on the effects of age, sex, and composite memory score. The SKAT, SPU, and aSPU tests also reject the null.

	Test Statistic	p-value	Rejection Threshold
Automatic (r=7)	38	$<$0.001	3.2
r=10	41	$<$0.001	3.4
r=20	50	$<$0.001	3.7
r=50	91	$<$0.001	4
Anatomical basis	179	0.04	3.4
SKAT	5×10^6	$<$0.001	
SPU Inf	47	$<$0.001	
aSPU		0.001	

Table 2.2: The χ^2 PST statistic and associated p-values for various basis dimensions; Automatic, 10, 20, and 50. "Anatomical" is a basis constructed from an anatomical atlas of dimension 148. The last column denotes the 5% familywise error rejection thresholds for the projected scores, i.e. the probability any projected score is above those values under the null is 5%. The thresholds are obtained using 10,000 simulations.

Given the results of the PST we are then interested in investigating how the scores contribute to the significant test statistic. To investigate the contributions of the scores to the PST statistic we perform *post hoc* inference on the projected scores. We use 10,000 simulations to obtain rejection regions for each of the basis dimensions. The simulations ran for all bases in less than 2 minutes.

Results suggest that thinner cortex in bilateral temporal and frontal lobes and right precuneus is associated with an increased risk of MCI (Figure 2.1 C & D). Results are given as $-\log_{10}(p)$ where the p-value p is obtained using the simulated distribution (2.23). These locations are known to be thinner in AD versus HC as well as in AD versus MCI (Singh et al., 2006) and the results here demonstrate that there are significant differences between MCI and HC in the same region. The results indicate that the degree of frontal and temporal lobe thinning is correlated with diagnostic severity, and suggest that measurements of cortical thickness may provide useful information over and above neurospsychological scales in identifying people at risk for AD. Differences in these regions between MCI and HC were previously shown by Wang et al. (2009); however the authors did not control for multiple comparisons or adjust for covariates.

To reiterate, the blue areas in Figure 2.1 C & D are based on low-rank inference and control the FWER of the projected scores. The procedure finds more significant locations over standard correction methods seen in Figure 2.1 A & B by performing inference in a lower dimensional space. The p-values obtained in Figures 2.1 and 2.2 use (2.23) and indicate the probability of observing a projected score statistic as or more extreme than the observed value under the global null $H_0 : \beta = 0$. Though interpretation is restricted to the projected scores, the results align with previous reports

(Singh et al., 2006; Wang et al., 2009).

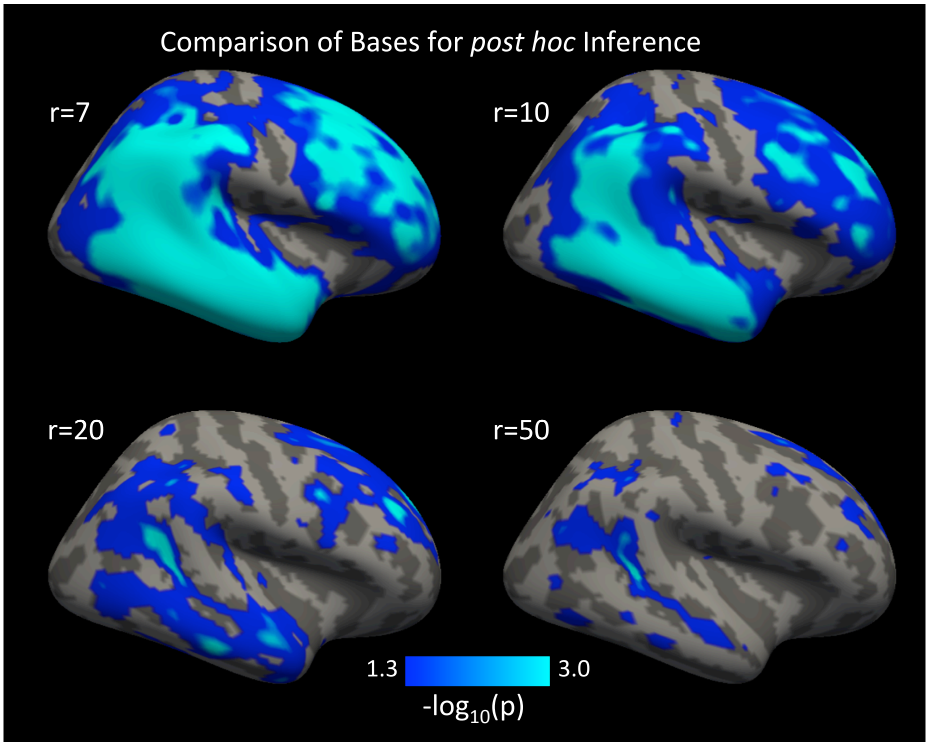

Figure 2.2: PST inference for PCA bases of various rank; Automatic (7), 10, 20, 50. Increasing the dimensionality of the basis increases spatial specificity, but comes with the cost of more conservative inference (see e.g. Table 2.2).

To demonstrate the impact of the choice of r, we performed *post hoc* inference on the scores for 4 different PCA bases (Figure 2.2). It is clear from Figure 2.2 that increasing the dimension of the basis increases the spatial specificity of the results. However, the larger bases also come with the cost of reduced power due to the larger degrees of freedom of the basis. This is also illustrated in Table 2.2, where the larger bases have a higher rejection threshold.

2.6. Neuroimaging-based Simulation Study

In this section, we perform a simulation study using data generated for the right hemisphere of the cortical thickness data from the ADNI dataset measured at p =9,361 locations, called vertices. We simulate a binary outcome of interest, as in the ADNI analyses presented above. We select two anatomical regions (superior temporal sulcus and superior frontal sulcus) of 669 vertices total to have a negative association with the outcome and one region (anterior part of the cingulate gyrus and sulcus) of 191 vertices to have a positive association. The first two of these regions were selected because of their association in the ADNI data set. The third region was selected to compare the performance of the tests when there are different locations with positive and negative associations with the outcome. To create a mean and covariance structure similar to real data within the regions of association, we create the mean vectors and covariance matrices for the simulations from the full sample of subjects used in the ADNI Freesurfer analysis above, yielding two full rank covariance matrices, Σ_- and Σ_+ and mean vectors μ_+ and μ_-.

For each simulation, we select a random subset of size n without replacement from the subset of control subjects used in the ADNI neuroimaging analysis. Data within the negatively and positively associated regions are generated as independent multivariate normal distributions for each subject, with covariance structures $G_{i,-} \sim N(\mu_-, \Sigma_-)$ and $G_{i,+} \sim N(\mu_+, \Sigma_+)$, respectively. We centered the imaging data prior to analysis.

In each simulation the outcome is generated under a logistic model

$$\text{logit}(\mathbb{E}Y_i) = \alpha_0 - \beta \mathbf{1}^T G_{i,-} + 2\beta \mathbf{1}^T G_{i,+}, \tag{2.25}$$

where α_0 is set to the log ratio of MCI to controls in the neuroimaging analyses section. $\mathbf{1}$, denotes a vector of ones, and β is an unknown parameter that we vary from 0 to 0.005. We multiply the values in the positive region by 2 to increase signal because it is a spatially smaller cluster than the two negative regions. In addition to simulations where the coefficients are constant across each region, in the Supplement we perform simulations generating the parameters from a uniform distribution.

We construct the subspace $\mathbb{L}$ in three ways. The first is to use the automatic procedure (Section 2.3.2) in each sample conditioning on the estimate $\hat{\alpha}_0$. The second basis type is constructed in each sample from the first $r = 10, 20, 50$ principal components from a PCA of $G(I - H)$, where H is the projection onto the intercept. The third basis is constructed from regions in the anatomical atlas of Destrieux et al. (2010), by randomly grouping the 74 regions into r groups and using normed indicator vectors for each group as the basis.

If we denote the set of indices with a nonzero association with the outcome by J, then the expectation of score j, μ_j, is nonzero only for $j \in J$, where μ is as defined in (2.20). Thus, for indices with $j \notin J$ we report type 1 error and for indices with $j \in J$ we report power.

Similarly, the mean of the standardized projected scores, $\Delta P\mu$, determines type 1 error and power for the projected scores $\Delta P\mathrm{S}_{np}$. The FWER and FDR of the projected scores are reported for the basis constructed from the anatomical atlas and the PCA bases. In general, no element of the standardized projected mean is exactly zero, so type 1 error is assessed by thresholding the standardized projected parameter vector at the 0.2 quantile and reporting the rejection rate for vertices with projected parameter values below that threshold.

We perform 1000 simulations for sample sizes of $n = 100, 200$ and compare the PST for the automatic procedure and fixed bases with dimensions of $r = 10, 20, 50$. In addition, we compare the PST to the sequence kernel association test (SKAT) (Wu et al., 2011), the sum of powered scores (SPU) test using the infinity norm, and the adaptive sum of powered scores test (aSPU) (Pan et al., 2014). We assess pointwise power and type 1 error of the PST inference procedure with uncorrected, Bonferroni-corrected, and BH-corrected results. We also compare FWER and FDR between methods. For the point-wise results we assess the power and type 1 error for the unprojected scores using inference designed for the projected scores.

Several types of bases for the PST demonstrate superior power to the other tests (Figure 2.3) due to their ability to remove the influence of unassociated scores from the test by maximizing over the basis. If the basis vectors are not informative about the structure of the signal, then the PST would not perform well. The PCA and aPCA PSTs leverage the spatial covariance of the data by using a basis constructed from the matrix G, which is an argument of the covariance of the scores as in equation (2.19). The tests with rank $r = 50$ do not perform well with $n = 100$ due to the error in the asymptotic approximation; the error is evident in the conservative type 1 error of these tests, which is well below the expected 0.05 threshold when $\beta = 0$. The aPCA has better power than the other PCA bases because a low rank basis suffices to capture the signal in the data, while the higher dimensional bases have low power due to the inclusion of basis vectors that do not capture signal. aSPU does not make any assumption about the structure of the signal but is adaptive to the sparsity of the signal, thus it performs better than the SKAT and PSTs with bases of higher dimension. In this simulation, the aSPU test does not perform as well as several of the PSTs because it does not leverage the spatial information in the covariance of the scores, which is informative in this case. However, if the covariance among predictors were uninformative about the signal, it is likely that the aSPU test would be superior.

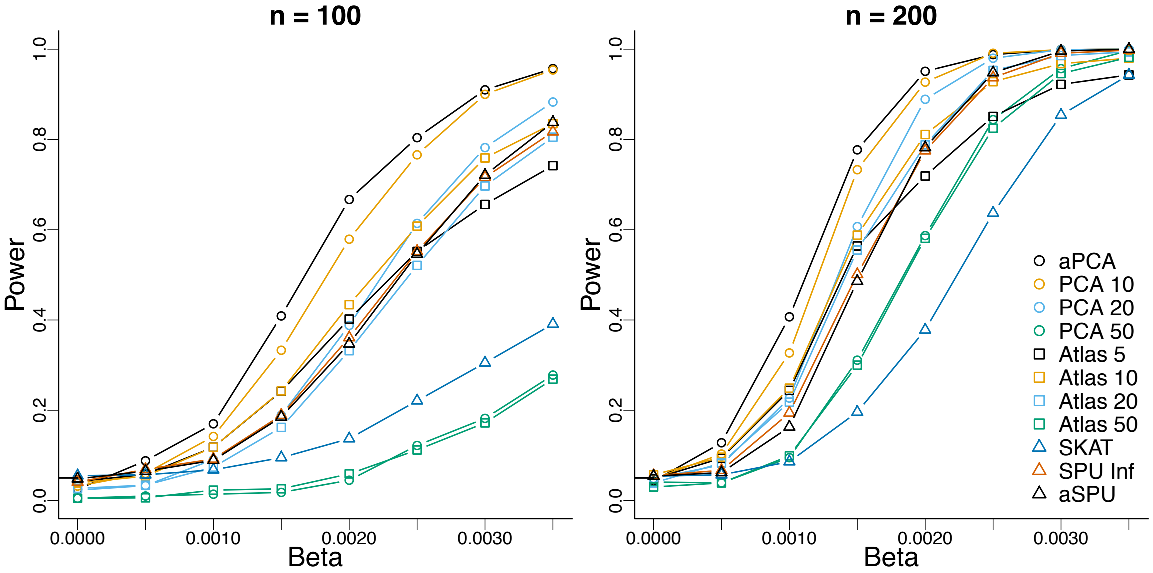

Figure 2.3: Power results for the PST with various bases compared to aSPU and the SKAT. "aPCA" is the automatic basis and "PCA r" indicates the basis formed from the first r components of the PCA of the design matrix G. "Atlas r" indicates the bases formed from the anatomical atlas with r regions covering the cortex.

	FWER	FDR
aPCA	0.06	< 0.01
PCA 10	0.04	< 0.01
PCA 20	0.03	< 0.01
PCA 50	0.02	< 0.01
Anatomical 5	0.03	0.02
Anatomical 10	0.04	0.02
Anatomical 20	0.04	0.02
Anatomical 50	0.06	0.02

Table 2.3: Error rates for the projected scores for the automatic PCA bases and anatomical bases for $n = 200$ and $\beta = 0.002$.

As expected, the *post hoc* inference procedure controls the FWER of the projected scores for all basis dimensions (Table 2.3). In general, the *post hoc* inference procedure does not control the FWER or FDR of the unprojected scores (Table 2.4) as the inference is intended for the projected scores. However, for larger PCA bases our procedure does control the FDR (bold rows in Table 2.4). This is likely because the projection captures most of the variation in μ, so that the projection $\Delta P\mu$ is close to μ. Future investigation of whether inference for the projected scores will control any error rate for unprojected score vector is warranted.

The vertexwise error rate describes how effective a procedure is at controlling the error rate for the

	HR	FDR	FWER
aPCA	0.69	0.13	0.70
PCA 10	**0.55**	**0.05**	**0.45**
PCA 20	**0.34**	**0.02**	**0.23**
PCA 50	**0.12**	**0.01**	**0.10**
Anatomical 5	0.16	0.79	0.27
Anatomical 10	0.34	0.64	0.57
Anatomical 20	0.53	0.45	0.84
Anatomical 50	0.63	0.17	0.70
Uncorrected	0.61	0.44	1.00
Holm	$\mathbf{< 0.01}$	$\mathbf{< 0.01}$	$\mathbf{< 0.01}$
BH	**0.10**	**0.05**	**0.22**

Table 2.4: Hit (HR), false discovery (FDR), and family-wise error (FWER) rates for the unprojected scores. Higher dimensions of the PCA basis control the FDR while maintaining a higher hit rate for the unprojected scores. Note, however, that in practice it is not possible to know the dimension of the basis required to control the FDR of the unprojected scores. Bold rows control the FDR at $q = 0.05$. BH=Benjamini-Hochberg.

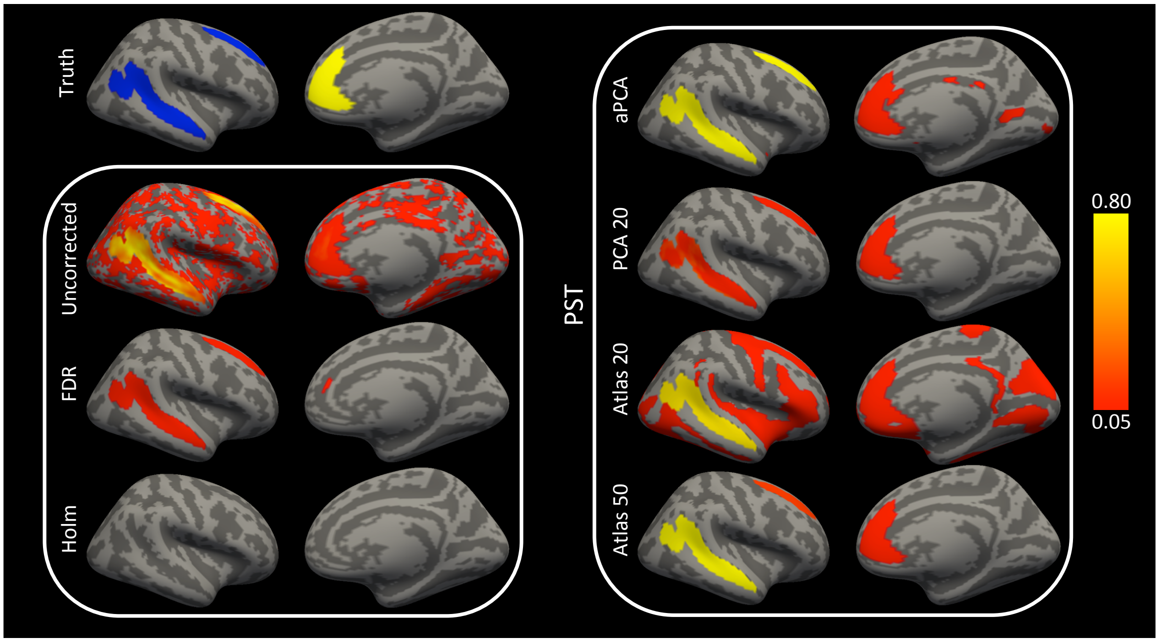

Figure 2.4: Vertexwise power or type 1 error is measured as the proportion of simulated samples where each of the testing procedures rejects the null at a given location. Results are shown for $\beta = 0.002$. "Truth" indicates locations where signal was simulated according to the model (2.25).

unprojected scores at each location. The vertexwise error rate of the PST inference procedure for the unprojected scores using the PCA 10 basis is low while maintaining better vertexwise power than BH (Figure 2.4; PCA 10). This is because in any given sample there may be a high false positive rate, but the errors across samples do not appear in the same locations. The BH and Bonferroni corrections both work well at controlling the vertexwise type 1 error rate but have lower power than the PCA-based PST procedure (Figure 2.4). The bases constructed from the anatomical atlas tend to have large regions of vertexwise type 1 error for the unprojected scores. At the largest basis dimension the atlas allows for enough specificity to reduce the vertexwise error. All methods have lower power to detect the positive cluster, than the two negative clusters. This may be due to the characteristics of the covariance structure in the positive cluster which overlaps gyral and sulcal regions.

2.7. Discussion

We have proposed the PST, a modification of the score test for high dimensional data that works by projecting the scores to a lower dimensional linear subspace. The procedure offers a novel *post hoc* inference on the projected scores by performing inference in the subspace where the test statistic was estimated. Because the posthoc inference is based on the same model and degrees of freedom as the PST statistic, the interpretation of high-dimensional results agree closely with the results from the PST.

The ability to choose a subspace $\mathbb{L}$ makes the procedure very flexible. For example, in medical imaging the basis for the space can be chosen based on anatomical or functional labels, or from data acquired in another imaging modality. Particular hypotheses can be targeted by selecting a basis that includes indicators of certain regions or weights particular locations to target specific spatial patterns. If orthogonal indicator vectors are used as the basis, then the approach can be seen as testing averages of subregions of the data as in Section 2.5. In this case, the PST procedure can be seen as a maxT multiple testing procedure of the regional averages that accounts for the correlation structure of the tests.

There are several limitations of the proposed procedure. First, the success of the procedure depends critically on the projection chosen. If a projection is chosen that is orthogonal to the mean vector, the PST will fail to capture any signal in the data. This is a limitation of any dimension-reducing procedure. Further research could investigate whether maximization of the score test with regularization can yield a test statistic whose distribution is tractable. Regularization may remove the subjectivity of selecting a basis and make the procedure more robust. Second, while the dimension reduction procedure preserves power and the results align closely with those from previous research, the inference does not guarantee control of the FWER or FDR of the original score vector. Future research will investigate how inference of the original score vector can be made by thresholding the projected score vector. This is similar in concept to the dependence-adjusted procedure discussed by Fan, Han, and Gu (2012) for controlling the FDP and may offer increased power by leveraging the covariance of the test statistics. These limitations notwithstanding, our procedure generalizes Rao's score test to the high- and infinite- dimensional settings and introduces a new inference approach based on projecting the test statistics to a lower-dimensional space where inference can be made on fewer degrees of freedom.

CHAPTER 3

Faster family-wise error control for neuroimaging with a parametric bootstrap

3.1. Introduction

Magnetic resonance imaging (MRI) is a widely used tool for studying the neurological correlates of human cognition, psychiatric disorders, and neurological diseases. This is due to the flexibility of MRI to noninvasively study various functional and physiological properties of the human brain. Often, many hypothesis tests are performed at every voxel or at anatomically defined brain regions in order to identify locations that are associated with a cognitive or diagnostic variable. Multiple testing procedures (MTPs) are crucial for controlling the number of false positive findings within a statistical parametric map or across a set of brain regions being investigated. Typically the family-wise error rate (FWER) is controlled at a level $0 < \alpha < 1$, meaning that the probability one or more null hypotheses is falsely rejected is less than or equal to α.

Recently, several studies have demonstrated that commonly used FWER controlling procedures yield incorrect false positive rates (Eklund, Nichols, and Knutsson, 2016; Eklund et al., 2012; Silver, Montana, and Nichols, 2011). Cluster-based spatial inference procedures (Friston et al., 1994b) that rely on Gaussian random field (GRF) theory can have hugely inflated false positive rates, while voxel-wise GRF MTPs (Friston et al., 1994a) tend to have exceedingly small FWERs that are far below the nominal level. The failure of GRF procedures is due to the fact that the spatial assumptions of Gaussian random field approaches are often violated in neuroimaging data sets (Eklund, Nichols, and Knutsson, 2016). The small type 1 error rate of voxel-wise procedures is due to the reliance on classical FWER procedures (such as the Bonferroni procedure) that do not account for the strong dependence between hypothesis tests in voxel-wise and region-wise analyses. This small type 1 error rate leads to an inflated type 2 error rate and loss of power. These recent studies demonstrate a dire need for robust and powerful inference procedures.

To our knowledge, the only methods used in neuroimaging that reliably control the FWER are permutation-based joint testing procedures (Dudoit and Laan, 2008; Eklund, Nichols, and Knutsson, 2016; Winkler et al., 2014). These methods maintain the nominal FWER because they appropriately reproduce the joint distribution of the imaging data, thereby overcoming the limitations of methods typically used in imaging. Unfortunately, permutation testing is computationally intensive, especially in modern imaging data sets that have large sample sizes and hundreds of thousands of voxels or hundreds of brain regions. Moreover, currently available neuroimaging software only performs single-step testing procedures, although step-down procedures are uniformly more powerful. The extensive computation time means it can take days to perform statistical tests, or can even lead investigators to reduce the number of permutations to an extent that adjusted *p*-values have large error.

As a solution, we design a parametric-bootstrap joint (PBJ) testing procedure for hypothesis testing in region- and voxel-wise analyses. Region-wise analyses are performed by averaging all voxel values within anatomically defined regions and fitting a model at each region. Voxel-wise analyses fit a model at each of hundreds of thousands of voxels in a brain image. As the parametric bootstrap does not require resampling and refitting the model for every iteration, it is faster than permutation testing procedures. In addition, our procedure allows the generated null distribution to be applied to multiple tests of statistical parameters. This drastically reduces computing time as the null distribution can be estimated from one bootstrap procedure and applied for many tests. We demonstrate the efficacy of our procedure by investigating sex differences in development-related changes of cerebral blood flow (CBF) measured using arterial spin labeled MRI (Satterthwaite et al., 2014a).

In Section 3.2 we discuss several FWER controlling procedures used in neuroimaging and classify them with regard to single-step/step-down and marginal/joint procedures. In Section 3.3 we present the new PBJ procedure. We summarize the data and analyses in Section 3.4 and the simulation methods in Section B.1.2 of the Supplement. In Section 3.5 we use simulations to investigate when joint MTPs maintain the nominal type 1 error rate, and we compare the power and FWER of the PBJ to commonly used MTPs using simulations of region- and voxel-wise data analyses. Finally, in Section 3.6 we perform region- and voxel-wise analyses of the CBF data.

3.2. Overview of Multiple Testing Procedures

Throughout, we will assume the image intensity for n subjects, $Y_v \in \mathbb{R}^n$, for voxels or regions, $v = 1 \ldots, V$, can be expressed as the linear model

$$Y_v = X_0\alpha_v + X_1\beta_v + \epsilon_v = X\zeta_v + \epsilon_v, \tag{3.1}$$

where X_0 is an $n \times m_0$ matrix of nuisance covariates, X_1 is an $n \times m_1$ matrix of variables to be tested, $m = m_0 + m_1$, $X = [X_0, X_1]$, parameters $\alpha_v \in \mathbb{R}^{m_0}$, $\beta_v \in \mathbb{R}^{m_1}$, and $\zeta_v = [\alpha_v^T, \beta_v^T]^T$. Let $Y = [Y_1, \ldots, Y_V]$ and let Y_i denote an arbitrary row vector of Y. Assume that the $V \times V$ covariance matrix is the same for each subject, cov$(Y_i) = \Psi$, and define the correlation matrix

$$\Sigma_{j,k} = \Psi_{j,k}/\sqrt{\Psi_{j,j}\Psi_{k,k}}. \tag{3.2}$$

We denote the observed test statistics by Z_{v0} for $v = 1, \ldots, V$, where we reject the null $H_0 : \beta_v = 0$ for large values of Z_{v0}. The notation Z is used (as opposed to F) as we consider transformed F-statistics in Section 3.3.

At each location we are interested in performing the test of the null hypothesis

$$H_{0v} : \beta_v = 0$$

using an F-statistic. The form of model (3.1) covers a wide-range of possible tests including tests of group differences, continuous covariates, analysis of variance, and interactions. V is typically in the hundreds for region-wise analyses or the hundreds of thousands for voxel-wise analyses.

The goal of all multiple testing procedures is to control some measure of the number of false positive findings in a family of hypothesis tests. We will assume the approach of controlling the FWER at some level $0 < \alpha < 1$. In most fields we would like to maintain control of the FWER even in the case that there are false null hypotheses. This is referred to as strong control of the FWER (Hochberg, 1988).

Definition 3.2.1. *Let $\{H_1, \ldots, H_V\} = H$ denote a set of hypotheses. A correction procedure has*

α level strong control of the FWER if for all $H' \subset H$ of true null hypotheses

$$\mathbb{P}(\textit{retain } H_v \textit{ for all } H_v \in H') \geq 1 - \alpha. \tag{3.3}$$

In neuroimaging, strong control of the FWER corresponds to maintaining the correct FWER control even if there is a set of regions or voxels where there is a true effect. The strong FWER controlling procedures discussed in this manuscript are given in Table B.1.

3.2.1. Single-step and Step-down procedures

Due to the complex dependence structure of the test statistics in neuroimaging data it is necessary to use testing procedures that are appropriate for any type of dependence among tests. Single-step and step-down procedures are two classes of MTPs that have strong control of the FWER for any dependence structure (Dudoit, Shaffer, and Boldrick, 2003). These are in contrast to step-up procedures, which make explicit assumptions about the dependence structure of the test statistics (Sarkar and Chang, 1997). When testing V hypotheses $H = \{H_1, \ldots, H_V\}$ in a family of tests, the single-step procedures use a more stringent common threshold $\alpha^* \leq \alpha$ such that the inequality (3.3) is guaranteed. While this procedure is simple, in most cases it is uniformly more powerful to use a step-down procedure.

