Bayesian Analysis of Neuroimage Data Using Gaussian Process Priors

by

Andrew S. Whiteman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2022

Doctoral Committee:

Professor Timothy D. Johnson, Co-Chair
Professor Jian Kang, Co-Chair
Professor Jeffrey A. Fessler
Professor Peter X.-K. Song

Andrew S. Whiteman
awhitem@umich.edu
ORCID iD: 0000-0002-5107-0506

DEDICATION

To those of my family who have supported me in this process.
I would not have grown without roots.

ACKNOWLEDGMENTS

I would like to thank the many people that have been so influential in my graduate career. Foremost, my advisors Drs. Jian Kang and Timothy D. Johnson have given me outstanding mentorship and support. Throughout your graduate studies, you collect the voices of your mentors in your mind, and theirs' will continue to inspire and shape the way I think about statistics and statistical computation for years to come. Once during a meeting with Tim and Jian I started to describe a numerical issue I seemed to be having with evaluation of a very high-dimensional log density. This calculation involved a sum over millions of variables, and I was loosing precision in the total. Tim—of course—had encountered this problem before and immediately pointed me to a fast and elegant algorithm to correct the accumulation of decimal error [89]. After the meeting I applied the correction and researched several alternative methods, profiling these for accuracy and speed. In our next meeting, I had just started to present and compare these various correction algorithms when Jian began to freestyle. Off the top of his head, he came very near to inventing what I was about to describe as the most numerically accurate method I had come across. I cannot help but smile and think how well this story summarizes my experience working with Tim and Jian.

I must greatly thank my committee members Drs. Peter X.-K. Song and Jeffrey A. Fessler for their interest in and support of my work. Having also known Peter in his role at the head of the classroom, his enthusiasm for statistics and teaching is infectious. This work also would not have been possible without the guidance and deep knowledge of Dr. Andreas J. Bartsch, who envisioned the use of data that inspired Chapter 2. Andreas is exceptionally kind and always quick to detail his insights and share his extensive neuroradiological expertise. I similarly thank Mike Angstadt, Dr. Chandra Sripada, and Dr. Mary Heitzeg, who oversaw preprocessing of the data we used to illustrate our method in Chapter 4, and who provided helpful feedback on our work.

I would also like to acknowledge a few members of the Michigan Biostatistics community who especially shaped my experience as a student. My many thanks go to Drs. Thomas Braun, Rod Little, Bhramar Mukherjee, Brisa Sanchez, and Ananda Sen for their wonderful instruction; and to Dr. Michael Boehnke and Nicole Fenech for their amazing support of students in the department.

I am very grateful to have spent time in the company of so many talented Michigan students. To Abhay, David, Elizabeth, Emily M., Emily R., Fatema, Holly, Jon, Josh, Meg, Nicky, Pedro, Stephen, Tianwen, members of the Kang lab, and to my other peers: I have learned a lot from all

of you and value your friendship. To Adam; Allen; Kelly, Juan, and MB; Leo: you all have been constants over the past few, sometimes stranger-than-fiction years. Thank you for being there. And of course my heartfelt gratitude to Emily *et al.* without whom I likely would have finished writing this dissertation either more quickly or not at all.

Last but not least, I credit my undergraduate mentors Dr. Chantal Stern and Dr. Karin Schon, who sparked my interest in imaging. Without their early encouragement it is doubtful I would have found my way to biostatistics in the first place.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ACRONYMS

ABCD Adolescent Brain and Cognitive Development (Study)

ABIDE Autism Brain Imaging Data Exchange

ADI-R Autism Diagnostic Interview-Revised

ADOS Autism Diagnostic Observation Schedule

ASD Autism Spectrum Disorder

BOLD Blood Oxygen Level Dependent (Signal)

C-PAC Configurable Pipeline for Analysis of Connectomes

DFT Discrete Fourier Transform

DIC Deviance Information Criterion

fMRI Functional Magnetic Resonance Imaging

FWHM Full-Width-at-Half-Maximum

GE-EPI Gradient Echo-Echo Planar Imaging

GLM General Linear Model

HMC Hamiltonian Monte Carlo

KDE Kernel Density Estimate

LSBP Logistic Stick-Breaking Process

MAP Maximum a Posteriori

MCE Minimum Contrast Estimation

MCMC Markov Chain Monte Carlo

MI Mutual Information

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

RMSE Root Mean Squared Error

SNR Signal-to-Noise Ratio

SPM Statistical Parametric Map

SVC Spatially Varying Coefficient

ABSTRACT

Magnetic Resonance Imaging (MRI) is a foundational tool for medical and academic research. Functional MRI (fMRI) and human brain research, for example, have become nearly synonymous phrases. MRI results in a dense, high-dimensional, highly correlated 3D or 4D datatype only digestible with concerted statistical effort. This dissertation focuses on developing new semiparametric Bayesian models and computational techniques to cope with some of the challenges that arise with fMRI data.