Procedure 3.2.1 (Step-down procedure)**.** Let $p_{(1)}, \ldots, p_{(V)}$ be the increasingly ordered *p*-values, for hypotheses $H_{(1)}, \ldots, H_{(V)}$. The step-down procedure uses thresholds $\alpha^*_{(1)} \leq \ldots \leq \alpha^*_{(V)} \leq \alpha$ to find the largest value of K such that

$$p_{(k)} < \alpha^*_{(k)} \text{ for all } k \leq K,$$

and rejects all hypotheses $H_{(1)}, \ldots H_{(K)}$.

The single-step procedure can usually be improved by modifying it to be step-down procedure while still maintaining strong control of the FWER (Holm, 1979; Hommel, 1988; Marcus, Eric, and Gabriel, 1976). The canonical example of this modification involves the Bonferroni and Holm procedures (Dunn, 1961; Holm, 1979). The Bonferroni procedure uses the common threshold $\alpha^* = \alpha/V$ for all hypotheses, and rejects all H_k such that $p_k < \alpha/V$. The Holm procedure instead uses the thresholds $\alpha^*_{(k)} = \alpha/(V + 1 - k)$ and rejects using Procedure 3.2.1. Holm's procedure is always more powerful than Bonferroni and still controls the FWER strongly (Holm, 1979).

3.2.2. Joint testing procedures

Multiple testing procedures can further be classified into marginal and joint testing procedures. The Bonferroni and Holm approaches are called marginal procedures because they do not make use of the dependence of the test statistics. As they must be able to control any dependence structure they are more conservative than joint testing procedures that cater exactly to the distribution of the test statistics (Dudoit and Laan, 2008). The benefit of accounting for the dependence of test statistics is critical in neuroimaging, where the test statistics are highly dependent due to spatial, anatomical, and functional dependence. Joint MTPs differ in how the null distribution is estimated from the sample, but use the same procedure to compute single-step or step-down adjusted *p*-values. For this reason we will first discuss the estimation of the null distribution and then discuss how adjusted *p*-values are computed from the estimate of the null.

Estimating the null distribution with permutations

The "Randomise" procedure proposed by Winkler et al. (2014) is a single-step permutation joint (PJ) MTP widely used in neuroimaging. The PJ MTP procedure is a modification of the Freedman-Lane procedure (Freedman and Lane, 1983) implemented by Winkler et al. (2014) that estimates the null distribution of the test statistics by permuting the residuals of the reduced model to obtain estimates of the parameters under the null hypothesis that there is no association between the variables in X_1 and the outcome. Though only the single-step procedure has been proposed for use in neuroimaging, for completeness we include null estimation of the test statistics for the step-down procedure.

Procedure 3.2.2 (Permutation Null Estimation)**.** Assuming the model (3.1):

1. Regress Y_v against the reduced model $Y_v = X_0\alpha_v + \epsilon_v$ to obtain the residuals $\hat\epsilon_v$ and the test statistics Z_{v0} for all regions or voxels $v = 1, \ldots, V$.

2. Order the test statistics $Z_{(1)0} < Z_{(2)0} < \ldots Z_{(V)0}$ and let $\epsilon_{(v)}$ be the corresponding residuals.

3. For $b = 1, \ldots, B$, randomly generate a permutation matrix P_b, permute the residuals $\hat\epsilon_{(v)b} = P_b\hat\epsilon_{(v)}$, and define the permuted data at each voxel as $Y_{(v)b} = \hat\epsilon_{(v)b}$.

4. For $v = 1, \dots, V$ and $b = 1, \dots, B$ regress $Y_{(v)b}$ onto the full model (3.1) to obtain the test statistic $Z_{(v)b}$ to be used as an estimate of the null distribution.

Ordering the test statistics is not necessary for the single-step procedure, but is required for computing step-down adjusted *p*-values. Note that for any given b the generated test statistics $Z_{(v)b}$ may not be increasing in v. Strong control of the FWER for this permutation procedure relies on the assumption of subset pivotality (see Supplement) (Westfall and Young, 1993).

This null distribution can be used to compute rejection regions or adjusted *p*-values. Because all joint testing procedures rely on estimates of the null distribution they are approximate in finite samples. The permutation methodology and our proposed PBJ procedure (Section 3.3) only differ in how the null distribution is estimated.

Computing adjusted p-values

The following procedures describe how to obtain single-step and step-down adjusted *p*-values using any estimate of the null distribution generated by permutation or bootstrapping.

Procedure 3.2.3 (Single-Step Joint Adjusted *p*-values)**.** Assuming the model (3.1) and an empirical distribution of null statistics $Z_{(v)b}$ for $v = 1, \dots, V$ and $b = 1, \dots, B$:

1. Compute $Z_{\text{max},b} = \max_{v \le V} Z_{(v)b}$.

2. Compute the voxel-wise corrected *p*-value as $\tilde{p}_v = 1/B \sum_{b=1}^{b} I(Z_{\text{max},b} \ge Z_{v0})$, where $I(\cdot)$ is the indicator function.

Procedure 3.2.4 (Step-down Joint Adjusted *p*-values)**.** Assuming the model (3.1) and an empirical distribution of null statistics $Z_{(v)b}$ for $v = 1, \dots, V$ and $b = 1, \dots, B$:

1. Compute the statistics $Z_{\max v,b} = \max_{k \le v} Z_{(k)b}$.

2. Compute the the intermediate value $p^*_{(v)} = 1/B \sum_{b=1}^{B} I(Z_{\max v,b} \ge Z_{(v)0})$.

3. The adjusted *p*-values are $\tilde{p}_{(v)} = \max_{k \le v} p^*_{(k)}$.

The single-step procedure is less powerful than the step-down counterpart (Dudoit and Laan, 2008). This is evident in comparing the procedures, as the adjusted *p*-values for the step-down procedure are at least as small as the adjusted *p*-values from the single-step. The adjusted *p*-

values obtained using the step-down approach correspond to using Procedure 3.2.1. The key feature of Procedure 3.2.3 is that the estimate of the joint distribution is used to compute adjusted *p*-values, where as Holm's procedure is a version of Procedure 3.2.1 that only uses the marginal distribution of the test statistics. So far, we have described existing MTPs used in neuroimaging.

3.3. Parametric-Bootstrap

In this section we propose single-step and step-down PBJ approaches that are conceptually identical to the PJ procedure, but differ in how the null distribution of the statistics is generated. The PBJ is based on the theory developed by Dudoit and Laan (2008) and therefore does not rely on the assumption of subset pivotality. We will allow the additional assumption that under the null the test statistics are approximately chi-squared. The chi-squared approximation can rely on asymptotic results, or as we show in Section 3.4.2, a transformation can be used so that the test statistics are approximately chi-squared. As with the PJ procedure discussed above this implies that the *p*-values are approximations that become more accurate as $n \to \infty$. In Section 3.5 we use simulations to show that the procedures control the FWER in sample sizes available in many neuroimaging studies.

3.3.1. Asymptotic control of the FWER

Here, we give a brief overview of the underlying assumptions sufficient to prove that the adjusted *p*-values from the PBJ control the FWER asymptotically. Details are given in the Supplement. We require that the test statistics' null distribution satisfies the null domination condition (Dudoit and Laan, 2008 p. 203) and need a consistent estimate of the null distribution.

Definition 3.3.1 (Asymptotic null domination)**.** *Let H_0 denote the indices of M true null hypotheses in the set of V hypotheses $H = \{H_1, \dots, H_V\}$, with corresponding test statistics $Z_{1n}, \dots, Z_{Vn}$. The V-dimensional null distribution Q_0 satisfies the asymptotic null domination condition if for all $x \in \mathbb{R}$*

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_{m\in H_0} Z_{mn} > x\right) \le \mathbb{P}\left(\max_{m\in H_0} Z_m > x\right),$$

where $Z_n \sim Q_n$ is distributed according to a finite sample null joint distribution Q_n and $Z \sim Q_0$.

The joint null distribution Q_0 for the test statistics can be used to compute asymptotically accurate adjusted *p*-values if the null domination condition holds. We use a diagonal singular Wishart distribution as the null because it is proportional to the asymptotic distribution of a vector of F-statistics. We also transform the F-statistics marginally to chi-squared statistics if

$$\epsilon_v \sim \mathcal{N}(0, \sigma_v I), \tag{3.4}$$

for the error term in model (3.1). The assumption of asymptotic null domination of Definition 3.3.1 is satisfied when using our transformated F-statistics even if the error distribution is not normal (See Theorem B.2.2 in the Supplement). Thus, the PBJ provides approximate *p*-values regardless of the error distribution.

In practice the joint distribution of the test statistics, Q_0, is not known *a priori* and must be estimated from the data. In order to obtain asymptotically valid adjusted *p*-values, the estimate for Q_0, $\hat{Q}_0$, must be consistent. Because the distribution $Q_0 = Q_0(\Sigma)$ is a function of the covariance matrix Σ, a consistent estimator for Q_0 can be obtained from a consistent estimator for Σ. The choice of the estimator is critical because $\hat{\Sigma}$ must be consistent for Σ under the alternative distribution. That is, even if some null hypotheses are false $\hat{\Sigma}$ must be consistent. The PBJ procedure uses a consistent estimator for Σ based on the residuals of the full design X. The consistency of $\hat{\Sigma}$ under the alternative guarantees asymptotic control of the FWER (see supplementary material). Note that, in general, the PJ procedure may not yield a consistent estimator for the joint distribution Q_0 if the alternative is true at more than one location because the estimates of covariances are biased. The covariance estimates are biased due to the fact that, for the reduced model used by the PJ MTP (see Procedure 3.2.2), the mean is incorrectly specified in locations where the alternative is true. If the assumption of subset pivotality is satisfied, then the permutation estimator will be consistent.

3.3.2. Parametric bootstrap null distribution

For the parametric bootstrap we assume model (3.1) and use F-statistics for the test $H_{0v} : \beta_v = 0$.

$$F_{vn} = \frac{(n-m) Y_v^T (R_{X_0} - R_X) Y_v}{m_1 Y_v^T R_X Y_v}, \tag{3.5}$$

where R_A denotes the residual forming matrix for A. When the errors are normally distributed (3.4), F_{vn} is an F-distributed random variable with m_1 and $n - m$ degrees of freedom. When the errors are not normal the statistics (3.5) are asymptotically $m_1^{-1}\chi^2_{m_1}$. The following theorem gives the asymptotic joint distribution of the statistics.

Theorem 3.3.2. *Assume model* (3.1)*, let F_{vn} be as defined in* (3.5)*, and define the $p \times V$ matrix $\alpha = [\alpha_1, \ldots, \alpha_V]$. Further assume that, under the null,*

$$R_{X_0}\mathbb{E}Y = \mathbb{E}Y - X_0\alpha = 0 \tag{3.6}$$

$$\|\Psi\|_M < \infty, \tag{3.7}$$

where $\|\Psi\|_M = \sup_x \|\Psi x\|/\|x\|$ is the induced norm.

Then the following hold:

1. *Define the matrix $\Phi_{i,i} = 1/\sqrt{\Psi_{i,i}}$ and $\Phi_{i,j} = 0$ for $i \neq j$. When* (3.4) *holds,*

$$\begin{aligned}\Phi Y^T(R_{X_0} - R_X)Y\Phi &\sim \mathcal{W}_V(m_1, \Sigma)\\ \Phi Y^T R_X Y\Phi &\sim \mathcal{W}_V(n-m, \Sigma)\end{aligned} \tag{3.8}$$

   where $\mathcal{W}_p(d, \Sigma)$ denotes a singular Wishart distribution with degrees of freedom $d < p$ and matrix $\Sigma \in \mathbb{R}^{p\times p}$ (Srivastava, 2003).

2. *The F-statistics converge in law to the diagonal of a singular Wishart distribution, that is,*

$$m_1 F_n = m_1[F_{1n}, \ldots, F_{Vn}] \to_L \operatorname{diag}\{\mathcal{W}_V(m_1, \Sigma)\}.$$

as $n \to \infty$.

In order to make the statistics robust regardless of whether the errors are normal we use the transformation

$$Z_{vn} = \Phi^{-1}\{\Phi_n(F_{vn})\}, \tag{3.9}$$

where Φ_n is the cumulative distribution function (CDF) for a $\mathcal{F}(m_1, n-m)$ random variable and Φ^{-1} is the inverse CDF for a $\chi^2_{m_1}$ random variable. If the errors are normal (3.4), then the transformed statistics (3.9) are marginally χ^2_m statistics. When the errors are nonnormal then Z_{vn} is asymptotically $\chi^2_{m_1}$ (see Theorem B.2.2 in the Supplement). The asymptotic joint distribution of the statistics given in Theorem 3.3.2 allows us to use a diagonal singular Wishart to compute approximate adjusted *p*-values. To compute probabilities we do not have to sample the full matrix (3.8), since only the diagonal elements, diag$\{Y^T(R_X - R_{X_F})Y\}$, are required.

To find adjusted *p*-values, $\tilde{p}_v$, for the single-step procedure we compute the probability

$$\tilde{p}_v = \mathbb{P}\left(\max_{k \leq V} |Z_k| > |Z_{v0}|\right), \tag{3.10}$$

where

$$Z = (Z_1, \ldots, Z_V) \sim \operatorname{diag}\left\{\mathcal{W}_V(m_1, \Sigma)\right\}, \tag{3.11}$$

and Z_{v0} is the observed statistic at location v. Theorems B.2.2 and B.2.4 in the Supplement guarantee asymptotic control of the FWER when (3.11) is used as the null distribution.

In practice the joint distribution (3.11) is unknown due to the fact that Σ is unobserved, so the probability (3.10) cannot be computed. We must obtain an estimate for Σ in order to compute estimates of these probabilities. Since the diagonal, $\Sigma_{v,v} = 1$, we only need to estimate the off-diagonal elements. By estimating ρ_{ij} with the consistent estimator

$$\hat{\rho}_{jk} = \frac{Y_j^T R_X Y_k}{\hat{\sigma}_j \hat{\sigma}_k},$$

we are guaranteed asymptotic control of the FWER (see the Supplement). This estimator is biased toward zero in finite samples, and yields conservative estimates of the correlation. Note that using the residuals of the full model is crucial here as that estimator yields consistent estimates of the correlation regardless of whether the alternative is true in each location. Instead of using Σ in (3.11) we use the estimated covariance matrix of the test statistics

$$\hat{\Sigma}_{jk} = \begin{cases} 1 & : j = k \\ \hat{\rho}_{jk} & : j \neq k \end{cases}. \tag{3.12}$$

Importantly, the covariance of the tests statistics does not depend on what model parameter is being tested, so a single null distribution can be used for tests of all parameters provided the tests are on the same degrees of freedom. This conserves computing time relative to the permutation procedure which must estimate a null distribution for each test.

3.3.3. The parametric-bootstrap procedure

We compute *p*-values using a parametric bootstrap: We use the estimate of $\hat{\Sigma}$ to generate B diagonal singular Wishart statistics. Because the rank of $\hat{\Sigma}$ is at most $\min\{(n-m), V\}$ it does not require the storage of the full $V \times V$ covariance matrix if $V > (n-m)$. This gives the following procedure for estimating the null distribution using the parametric bootstrap.

Procedure 3.3.1 (Parametric bootstrap null estimation)**.** Assuming the model (3.1):

1. Regress Y_v onto X to obtain the test statistics for $H_{v0} : \beta_v = \beta_{v0}$ using (3.9). Let $Z_{(1)0} < Z_{(2)0} <, \ldots, < Z_{(V)0}$ denote the ascending test statistics and $\tilde{E} = [\hat{\epsilon}_{(1),0}, \ldots, \hat{\epsilon}_{(V)0}]$, the $n \times V$ matrix of their associated residuals from model (3.1).

2. Standardize $\tilde{E}$ so that the column norms are 1. Denote the standardized matrix by E. Let $r = \text{rank}(E) = \min\{n-m, V\}$.

3. Use E, m_1, and r to generate the null test statistics $Z_b = (Z_{(1)0}, \ldots, Z_{(V)b})$ for $b = 1, \ldots, B$.

   (a) Perform the singular value decomposition of $E = UD\tilde{M}^T$ where D is an $r \times r$ diagonal matrix and U and M have orthonormal columns. Let $M = \tilde{M}D$.

   (b) Generate an $r \times m_1$ matrix S_b with independent standard normal entries.

   (c) Obtain the null statistics $Z_b = \text{diag}(MS_bS_b^TM^T)$. Z_b are distributed according to a diagonal singular Wishart distribution.

The singular value decomposition only needs to be performed once for the entire procedure. Computing the statistics does not require multiplying $MS_bS_b^TM^T$, because we only need the diagonal entries. Procedures 3.2.3 and 3.2.4 are used to compute single-step and step-down adjusted *p*-values from the bootstrap sample.

3.4. Methods

Code to perform the analyses presented in this manuscript is available at `https://bitbucket.org/simonvandekar/param-boot`. While the processed data are not publicly available, unprocessed data are available for download through the Database of Genotypes and Phenotypes (dbGaP) (Satterthwaite et al., 2016). We provide simulated region-wise data with the code so that readers can perform the region-wise analyses presented here. Synthetic simulations are used to assess finite sample properties of each testing procedure. Simulations that resample from the CBF data are used to assess the FWER in real data. Methods describing the simulation procedures are given in Section B.1.2 of the online supplement.

3.4.1. Cerebral blood flow data description

The Philadelphia Neurodevelopmental Cohort (PNC) is a large initiative to understand how brain maturation mediates cognitive development and vulnerability to psychiatric illness (Satterthwaite et al., 2014b). In this study, we investigate image and regional measurements of CBF using an arterial spin labeling (ASL) sequence collected on 1,578 subjects, ages 8-21, from the PNC (see Satterthwaite et al. (2014b) for details). Abnormalities in CBF have been documented in several psychiatric disorders including schizophrenia (Pinkham et al., 2011) and mood and anxiety disorders (Kaczkurkin et al., 2016). Establishing normative trajectories in CBF is critical to characterizing neurodevelopmental psychopathology (Satterthwaite et al., 2014a).

Image preprocessing steps are described in Section B.1.1 of the Supplement. In order to parcellate the brain into anatomically defined regions, we use an advanced multi-atlas labelling approach which creates an anatomically labeled image in subject space (Avants et al., 2011). The ASL data are pre-processed using standard tools included with FSL (Jenkinson et al., 2002; Satterthwaite et al., 2014a). The T1 image is then co-registered to the CBF image using boundary-based registration (Greve and Fischl, 2009), the transformation is applied to the label image, and then CBF values are averaged within each anatomically-defined parcel.

Of the 1,601 subjects with imaging data 23 did not have CBF data. Three hundred thirty two were excluded due to clinical criteria, which include a history of medical disorders that affect the brain, a history of inpatient psychiatric hospitalization, or current use of a psychotropic medication. An additional 274 subjects were excluded by an imaging data quality assurance procedure which included automatic and manual assessments of data quality and removal of subjects with negative mean CBF values in any of the anatomical parcels. These exclusions yielded a total of 972 subjects used for the imaging simulations and analysis.

3.4.2. Cerebral blood flow statistical analysis

We perform region- and voxel-wise analyses of CBF in order to identify locations where there are sex differences in development-related changes of CBF. For the region-wise analysis, we test the sex by age interaction on the average CBF trajectories in the $V = 112$ regions using an F-statistic from an unpenalized spline model. For the voxel-wise analysis, we perform the same test in all $V =$127,756 gray matter voxels.

For the region-wise data we fit the age terms with thin plate splines with 10 degrees of freedom. Thus, the numerator of the F-statistic has 9 degrees of freedom using the `mgcv` package in `R` (R Core Team, 2016; Wood, 2011). Results from the simulation analyses and previous theoretical results (Gotze, 1991) demonstrate that multivariate convergence rates depends on the dimension of the vector of statistics (Tables 3.1 and B.2), and the degrees of freedom of the test (Figure 3.2). For this reason, we use the Yeo-Johnson transformation for the PBJ procedure so that the transformed CBF data are approximately normal (Yeo and Johnson, 2000). We estimate the age terms for the voxelwise data on 5 degrees of freedom, so that the test for the interaction is on 4 degrees of freedom. Race and motion (mean relative displacement; MRD) are included as covariates at each location for both analyses. Similar results were presented in a previous report (Satterthwaite et al., 2014a) using Bonferroni adjustment. We also perform the voxel-wise analyses using a 10 degrees of freedom spline basis in the Supplement.

The same MTPs are compared for the region- and voxel-wise CBF analysis as used in simulations. For the region- and voxel-wise analyses 10,000 samples are used for the PBJ and PJ procedures. We present the corrected results with the FWER controlled at $\alpha = 0.01$ for the region-wise analysis and $\alpha = 0.05$ for the voxel-wise analysis. A more conservative threshold is used for the region-wise data as the smaller number of comparisons and noise reduction from averaging within regions increases power considerably.

3.5. Simulation Results

3.5.1. Synthetic simulation results

We use simulations to explore how the FWER is affected by using asymptotic approximations and estimating the covariance matrix in finite sample sizes. Results are shown for sample sizes of

	$n = 100$	T_n | Holm	Z_n | Holm	$T_n \mid \Sigma$	$T_n \mid \hat{\Sigma}$	$Z_n \mid \hat{\Sigma}$	T_n | Perm
Indep	$m = 100$	6	6	6	8	6	4
	$m = 200$	6	4	7	8	5	5
	$m = 500$	7	6	8	8	6	5
	$m = 1000$	10	4	10	10	5	6
	$m = 5000$	10	4	11	10	4	5
	$m = 10000$	14	4	15	14	4	5
Pos AR(1)	$m = 100$	2	1	6	4	4	3
	$m = 200$	4	3	6	8	5	4
	$m = 500$	5	3	7	8	4	5
	$m = 1000$	7	2	8	9	5	6
	$m = 5000$	6	3	9	8	4	4
	$m = 10000$	9	3	11	10	4	3
Neg AR(1)	$m = 100$	5	3	7	9	6	5
	$m = 200$	3	2	5	6	4	3
	$m = 500$	4	2	8	8	5	4
	$m = 1000$	5	3	8	8	4	4
	$m = 5000$	8	4	10	9	6	5
	$m = 10000$	8	3	11	9	6	5

Table 3.1: Type 1 error results for $n = 100$ to assess convergence rates. See Table B.2 for details.

$n = 100$ and $n = 40$ (Tables 3.1 and B.2). Two hypotheses are tested on one (B.1) and three (B.2) degrees of freedom. The results demonstrate that by using the sample covariance estimate in the PBJ MTP the FWER is only slightly inflated for $n = 40$ and when $n = 100$ the FWER is controlled at the nominal level for all the dimensions considered (Column $Z_n \mid \hat{\Sigma}$). Note, that when the transformation (3.9) is not used all procedures have inflated FWERs (Columns T_n in Tables 3.1 and B.2) due to the fact that the multivariate normal approximation is not accurate and is worse for a larger number of tests (Gotze, 1991). We can conclude that most of the error in estimating the FWER comes from the normal approximation. Even when $\hat{\Sigma}$ is rank deficient it still provides nominal FWER control. Interestingly, the PJ procedure controls the FWER for all the sample sizes considered. While permutation tests are exact for univariate distributions (Lehmann and Romano, 2006), to our knowledge there is no theoretical justification that multivariate permutations are accurate when the number of statistics exceeds the sample size. Finally, the PBJ and PJ MTPs control the FWER at the nominal level, while Holm's procedure is conservative for correlated test statistics.

3.5.2. Region-wise FWER and power

We use simulations that sample from real imaging data to assess the FWER in finite samples for region-wise analyses. For the test of hypothesis (B.1), the PBJ and PJ procedures maintain the

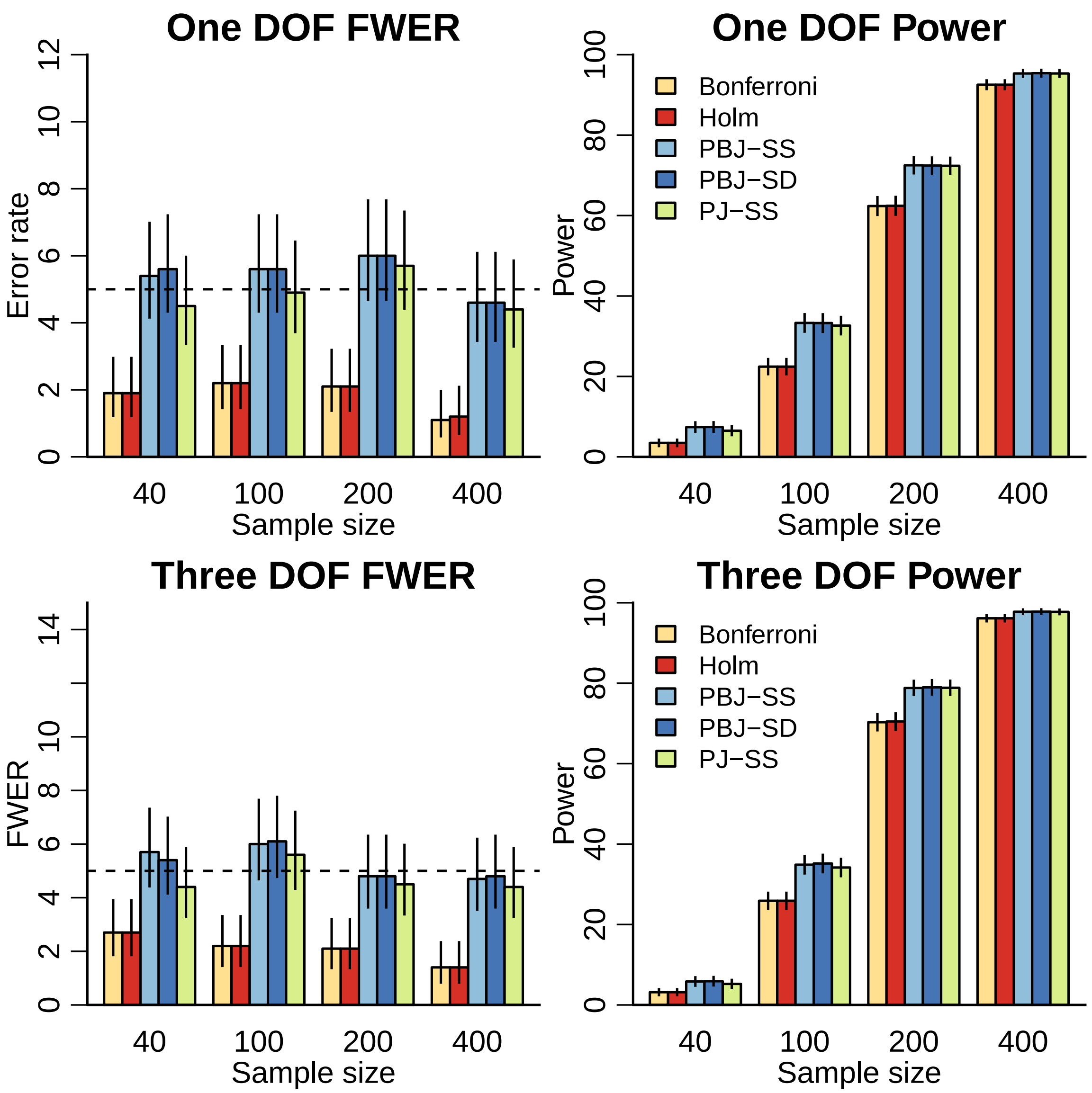

Figure 3.1: FWER and power for the F-statistic of (B.1) on one degree of freedom (DOF) and F-statistic of (B.2) on three DOF for $V = 112$ brain regions. Bonferroni and Holm both have conservative control. The PBJ maintains accurate FWE control even when $n < p$. The power for the PBJ and PJ procedures are equal. SS=Single-step; SD=Step-down. Lines indicate 95% confidence intervals.

nominal FWER for all samples (Figure 3.1). As expected, the FWER for the Bonferroni and Holm procedure are conservative. The power of the joint testing procedures is higher than the marginal testing procedures (Figure 3.1). The PBJ has equal power to the PJ procedure. The step-down procedure confers almost no benefit.

For testing the hypothesis (B.2) the PBJ and PJ procedures control the FWER at the nominal level for all sample sizes (Figure 3.1). The Bonferroni and Holm procedures give similarly conservative FWERs as the single degree of freedom test. Power analyses demonstrate that the PBJ and PJ procedures have the same power. As with testing (B.1) the step-down procedure shows little improvement over the single-step.

3.5.3. Voxel-wise FWER and power

We use simulations that sample from real imaging data to assess the type 1 error and power for voxel-wise analyses. For the test of (B.1), the PBJ maintains the nominal FWER for samples sizes greater than 200. The PJ procedure maintains control for all sample sizes considered (Figure 3.2). As expected, Bonferroni and Holm procedures have conservative FWER. This conservative FWER leads to a reduction in power for these methods (Figure 3.2). The power for PBJ was comparable to the PJ procedure.

For the test of (B.2) on 3 degrees of freedom the PBJ controls the FWER for sample sizes larger than 200 (Figure 3.2). The PBJ procedure does not control the FWER for smaller sample sizes because the estimate of the covariance structure (3.2) is not accurate enough to work as a plug-in estimator for the full rank covariance matrix. The PJ procedure maintains nominal control of the FWER for all sample sizes. The marginal testing procedures have the same conservative FWER control as above. Both joint testing procedures have greater power than the marginal procedures.

3.5.4. Computing time

While the PJ and PBJ procedures are both linear in the sample size, as the PBJ procedure works directly with the distribution of the test statistics we expect it to be faster than permutations. To compare observed computing times we took the ratio of the time to evaluate both tests for PJ and PBJ procedures. The computing times for the sample sizes simulated for the region-wise data are given in Figure B.1. The PBJ is at least twice as fast as the PJ procedure and increases with the sample size so that the computing time reduction is larger for larger sample sizes.

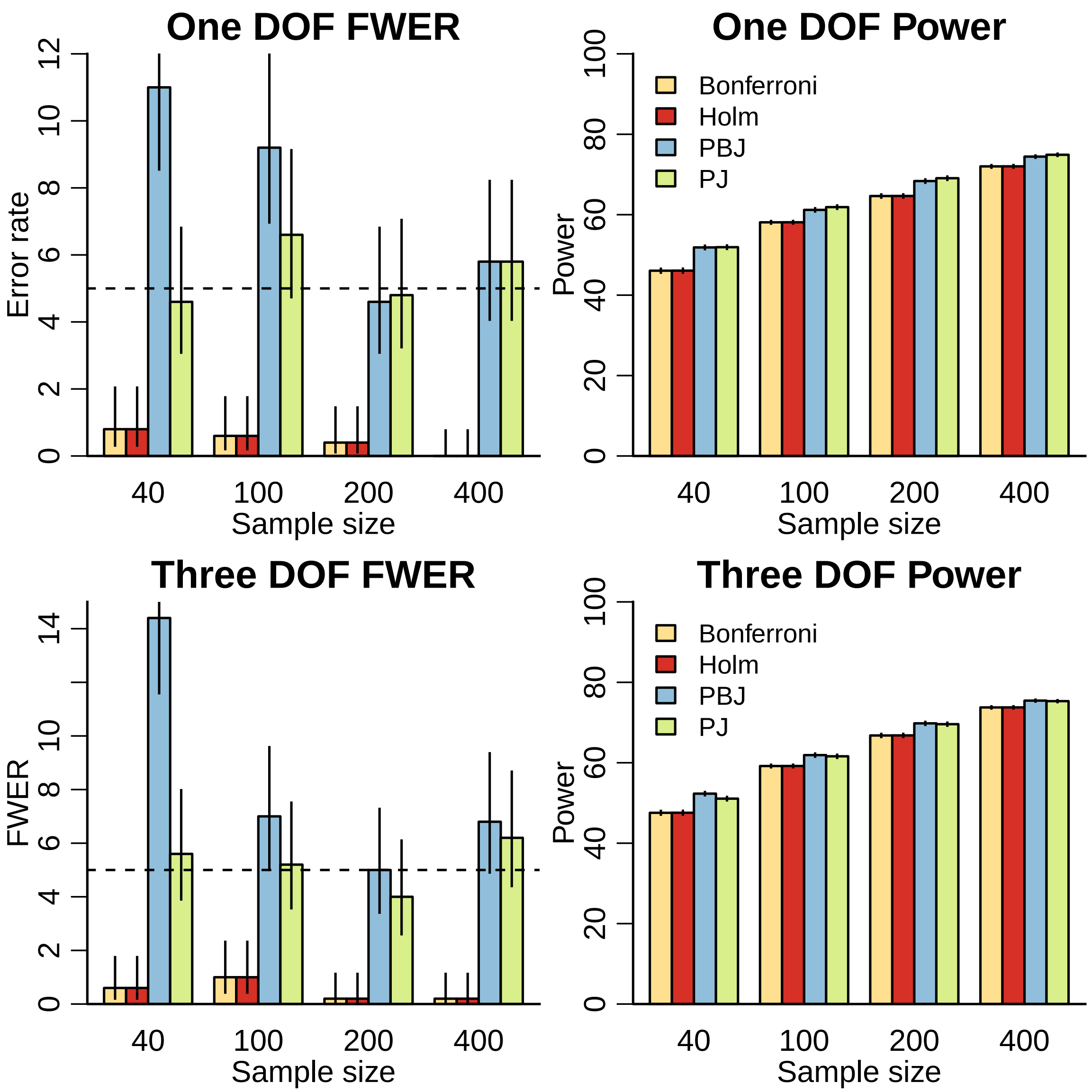

Figure 3.2: FWER and power for the F-statistic of (B.1) on one degree of freedom (DOF) and F-statistic of (B.2) on three DOF for the CBF image with V =127,756 voxels. Bonferroni and Holm both have conservative control. The PBJ controls the FWER for samples sizes greater than 200. The power for the PBJ is approximately equal to the PJ tests. Lines indicate 95% confidence intervals.

3.6. Cerebral Blood Flow Results

To compare the testing procedures in the CBF data we perform a test for a nonlinear age-by-sex interaction on CBF trajectories for region- and voxel-wise analyses. For the voxel-wise analysis the PBJ and PJ MTPs take 5.9 and 69.2 hours to run, respectively. We perform an F-test for the interaction at each of the 112 regions (Figure 3.3). The Bonferroni, Holm, PBJ, and PJ procedures reject 56, 71, 78, and 61 regions, respectively. The PJ procedure is more conservative for two reasons. The first is that it is a single-step procedure; when there is a relatively large number of rejected tests then using a step-down procedure is more likely to improve power. The second reason is that the finite sample distribution of each of the regions is different. Regions near the edge of the brain are likely to be more heavily skewed due to imperfections in the image registration. By comparing all regions to the distribution of the maximum the PJ procedure is necessarily conservative because it compares to the most heavily skewed regions. In contrast, by transforming the data prior to using the PBJ procedure the marginal distribution of the test statistics are approximately equal across regions.

For the voxel-wise analysis we perform the F-test for the nonlinear interaction on 4 degrees of freedom. The single-step PBJ offers improved power over the Holm and PJ procedures (Figure 3.4). The Holm procedure ignores the covariance structure of the test statistics so yields conservative results. The PJ procedure is more conservative even than the Holm procedure. As with the region-wise analysis this is likely because the finite sample distribution of the test statistics is different: voxels near the edge of the brain tend to have higher variance and are likely heavily skewed. If there is a subset of voxels with a heavily skewed distribution then taking the maximum test statistic will yield conservative inference for all locations that have tighter distribution. By transforming the distribution of the voxels to be approximately normal the PBJ procedure offers improved power and speed.

3.7. Discussion

We introduced a fast parametric bootstrap joint testing procedure as a new tool for multiple comparisons in neuroimaging. The PBJ procedure improves computing time by generating the test statistics directly instead of permuting the original data. If normality assumptions about the data generating distribution do not hold, then the Yeo-Johnson transformation can be used to obtain

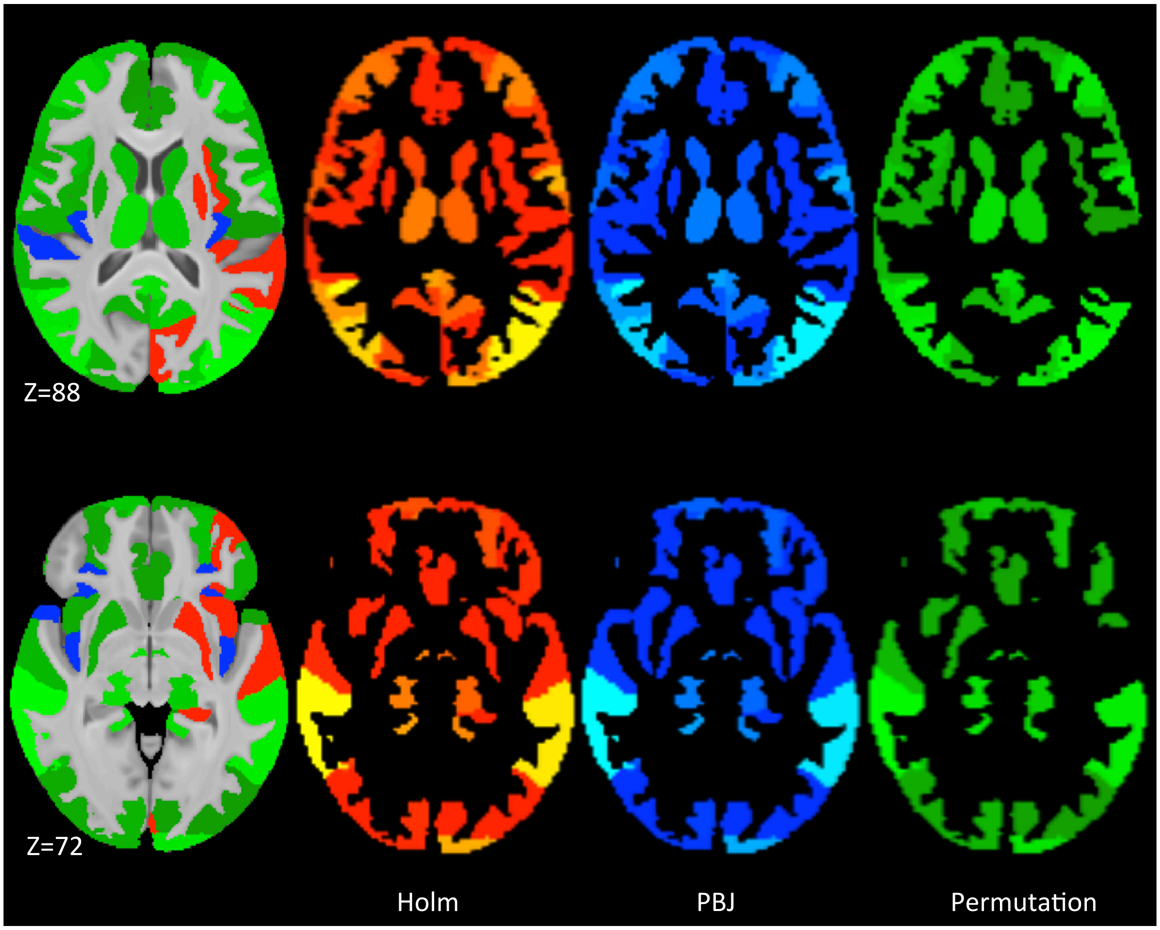

Figure 3.3: FWER controlled results at $\alpha = 0.01$ for Holm (red), PBJ step-down (blue), and PJ single-step (green) for the region-wise analysis. Color scale is $-\log_{10}(p)$ and shows results greater than 2. The left-most images show the overlay of PBJ, Holm, and PJ in that order. The color images show regions identified by Holm, PBJ, and NPBJ.

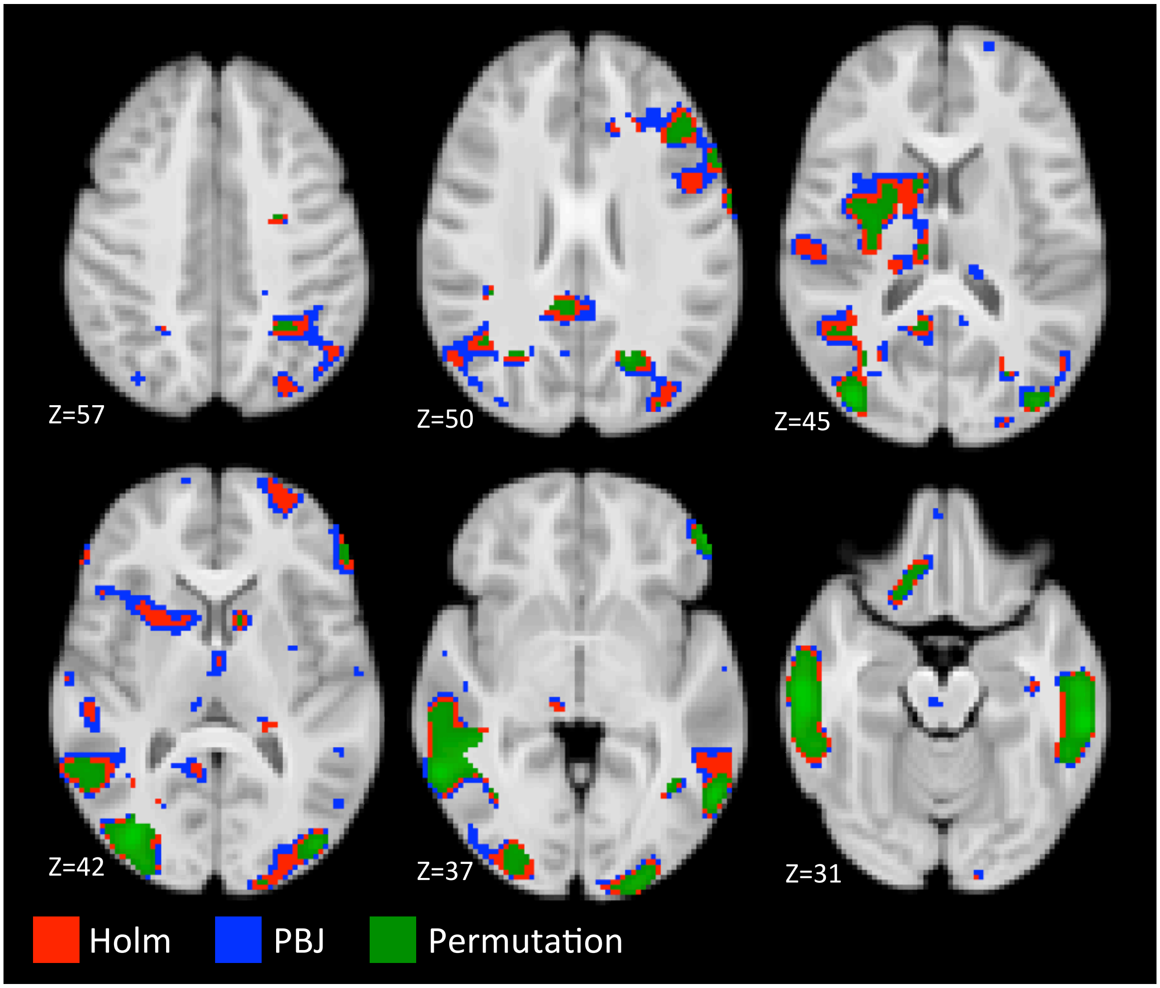

Figure 3.4: FWER controlled results at $\alpha = 0.05$ for Holm (red), PBJ single-step (blue), and PJ single-step (green) for the voxel-wise analysis. Color scale is $-\log_{10}(p)$ for the adjusted *p*-values and shows results greater than 1.3. The overlay order is the PBJ, Holm, and PJ procedures, so that green indicates regions where all three regions reject the null, red and green indicate regions where Holm and PJ reject, and the union of all colors is where PBJ rejects. Blue indicates locations where only the PBJ procedure rejects.

statistics that are approximately normal to improve the finite sample performance of the procedure.

In the CBF data analysis the PBJ is more powerful than the PJ MTP because the PJ MTP does not account for the fact that the finite sample distribution of the test statistics can be different. Differences in the finite sample distribution of the statistics are attributable to certain regions near the edge of the brain having larger variance and skew. For this reason taking the maximum across locations leads to conservative inference in locations that actually have tighter tails. While the PBJ generates from a chi-squared distribution this ensures that a few heavy-tailed locations do not affect the distribution of the maximum.

In simulations, the step-down procedures provide little improvement in power over the single-step procedures. However, in the regionwise analyses Holm rejected 15 more regions than the Bonferroni procedure. The reason for the difference is that step-down procedures offer little benefit when there is a small number of false null hypotheses and a large number of tests.

Using simulations we found that both joint procedures perform well, in some cases even when the number of tests exceeds the sample size. This is quite surprising as it seems impossible that any given estimate of the joint distribution will satisfactorily reproduce the true joint distribution of the test statistics. For example, if we consider the case of normal test statisics, $Z_n = (Z_{1n}, \ldots, Z_{Vn}) \sim N(0, \Sigma)$ with full rank covariance, then the sufficient statistic is $\hat{\Sigma}_n$, which can be of rank n at most. So, the probabilities generated conditioning on $\hat{\Sigma}_n$ assume Z_n is restricted to a linear subspace of $\mathbb{R}^V$. With nonnormal test statistics and more complex dependence structures it can only be more difficult to reproduce the null distribution.

MRI is a flexible noninvasive tool for studying neural aberrations related to psychiatric disorders such as schizophrenia (Pinkham et al., 2011) and mood and anxiety disorders (Kaczkurkin et al., 2016). However, recent studies have shown that the PJ MTP is the only inference methodology to reliably control the FWER in neuroimaging data (Eklund, Nichols, and Knutsson, 2016; Silver, Montana, and Nichols, 2011). We have shown that inference using the currently available permutation procedure can take days and lead to conservative inference. Our proposed PBJ MTP is a reliable and fast testing procedure that will be a critical tool in studying functional and physiological features that can improve our understanding of the brain and its relation to behavior.

CHAPTER 4

Subject-level Measurement of Local Cortical Coupling

4.1. Introduction

To accommodate a marked expansion of surface area, the cortical sheet of the human brain is highly folded (Van Essen, 1997; Zilles, Palomero-Gallagher, and Amunts, 2013). Such folding has important functional implications and may increase computational efficiency due to reduced axonal distance within a gyrus (Van Essen, 1997). Moreover, folding patterns and local cortical thickness are closely inter-related (Economo, 1925; Fischl, 2013). Economo (1925) described that the thickness of the cortex is significantly reduced in passing from the crown of a gyrus to the floor of a sulcus. This striking relationship was subsequently verified in vivo using structural neuroimaging techniques (Fischl and Dale, 2000)(Fischl and Dale, 2000).

It was long posited that this relationship between sulcal depth (SD) and cortical thickness (CT) is established in utero or in the perinatal period simultaneously with the development of cortical convolution (Economo, 1925; Toro and Burnod, 2005; Zilles, Palomero-Gallagher, and Amunts, 2013). However, we recently demonstrated that the relationship between CT and SD evolves dynamically throughout youth (Vandekar et al., 2015). Notably, topological position influences cortical maturation throughout this critical period. We found that linear thinning is widespread across the cortex but is maximal in the depths of the sulcus, whereas circumscribed areas of gyral cortex demonstrate marked nonlinear thickening between the ages of 8 and 14 years (Vandekar et al., 2015).

This work established that understanding relationships between cortical measures (e.g., thickness and depth) is critical for an accurate characterization of the plastic remodeling that occurs during youth. Furthermore, it suggests that such brain phenotypes may be important to understanding neuropsychiatric disorders (schizophrenia, autism, ADHD) that are increasingly conceptualized as disorders of brain development (Insel, 2010; Krain and Castellanos, 2006; Rapoport, Giedd, and Gogtay, 2012; Shaw et al., 2012; Steen et al., 2006; White et al., 2003) and may occur on a localized scale (Ronan et al., 2012; Wagstyl, 2015).

There are important methodological considerations in trying to relate the properties of two cortical surfaces or volumetric brain images: cortical surfaces and brain images are highly autocorrelated, non-independent measures that can be re-sampled to an arbitrary number of observations. Thus, standard parametric statistics are not applicable. Prior work used canonical correlation analysis (Avants et al., 2010; Ouyang et al., 2015) to describe the relationship between two measures in volumetric space. In order to assess the significance of the observed correlation, (Avants et al., 2010) relied on permutation tests that relabeled subjects. This approach seeks to understand the relationship between the two measures across subjects. In our recent work (Vandekar et al., 2015), we introduced a novel spatial permutation testing procedure, adapted from the field of microscopy research, that evaluates the relationship between cortical measures in a statistically rigorous framework. Our approach differed from Avants et al. (2010) in that the interest was in assessing the spatial relationship between two variables within two averaged cortical surfaces. While this approach successfully delineated robust spatial effects associated with development in youth, it was limited in that analyses studied only global effects across subjects. For studies of more subtle individual and group differences in spatial relationships such an approach is not ideal, as it does not allow exploration of local regional effects. A within-subject map describing the spatial relationship among cortical measures would be a substantial advance and allow the application of standard tools for group-level analysis (e.g., GLM, MVPA, etc.) across subjects.

Here we introduce a method for describing local cortical coupling on a within-subject basis using a surface-based, locally-weighted regression procedure. Although this procedure could be used for any two surface maps, we apply it to extend our prior work describing how changes in CT are coupled to SD in development. We capitalize on the Philadelphia Neurodevelopmental Cohort (PNC), a large-scale, single-site study of brain development (Calkins et al., 2015; Satterthwaite et al., 2014b, 2016). Results demonstrate that local coupling is developmentally relevant and specific to regions where the relationship between CT and SD is evolving. We further illustrate the measure's applicability to studies of individual and group differences by demonstrating the presence of substantial sex differences in cortical coupling. Finally, we provide publically available R code (`https://bitbucket.org/simonvandekar/coupling`) for the estimation of coupling of Freesurfer surfaces as a resource for the neuroimaging community.

4.2. Methods

4.2.1. Subjects

Subjects included 932 youths (504 females) aged 8-22 (mean=14.8; sd=3.6) who completed neuroimaging as part of the PNC. The Institutional Review Boards of Penn and the Children's Hospital of Philadelphia approved all study procedures. All study participants provided informed consent; minors under age 18 provided assent and the parent or guardian provided consent. The sample, screening, and quality assurance procedures were previously detailed (Satterthwaite et al., 2014b; Vandekar et al., 2015).

Briefly, this study considered 1,445 subjects imaged as part of the PNC. Of this sample, 332 subjects met exclusionary criteria due to a history of potential abnormalities in brain development: medical problems that may affect the brain (n=166), inpatient psychiatric hospitalization (n=51), or current use of psychotropic medication (n=165). Two hundred thirty-nine subjects met image quality assurance exclusionary criteria that included an automated and manual screening. Many subjects were excluded due to multiple of the listed criteria, yielding the 932 subjects included in the present analysis and also used in our prior report (Vandekar et al., 2015). Quality assurance screening was based only on the standard Freesurfer output; no additional screening was made on the estimated cortical coupling maps. That is, if the data quality is suitable for the analysis of standard cortical measures (e.g. thickness and sulcal depth), then it should be suitable for analyses of coupling estimated from these measures.

4.2.2. Image acquisition and preprocessing

Image acquisition, preprocessing, and cortical reconstruction steps are as described in Vandekar et al. (2015). A magnetization-prepared, rapid acquisition gradient-echo (MPRAGE) T1-weighted structural image was acquired, using the following parameters: TR, 1810 ms; TE, 3.51 ms; FOV, 180x240 mm; matrix, 192x256; 160 slices; TI, 1100 ms; flip angle, 9; effective voxel resolution, 0.9375x0.9375x1mm; axial acquisition plane (Satterthwaite et al., 2014b). Cortical reconstruction was performed using Freesurfer 5.3.0. Freesurfer processing includes intensity normalization, gray and white matter segmentation, tessellation of the pial and gray/white matter boundaries, and spherical registration to a template (Dale, Fischl, and Sereno, 1999), ultimately producing CT and

SD surface maps (Dale, Fischl, and Sereno, 1999; Fischl, Sereno, and Dale, 1999) which were used to estimate coupling for each subject. Cortical thickness is estimated in Freesurfer as the shortest distance between the estimated gray/white and pial surfaces. SD is estimated by the formula

$$\int n(k)^T \frac{\partial J}{\partial x_k^t} dt,$$

where $n(k)^T$ is the surface normal vector at vertex k, $\partial J/\partial x_k^t$ is the gradient of J with respect to x_k at time t, and J is a cost function for the inflation of the cortical surface that is based on the distance of each vertex from its neighbors. SD measures the distance a given vertex moves outward during the inflation process. Prior to Freesurfer 6 the units of sulcal depth are in arbitrary units. Other proposed methods for investigating SD use different geometry for estimating SD (Yun et al., 2013), or allow for comparison of SD in subject space using automatically labeled neuroanatomical regions of interest (Mangin et al., 2004). The pointwise calculation of SD in Freesurfer allows for estimation of coupling in template space. After estimation of CT and SD, surface maps were registered to the fsaverage5 template using the standard spherical registration procedure in Freesurfer (Fischl, Sereno, and Dale, 1999). The fsaverage5 template was used to decrease the time to estimate coupling and reduce the number of comparisons conducted.

4.2.3. Estimation of coupling maps

Local CT-SD coupling is a subject specific measure that is estimated at each vertex on the cortical surface in template space. The measure is intended to capture one aspect of the multivariate nature of cortical measurements; specifically, coupling describes the localized relationship between CT and SD in a neighborhood of a given vertex.