The first project (Chapter 2) presents a model designed to integrate presurgical fMRI data collected at two different spatial resolutions. Modern neuroradiologists use fMRI to map patient-specific functional neuroanatomy to assist in presurgical planning. This application requires a high degree of spatial precision, but in practice the fMRI signal-to-noise ratio decreases with increasing spatial resolution. To mitigate this issue, our collaborator collected functional scans of preoperative patients at high and low spatial resolutions. The data inherently exhibit different levels of noise and lack a common spatial support, rendering them difficult to combine in a straightforward manner. We solve this problem by modeling the mean image intensity function of both data sources using a Gaussian process and develop a scalable posterior computation algorithm based on Riemann manifold Hamiltonian Monte Carlo methods. We show in simulation our method enables more accurate inference on image mean intensity than single-resolution alternatives, and further illustrate our approach in analyses of preoperative patient images.

The second project (Chapter 3) is motivated by studies where heterogeneous latent imaging subgroup effects may be present in the study population. We propose a Bayesian semiparametric hierarchical model for image-on-scalar regression with subgroup detection. We model the mean intensity of imaging outcomes with a mixture of spatially varying coefficient (SVC) regression models, and take into account spatial dependence in the SVCs with Gaussian processes. Additional individual-level covariates are used to inform the mixing distribution via a logistic stick-breaking process prior. This class of prior admits individual-specific mixture weights and induces correlation in mixture component assignments between individuals with similar covariate profiles. We show through simulation our model can lead to superior clustering and feature estimation compared to common unsupervised methods. Further, we illustrate our method via analysis of resting-state fMRI data from the Autism Brain Imaging Data Exchange (ABIDE) study.

In the third project (Chapter 4), we address an important issue in neuroimaging research: improving spatial modeling of group-level effects of interest on the cortical surface. A state-of-the-art image preprocessing tool computes cross-subject alignment of cortical features by first mapping each hemisphere of the brain onto a sphere. Critically, this procedure enables a measure of great-circle distance between cortical points. Geodesic distances along the cortical surface are more biologically meaningful than the classically used Euclidean distance in 3D space. We propose a Bayesian spatially varying coefficient model for imaging outcome data observed at locations on a sphere, and use Gaussian processes to model the probability law governing the regression coefficient functions. We consider different approaches to approximate posterior inference with our model and compare performance against standard vertex-wise analyses. Finally, we illustrate our method in an analysis of fMRI task contrast data from a large cohort of children in the Adolescent Brain Cognitive Development (ABCD) study.

CHAPTER 1

Introduction

1.1 Functional Magnetic Resonance Imaging (fMRI): Data and Standard Analysis Methods

1.1.1 FMRI data

Since its invention in the early 1970's [119], magnetic resonance imaging (MRI) has been widely used not only as clinical tool, but also as a foundational instrument of neurobiological and neuropsychological research. MR images are captured using a series of magnetic pulses alternating at precise frequencies. Briefly, these magnetic pulses are carefully engineered to interact with the inherent spin of protons present in biological tissue [125]. Hydrogen-1 nuclei, for example, are abundant in water and exhibit a natural spin. In different tissues, some proportion of present hydrogen-1 nuclei can be induced to align their spin parallel to the direction of a strong magnetic field. Synchronization of these nuclear spins then contributes to a bulk magnetization effect in the tissue. If the initial external magnetic field is then perturbed by a second, spatially orthogonal magnetic field, proton spin will gradually fall out of alignment with the initial field. This process is termed "relaxation," and forms the basis of a nuclear magnetic signal that can be detected by special receiver coils. Measurement of this signal can be used to reconstruct three-dimensional images that relate to proton densities and the magnetic susceptibility of different tissue types. Spatial image reconstruction is made possible through application of varied magnetic field gradients. These magnetic gradients are designed such that induced spatio-temporal fluctuations in the nuclear relaxation-related signal can be processed and recorded as the Fourier transform of a recognizable brain image, for example. Interested readers can find a much more comprehensive introduction to MR physics in [125]. Modern scanners are engineered to produce powerful magnetic fields and can measure and reconstruct such images with a high signal-to-noise ratio (SNR) [92]. MR scanners can in practice turn this SNR to the advantage of images with higher spatial and/or temporal resolution. For example, high-field seven Tesla magnets have been used to capture

structural images of the *in vivo* human brain with amazingly high spatial resolution. In this context, image voxel (volumetric pixel) sizes have been achieved as small as $250 \times 250 \times 250$ microns [e.g., 106].