For CT and SD, the coupling measure at a given vertex, v_0, is defined as the slope parameter estimate of a weighted regression of CT onto SD in a neighborhood, N_{v_0}, of v_0 (Figure 4.1). The weights in the regression are related to the neighboring vertices' distance from the central vertex. Distance is defined as the Euclidean distance between points on the surface. We allow neighbors up to 15 degrees of separation. Weights are all less than or equal to one and rounded to three decimal places. To minimize interpolation through the surface, weights for neighbors of the same order as a neighbor with a weight of zero are all set to zero. The weights, w_j, used in the regression are proportional to the standard normal probability density function $w_j = e^{-1/(2\sigma^2)d(v_0,v_j)^2}$, where σ is a

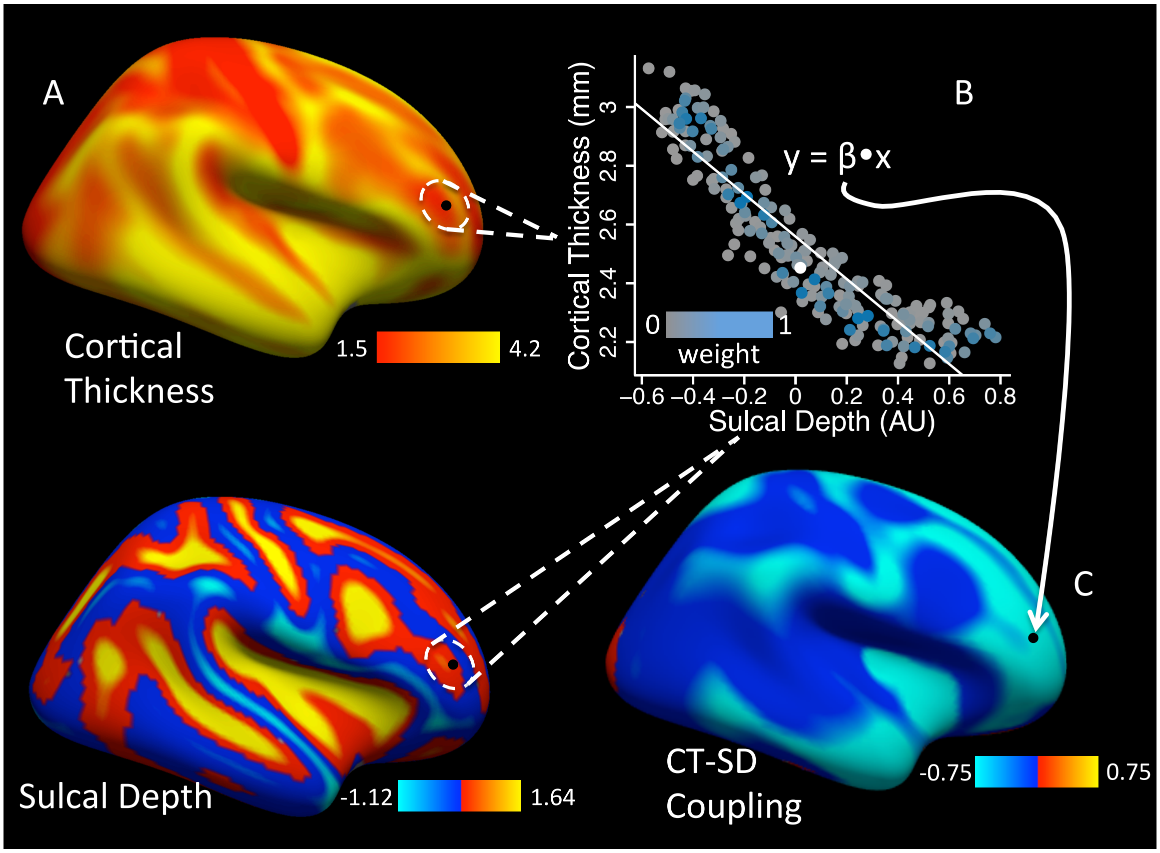

Figure 4.1: Estimation of local coupling using cortical thickness and sulcal depth for FWHM=15. (A) For each subject a neighborhood for each vertex is used to estimate the slope of a weighted regression of thickness onto sulcal depth. (B) Weights in the regression are inversely related to the distance from the central vertex. The central vertex is indicated with a white point. (C) The estimated slope parameter from the regression is assigned to the central vertex. The white point indicates the central vertex, which is shown in black on the surface.

parameter that can be changed to modify the smoothness of the surface by adjusting the weights in the regression. A larger value of σ gives a smoother coupling surface. The units specified by σ are millimeters, measured from the center vertex. The R code provided with the manuscript gives the smoothness parameter in FWHM for consistency with Freesurfer's convention. After estimating weights, a regression in the local neighborhood for each vertex is performed and the estimated slope parameter is the measure of coupling for that vertex.

The units of coupling are in (mm thickness)/(1 unit sulcal depth) where negative values indicate thinner cortex within a sulcus and positive values indicate thicker cortex within a sulcus. For the current analyses coupling was estimated for each subject within a FWHM kernel of 5, 10, and

15mm in fsaverage5 space. After estimation, mean maps were created to summarize the measure. After review of summary statistics, coupling maps with FWHM=15 were analyzed using the group-level statistical analyses discussed below. Linear analyses with FWHM=10 are included as supplementary analyses (Figure C.1). In order to assess the fit of the coupling maps, weighted correlation coupling (WC-coupling) maps were estimated for each subject. By squaring these maps, a vertexwise R^2 map can be estimated that allows for a visual assessment of the quality of coupling estimation. The correlation maps can also be used as an outcome in analyses.

Notably, use of coupling is not restricted to relating CT and SD, or even to analyses on the cortical surface. Coupling could be estimated from registered volumetric brain images of different modalities. At present, the R code provided allows for estimation of coupling between any two FreeSurfer surfaces in template space.

4.2.4. Mean coupling surface maps

Mean coupling surface maps were created by averaging the surfaces of all 932 subjects. Local plots of CT and SD were created by averaging across all subjects and then plotting the local neighborhood of each vertex. To assess the fit of the coupling estimates, R^2 maps were created by squaring the WC-coupling maps. Though R^2 does not explain whether the linear assumption of the coupling model is correct, it does serve as a useful scalar assessment of model fit. Local coupling relationships were investigated visually, (as in Figure 4.3 C & D) using the Desikan-Killiany and Destrieux atlases (Desikan et al., 2006; Destrieux et al., 2010).

4.2.5. Mass-univariate statistical analyses

To demonstrate the biological importance of coupling we investigated linear and nonlinear age effects, as well as sex differences on the coupling at each vertex. Race and intracranial volume (ICV) were included as covariates in all analyses. Additionally, we assess the effect of ICV as a covariate with respect to sex differences. No smoothing was applied to the surfaces as the local weighting in the estimation of coupling maps acts as a smoothing kernel. Separate analyses of CT and SD are presented in the supplement. These maps were smoothed with a kernel of 10mm instead of the 15mm used for coupling. The smaller kernel was used to preserve localized differences in CT in order to demonstrate that the results are sensitive to gyral and sulcal regions. We use a larger kernel for coupling in order to characterize differences across sulcal and gyral boundaries. All statistical maps were false discovery rate thresholded at q=0.01.

Linear models were fit within Freesurfer. Shapiro-Wilk maps were created to assess the normality of the residuals from the linear models for coupling as well as thickness. If the Shapiro-Wilk test is significant it indicates deviance from normality. Flexible nonlinear functions were estimated with thin plate splines using Generalize Additive Models (GAMs) with the mgcv package in R (R Core Team, 2016; Satterthwaite et al., 2014a; Wood, 2011). The penalty parameters for the nonlinear spline terms were fit as random effects and tested using restricted likelihood ratio tests with RLRsim (Scheipl, Greven, and Kchenhoff, 2008). Note that these tests of nonlinearity are for nonlinear effects over and above any linear effects that may be present. In addition to a nonlinear main effect of age we also tested for nonlinear and linear age by sex interactions. This is a single test at each vertex for linear or nonlinear differences in the age trajectory between sexes.

Significant differences in coupling indicate regions where the independent variable (e.g. age) affects the topological relationship between CT and SD. Thus, analyses of coupling allow direct assessment and hypothesis testing of the spatial relationship between the two variables that is not possible by analyzing the two measures separately. However, to understand if variation in one variable is driving the coupling results, it is necessary to investigate the relationship between CT and SD. This can be done by investigating CT and SD surface maps separately to identify which variable has significant associations in the region where coupling results are observed.

Alternatively, we can understand the relationship between the two measures better by stratifying each cluster of significant coupling with respect to one of the imaging measures and plotting the relationship of the other with the independent variable. We do this for the nonlinear age results and the sex differences: to explore nonlinear age results, in each cluster we averaged CT over regions where SD was less than and greater than the median SD separately. We then fit age trajectories to the mean CT values to demonstrate how CT trajectories differ in regions of high and low SD (Figure 4.5). To understand the sex differences we also averaged CT over regions where SD was less than and greater than the median of SD. Then we create scatterplots of CT for these subregions of each cluster (Figure 4.7).

4.3. Results

4.3.1. Cortical coupling mean characteristics

Mean maps were created to summarize the spatial distribution of local coupling between CT and SD. Coupling maps showed the expected global negative relationship between CT and SD, indicating that over most of the cortex the cortical sheet is thinner in sulcal compared to gyral locations (Figure 4.2 A, B, & C). However, there was also marked spatial heterogeneity, suggesting that the method detects effects that are localized to specific regions compared to the whole brain spatial correlations we previously documented (Vandekar et al., 2015). The strongest negative coupling was observed on gyral locations throughout parietal, temporal, and frontal cortex.

Notably, in contrast to the negative coupling seen in most regions, clusters within inferior temporal, primary visual, and somatomotor cortex (Figure 4.2 A) show positive correlations between CT and SD for some kernels. Detailed examination of primary visual and somatomotor clusters revealed that this is due to complex nonlinear relationships in local neighborhoods that have unique topological characteristics, where regions of different thickness coalesce (Figure 4.3). This effect is clearly seen in the central sulcus, where precentral cortex is substantially thicker than postcentral cortex at a similar SD (Figure 4.3 B). The sign of the relationship changes in the central sulcus based on the size of the kernel used because in a localized neighborhood in the sulcus (FWHM=5) there is a positive association, however at a broader scale (FWHM=15) there is a negative association. Similarly, for visual cortex there is a gradient in thickness in the regions around the calcarine fissure, with parieto-occipital cortex being thicker than calcarine cortex, which in turn is thicker than cuneal cortex. The nonlinear relationship in these regions represents a violation in the linearity assumption of coupling. This is indicated by a relative reduction in R^2 in this region (Figure C.2). Significant results in regions with a low R2 must be interpreted carefully with scatterplots of the data. Though the larger kernels lose some information about the complexity of the relationship, they provide information about the general trend that occurs in larger neighborhoods that encompass gyral and sulcal regions. Because differences across gyral/sulcal boundaries are of interest, we selected a kernel of 15mm for analyses.

Shapiro-Wilk tests of the residuals from the group level linear models indicated significant deviation from normality in much of the cortex for coupling and thickness (Figure C.3). These results suggest

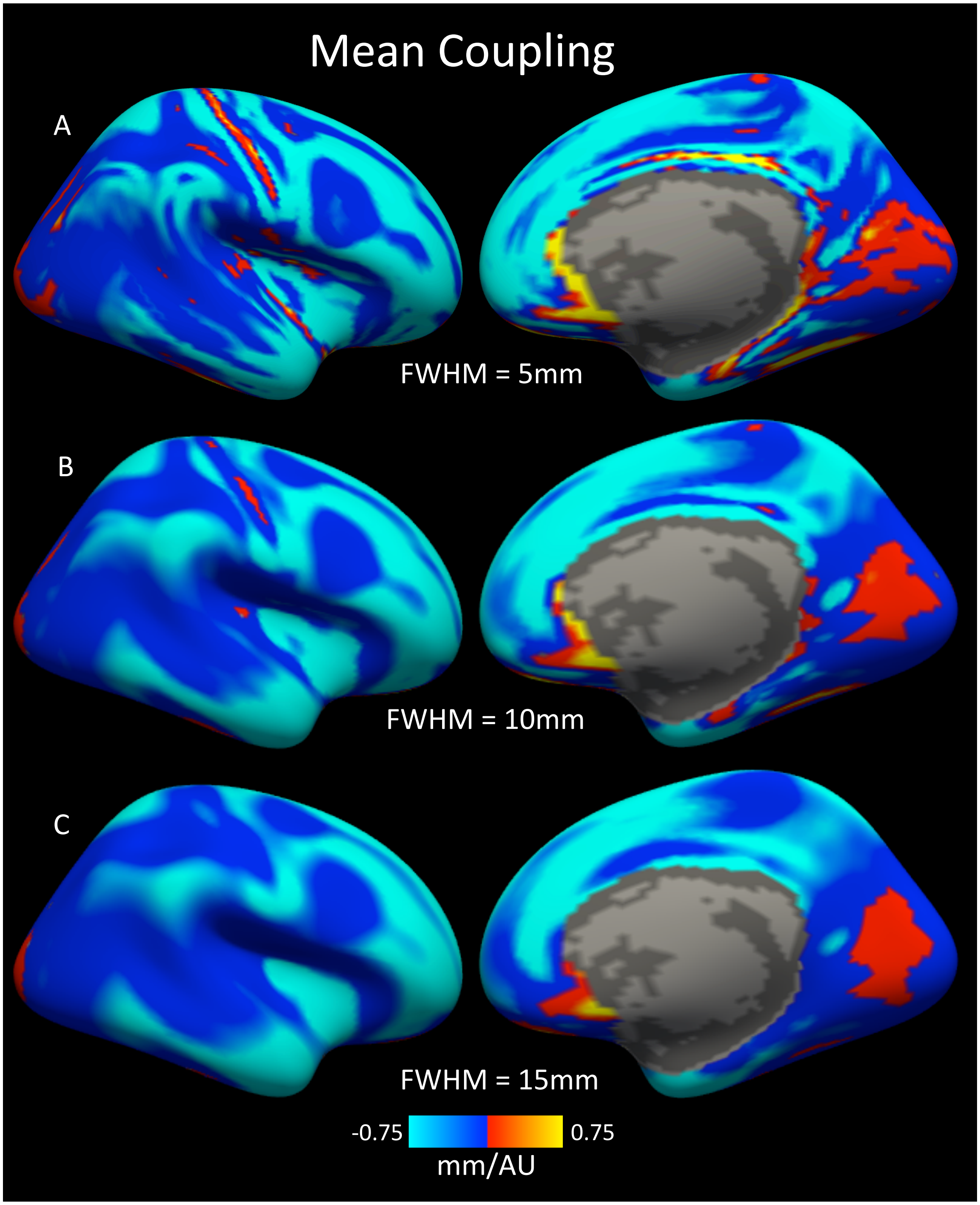

Figure 4.2: Mean maps of coupling for healthy adolescents show substantial spatial heterogeneity with FWHM of 5, 10, and 15.

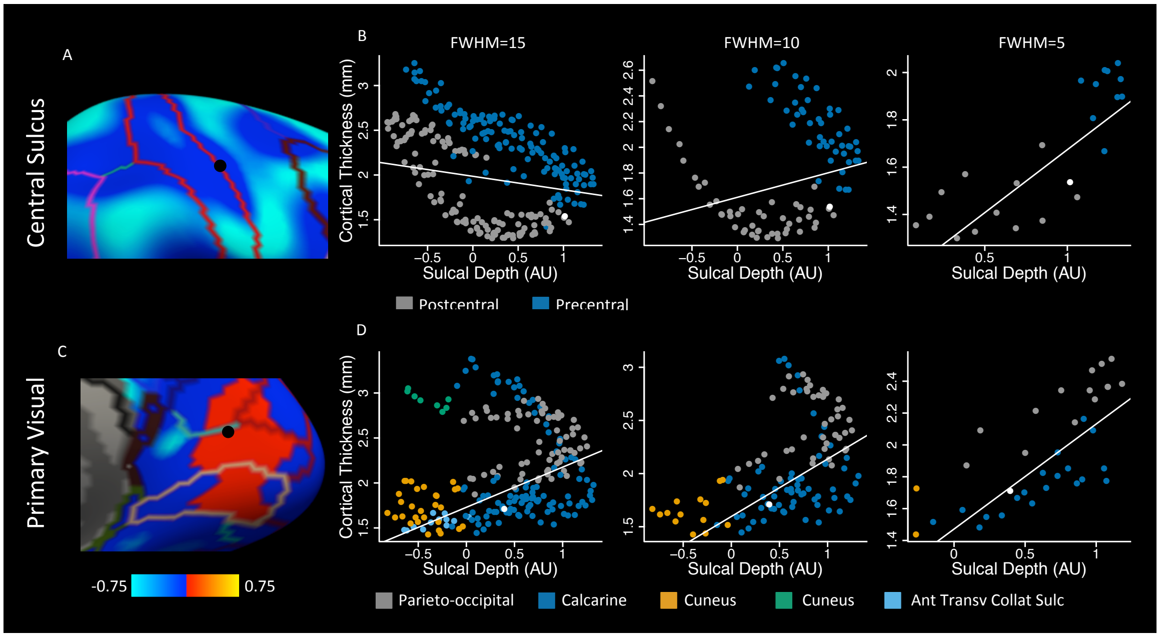

Figure 4.3: An illustration of coupling over kernels of FWHM=15, 10, and 5mm in (A) the central sulcus, where coupling changes between kernels and (C) primary visual cortex, where coupling is positive. Vertices plotted were selected based on their location in a junction of anatomical regions. Plotting the local neighborhood for a vertex in the central sulcus shows the juxtaposition of cortex with very different characteristics: the postcentral gyrus is much thinner than the precentral gyrus. (B) The larger kernels capture the more general trend, however within a small neighborhood within the sulcus coupling is positive. (D) The local neighborhood of a vertex in primary visual cortex shows that positive coupling is the result of the vertex lying on the border of three anatomically distinct regions. Colors in the scatter plots indicate regions from the (A) Desikan-Killiany and (B) Destrieux atlases. White lines are the fit of a weighted regression on average cortical thickness and sulcal depth maps. The white point indicates the central vertex in the scatterplots, which is shown in black on the surfaces in (A) and (C).

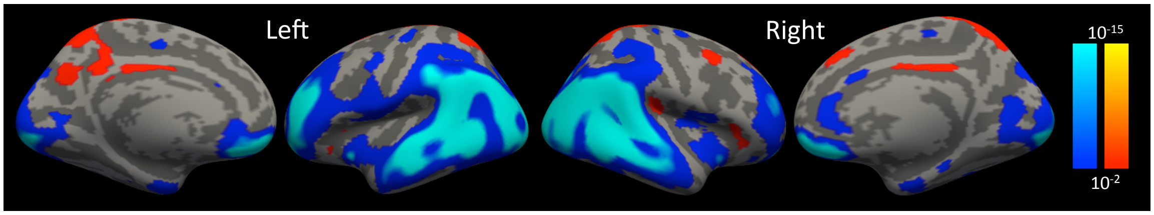

Figure 4.4: Coupling shows linear age-related decreases in occipital, parietal, temporal, and anterior frontal lobes. Increases are observed in left cuneus and bilateral anterior parietal cortex. Statistical maps are FDR thresholded at $q = 0.01$. Color bars show signed p-values where blue indicates decreases with age and red indicates increases with age.

the assumptions of the regression model used are violated. However, linear regression estimators are known to be asymptotically normal by the central limit theorem (Boos and Stefanski, 2013). This justifies the use of the normal distribution for inference in large samples. The default in Freesurfer is to use the t-distribution, which is asymptotically equivalent to the normal distribution as the degrees of freedom go to infinity, so inference in large samples using this distribution will be approximately equivalent to inference using the normal distribution.

4.3.2. Local cortical coupling evolves markedly with age in youth

Having established that local cortical coupling is spatially heterogeneous, we next examined how coupling evolves in development. Robust linear declines in coupling with age were observed in bilateral middle and superior temporal gyrus, parietal cortex, occipital cortex, and frontopolar cortex (Figure 4.4). The changes in coupling are due primarily to thinning in the surrounding sulcal regions (Figure C.4 A) rather than changes in SD itself (Figure C.4 B). While there is significant linear thinning throughout the cortex, coupling is sensitive to regions where thinning occurs differentially between gyral and sulcal regions, highlighting bivariate patterns of change. In addition to age-related declines, we observed age-related increases in the coupling coefficient (a negative slope becoming more positive) in the left precuneus. This region also exhibited decreasing thickness overall, but the positive change in coupling reflects relatively greater age associated thinning in gyral cortex.

4.3.3. Nonlinear developmental effects in local cortical coupling

In order to assess whether regions exhibited nonlinear developmental differences over and above linear effects, cortical coupling was evaluated using penalized splines with generalized additive models (GAMs). Results demonstrated significant nonlinearities in the temporal lobe where linear age related differences were also observed (Figure 4.5 A). Plotted mean trajectories within significant clusters indicate that nonlinear effects are characterized by a decline in coupling (becoming more negative) until age 14, at which point coupling stabilizes (Figure 4.5 B, C, & D). To investigate the effects driving coupling differences we created mean gyral and sulcal CT trajectories by stratifying CT by the median SD in each cluster. Examination of gyral and sulcal CT changes in these regions revealed that this effect is driven principally by nonlinear age related differences in CT, with effects in CT following an inverse trajectory to coupling in gyral regions until age 14 (Figure 4.5 B, C, & D). After age 14 coupling is stable because cortical thinning in gyral regions occurs in tandem with the surrounding sulcal regions (Figure 4.5 B, C, & D). These results suggest that nonlinear age related differences in coupling are principally driven by nonlinear thickening in gyral cortex prior to age 14. These results again indicate that the effects in CT occur differentially in sulcal and gyral locations, and demonstrate coupling's sensitivity to nonlinear, bivariate patterns of age related effects. Plots of SD by stratified by CT were stable with respect to age and are not shown. The test for a nonlinear age by sex interaction in coupling yielded no significant results.

4.3.4. Evidence for greater coupling in females than males

As a final step, in order to further examine local coupling's sensitivity to studies of group and individual differences, a linear model was used to assess sex differences while controlling for age, ICV, and race. Sex differences were found in bilateral inferior parietal cortex and frontal lobe. Females have stronger negative CT-SD coupling than males in this region (Figure 4.6). By stratifying CT by the median SD within significant clusters we see that these coupling results are driven by thicker cortex in females than males in gyral locations (Figure 4.7). Vertexwise CT and SD analyses corroborate this finding (Figure C.5).

Because sex is associated with ICV we investigated the effect of ICV and the results after removing it from the model. With sex and age included in the model, the effect of ICV on coupling was spatially conservative with significant associations in the frontal pole, precuneus, and anterior cingulate (Figure C.6 A). However, sex differences were more robust in the frontal and temporal lobes when ICV was not included as a covariate (Figure C.6 B), suggesting it may be an important covariate to include in analyses of coupling.

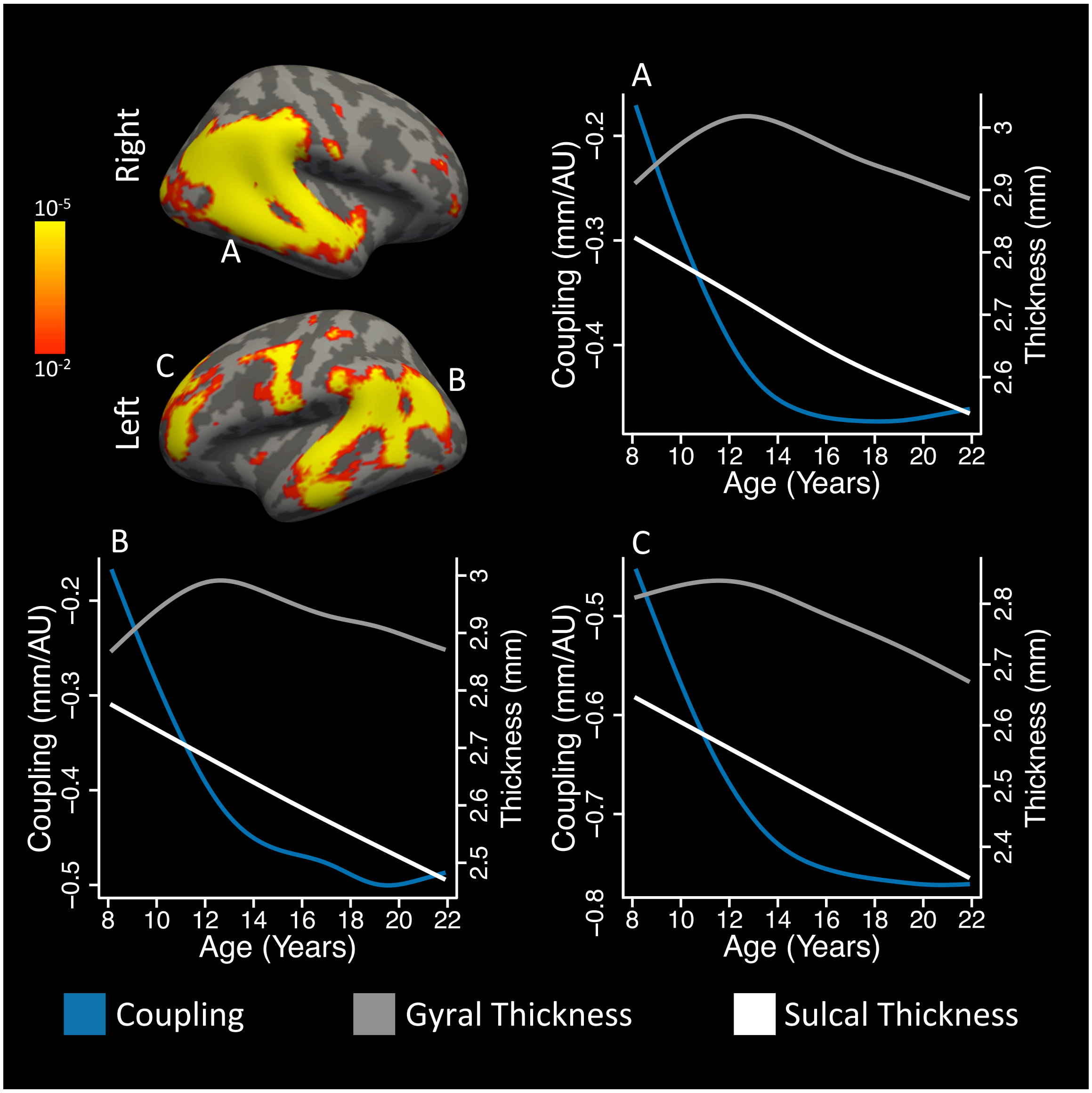

Figure 4.5: Nonlinear age-related differences in coupling are driven by nonlinear differences in cortical thickness that are spatially related to sulcal depth. Nonlinear development-related effects were observed bilaterally in temporal and parietal lobes and left frontal pole. (A, B, C) Mean trajectories of coupling (blue curve) in significant regions were characterized by a smooth decline until age 14, when coupling stabilized. In the same regions, age-related increases in thickness occur in regions of gyral cortex until age 14 (gray curve), when thinning begins to occur at a rate similar to surrounding sulcal regions (white curve). Trajectories are fit using generalized additive models. Gyral and sulcal cortex are defined as above and below the median in each region. Statistical maps are FDR thresholded at $q = 0.01$. Color bars show uncorrected p-values.

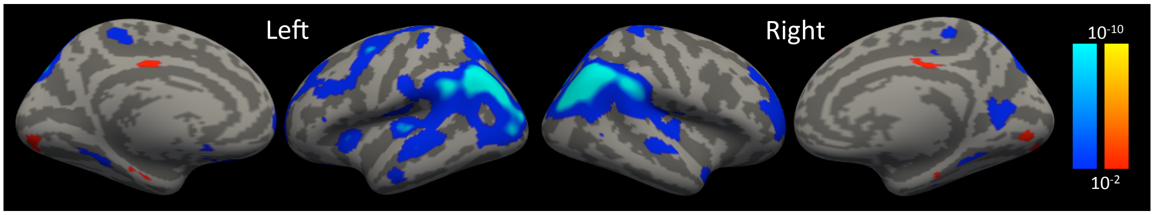

Figure 4.6: Females show stronger coupling between cortical thickness and sulcal depth. This effect is present in bilateral occipital and parietal cortex, and indicates a stronger negative correlation between thickness and sulcal depth in this region. Statistical maps are FDR thresholded at $q = 0.01$. Color bars show signed p-values where blue represents regions that are greater in males.

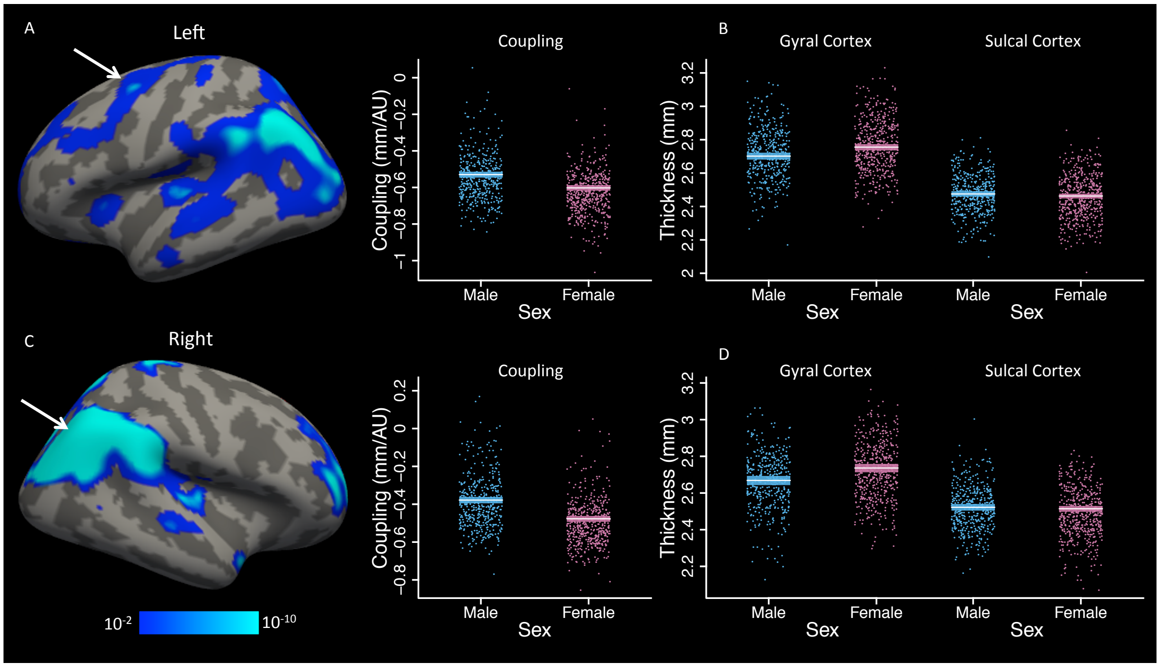

Figure 4.7: Sex differences in coupling are driven by thicker gyral cortex in females. Mean cortical thickness data is plotted for gyral and sulcal cortex separately in clusters of significant sex differences in coupling. (A) Females show lower coupling in left frontal lobe. (B) Within this cluster gyral cortex is thicker for females than males, while sulcal cortex is similar between the two sexes. (C) Similarly, in right parieto-occipital junction females have stronger negative coupling due to (D) thicker gyral cortex than males in this region. Gyral and sulcal cortex are defined as above and below the median in each region. Color bars show signed p-values where blue represents more negative coupling in females than males.

4.3.5. Discussion

Neuroanatomists discovered almost 100 years ago that CT varies substantially with cortical topology (Economo, 1925), wherein gyral cortex is thicker than sulcal cortex. While it was previously thought that this relationship was fully established in the pre- or perinatal period, we recently demonstrated that coupling continues to evolve during adolescence (Vandekar et al., 2015). Because many major neuropsychiatric disorders are linked to abnormalities of cortical development and folding (Insel, 2010; Krain and Castellanos, 2006; Rapoport, Giedd, and Gogtay, 2012; Shaw et al., 2012; Steen et al., 2006; White et al., 2003), quantitative measurement of this relationship between CT and SD may be useful for studies of developmental psychopathology.

Here we introduced a new measure of local cortical coupling to describe the relationship between CT and SD, which was historically examined manually in post-mortem brains by early investigators. Local cortical coupling is a subject-level, surface-based measure that can be calculated using the R code made publically available (`https://bitbucket.org/simonvandekar/coupling`), and can be readily applied to studies of development or individual difference using standard analysis tools. We demonstrate that local cortical coupling is spatially heterogeneous and evolves with age. Importantly, age-related differences in coupling demonstrate that local cortical topology is not fixed, but continues to develop in specific regions through adolescence. Moreover, significant sex differences in coupling are present, with females exhibiting more robust coupling than males in inferior parietal and posterior temporal cortex.

4.3.6. Local regression provides subject-level estimate of cortical coupling

Given the highly complex nature of the human cortex, there is increasing interest in describing how different cortical measures inter-relate. Local cortical coupling is an approach that can be used to investigate the localized bivariate relationship between two cortical measures. Previously, using PCA and spatial permutation tests, we showed that age-related thinning was concentrated primarily in sulcal cortex and that age-related thickening occurred in surrounding gyral cortex (Vandekar et al., 2015). Similar analyses have considered the topological relationship between measures on a regional basis (Alemn-Gmez et al., 2013; Klein et al., 2014). However, in all previous studies, these associations were identified across the entire cortex or averaged across large regions. We developed coupling as a vertexwise measure that describes localized relationships between cortical measures. Notably, CT-SD coupling is spatially heterogeneous, providing a fine level of resolution that could not be captured in prior studies.

Calculation of local cortical coupling produces a subject-level measure that allows it to be easily integrated into standard analysis tools as an outcome variable or high dimensional predictor. Here, coupling was used to describe the relationship between CT and SD. However, it should be noted that any surface based map in Freesurfer could be used; indeed, coupling could also be estimated in registered volumetric images. Coupling therefore may represent a valuable new subject-level imaging phenotype for studies of development that is sensitive to individual differences, and potentially to the presence of neuropsychiatric disorders.

4.3.7. Coupling identifies topologically heterogeneous regions

As documented by early anatomists, CT-SD coupling is negative across most of the cortical surface. However, our local coupling measure provides novel evidence of spatial heterogeneity in this relationship. Notably, coupling is most robust in gyral heteromodal association cortex and large sulci such as the insula. Local variations in coupling delineate developmentally relevant regions that are not well differentiated by cortical measures evaluated on their own.

In contrast to the negative coupling seen throughout the cortex, clusters of positive coupling for larger kernels are quite sparse. Positive coupling is not driven by an inversion of the normal relationship between CT and SD, but rather is seen at junctions of anatomic regions where there are systematic differences in thickness between gyri on opposite sides of a sulcus (Economo, 1925; Fischl and Dale, 2000). Larger kernels detect more general changes in the CT-SD relationship at the expense of having complex nonlinear relationships go undetected. Low R^2 in these regions is indicative of a failure of coupling to capture these complex relationships. This relationship is particularly evident in visual cortex and central sulcus. Significant coupling results in these regions should be interpreted carefully.

4.3.8. Age and sex differences in coupling

Human adolescent development is characterized by a decrease in gray matter volume and an increase in white matter volume (Huttenlocher and Dabholkar, 1997; Levitt, 2003; Matsuzawa et al., 2001). However, such remodeling has substantial local variation that occurs at multiple spatial scales. The results shown here and in our prior report (Vandekar et al., 2015) indicate that cortical remodeling occurs at a fine scale that is strongly related to cortical topology. Within-subject measurement of local cortical coupling allows us to evaluate such developmental effects in much greater detail. Coupling changes robustly during youth, and follows both linear and non-linear developmental patterns. Such age related differences in coupling are driven by cortical thinning that occurs more robustly in sulci, as well as in gyral cortex that shows nonlinear increases in CT. Our findings therefore emphasize that changes in gray matter development are topologically sensitive to SD and also that continued development of CT-SD coupling occurs in specific regions during youth, contributing to the well-established relationship between CT and SD observed in adults.

In addition to such developmental changes, we found evidence for sex differences in cortical coupling. Specifically, females had stronger negative coupling in the temporal and parietal lobes, which was driven by greater gyral CT in females. As substantial developmental strengthening in coupling occurs in these regions, sex differences suggest that coupling may advance more rapidly in females from an early age. Greater CT in females in parietal and posterior temporal lobes has been reported across various age ranges (Im et al., 2006; Luders et al., 2006; Sowell et al., 2007). The results reported here thus accord with prior literature, and also demonstrate that sex differences in CT are topologically related to SD, producing stronger coupling in females. These results further underline the utility of local cortical coupling as a useful subject-level measure for studies of individual differences and, potentially, developmental psychopathology.

4.3.9. Limitations

As noted above, nonlinear spatial relationships of CT and SD are not detected with this methodology. Instead coupling will take a positive or negative value based on the general direction of the relationship between the two variables. This has the advantage of being easily interpretable at the cost of missing complex nonlinear patterns that may be of interest.

Because coupling maps are slope estimates they represent the anatomical strength of the relationship, but do not capture the statistical strength or give any representation of how much noise is in the estimate of the slope. Two vertices with similar values of coupling can have different correlations. If the statistical strength of the relationship is of interest then WC-coupling maps can be analyzed instead. Due to the computational burden of estimating geodesic distance across the folded pial surface, Euclidean distance on the inflated surface was used to estimate neighborhoods and weights in the regression. Using Euclidean distance on the inflated cortical sheet serves only as an approximation to the true anatomical distances measured across the surface. Because of this approximation the number of points used in each local regression is variable depending on the folding of the cortex at that location. Euclidean distances on the inflated surface are shorter than the geodesic distances in places where the cortex is highly folded, so estimates of coupling in these regions are likely smoother than in other regions that are less folded.

4.3.10. Future Directions

Examining the bivariate relationship of brain phenotypes through local cortical coupling may have several potentially fruitful applications moving forward. First, several studies have reported correlations of MRI derived measures, including folding and thickness, with cytoarchitecture (Fischl, 2013; Fischl et al., 2008; Wagstyl, 2015). These prior reports suggest coupling may be a valuable tool to combine measures to predict cytoarchitecture using MRI. Second, developmental and neuropsychiatric differences in thickness (Raznahan et al., 2011; Shaw et al., 2008; Sowell et al., 2007) and gyrification (Cachia et al., 2008; Csernansky et al., 2008; Klein et al., 2014; Shaw et al., 2012) suggest the promising potential of coupling to identify regions where these measures interact. For example, thickness differences that are spatially related to SD have been reported in development and neuropsychiatric disorders (Goghari et al., 2007; Selemon, Rajkowska, and Goldman-Rakic, 1998; Vandekar et al., 2015). Goghari et al. (2007) identified regions of sulcal cortex in the temporal lobe where patients with schizophrenia showed significant reductions in thickness compared to controls. Their findings suggest that coupling should be more negative in the temporal lobe in schizophrenia. Reductions in the CT of sulci in schizophrenia have been attributed to thinning of Layer 2, which is relatively thicker in sulci than gyri (Selemon, Rajkowska, and Goldman-Rakic, 1998; Wagstyl, 2015). Coupling may be a useful tool to easily identify regions where such complex relationships occur. Third and finally, spatial relationships with other cortical characteristics, such as underlying white matter (Vandekar et al., 2015), can also be investigated using the coupling approach. More generally, coupling is not restricted to surface based analyses and can be used to describe the relationship of volumetric images of different modalities.

4.3.11. Conclusions

We introduced a novel method for measurement of local cortical coupling, which captures the bivariate relationship between CT and SD. This measure is spatially heterogeneous, evolves conspicuously in youth, and is different between males and females. Coupling is not restricted to surface based analyses and can be used to describe the relationship of volumetric images of different modalities. Such measures can facilitate investigation of local individual differences in cortical topology, and may offer a valuable brain phenotype for identification of abnormal brain development related to neuropsychiatric illness.

CHAPTER 5

Discussion

5.1. The Projected Score Test

In Chapter 2 proposed the PST, a modification of the score test for high dimensional data that works by projecting the scores to a lower dimensional linear subspace. The procedure offers a novel *post hoc* inference on the projected scores by performing inference in the subspace where the test statistic was estimated. Because the posthoc inference is based on the same model and degrees of freedom as the PST statistic, the interpretation of high-dimensional results agree closely with the results from the PST.

The ability to choose a subspace $\mathbb{L}$ makes the procedure very flexible. For example, in medical imaging the basis for the space can be chosen based on anatomical or functional labels, or from data acquired in another imaging modality. Particular hypotheses can be targeted by selecting a basis that includes indicators of certain regions or weights particular locations to target specific spatial patterns. If orthogonal indicator vectors are used as the basis, then the approach can be seen as testing averages of subregions of the data as in Section 2.5. In this case, the PST procedure can be seen as a maxT multiple testing procedure of the regional averages that accounts for the correlation structure of the tests.

There are several limitations of the proposed procedure. First, the success of the procedure depends critically on the projection chosen. If a projection is chosen that is orthogonal to the mean vector, the PST will fail to capture any signal in the data. This is a limitation of any dimension-reducing procedure. Further research could investigate whether maximization of the score test with regularization can yield a test statistic whose distribution is tractable. Regularization may remove the subjectivity of selecting a basis and make the procedure more robust. Second, while the dimension reduction procedure preserves power and the results align closely with those from previous research, the inference does not guarantee control of the FWER or FDR of the original score vector. Future research will investigate how inference of the original score vector can be made by thresholding the projected score vector. This is similar in concept to the dependence-adjusted procedure discussed by Fan, Han, and Gu (2012) for controlling the FDP and may offer increased power by leveraging the covariance of the test statistics. These limitations notwithstanding, our procedure generalizes Rao's score test to the high- and infinite- dimensional settings and introduces a new inference approach based on projecting the test statistics to a lower-dimensional space where inference can be made on fewer degrees of freedom.

5.2. The Parametric Bootstrap Joint Testing Procedure

In Chapter 3 we introduced a fast parametric bootstrap joint testing procedure as a new tool for multiple comparisons in neuroimaging. The PBJ procedure improves computing time by generating the test statistics directly instead of permuting the original data. If normality assumptions about the data generating distribution do not hold, then the Yeo-Johnson transformation can be used to obtain statistics that are approximately normal to improve the finite sample performance of the procedure.

In the CBF data analysis the PBJ is more powerful than the PJ MTP because the PJ MTP does not account for the fact that the finite sample distribution of the test statistics can be different. Differences in the finite sample distribution of the statistics are attributable to certain regions near the edge of the brain having larger variance and skew. For this reason taking the maximum across locations leads to conservative inference in locations that actually have tighter tails. While the PBJ generates from a chi-squared distribution this ensures that a few heavy-tailed locations do not affect the distribution of the maximum.

In simulations, the step-down procedures provide little improvement in power over the single-step procedures. However, in the regionwise analyses Holm rejected 15 more regions than the Bonferroni procedure. The reason for the difference is that step-down procedures offer little benefit when there is a small number of false null hypotheses and a large number of tests.

Cluster correction is used in 70% of studies to perform hypothesis testing in neuroimaging studies. Future work will use the PBJ to perform cluster and assess performance in small samples. In addition we will use robust test statistics and covariance estimates in an estimating equations to make the method robust to model variance misspecification.

5.3. Coupling

In Chapter 4 we introduced a new measure of local cortical coupling to describe the relationship between CT and SD, which was historically examined manually in post-mortem brains by early investigators. Local cortical coupling is a subject-level, surface-based measure that can be calculated using the R code made publically available (`https://bitbucket.org/simonvandekar/coupling`), and can be readily applied to studies of development or individual difference using standard analysis tools. We demonstrate that local cortical coupling is spatially heterogeneous and evolves with age. Importantly, age-related differences in coupling demonstrate that local cortical topology is not fixed, but continues to develop in specific regions through adolescence. Moreover, significant sex differences in coupling are present, with females exhibiting more robust coupling than males in inferior parietal and posterior temporal cortex.

APPENDIX A

THEORETICAL RESULTS FOR CHAPTER 2

A.1. Theoretical framework

We assume the observed imaging data are finite-dimensional representations that are generated from an underlying unknown function. To be more specific, we assume there are 3 components to each observation: the random outcome $\mathrm{Y}_i \in \mathbb{R}$, a set of covariates $X_i \in \mathbb{R}^k$ and a function $G_i \in \mathcal{L}^2(\mathbb{V})$, where $\mathbb{V}$ is a nonempty compact subset of $\mathbb{R}^3$ and $\mathcal{L}^2(\mathbb{V})$ is the space of square integrable functions from $\mathbb{V}$ to $\mathbb{R}$. $\mathbb{V}$ represents the space on which data can be observed; in neuroimaging this space is the volume of the brain. Our goal is to use the conditional distribution of Y_i given X_i and G_i to test the association between G_i and Y_i. We denote the collection of independent and identically distributed data by $\mathrm{Y} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_n)$, $X = (X_1, \ldots, X_n)$, and $G = (G_1, \ldots, G_n)$. In practice G_i are unobservable and we only observe discretized functions at a finite number of locations that are voxels in the image.

We define a parameter space $\Theta = \mathrm{A} \times \mathrm{B}$ that includes a finite-dimensional parameter $\alpha \in \mathrm{A} \subset \mathbb{R}^m$ and an infinite-dimensional parameter on $\beta \in \mathrm{B}$. Together these parameters describe the conditional distribution of the imaging and nonimaging data. Throughout, we assume $\mathrm{B} = \mathcal{L}^2(\mathbb{V})$, so that the infinite-dimensional parameter and infinite-dimensional data are defined on the same space, but this assumption is not required. The distribution of the observed data is defined by a p-dimensional discretization of the infinite-dimensional parameter. We prove that under regularity assumptions, as $n, p \to \infty$ the test statistic for the discretized data approaches the statistic for the infinite-dimensional parameter.

To relate the unobserved data Y to the parameters, we further assume that the distribution function of Y_i conditional on X_i and G_i is in a family of probability models, $\{F_\theta : \theta \in \Theta\}$, indexed by the parameter

$$\theta = (\alpha, \beta). \tag{A.1}$$

That is, there exists an interior point $\theta_0 \in \Theta$ such that for all sets $\rho \in \mathcal{R}$

$$\mathbb{P}(\mathrm{Y}_i \in \rho \mid X_i, G_i) = F_{\theta_0}(\rho).$$

We define the density function f_θ as the Radon-Nikodym derivative of F_θ with respect to the Lebesgue measure,

$$\mathbb{P}\left(\mathrm{Y}_i \in \rho\right) = \int_\rho \frac{dF_\theta}{d\nu} d\nu = \int_\rho f_\theta d\nu.$$

Let $\ell(\theta; \mathrm{Y}) = n^{-1} \sum_{i=1}^n \log f_\theta(\mathrm{Y}_i)$ be the log-likelihood function for θ.

In order to define the finite-dimensional data and parameter space we must partition $\mathbb{V}$ into finitely many sets and define the observable imaging data as a realization from the partitioned space. For any integer p, the space $\mathbb{V}$ can be partitioned into p nonempty sets. Let $\mathcal{V}_p = \{\mathbb{V}_{1p}, \ldots, \mathbb{V}_{pp}\}$ be such a partition, i.e. $\bigcup_{j=1}^p \mathbb{V}_{jp} = \mathbb{V}$ and $\mathbb{V}_{jp} \cap \mathbb{V}_{kp} = \varnothing$, for $j \neq k$. For $j = 1, \ldots, p$, let v_j be an arbitrary interior point of $\mathbb{V}_{jp} \in \mathcal{V}_p$. As $\mathcal{V}_p$ is a partition, each v is in only one $\mathbb{V}_{jp}$. Let the ith discretized image observation be $G_{ip} = (G_i(v_1), \ldots G_i(v_p)) \in \mathbb{R}^p$. The conditional distribution of Y_i given X_i and G_{ip} is determined by the finite parameter $\theta_p = (\alpha, \beta(v_1), \ldots, \beta(v_p)) \in \mathbb{R}^{m+p}$. In order to define a finite-dimensional likelihood from the likelihood for the infinite-dimensional parameters we define the function $\beta_p \in \mathcal{L}^2(\mathbb{V})$ by

$$\beta_p(v) = \beta(v_j) \text{ for all } v \in \mathbb{V}_{jp}$$

and the function

$$G_{ip}(v) = G_i(v_j) \text{ for all } v \in \mathbb{V}_{jp}.$$

This allows us to define the log-likelihood from the function

$$\ell(\theta_p; \mathrm{Y}_p) = n^{-1} \sum_{i=1}^n \log f_{\theta_p}(\mathrm{Y}_{ip}), \tag{A.2}$$

where $f_{\theta_p}(y) = f(y; (\alpha, \beta_p), X_i, G_{ip})$ is the finite-dimensional density.

Following Vaart (2000), we define the Fréchet derivative with respect to β.

Definition A.1.1. *For B as defined in Section 2.2.1, a function $f : \mathrm{B} \to \mathbb{R}$ is called Fréchet differentiable at β if there exists a bounded linear map $L_\beta : \mathrm{B} \to \mathbb{R}$ such that*

$$\lim_{\|h\| \to 0} \frac{\|f(\theta + h) - f(\theta) - L_\beta h\|}{\|h\|},$$

for $h \in \mathrm{B}$.

Assuming ℓ is Fréchet differentiable with respect to β, we define the scores

$$\mathrm{U}_n = \mathrm{U}_n(v) = \frac{\partial \ell}{\partial \beta}\{(\alpha, \beta(v)); \mathrm{Y}(v)\} \tag{A.3}$$

$$\mathrm{U}_{np} = \mathrm{U}_{np}(v) = \frac{\partial \ell}{\partial \beta_p}\{(\alpha, \beta_p(v)); \mathrm{Y}_p(v)\}. \tag{A.4}$$

Let

$$\mathrm{S}_n = \mathrm{U}_n(\cdot; (\hat{\alpha}, \beta_0)) \in \mathcal{L}^2(\mathbb{V})$$

$$\mathrm{S}_{np} = \mathrm{U}_{np}(\hat{\alpha}, \beta_{p0}) \in \mathbb{R}^p$$

where β_0 denotes the value of the parameter under the null $H_0 : \beta = \beta_0$ and $\hat{\alpha}$ is the maximum likelihood estimator for α under the null.

A.2. Conditions for Theorem 2.2.2

The conclusion of Theorem 2.2.2 requires the asymptotic normality of the scores, which holds under the following conditions (Boos and Stefanski, 2013):

1. Identifiability of the parameters.
2. The support of $f(y \mid \theta)$ does not depend on θ
3. The true parameter lies on the interior of the parameter space.
4. $\left|(\partial^3/\partial\theta_j\partial\theta_k\partial\theta_\ell)\log f(y \mid \theta)\right| \leq g(y)$, with $\mathbb{E}g(y) < \infty$
5. $\mathbb{E}\hat{\Omega}_F = \Omega_F(\theta_0)$, for $\hat{\Omega}_F$ in (2.7), and Ω_F is nonsingular and continuous with respect to θ.

A.3. Proof of Theorem 2.2.2

Let $\phi = Q^T\zeta$. Then the PST statistic is

$$\begin{aligned}
\mathrm{R}^{\mathbb{L}} &= \max_{\zeta\in\mathbb{R}^p\setminus\{0\}} n\frac{(\zeta^T P\mathrm{S}_{np})^2}{\zeta^T P\hat{\Omega}P\zeta} \\
&= \max_{\phi\in\mathbb{R}^r\setminus\{0\}} n\frac{(\phi^T Q^T\mathrm{S}_{np})^2}{\phi^T\hat{\mathrm{V}}_{np}\phi} \\
&= n\mathrm{S}_{np}^T Q\hat{\mathrm{V}}_{np}^{-1}Q^T\mathrm{S}_{np}, \qquad \text{(A.5)}
\end{aligned}$$

where the last line follows from a standard maximization lemma (Johnson and Wichern, 2007 p. 80).

Equation (2.9) holds by the multivariate central limit theorem and the variance estimate of $n^{1/2}Q^T\mathrm{S}_{np}$ is

$$\begin{aligned}
\hat{\mathrm{V}}_{np}(\theta_0) &= Q^T\hat{\Omega}(\theta_0)Q \\
&= (Q^T\hat{\Omega}_\beta Q - Q^T\hat{\Omega}_{\beta\alpha}\hat{\Omega}_\alpha^{-1}\hat{\Omega}_{\alpha\beta}Q),
\end{aligned}$$

which converges to $V(\theta_0)$ by the continuous mapping theorem because $\hat{\Omega}_F \to_P \Omega_F$. Thus $n\mathrm{S}_{np}^T Q\widehat{\mathrm{V}_{np}}^{-1}Q^T\mathrm{S}_{np} \to_L \chi_r^2$ by the continuous mapping theorem.

Remark 1. The conclusion of Theorem 2.2.2 implies that expression (2.10) does not depend on the choice of Q. This fact can also be shown directly, as follows. Consider another matrix Q_* with orthonormal columns such that $P = Q_*Q_*^T$, and accordingly define $\hat{V}_* = Q_*^T\hat{\Omega}(\theta_0)Q_*$. Then $Q_* = QM$ where $M = Q^TQ_*$. Since $P = QMQ_*^T$ is of rank r, M is of rank r and hence invertible, so

$$Q_*\hat{V}_*^{-1}Q_*^T = QM(M^T\hat{V}M)^{-1}M^TQ^T = Q\hat{V}^{-1}Q^T,$$

and thus formula (2.10) for the PST statistic is unchanged by substituting $Q_*, \hat{V}_*$ for $Q, \hat{V}$.

A.4. Proof of Theorem 2.2.4

In Section 2.2.2 we assumed $n^{1/2}\mathrm{S}_n \to_P \mathrm{S}$ where S is a mean zero Gaussian process. The following theorem (Vaart, 2000 Thm 18.14) gives conditions that guarantee the convergence of S_n.

Theorem A.4.1. *The sequence of elements $\sqrt{n}\mathrm{S}_n$ converges weakly to a tight random element if and only if* *a. The sequence $n^{1/2}(\mathrm{S}_n(v_1), \ldots, \mathrm{S}_n(v_p))$ converges in distribution in $\mathbb{R}^p$ for every finite set of points $v_1, \ldots, v_p \in \mathbb{V}$.*

b. For every $\epsilon, \eta > 0$ there exists a partition of $\mathbb{V}$ into finitely many sets $\mathbb{V}_1, \ldots, \mathbb{V}_p$ such that

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_j \sup_{v_1, v_2 \in \mathbb{V}_j} n^{1/2} |\mathrm{S}_n(v_1) - \mathrm{S}_n(v_2)| \geq \epsilon \right) \leq \eta.$$

Condition (a) of Theorem A.4.1 is satisfied under the assumptions in Appendix A.2.

Proof of Theorem 2.2.4. It suffices to show $\mathrm{S}^Q_{np} \to_P S^Q$ and $\hat{\mathrm{V}}_{np} \to_P V$. We will give only the proof for S^Q_{np} as the proof for $\hat{\mathrm{V}}_{np}$ is similar. Let $\mathrm{S}^{q_k}_{np}$ and S^{q_k} denote the kth element of the random vectors S^Q_{np} and S^Q, respectively. Note that for any $\epsilon > 0$

$$\begin{aligned} \mathbb{P}\left(|n^{1/2}\mathrm{S}^{q_k}_{np} - \mathrm{S}^{q_k}| > \epsilon \right) &\leq \mathbb{P}\left(|n^{1/2}\mathrm{S}^{q_k}_{np} - \mathrm{S}^{q_k}_p| + |\mathrm{S}^{q_k}_p - \mathrm{S}^{q_k}| > \epsilon \right) \\ &\leq \mathbb{P}\left(|n^{1/2}\mathrm{S}^{q_k}_{np} - \mathrm{S}^{q_k}_p| > \epsilon/2 \right) + \mathbb{P}\left(|\mathrm{S}^{q_k}_p - \mathrm{S}^{q_k}| > \epsilon/2 \right). \end{aligned} \tag{A.6}$$

Thus, it suffices to show that the two terms on the right hand side converge to zero. Let $\epsilon, \delta > 0$ be given, then there exists a p_0 such that for $p \geq p_0$ there is a partition $\mathcal{V}_p$ with

$$\mathbb{P}\left(|\mathrm{S}^{q_k}_p - \mathrm{S}^{q_k}| < \epsilon/2 \right) < \delta/2, \tag{A.7}$$

because S has continuous sample paths.

For the first term of (A.6)

$$\begin{aligned} \mathbb{P}\left(|n^{1/2}\mathrm{S}^{q_k}_{np_0} - \mathrm{S}^{q_k}_{p_0}| > \epsilon/2 \right) &= \mathbb{P}\left(\left| \sum_{j=1}^{p_0} \{ n^{1/2}\mathrm{S}_n(v_j) - \mathrm{S}(v_j) \} q_k(v_j) \nu(\mathbb{V}_j) \right| > \epsilon/2 \right) \\ &\leq \mathbb{P}\left(\sup_{j \leq p_0} \left| n^{1/2}\mathrm{S}_n(v_j) - \mathrm{S}(v_j) \right| \times \sum_{j=1}^{p_0} |q_k(v_j)| \, \nu(\mathbb{V}_j) > \epsilon/2 \right), \end{aligned} \tag{A.8}$$

where the second line follows by Hölder's inequality using the infinity and 1 norms. Let $M =$

$\sup_{p \geq p_0} \sum_{j=1}^{p} |q_k(v_j)|\nu(\mathbb{V}_j)$, which is finite because

$$\lim_{p\to\infty} \sum_{j=1}^{p} |q_k(v_j)|\nu(\mathbb{V}_j) = \int_{\mathbb{V}} |q_k(v)| dv < \infty.$$

For the given ϵ, δ there exists N such that for all $n \geq N$

$$\mathbb{P}\left(\sup_v \left| n^{1/2}\mathrm{S}_n(v) - \mathrm{S}(v) \right| > \epsilon/2M \right) < \delta/2 \tag{A.9}$$

by the continuous mapping theorem, because $n^{1/2}\mathrm{S}_n$ and S have continuous sample paths, and by Theorem A.4.1. The left side of (A.9) bounds (A.8), which with (A.7) implies that $n^{1/2}\mathrm{S}^Q_{np} \to_P \mathrm{S}^Q$.

□

A.5. Proof of Theorem 2.2.5

By (A.5) applied to model (2.17)

$$\mathrm{R}_{\mathbb{L}} = \mathrm{Y}^T(I - H)GQ^T\{\hat{\sigma}^2 QG^T(I - H)GQ^T\}^{-1}QG^T(I - H)\mathrm{Y}$$

where $H = X(X^TX)^{-1}X^T$, Q is as defined in Theorem (2.2.2) and $\hat{\sigma}^2 = (n - m)^{-1}\mathrm{Y}^T(I - H)\mathrm{Y}$. Thus,

$$\begin{aligned} \mathrm{R}_{\mathbb{L}} &= \hat{\sigma}^{-2}\mathrm{Y}^T W \mathrm{Y} \\ &= (n - m)\frac{\mathrm{Y}^T W \mathrm{Y}}{\mathrm{Y}^T(I - H)\mathrm{Y}}, \end{aligned}$$

where $W = (I - H)GQ^T\{QG^T(I - H)GQ^T\}^{-1}QG^T(I - H)$. To derive (2.18), note

$$\frac{\mathrm{Y}^T(I - H)\mathrm{Y}}{\mathrm{Y}^T W \mathrm{Y}} = 1 + \frac{\mathrm{Y}^T\{(I - H) - W\}\mathrm{Y}}{\mathrm{Y}^T W \mathrm{Y}} \tag{A.10}$$

because W and $I - H$ are idempotent. The numerator and denominator of the random term on the right hand side are distributed as χ^2_r and χ^2_{n-m-r}, respectively, and independent by Theorem 11.14 of Schott (2016) since $W\{(I - H) - W\} = W(I - W) = 0$. So (A.10) is distributed as

$1 + \frac{n-m-r}{r} F_{(n-m-r),r}$, and

$$\begin{aligned} \mathrm{R}^{\mathbb{L}} &=_L \frac{(n-m)}{1 + \frac{(n-m-r)}{r} F_{(n-m-r),r}} \\ &= \frac{r(n-m)}{r + (n-m-r) F_{(n-m-r),r}}. \end{aligned}$$

APPENDIX B

Supplementary Materials for Chapter 3

The code to execute the simulations and analyses for this manuscript is available at `https://bitbucket.org/simonvandekar/param-boot`.

B.1. Supplementary Methods

B.1.1. Image preprocessing

T1-weighted structural images are processed using tools included in ANTs (Tustison et al., 2010). Voxel-wise analyses and simulations are restricted to gray matter locations. For region-wise analyses CBF is averaged within 112 anatomically defined gray matter regions.

The CBF image is co-registered to the T1 image using boundary-based registration (Greve and Fischl, 2009), and normalized to the custom PNC adolescent template using the top-performing diffeomorphic SyN registration included in ANTs (Avants et al., 2011; Klein et al., 2009). Images were down-sampled to 2mm resolution and smoothed with a Gaussian kernel at a FWHM of 6mm prior to group-level analysis.

B.1.2. Simulation Methods

Synthetic simulations

In order to better understand convergence rates for joint MTP procedures we perform simulations where the data generating covariance structure is known. The parametric procedures like Holm and PBJ rely on approximations due to estimating the covariance structure and using a multivariate normal approximation. We assume a normal data generating distribution so that the test statistics are T-distributed. Specifically, we assume two samples

$$X_i \sim \mathcal{N}(\mu_0, \Sigma) \text{ for } i \le n/2$$

$$X_i \sim \mathcal{N}(\mu_1, \Sigma) \text{ for } i > n/2,$$

where $\mu_k \in \mathbb{R}^V$, and we perform the tests $H_{0v} : \hat{\mu}_{1v} - \hat{\mu}_{0v} = 0$ for $v = 1, \ldots, V$. We vary the values $n \in \{40, 80, 100, 200\}$, $p \in \{100, 200, 500, 1000, 10000\}$, $\Sigma_{jj} = 1$, and $\Sigma_{jk} \in \{0, 0.9^{|j-k|}, (-0.9)^{|j-k|}\}$ for $j \neq k$. We make the first 10% of the components of μ_1 nonzero with parameter $\mu_{1v} = 0.4$ and all other mean parameters zero. We perform 500 simulations to estimate FWER and power.

Several step-down MTPs are considered. For insight into convergence properties, we let some MTPs rely on the unobserved covariance matrix, so they are not possible in practice. We use the following notation: $(T_n \mid \text{Holm})$ denotes Holm's procedure where a standard normal distribution is used to compute *p*-values for the T-statistics. $(Z_n \mid \text{Holm})$ is Holm's procedure where the T-statistics are transformed using (3.9), that is, the T-distribution is used to compute *p*-values. $(T_n \mid \text{Holm})$ and $(Z_n \mid \text{Holm})$ both demonstrate the how conservative using Holm procedure is and compare the effect of a normal approximation to the marginal densities. $(T_n \mid \Sigma)$ denotes adjusted *p*-values computed using PBJ with Procedure 3.2.4 using the true covariance Σ and the raw T-statistics. The FWER for $(T_n \mid \Sigma)$ gives us an idea of the sample size required for approximating a multivariate T-statistic with a normal distribution. $(T_n \mid \hat{\Sigma})$ uses PBJ with the sample estimate $\hat{\Sigma}$ using the raw T-statistics. $(Z_n \mid \hat{\Sigma})$ uses PBJ with the sample estimate $\hat{\Sigma}$ and the transformation (3.9). $(T_n \mid \text{Perm})$, uses permutations to generate the joint distribution and untransformed T-statistics. Within each simulation 1000 bootstraps and permutations are used to compute *p*-values for the PBJ and PJ MTPs.

Real data simulations

To create realistic simulated data sets we use samples generated from real data to compare the FWER and power of the MTPs for region-wise ($V = 112$) and voxel-wise analyses (V =127,756). For each simulation we draw subsamples without replacement from the CBF data for sample sizes of $n = 40, 100, 200, 400$. The region-wise simulations cover the case where $p > n$ and $p \approx n$. For the voxel-wise simulations we smooth at FWHM$= 6$. We present voxel-wise results with different smoothing kernels in the Supplement.

Results from the the synthetic simulations demonstrate that deviations from normality increases the FWER above the nominal level (see Section 3.5.1). For this reason, we perform the Yeo-Johnson transformation prior to performing hypothesis tests in the real data simulation and CBF analyses (Yeo and Johnson, 2000). The Yeo-Johnson transformation is a single parameter transformation similar to the Box-Cox that allows negative values for the outcome variable. We estimate the parameter at each location in the image using a profile-likelihood approach (Yeo and Johnson, 2000). Inference is performed conditional on the estimated parameter.

In each simulation we fit the following model with real covariates including age, sex, race and motion (mean relative displacement; MRD) as well as artificially generated covariates

$$\mathbb{E}Y_{iv} = \alpha_0 + \alpha_1\text{age}_i + \alpha_2\text{sex}_i + \alpha_3\text{race}_{i1} + \alpha_4\text{race}_{i2} + \alpha_5\text{MRD}_i + \sum_{j=1}^{3}\beta_{jv}g_j$$

where g_j are indicators for an artificial factor with 4 levels to represent different clinical groups in equal proportions and $\mathbb{E}Y_{iv}$ represents the conditional expectation of the transformed outcome. Multiple groups were generated so that we could perform a test of the parameter for the second group indicator

$$H_{0v} : \beta_{1v} = 0 \tag{B.1}$$

on one degree of freedom, and the test of

$$H_{0v} : \beta_{jv} = 0 \text{ for all } j \tag{B.2}$$

on 3 degrees of freedom. To assess power, in each simulation we generate signal in randomly selected locations. For the region-wise simulation we randomly selected 3 brain regions $v_k \in \{1, \ldots, 112\}$ for $k = 1, 2, 3$, by setting $\beta_{1v_k} = 10$ and $\beta_{jv} = 0$ otherwise. For the voxel-wise simulations we first select a random gray matter voxel v_0 and create a cube with a radius of 6 voxels centered at v_0. Let $N_{v_0} \subset \{1, \ldots, 127756\}$ denote the gray matter voxels within the cube. We create a parameter image where $\beta_{1v} = 120$ for all $v \in N_{v_0}$ and $\beta_{1v} = 0$ otherwise. The generated parameter image is smoothed at FWHM$=6$mm and is added to the CBF images smoothed with the same kernel. All other artificial parameters were set to 0.

We use 1000 simulations to estimate FWER and power for the region-wise data and 500 simulations for the voxel-wise data, which take considerably longer to run. All results are presented for the rejection level $\alpha = 0.05$. FWER is defined as the proportion of simulations where any true null hypothesis was rejected and power is estimated by the mean number of rejected hypotheses amongst the false null hypotheses. For the voxel-wise simulations the false null hypotheses were defined as all voxels in the smoothed parameter image where $\beta_{v1} > 1$. We use Wilson confidence intervals for the FWER estimated from the 500 simulations implemented within the `binom` package in R (Dorai-Raj, 2014; Wilson, 1927). For power results we use a normal approximation for confidence intervals. Note however, that the variance estimate for the power results will be biased downward since the generated subsamples are dependent.

We compare the FWER and power of the Bonferroni, Holm, PBJ, and PJ MTPs. Permutation tests are performed using Randomise (Winkler et al., 2014), which supports a wide array of possible tests, but only performs the single-step method. In the region-wise simulations we assess the single-step and step-down PBJ MTP. For the voxel-wise data we only compare single-step methods for the PBJ and PJ procedures because accurately estimating V cutoffs when V is large requires an infeasibly large number of samples. For the region-wise simulations we use 5000 samples for both the PBJ and PJ MTPs. For the voxel-wise simulations we use 5000 bootstraps for the PBJ, and 500 permutations with Randomise. We use 500 permutations with Randomise because the procedure takes considerable time to run. Mean run times to perform 5000 simulations are given in Figure B.1. Estimates of the *p*-values are unbiased for both procedures, but the *p*-values for the permutation procedure will have higher variance as fewer permutations are used (see Winkler et al., 2016). Wilson confidence intervals for a true *p*-value of $p = 0.05$ for 500 and 5000 permutations are $[0.03, 0.07]$ and $[0.04, 0.06]$.

In Section 3.5 we discuss the computational complexity of the PJ procedure versus the parametric-bootstrap. To compare actual computing time we take the mean of the time to perform the 1 and 3 degree of freedom tests with 5000 simulations for each method for the region-wise analyses. For the computing times of the voxel-wise analyses we multiply the computing time for the PJ procedure by 10 because 10 times fewer permutations were used than bootstraps. Note that this slightly over estimates the PJ computing time as it also multiplies the image load time, which is 1-2 minutes depending on the sample size.

B.2. Supplementary proofs

The proof of Theorem 3.3.2 requires defining a matrix normal distribution and identifying some useful properties of matrix-variate random variables (Dawid, 1981).

Theorem B.2.1 (Properties of matrix-variate random variables)**.** *The following are properties for*

Procedure	Analysis	Marg/Joint	Null estimation	Step proc
Bonferroni	Region-wise	Marg	Theor	Single-step
Holm	Region-wise	Marg	Theor	Step-down
PBJ	Region-wise	Joint	Theor; Boot	Single-step/Step-down
Permutation	Region-wise	Joint	Boot	Single-step/Step-down
Bonferroni	Voxel-wise	Marg	Theor	Single-step
Holm	Voxel-wise	Marg	Theor	Step-down
PBJ	Voxel-wise	Joint	Theor; Boot	Single-step
Permutation	Voxel-wise	Joint	Perm	Single-step

Table B.1: A summary of hypothesis correction procedures for neuroimaging. Joint methods are more powerful than marginal methods. Step-down procedures are more powerful than single-step. Marg= marginal; Step proc= step procedure; Speed= computing speed; Theor= Assumes theoretical distribution; Perm= distribution obtain using permutations.

	$n = 40$	$T_n \mid$ Holm	$Z_n \mid$ Holm	$T_n \mid \Sigma$	$T_n \mid \hat{\Sigma}$	$Z_n \mid \hat{\Sigma}$	$T_n \mid$ Perm
Indep	$m = 100$	12	6	12	16	6	5
	$m = 200$	13	4	13	17	4	4
	$m = 500$	16	6	17	21	7	5
	$m = 1000$	20	4	20	23	5	3
	$m = 5000$	27	5	28	29	6	3
	$m = 10000$	38	5	39	40	9	5
Pos AR(1)	$m = 100$	6	2	10	12	5	5
	$m = 200$	8	3	13	13	5	6
	$m = 500$	9	3	14	18	6	4
	$m = 1000$	13	2	17	22	5	5
	$m = 5000$	22	2	27	32	4	3
	$m = 10000$	26	4	33	34	7	6
Neg AR(1)	$m = 100$	7	2	11	12	6	5
	$m = 200$	6	3	11	16	5	4
	$m = 500$	7	2	11	18	4	3
	$m = 1000$	12	3	17	19	6	4
	$m = 5000$	20	3	25	29	7	4
	$m = 10000$	27	4	34	35	6	5

Table B.2: Type 1 error results for $n = 40$ to assess convergence rates. Values are mean percentage of correctly rejected tests across 500 simulations. Test statistics were simulated as normal with independent (Indep), positive autoregressive (Pos AR(1)), and negative autoregressive (Neg AR(1)) correlation structures with $\rho = 0.9$ and $\rho = -0.9$. The number of tests was varied within $m = (200, 500, 1{,}000, 5{,}000, 10{,}000)$ with 10% non-null test statistics. Detailed descriptions of the column names are given in Section 3.4. Results demonstrate that the majority of error in the PBJ procedure is due to the convergence of the T-statistics to a normal distribution.

matrix-variate normal random variables.

1. *Let the $n \times p$ matrix Z have independent standard normal entries, Then the matrix $A + CZB \sim \mathcal{MN}(A, CC^T, B^TB)$, is matrix-variate normal with mean matrix A, row covariance matrix CC^T, and column covariance B^TB.*

2. *Let the $n \times p$ matrix $X \sim \mathcal{MN}(0, I_{n\times n}, \Sigma)$, then $X^TX \sim \mathcal{W}_p(n, \Sigma)$. If $n < p$ then $\mathcal{W}_p(n, \Sigma)$ is a singular Wishart distribution.*

The following are properties of matrix-variate random variables

1. *Let Z be an $n \times p$ matrix-variate random variable. For positive semi-definite matrices Ψ ($n \times n$) and Φ ($p \times p$), if the row covariance $\mathrm{cov}(Z_i^T) = \Psi\Phi_{ii}$ and the column covariance $\mathrm{cov}(Z_j) = \Phi\Psi_{ii}$, then we write $\mathrm{cov}(Z) = (\Psi, \Phi)$.*

2. *If Z is an $n\times p$ matrix-variate random variable with $\mathrm{cov}(Z) = (\Psi, \Phi)$, then $\mathrm{cov}(\mathrm{vec}(Z)) = \Phi \otimes \Psi$.*

Proof of Theorem 3.3.2. For the first property write

$$\begin{aligned} (R_{X_0} - R_X) &= AA^T \\ R_X &= BB^T, \end{aligned} \tag{B.3}$$

where A is an $n \times m_1$ matrix and B is and $n \times (n - m)$ matrix both with with orthonormal columns. Then, under the assumption (3.6), $A^TX_0\alpha = (A^TA)A^TX_0\alpha = A^T(R_{X_0} - R_X)X_0\alpha = 0$, since X_0 is orthogonal to the column space of R_{X_0} and R_X. Then with normal errors (3.4), Theorem B.2.1 implies $A^TY \sim \mathcal{MN}(0, I_{m_1\times m_1}, \Psi)$. Similarly, $B^TX_0\alpha = 0$ and we obtain $B^TY \sim \mathcal{MN}(0, I_{(n-m)\times(n-m)}, \Psi)$. Then equation (3.8) follows from Theorem B.2.1.

For the proof of the second property let A and B be as defined in (B.3). To prove the convergence of m_1F_n, we will invoke the central limit theorem for A^TY. To do this we use the Cramér-Wold device (Vaart, 2000) to prove that A^TY converges in law to a $\mathcal{MN}(0, I_{m_1\times m_1}, \Psi)$ distribution.

The $m_1 \times V$ matrix-variate random variable A^TY has vth row covariance cov($A^TY_v) = \Psi_{v,v}I_{m_1\times m_1}$ and ith column covariance cov($A_i^TY) = \Psi$. Theorem B.2.1 implies that the vectorized version has covariance cov(vec(A^TY)) $= \Psi \otimes I_{m_1\times m_1}$. Using the Cramér-Wold device we need only prove that for any vector t,

$$t^T \text{vec}(A^T Y) \to_L \mathcal{N}(0, t^T(\Psi \otimes I_{m_1 \times m_1})t). \tag{B.4}$$

Assumption (3.6) implies $t^T\text{vec}(A^T\mathbb{E}Y) = 0$, by the same argument as above. Assumption (3.7) implies $t^T(\Psi \otimes I_{m_1\times m_1})t < \infty$. So, by the central limit theorem (B.4) holds, which implies $A^TY \to_L \mathcal{MN}(0, I_{m_1\times m_1}, \Psi)$ by the Cramér-Wold device. From there, the continuous mapping theorem gives $\Phi^2\text{diag}\{Y^TAA^TY\} \to_L \text{diag}\{\mathcal{W}_V(m_1, \Sigma)\}$.

For the denominator let B be as defined in (B.3). Then, under $Y_{iv} \perp\!\!\!\perp Y_{jv}$ for all $i \neq j$ and letting $W_v = B^TY_v$,

$$\text{cov}(W_v) = B^T\text{cov}(Y_v)B = \Psi_{v,v}I_{(n-m)\times(n-m)}.$$

Assumption (3.6) means $\mathbb{E}W_v = 0$ by the same argument as for $A^T\mathbb{E}Y$, so

$$\frac{1}{n-m}\mathbb{E}Y_v^TR_XY_v = \mathbb{E}\frac{1}{n-m}W_v^TW_v = \frac{1}{n-m}\sum_{i=1}^{n-m}\mathbb{E}W_{iv}^2 = \Psi_{v,v}.$$

By the weak law of large numbers $Y_v^TR_XY_v/(n-m) \to_P \Psi_{v,v}$. The convergence of $m_1F_n \to_L \text{diag}\{\mathcal{W}_V(m_1, \Sigma)\}$ follows by Slutsky's theorem (Vaart, 2000).

□

The joint CDFs for the numerators can be used to estimate the asymptotic joint CDF of the transformed statistics (3.9). We use Monte Carlo simulations to generate the numerators using the distributions given in Theorem 3.3.2 to estimate the null distribution of Z_{vn} given in (3.9).

To show that the PBJ procedure guarantees asymptotic control of the FWER we must show that the joint distributions satisfy the null domination condition of Definition 3.3.1. When null domination holds then Theorem B.2.4 guarantees asymptotic control of the FWER.

Theorem B.2.2 (Null domination)**.** *Let $F_n = (F_{1n}, \dots, F_{Vn})$, and $Z \sim Q_0 = \text{diag}\{\mathcal{W}_V(m_1, \Sigma)\}$. Let, $g_n(x) = \Phi_n^{-1}(\Phi(x))$ where Φ and Φ_n are as defined in* (3.9)*. Then, the joint distribution Q_n of Z_n, defined by the transformation* (3.9)*, is asymptotically dominated by the joint distribution Q_0 of Z.*

The following lemma is used to prove Theorem B.2.2 and is presented here without proof.

Lemma B.2.3. *Let $f_n : \mathbb{R} \mapsto [0, 1]$ converge uniformly to a continuous function f and $g_n : \mathbb{R} \mapsto \mathbb{R}$*

converge pointwise to g. Then $f_n(g_n(x))$ converges pointwise to $f(g(x))$.

Proof. Let x and ϵ be given. Because f is continuous there exists δ such that $|f(y) - f(g(x))| < \epsilon/2$ for all y such that $|g(x) - y| < \delta$. Choose $N_1 = N_1(\epsilon, x)$ such that for all $n \geq N$, $|g_n(x) - g(x)| < \delta$, which is possible due to the pointwise convergence of g_n. Because f_n converges uniformly, there exists $N_2 = N_2(\epsilon)$ such that $|f_n(y) - f(y)| < \epsilon/2$ for all $y \in \mathbb{R}$. Thus, it follows that for all $n \geq N = \max\{N_1, N_2\}$.

$$|f_n(g_n(x)) - f(g(x))| \leq |f(g_n(x)) - f(g(x))| + |f_n(g_n(x)) - f(g_n(x))| < \epsilon$$

□

Proof of Theorem B.2.2. For any V dimensional vector of random variables Z let $F_Z(x) = \mathbb{P}(\max_{v \leq V} Z_j < x)$. We will show that $F_{Z_n}(x) \to F_Z(x)$, for all x as $n \to \infty$. This implies that the null domination condition holds because then $\limsup_{n\to\infty} 1 - F_{Z_n}(x) \leq 1 - F_Z(x)$.

First note that F_n are continuous random variables, so taking the maximum of F_n is a continuous function. The continuous mapping theorem implies that

$$\max_v m_1 F_{vn} \to_L \max_v Z_v, \tag{B.5}$$

because $m_1 F_n \to_L Z$ by Theorem 3.3.2.

g_n is monotone in x, which implies $F_{Z_n}(x) = F_{F_n}\{g_n(x)\}$ because $Z_{vn} = g_n^{-1}(F_{vn})$. Thus for any n,

$$F_{Z_n}(x) - F_Z(x) = F_{F_n}\{g_n(x)\} - F_Z(x).$$

$F_{F_n}(m_1^{-1}x) - F_Z(x)$ converges to zero pointwise by (B.5) and since F_Z is continuous then convergence in law implies uniform convergence to $F_Z(x)$ (Vaart, 2000 Lemma 2.11). Since $g_n(x) \to m_1^{-1}x$, then uniform convergence of $F_{F_n}(m_1^{-1}x)$ to $F_Z(x)$ and the continuity of $F_Z(x)$ imply $F_{F_n}\{g_n(x)\} - F_Z(x) \to 0$ by Lemma B.2.3.

□

The assumption of convergence of Z_n to Z in Theorem B.2.2 holds for many statistics of interest by the central limit theorem. When Theorem B.2.2 holds then, the following theorem from Dudoit and Laan (2008 p. 205) ensures asymptotic control of the FWER.

Theorem B.2.4 (Asymptotic control of the FWER by step-down procedure)**.** *Let $Z_{(1)n} < \ldots < Z_{(V)n}$ denote the ordered test statistics, and $H_{(1)}, \ldots H_{(V)}$ their associated hypotheses. Assume that the distribution, Q_n, for the test statistics $Z_{(v)n}$ is dominated asymptotically by Q_0 and that $Z = Z_{(1)}, \ldots, Z_{(V)} \sim Q_0$. For a given level α define the thresholds C_{vn} as the smallest value that satisfies $\mathbb{P}\left(\max_{k\leq v} Z_{(k)} < C_{vn}\right) = 1 - \alpha$. Where C_{vn} depends on the sample through the order of the test statistics $Z_{(v)n}$. Let H_0 denote the set of true null hypotheses and define the number of type 1 errors as $E_n = \sum_{v=1}^{V} I\left(Z_{(v)n} > C_{vn}, H_{(v)} \in H_0\right)$.*

Then the PBJ procedure provides asymptotic control of the FWER at level α,

$$\limsup_{n\to\infty} \mathbb{P}(E_n > 0) \leq \alpha.$$

The proof of Theorem B.2.4 is given in Dudoit and Laan (2008 p. 206). Propostion 5.5 of Dudoit and Laan (2008 p. 211) gives the adjusted *p*-values used in Procedure 3.2.4. The proof of asymptotic FWER control for Procedure 3.2.3 is implied by Theorem B.2.4, because the single step procedure uses the single most conservative threshold, so leads to more conservative inference.

The joint distribution Q_0 is unavailable in practice and must be estimated from the sample. Theorem 5.12 of Dudoit and Laan (2008 p. 228) states that if our estimate $\hat{Q}_0 = Q_0(\hat{\Sigma})$ converges in probability to Q_0 then the thresholds C_{vn} in Theorem B.2.4 are consistent, which gives consistent adjusted *p*-values by the continuous mapping theorem. Because our estimate $\hat{\Sigma}$ in (3.12) is consistent, then $\hat{Q}_0$ converges in probability to Q_0. Thus, using this estimate of Σ gives valid asymptotic inference using the PBJ.

B.2.1. Supplementary simulation analyses

We also ran the simulation analyses with 3mm and 9mm smoothing. The parameter image for each simulation was smoothed with the same kernel as the imaging data. The results are presented in Figures B.2 and B.3.

Figure B.4 shows FWER controlled results fitting a fixed degree spline model with 10 knots. We presented results with 5 knots in the manuscript. The permutation procedure yielded no significant results because in the untransformed data there are voxels whose test statistics have heavily skewed null distributions. When taking the maximum across the image this leads to very conservative results.

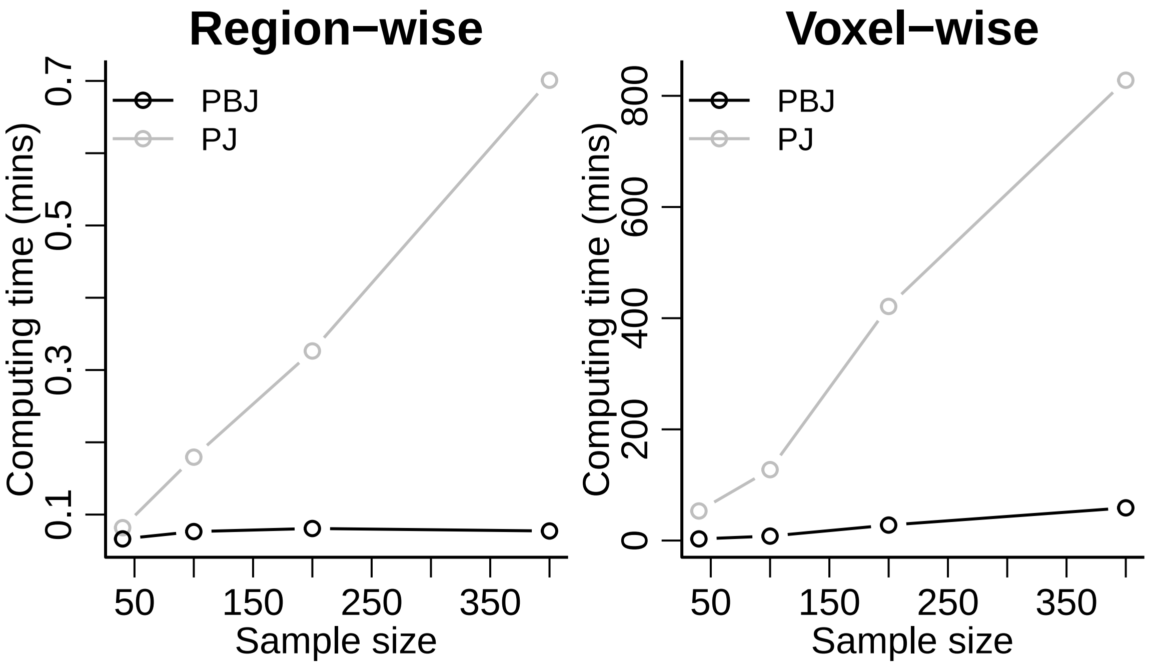

Figure B.1: Computing times by sample size for the PBJ and PJ testing procedures for the region- and voxel-wise simulation analyses. We multiply the PJ computing time by 10 for the voxel-wise times because 10 times fewer permutations were used for that procedure.

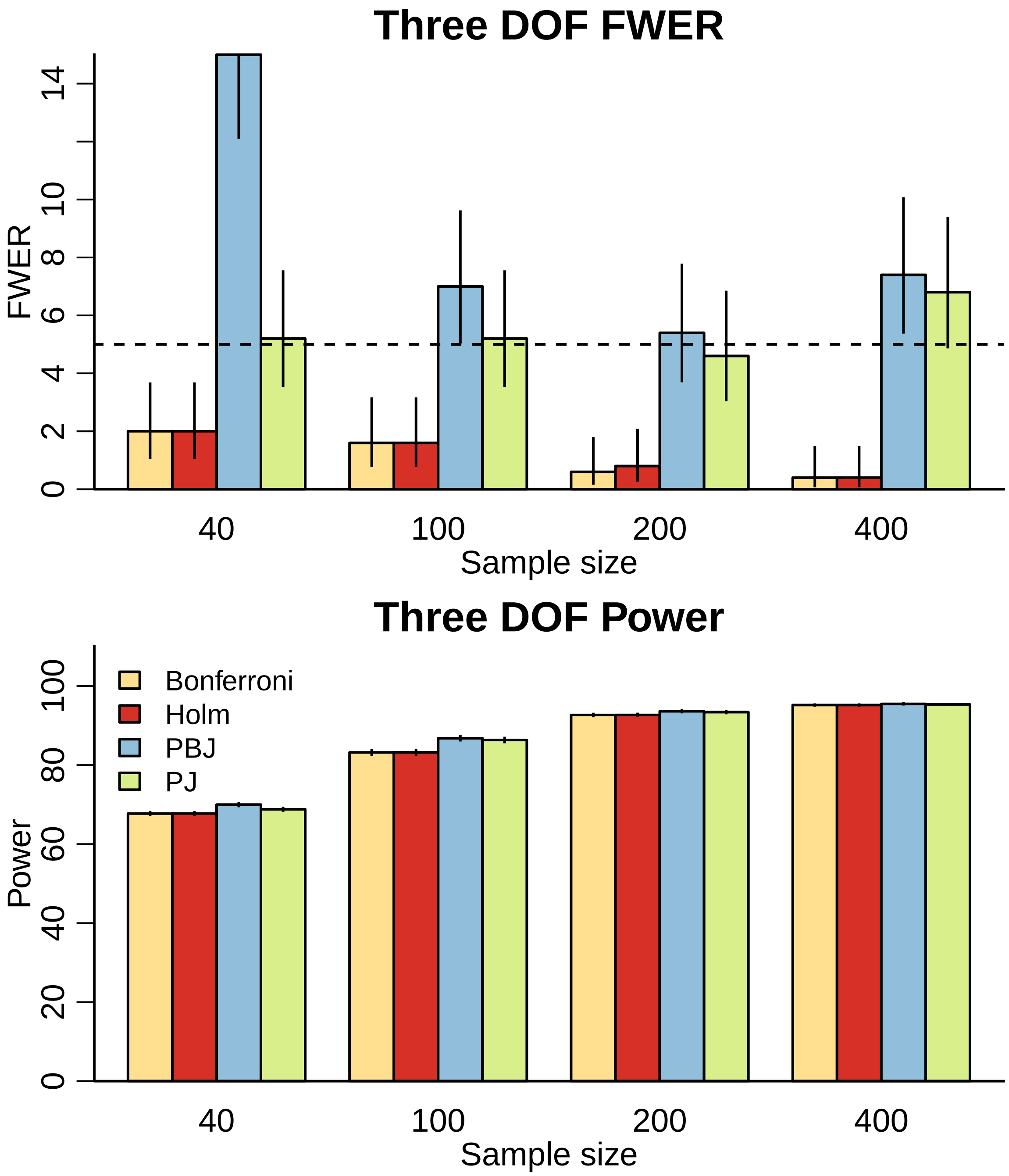

Figure B.2: FWER and power for simulations with a Gaussian smoothing kernel of FWHM=3mm for the F-statistic of (B.2) on three degrees of freedom (DOF) for the CBF image.

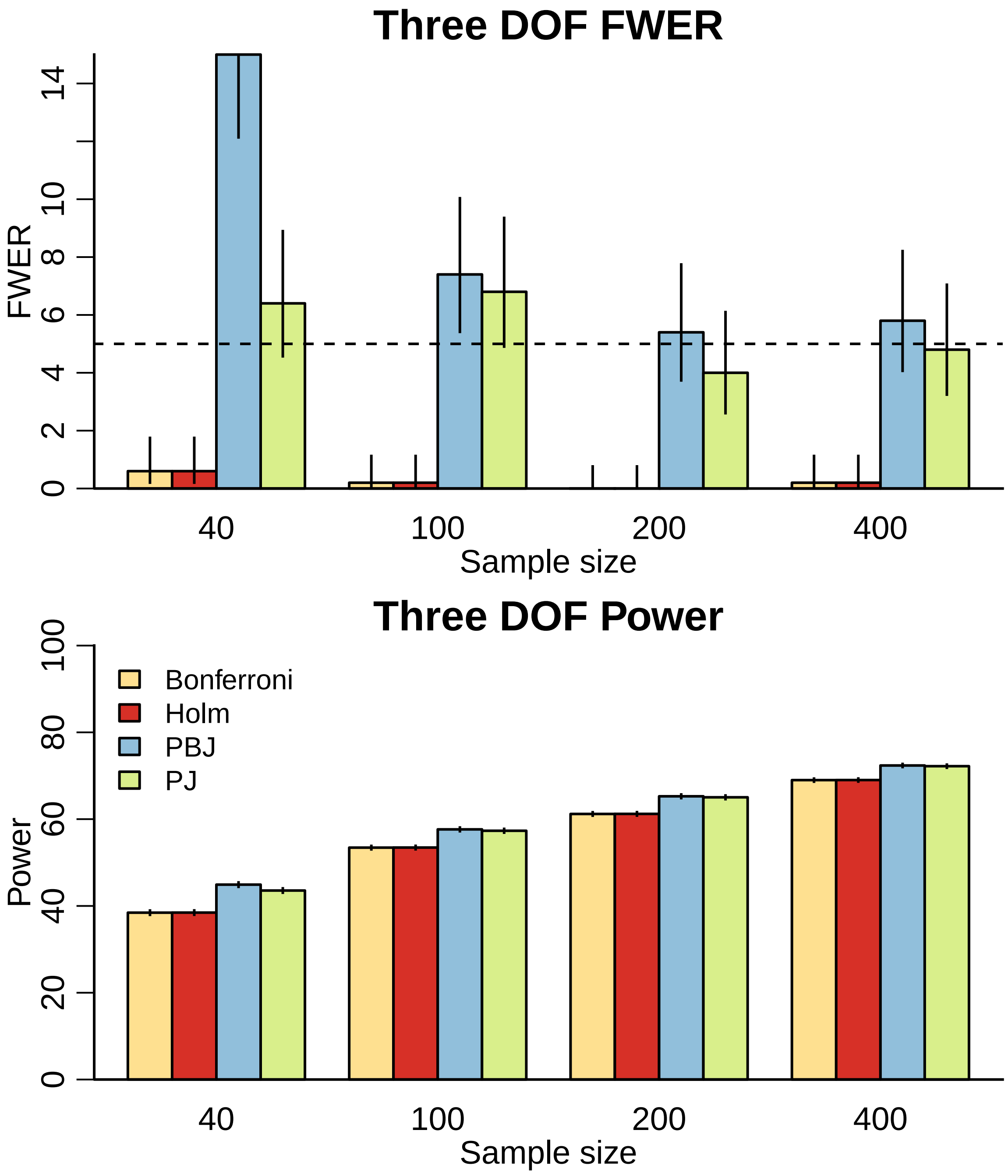

Figure B.3: FWER and power for simulations with a Gaussian smoothing kernel of FWHM=9mm for the F-statistic of (B.2) on three degrees of freedom (DOF) for the CBF image.

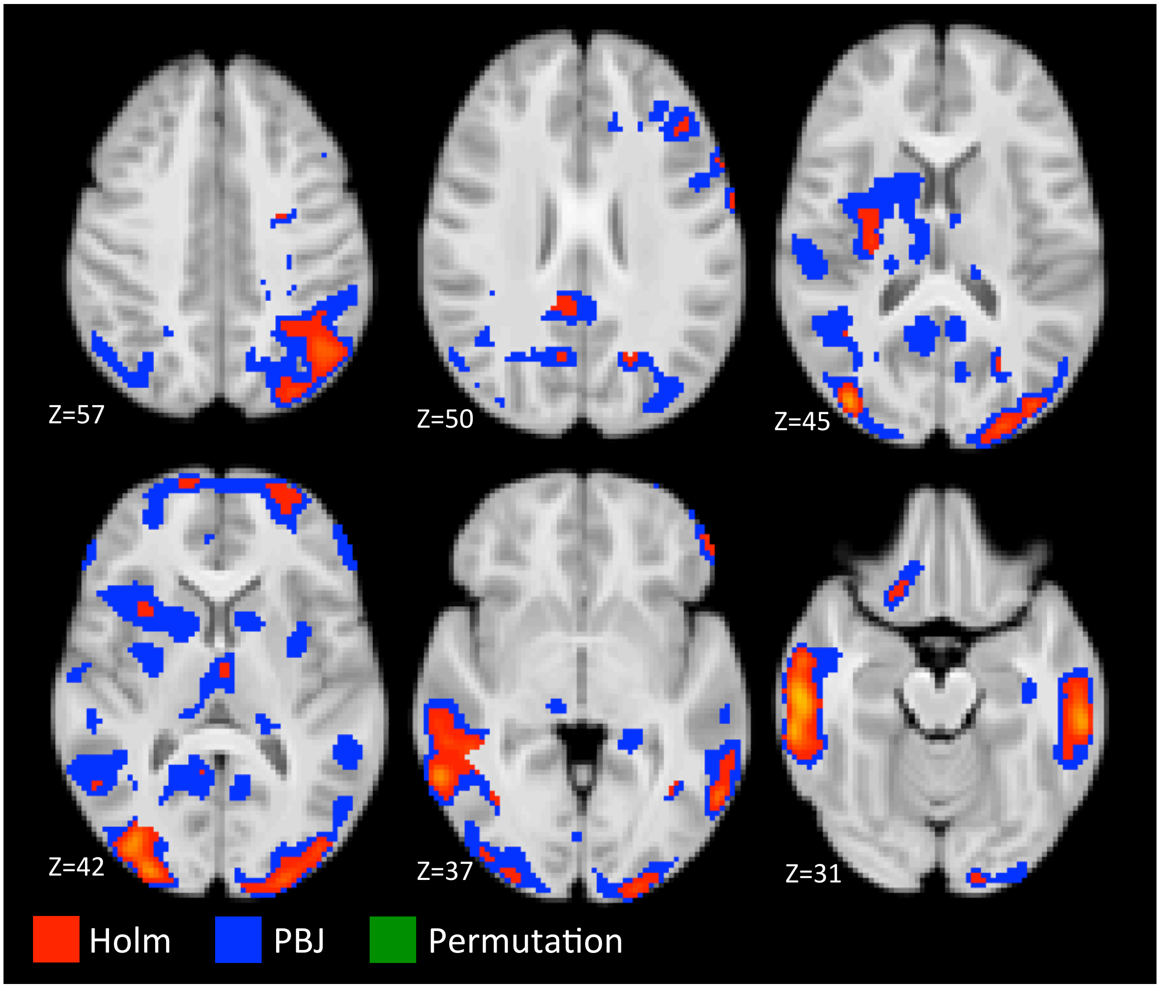

Figure B.4: FWER controlled results at $\alpha = 0.05$ for Holm (red), PBJ single-step (blue), and PJ single-step (green) for the spline model fit with 10 knots. Color scale is $-\log_{10}(p)$ for the adjusted *p*-values and shows results greater than 1.3. The overlay order is the PBJ, Holm, and PJ procedures, so that green indicates regions where all three regions reject the null, red and green indicate regions where Holm and PJ reject, and the union of all colors is where PBJ rejects. Blue indicates locations where only the PBJ procedure rejects.

APPENDIX C

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

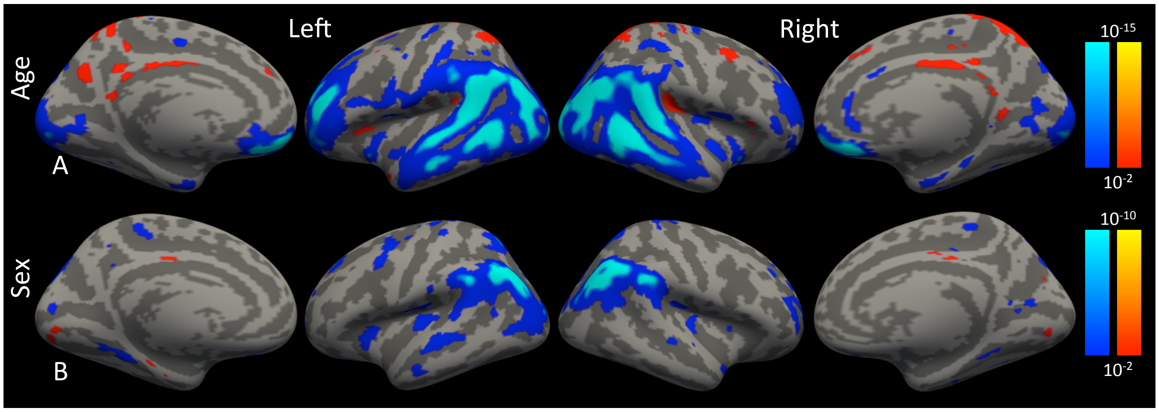

Figure C.1: Linear age related changes in coupling with FWHM=10 for comparison with those shown in Figures 4 and 6. Color bars show signed p-values; blue is negative, red is positive.

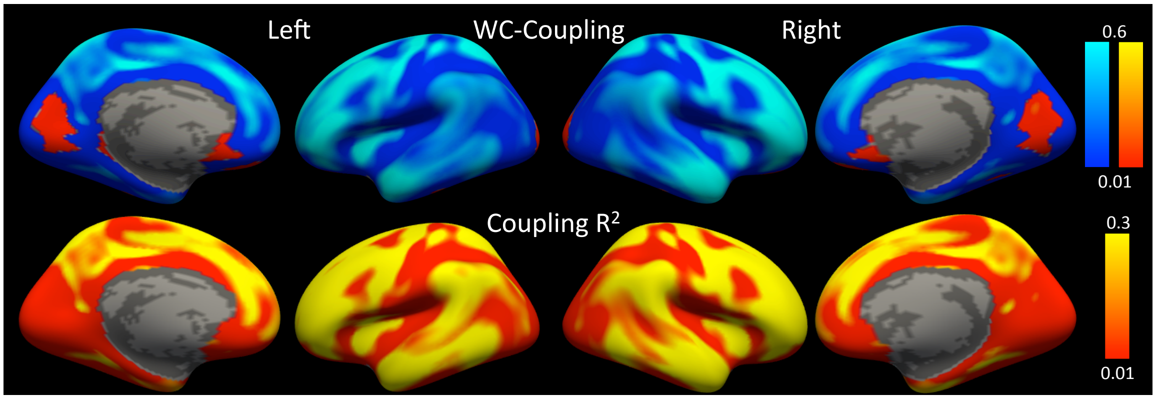

Figure C.2: Average weighted correlation (WC-Coupling) and R^2 maps for the fit of the weighted regression at each vertex. Blue and red indicate negative and positive correlations, respectively. Note some regions have low R^2 (e.g. primary visual and medial orbital frontal). Results in these regions must be interpreted carefully.

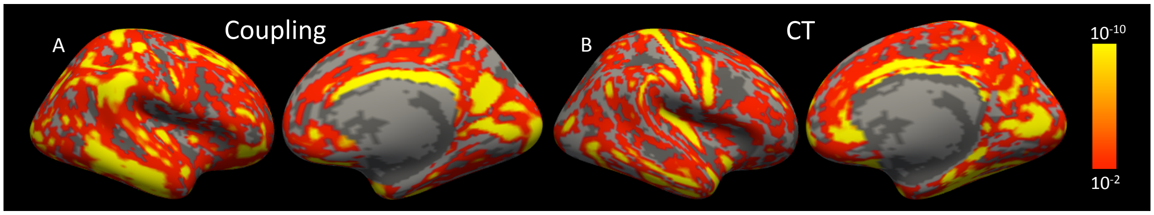

Figure C.3: Shapiro-Wilk statistical maps for the test of normality in (A) coupling and (B) cortical thickness (CT). Both data show significant deviation from normality over large portions of the cortex. Large samples are required for asymptotic normality of the parameter estimates to perform inference. Color bars show uncorrected p-values for the test.

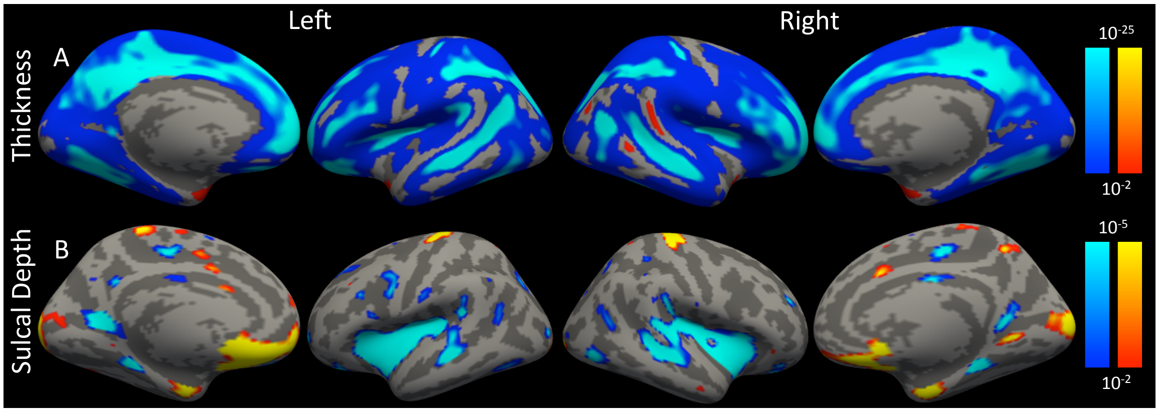

Figure C.4: Age related coupling differences (see Figure 4) are due in part to (A) robust thinning that occurs with age primarily in sulcal regions. Note, however, that there are age-related differences in sulcal depth in (B) the Sylvian Fissure and superior temporal sulcus that may contribute to the significant age-related differences in coupling. Statistical maps are FDR thresholded at $q = 0.01$. Color bars show signed uncorrected p-values; blue is negative, red is positive.

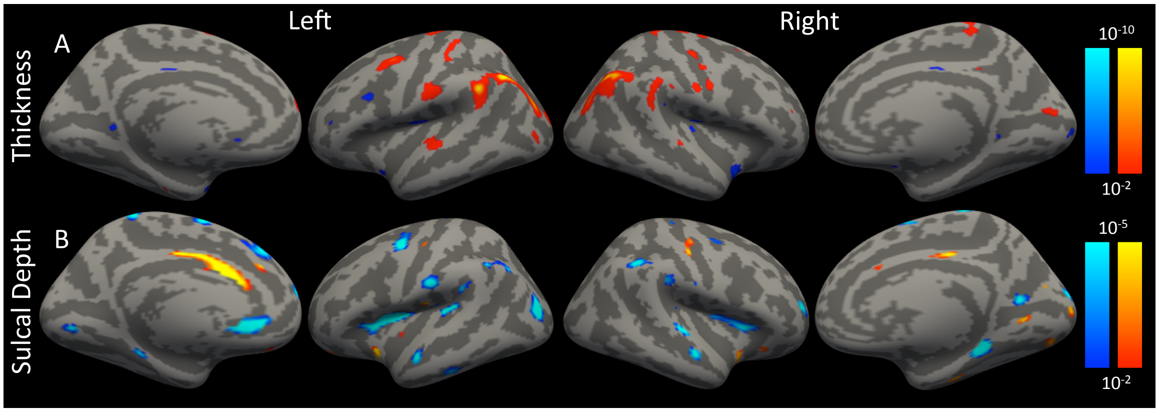

Figure C.5: Sex differences in coupling (see Figure 6) are driven by (A) thicker gyral cortex in females in the parietal cortex. No sex differences in sulcal depth (B) were observed in this region. Thus, sex differences in coupling are due to differences in thickness that are topologically related to sulcal depth. Statistical maps are FDR thresholded at $q = 0.01$. Color bars show signed uncorrected p-values, blue is greater in males, red is greater in females.

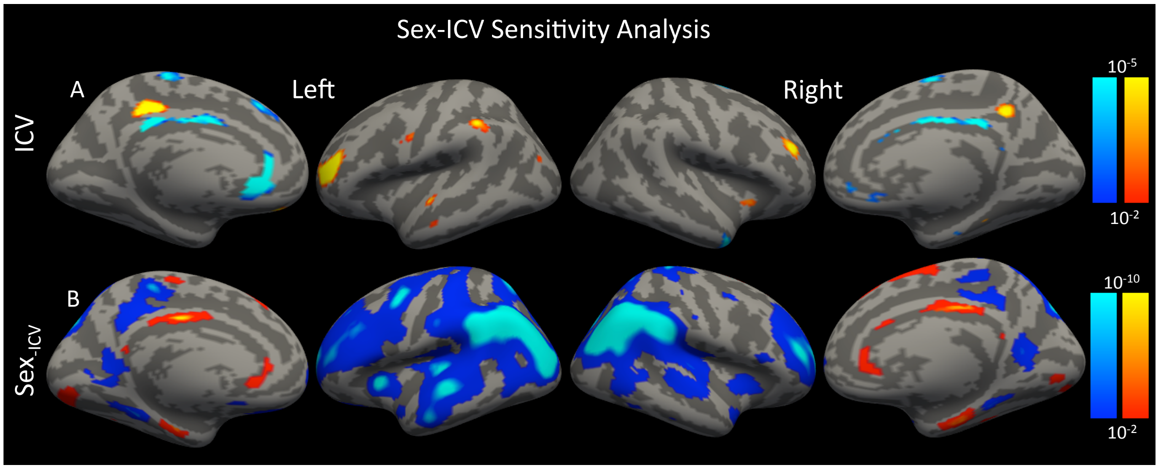

Figure C.6: (A) Significant effects of intracranial volume (ICV) on coupling are spatially sparse, with positive associations in the frontal pole and precuneus. (B) Sex differences without ICV have greater associations in the frontal and temporal poles (compare with Figure 6). Statistical maps are FDR thresholded at $q = 0.01$. Color bars show signed p-values; blue is negative, red is positive.

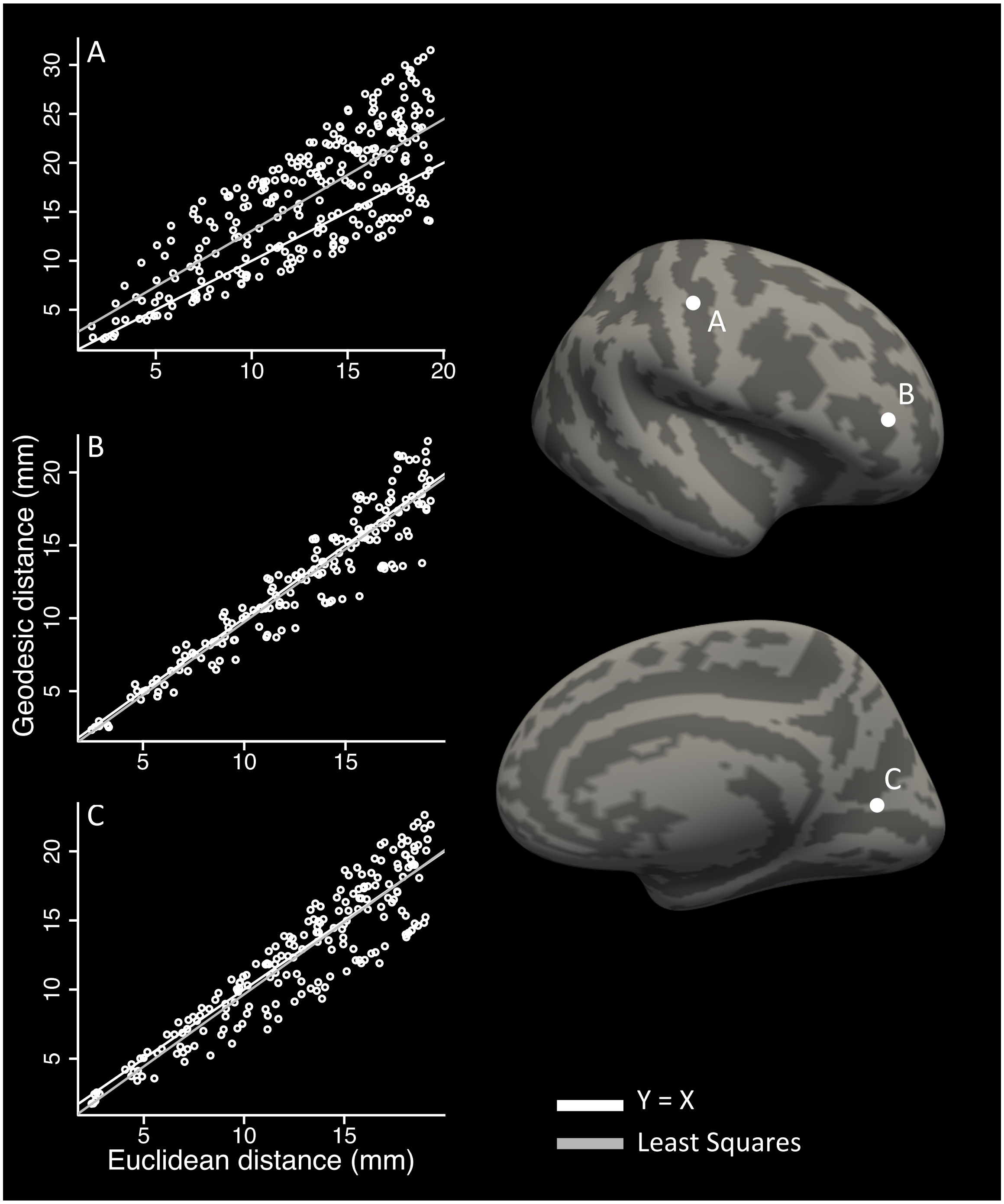

Figure C.7: A comparison of Euclidean distance on the fsaverage5 template inflated surface to geodesic distance on the template pial surface. (A) Geodesic distance is greater on average than Euclidean distance in highly folded regions like the central sulcus. (B, C) In regions with moderate folding the metrics are comparable. The scatter plots include vertices that contribute to the estimate of coupling more than 1% at FWHM=15. White points on the cortical surface indicate the central vertex. Geodesic distance to the center vertex is estimated using a Freesurfer command line tool.

BIBLIOGRAPHY

Alemn-Gmez, Y, Janssen, J, Schnack, H, Balaban, E, Pina-Camacho, L, Alfaro-Almagro, F, Castro-Fornieles, J, Otero, S, Baeza, I, Moreno, D, Bargall, N, Parellada, M, Arango, C, and Desco, M (Sept. 2013). The Human Cerebral Cortex Flattens during Adolescence. en. *The Journal of Neuroscience* 33.38, 15004–15010. ISSN: 0270-6474, 1529-2401.

Avants, BB, Cook, PA, Ungar, L, Gee, JC, and Grossman, M (Apr. 2010). Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis. *NeuroImage* 50.3, 1004–1016. ISSN: 1053-8119.

Avants, BB, Tustison, NJ, Song, G, Cook, PA, Klein, A, and Gee, JC (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 54.3, 2033–2044.

Boos, DD and Stefanski, LA (2013). *Essential Statistical Inference*. New York, NY: Springer. ISBN: 978-1-4614-4817-4 978-1-4614-4818-1.

Cachia, A, Paillre-Martinot, ML, Galinowski, A, Januel, D, Beaurepaire, R de, Bellivier, F, Artiges, E, Andoh, J, Bartrs-Faz, D, Duchesnay, E, Rivire, D, Plaze, M, Mangin, JF, and Martinot, JL (Feb. 2008). Cortical folding abnormalities in schizophrenia patients with resistant auditory hallucinations. *NeuroImage* 39.3, 927–935. ISSN: 1053-8119.

Cai, TT, Liu, W, and Xia, Y (Mar. 2014). Two-sample test of high dimensional means under dependence. en. *Journal of the Royal Statistical Society: Series B* 76.2, 349–372. ISSN: 1467-9868.

Calkins, ME, Merikangas, KR, Moore, TM, Burstein, M, Behr, MA, Satterthwaite, TD, Ruparel, K, Wolf, DH, Roalf, DR, Mentch, FD, Qiu, H, Chiavacci, R, Connolly, JJ, Sleiman, PM, Gur, RC, Hakonarson, H, and Gur, RE (Apr. 2015). The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. en. *Journal of Child Psychology and Psychiatry*, n/a–n/a. ISSN: 1469-7610.

Crane, PK, Carle, A, Gibbons, LE, Insel, P, Mackin, RS, Gross, A, Jones, RN, Mukherjee, S, Curtis, SM, Harvey, D, Weiner, M, and Mungas, D (Dec. 2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain imaging and behavior* 6.4, 502–516. ISSN: 1931-7557.

Csernansky, JG, Gillespie, SK, Dierker, DL, Anticevic, A, Wang, L, Barch, DM, and Van Essen, DC (Nov. 2008). Symmetric abnormalities in sulcal patterning in schizophrenia. *NeuroImage* 43.3, 440–446. ISSN: 1053-8119.

Dale, AM, Fischl, B, and Sereno, MI (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage* 9.2, 179–194.

Dawid, AP (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* 68.1, 265–274.

Desikan, RS, Sgonne, F, Fischl, B, Quinn, BT, Dickerson, BC, Blacker, D, Buckner, RL, Dale, AM, Maguire, RP, Hyman, BT, and others (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31.3, 968–980.

Destrieux, C, Fischl, B, Dale, A, and Halgren, E (Oct. 2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* 53.1, 1–15. ISSN: 1053-8119.

Dorai-Raj, S (2014). *binom: Binomial Confidence Intervals For Several Parameterizations*. R package version 1.1-1. URL: `https://CRAN.R-project.org/package=binom`.

Dudoit, S and Laan, MJ van der (2008). *Multiple Testing Procedures with Applications to Genomics*. New York, NY: Springer. ISBN: 978-0-387-49316-9 978-0-387-49317-6.

Dudoit, S, Shaffer, JP, and Boldrick, JC (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* 18.1, 71–103.

Dunn, OJ (1961). Multiple comparisons among means. *Journal of the American Statistical Association* 56.293, 52–64.

Economo, C von (1925). *The Cytoarchitectonics of the Human Cerebral Cortex (English translation by S. Parker in 1929)*. London: Oxford University Press.

Efron, B (2007). Correlation and Large-Scale Simultaneous Significance Testing. *Journal of the American Statistical Association* 102.477, 93–103. ISSN: 0162-1459.

Eklund, A, Nichols, TE, and Knutsson, H (June 2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. en. *Proceedings of the National Academy of Sciences* 113.28, 7900–7905. ISSN: 0027-8424, 1091-6490.

Eklund, A, Andersson, M, Josephson, C, Johannesson, M, and Knutsson, H (2012). Does parametric fMRI analysis with SPM yield valid results?An empirical study of 1484 rest datasets. *NeuroImage* 61.3, 565–578.

Fan, J, Han, X, and Gu, W (2012). Estimating False Discovery Proportion Under Arbitrary Covariance Dependence. *Journal of the American Statistical Association* 107.499, 1019–1035. ISSN: 0162-1459.

Fischl, B (Aug. 2012). FreeSurfer. *NeuroImage* 62.2, 774–781. ISSN: 1053-8119.

Fischl, B (Jan. 2013). Estimating the Location of Brodmann Areas from Cortical Folding Patterns Using Histology and Ex Vivo MRI. en. In: *Microstructural Parcellation of the Human Cerebral Cortex*. Ed. by S Geyer and R Turner. Springer Berlin Heidelberg, 129–156. ISBN: 978-3-642-37823-2 978-3-642-37824-9.

Fischl, B and Dale, AM (Sept. 2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. en. *Proceedings of the National Academy of Sciences* 97.20, 11050–11055. ISSN: 0027-8424, 1091-6490.

Fischl, B, Sereno, MI, and Dale, AM (1999). Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9.2, 195–207.

Fischl, B, Rajendran, N, Busa, E, Augustinack, J, Hinds, O, Yeo, BTT, Mohlberg, H, Amunts, K, and Zilles, K (Aug. 2008). Cortical Folding Patterns and Predicting Cytoarchitecture. en. *Cerebral Cortex* 18.8, 1973–1980. ISSN: 1047-3211, 1460-2199.

Freedman, D and Lane, D (Oct. 1983). A Nonstochastic Interpretation of Reported Significance Levels. *Journal of Business & Economic Statistics* 1.4, 292–298. ISSN: 0735-0015.

Friston, KJ, Worsley, KJ, Frackowiak, RSJ, Mazziotta, JC, and Evans, AC (1994a). Assessing the significance of focal activations using their spatial extent. en. *Human Brain Mapping* 1.3, 210–220. ISSN: 1097-0193.

Friston, KJ, Holmes, AP, Worsley, KJ, Poline, JP, Frith, CD, and Frackowiak, RS (1994b). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping* 2.4, 189–210.

Goghari, VM, Rehm, K, Carter, CS, and Macdonald, AW (Sept. 2007). Sulcal thickness as a vulnerability indicator for schizophrenia. en. *The British Journal of Psychiatry* 191.3, 229–233. ISSN: 0007-1250, 1472-1465.

Gotze, F (Apr. 1991). On the Rate of Convergence in the Multivariate CLT. EN. *The Annals of Probability* 19.2, 724–739. ISSN: 0091-1798, 2168-894X.

Greve, DN and Fischl, B (Oct. 2009). Accurate and Robust Brain Image Alignment using Boundary-based Registration. *NeuroImage* 48.1, 63–72. ISSN: 1053-8119.

Hochberg, Y (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75.4, 800–802.

Holm, S (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* 6.2, 65–70.

Hommel, G (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75.2, 383–386.

Huttenlocher, PR and Dabholkar, AS (Oct. 1997). Regional differences in synaptogenesis in human cerebral cortex. eng. *The Journal of Comparative Neurology* 387.2, 167–178. ISSN: 0021-9967.

Im, K, Lee, JM, Lee, J, Shin, YW, Kim, IY, Kwon, JS, and Kim, SI (2006). Gender difference analysis of cortical thickness in healthy young adults with surface-based methods. *Neuroimage* 31.1, 31–38.

Insel, TR (Nov. 2010). Rethinking schizophrenia. en. *Nature* 468.7321, 187–193. ISSN: 0028-0836.

Jack, CR, Bernstein, MA, Fox, NC, Thompson, P, Alexander, G, Harvey, D, Borowski, B, Britson, PJ, Whitwell, JL, Ward, C, Dale, AM, Felmlee, JP, Gunter, JL, Hill, DL, Killiany, R, Schuff, N, Fox-Bosetti, S, Lin, C, Studholme, C, DeCarli, CS, Krueger, G, Ward, HA, Metzger, GJ, Scott, KT, Mallozzi, R, Blezek, D, Levy, J, Debbins, JP, Fleisher, AS, Albert, M, Green, R, Bartzokis, G, Glover, G, Mugler, J, and Weiner, MW (Apr. 2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods. *Journal of magnetic resonance imaging : JMRI* 27.4, 685–691. ISSN: 1053-1807.

Jenkinson, M, Bannister, P, Brady, M, and Smith, S (Oct. 2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage* 17.2, 825–841. ISSN: 1053-8119.

Johnson, RA and Wichern, DW (2007). *Applied Multivariate Statistical Analysis*. 6th. Prentice Hall.

Kaczkurkin, AN, Moore, TM, Ruparel, K, Ciric, R, Calkins, ME, Shinohara, RT, Elliott, MA, Hopson, R, Roalf, DR, Vandekar, SN, and others (2016). Elevated Amygdala Perfusion Mediates Developmental Sex Differences in Trait Anxiety. *Biological Psychiatry* 80.10, 775–785.

Kim, J, Wozniak, JR, Mueller, BA, Shen, X, and Pan, W (Nov. 2014). Comparison of statistical tests for group differences in brain functional networks. *NeuroImage* 101, 681–694.

Klein, A, Andersson, J, Ardekani, BA, Ashburner, J, Avants, B, Chiang, MC, Christensen, GE, Collins, DL, Gee, J, Hellier, P, and others (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage* 46.3, 786–802.

Klein, D, Rotarska-Jagiela, A, Genc, E, Sritharan, S, Mohr, H, Roux, F, Han, CE, Kaiser, M, Singer, W, and Uhlhaas, PJ (Jan. 2014). Adolescent Brain Maturation and Cortical Folding: Evidence for Reductions in Gyrification. *PLoS ONE* 9.1, e84914.

Krain, AL and Castellanos, FX (Aug. 2006). Brain development and ADHD. *Clinical Psychology Review*. Attention Deficit Hyperactivity Disorder From A Neurosciences And Behavioral Approach 26.4, 433–444. ISSN: 0272-7358.

Lehmann, EL and Romano, JP (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Levitt, P (Oct. 2003). Structural and functional maturation of the developing primate brain. *The Journal of Pediatrics*. Mechanisms of Action of LCPUFA Effects on Infant Growth and Neurodevelopment 143.4, Supplement, 35–45. ISSN: 0022-3476.

Luders, E, Narr, K, Thompson, P, Rex, D, Woods, R, DeLuca, H, Jancke, L, and Toga, A (Apr. 2006). Gender effects on cortical thickness and the influence of scaling. en. *Human Brain Mapping* 27.4, 314–324. ISSN: 1097-0193.

Madsen, BE and Browning, SR (Feb. 2009). A groupwise association test for rare mutations using a weighted sum statistic. eng. *PLoS genetics* 5.2, e1000384. ISSN: 1553-7404.

Mangin, JF, Riviere, D, Cachia, A, Duchesnay, E, Cointepas, Y, Papadopoulos-Orfanos, D, Collins, D, Evans, A, and Regis, J (Aug. 2004). Object-Based Morphometry of the Cerebral Cortex. en. *IEEE Transactions on Medical Imaging* 23.8, 968–982. ISSN: 0278-0062.

Marcus, R, Eric, P, and Gabriel, KR (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63.3, 655–660.

Matsuzawa, J, Matsui, M, Konishi, T, Noguchi, K, Gur, RC, Bilker, W, and Miyawaki, T (Apr. 2001). Age-related Volumetric Changes of Brain Gray and White Matter in Healthy Infants and Children. en. *Cerebral Cortex* 11.4, 335–342. ISSN: 1047-3211, 1460-2199.

McCullagh, P and Nelder, JA (1989). *Generalized Linear Models*. English. 2nd. Boca Raton: Chapman and Hall/CRC. ISBN: 978-0-412-31760-6.

Ouyang, X, Chen, K, Yao, L, Hu, B, Wu, X, Ye, Q, and Guo, X (Aug. 2015). Simultaneous changes in gray matter volume and white matter fractional anisotropy in Alzheimers disease revealed by multimodal CCA and joint ICA. *Neuroscience* 301, 553–562. ISSN: 0306-4522.

Pan, W (Sept. 2009). Asymptotic Tests of Association with Multiple SNPs in Linkage Disequilibrium. *Genetic epidemiology* 33.6, 497–507. ISSN: 0741-0395.

Pan, W, Kim, J, Zhang, Y, Shen, X, and Wei, P (Aug. 2014). A Powerful and Adaptive Association Test for Rare Variants. en. *Genetics* 197.4, 1081–1095. ISSN: 0016-6731, 1943-2631.

Pinkham, A, Loughead, J, Ruparel, K, Wu, WC, Overton, E, Gur, R, and Gur, R (Oct. 2011). Resting quantitative cerebral blood flow in schizophrenia measured by pulsed arterial spin labeling perfusion MRI. eng. *Psychiatry Research* 194.1, 64–72. ISSN: 0165-1781.

Press, WH, Teukolsky, SA, Vetterling, WT, and Flannery, BP (2007). *Numerical recipes: the art of scientific computing*. Cambridge University Press, New York.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rao, CR (1948). “Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation”. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 44, 50–57.

Rapoport, JL, Giedd, JN, and Gogtay, N (Dec. 2012). Neurodevelopmental model of schizophrenia: update 2012. en. *Molecular Psychiatry* 17.12, 1228–1238. ISSN: 1359-4184.

Raznahan, A, Shaw, P, Lalonde, F, Stockman, M, Wallace, GL, Greenstein, D, Clasen, L, Gogtay, N, and Giedd, JN (May 2011). How Does Your Cortex Grow? en. *The Journal of Neuroscience* 31.19, 7174–7177. ISSN: 0270-6474, 1529-2401.

Reiss, PT and Ogden, RT (Mar. 2010). Functional Generalized Linear Models with Images as Predictors. en. *Biometrics* 66.1, 61–69. ISSN: 1541-0420.

Romano, JP and Wolf, M (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association* 100.469, 94–108.

Ronan, L, Voets, NL, Hough, M, Mackay, C, Roberts, N, Suckling, J, Bullmore, E, James, A, and Fletcher, PC (Oct. 2012). Consistency and interpretation of changes in millimeter-scale cortical intrinsic curvature across three independent datasets in schizophrenia. *NeuroImage* 63.1, 611–621. ISSN: 1053-8119.

Sarkar, SK and Chang, CK (1997). The Simes Method for Multiple Hypothesis Testing With Positively Dependent Test Statistics. *Journal of the American Statistical Association* 92.440, 1601–1608. ISSN: 0162-1459.

Satterthwaite, TD, Shinohara, RT, Wolf, DH, Hopson, RD, Elliott, MA, Vandekar, SN, Ruparel, K, Calkins, ME, Roalf, DR, Gennatas, ED, and others (2014a). Impact of puberty on the evolution of cerebral perfusion during adolescence. *Proceedings of the National Academy of Sciences* 111.23, 8643–8648.

Satterthwaite, TD, Elliott, MA, Ruparel, K, Loughead, J, Prabhakaran, K, Calkins, ME, Hopson, R, Jackson, C, Keefe, J, Riley, M, Mentch, FD, Sleiman, P, Verma, R, Davatzikos, C, Hakonarson, H, Gur, RC, and Gur, RE (Feb. 2014b). Neuroimaging of the Philadelphia Neurodevelopmental Cohort. *NeuroImage* 86, 544–553. ISSN: 1053-8119.

Satterthwaite, TD, Connolly, JJ, Ruparel, K, Calkins, ME, Jackson, C, Elliott, MA, Roalf, DR, Hopson, R, Prabhakaran, K, Behr, M, and others (2016). The Philadelphia Neurodevelopmental Cohort: a publicly available resource for the study of normal and abnormal brain development in youth. *Neuroimage* 124, 1115–1119.

Scheipl, F, Greven, S, and Kchenhoff, H (Mar. 2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis* 52.7, 3283–3299. ISSN: 0167-9473.

Schott, JR (2016). *Matrix analysis for statistics*. 3rd ed. Wiley series in probability and statistics. Hoboken, New Jersey: Wiley. ISBN: 1-119-09248-5.

Selemon, LD, Rajkowska, G, and Goldman-Rakic, PS (Mar. 1998). Elevated neuronal density in prefrontal area 46 in brains from schizophrenic patients: Application of a three-dimensional, stereologic counting method. en. *The Journal of Comparative Neurology* 392.3, 402–412. ISSN: 1096-9861.

Shaw, P, Kabani, NJ, Lerch, JP, Eckstrand, K, Lenroot, R, Gogtay, N, Greenstein, D, Clasen, L, Evans, A, Rapoport, JL, Giedd, JN, and Wise, SP (Apr. 2008). Neurodevelopmental trajectories of the human cerebral cortex. eng. *The Journal of neuroscience: the official journal of the Society for Neuroscience* 28.14, 3586–3594. ISSN: 1529-2401.

Shaw, P, Malek, M, Watson, B, Sharp, W, Evans, A, and Greenstein, D (Aug. 2012). Development of Cortical Surface Area and Gyrification in Attention-Deficit/Hyperactivity Disorder. *Biological Psychiatry*. Oxytocin and Social Bonds in Development 72.3, 191–197. ISSN: 0006-3223.

Silver, M, Montana, G, and Nichols, TE (Jan. 2011). False positives in neuroimaging genetics using voxel-based morphometry data. *NeuroImage* 54.2, 992–1000. ISSN: 1053-8119.

Singh, V, Chertkow, H, Lerch, JP, Evans, AC, Dorr, AE, and Kabani, NJ (2006). Spatial patterns of cortical thinning in mild cognitive impairment and Alzheimer's disease. *Brain* 129.11, 2885–2893.

Smith, M and Fahrmeir, L (June 2007). Spatial Bayesian Variable Selection With Application to Functional Magnetic Resonance Imaging. *Journal of the American Statistical Association* 102.478, 417–431. ISSN: 0162-1459.

Sowell, ER, Peterson, BS, Kan, E, Woods, RP, Yoshii, J, Bansal, R, Xu, D, Zhu, H, Thompson, PM, and Toga, AW (July 2007). Sex Differences in Cortical Thickness Mapped in 176 Healthy Individuals between 7 and 87 Years of Age. en. *Cerebral Cortex* 17.7, 1550–1560. ISSN: 1047-3211, 1460-2199.

Srivastava, MS (Oct. 2003). Singular Wishart and multivariate beta distributions. *The Annals of Statistics* 31.5, 1537–1560. ISSN: 0090-5364, 2168-8966.

Steen, RG, Mull, C, Mcclure, R, Hamer, RM, and Lieberman, JA (June 2006). Brain volume in first-episode schizophrenia Systematic review and meta-analysis of magnetic resonance imaging studies. en. *The British Journal of Psychiatry* 188.6, 510–518. ISSN: 0007-1250, 1472-1465.

Sun, W, Reich, BJ, Tony Cai, T, Guindani, M, and Schwartzman, A (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B* 77.1, 59–83.

Toro, R and Burnod, Y (Dec. 2005). A Morphogenetic Model for the Development of Cortical Convolutions. en. *Cerebral Cortex* 15.12, 1900–1913. ISSN: 1047-3211, 1460-2199.

Tustison, NJ, Avants, BB, Cook, PA, Zheng, Y, Egan, A, Yushkevich, PA, and Gee, JC (2010). N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging* 29.6, 1310–1320.

Vaart, AW Van der (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Van Essen, DC (Jan. 1997). A tension-based theory of morphogenesis and compact wiring in the central nervous system. *Nature* 385.6614, 313–318. ISSN: 0028-0836.

Vandekar, SN, Shinohara, RT, Raznahan, A, Roalf, DR, Ross, M, DeLeo, N, Ruparel, K, Verma, R, Wolf, DH, Gur, RC, Gur, RE, and Satterthwaite, TD (Jan. 2015). Topologically Dissociable Patterns of Development of the Human Cerebral Cortex. en. *The Journal of Neuroscience* 35.2, 599–609. ISSN: 0270-6474, 1529-2401.

Wagstyl, K (June 2015). *Non-uniform cortical thinning in schizophrenia: are sulci worse off?* Paper. Honolulu, HI, USA.

Wang, L, Goldstein, FC, Veledar, E, Levey, AI, Lah, JJ, Meltzer, CC, Holder, CA, and Mao, H (2009). Alterations in cortical thickness and white matter integrity in mild cognitive impairment measured by whole-brain cortical thickness mapping and diffusion tensor imaging. *American Journal of Neuroradiology* 30.5, 893–899.

Westfall, PH and Young, SS (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Vol. 279. John Wiley & Sons.

White, T, Andreasen, NC, Nopoulos, P, and Magnotta, V (Aug. 2003). Gyrification abnormalities in childhood- and adolescent-onset schizophrenia. *Biological Psychiatry* 54.4, 418–426. ISSN: 0006-3223.

Wilson, EB (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22.158, 209–212.

Winkler, AM, Ridgway, GR, Webster, MA, Smith, SM, and Nichols, TE (May 2014). Permutation inference for the general linear model. *NeuroImage* 92, 381–397. ISSN: 1053-8119.

Winkler, AM, Ridgway, GR, Douaud, G, Nichols, TE, and Smith, SM (Nov. 2016). Faster permutation inference in brain imaging. *NeuroImage* 141, 502–516. ISSN: 1053-8119. DOI: 10.1016/j.

neuroimage.2016.05.068. URL: http://www.sciencedirect.com/science/article/pii/S1053811916301902 (visited on 07/20/2017).

Wood, SN (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B* 73.1, 3–36.

Wu, MC, Lee, S, Cai, T, Li, Y, Boehnke, M, and Lin, X (July 2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics* 89.1, 82–93. ISSN: 0002-9297.

Xu, G, Lin, L, Wei, P, and Pan, W (Sept. 2016). An adaptive two-sample test for high-dimensional means. *Biometrika* 103.3, 609–624. ISSN: 0006-3444.

Yeo, IK and Johnson, RA (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* 87.4, 954–959.

Yun, HJ, Im, K, Jin-Ju Yang, Yoon, U, and Lee, JM (Feb. 2013). Automated Sulcal Depth Measurement on Cortical Surface Reflecting Geometrical Properties of Sulci. *PLoS ONE* 8.2, e55977.

Zilles, K, Palomero-Gallagher, N, and Amunts, K (Jan. 2013). Development of cortical folding during evolution and ontogeny. English. *Trends in Neurosciences* 36.5, 275–284. ISSN: 0166-2236.