Functional MRI (fMRI) constitutes a procedure for collecting a series of MR images of tissue over time, where a contrast agent can be used to enhance temporal changes in the signal [e.g. 9]. The blood oxygen level dependent (BOLD) signal commonly used in fMRI is an endogenous contrast that measures the ratio of oxygenated and deoxygenated hemoglobin in the bloodstream [120]. Measuring the BOLD signal comprises a minimally invasive technique researchers can use to study metabolic activity in brain tissue, for example, in response to patterns of stimuli [138]. In the brain, the BOLD signal is not directly related to neuronal activity, but rather acts as a correlate of local field potentials, or the concerted activity of many nearby neurons sending and receiving electrical currents (action potentials) within a particular frequency band [102]. Researchers may study the relationship between this measure of local brain metabolism (or "activity," loosely) and behavioral output by scanning participants instructed to perform some experimentalized task. An fMRI data set collected this way will typically consist of several hundred to several thousand time series images per participant. The temporal resolution of fMRI is such that a single brain image acquisition typically takes around two seconds or less to complete [see e.g., 3, 144]. How the image intensity, measured at tens of thousands to hundreds of thousands of voxels, varies across the timeseries can then be summarized statistically. Often, the analyst's goal is to infer regions of the brain that share common task-related activation patterns across a group of participants.

1.1.2 FMRI preprocessing

For group-level inference derived from fMRI data to make any sense in practice, the raw data must at minimum be aligned to a common coordinate system and normalized in intensity. It is not the object of this dissertation to make a full review of fMRI preprocessing methods. Rather we acknowledge, broadly, that there are several important limitations to the data collection process that must be accounted for via preprocessing, and that many methods have been developed to accomplish these tasks. In general, fMRI preprocessing methods are not universally agreed upon within the field, and altering the preprocessing protocol can have substantial impact on final results [see e.g., 98, for an analytically-driven review].

A tremendous amount of work has contributed to the development of algorithms for within and between participant image alignment and spatial feature normalization [e.g., 46, 47, 84, 137]. Intensity normalization, however, is a somewhat more open question in the scope of modern large scale multi-cohort, multi-site imaging studies [2, 153]. Data collected in stages, on different scanners, or with different scanning protocols can result in a mishmash of different baseline image

intensities, SNRs, etc. across participants, all of which may be artifactual and mask true signal. The current common practice for multi-site studies, for example, is simply to homogenize the mean and variance of the data across collection sites. Image intensity normalization in this context is still an area of active research, however [e.g., 25]. With these types of issues in mind, large-scale imaging collective studies sometimes coordinate their fMRI collection and preprocessing protocols [e.g., 67] so that the raw data ideally have less heterogeneity prior to preprocessing.

1.1.3 Standard applied methods for fMRI analysis

The canonical framework for fMRI analysis was created by Karl Friston and colleagues in the early 1990s. The authors christened their approach "statistical parametric mapping" [49]. The idea behind this framework is very straightforward: a series of statistical models are fit to the data associated with each voxel in the brain, reducing the data to a summary or test statistic map that characterizes the statistical evidence for a given hypothesis at a local level. In common use cases, these "statistical parametric maps" (SPMs) might be constructed with t or z-statistic values at every voxel, representing, loosely, the relative strength of evidence in favor of the experimental null hypothesis. In this case, the SPM might represent on the order of 100,000 null hypothesis tests, and so some multiplicity adjustment [e.g., 10, 117, 181] is usually necessary to make valid claims about voxels where the researcher rejects the null hypothesis.

Following the SPM framework, a typical task-based fMRI analysis (of preprocessed data) might proceed in two stages as follows. In the first stage, researchers construct a model for the within subject time series data. This is usually accomplished with a general linear regression model with some autoregressive assumption on the model errors (AR(1), say). For participant i, the mean model design matrix, $\boldsymbol{Z}_i$, may denote a $(T \times Q)$ matrix with rows corresponding to T time points and columns corresponding to Q predictors. The $\boldsymbol{Z}_i$ will typically contain a set of indicators that mark task-on/task-off blocks in the time series convolved with a model for the hemodynamic response function [e.g., 50]. Design matrices will also usually contain a set of participant-specific nuisance terms to help regress out to potential sources of artifactual signal such as spikes due to head motion. Typically, the models for the time series data will be fit in "embarrassingly parallel" fashion by considering both voxels and patients as independent units of analysis. Models are then fit to the data within each unit marginally.

Throughout this first stage, the coefficients on the task-based regressors are of primary interest. Spatial regularization of the effects of interest is usually accomplished indirectly by applying a spatial smoothing kernel to the timeseries data as a preprocessing step. For the sake of further explanation, let $\gamma_i(\mathbf{v})$ denote the regression coefficient for a task-on term for patient i at voxel $\mathbf{v}$. The $\gamma_i(\mathbf{v})$ can be interpreted as the average relative response in brain activity related to task per-

formance. In a second stage of analysis, researchers then gather the $[\gamma_i(\mathbf{v})]_{i=1}^N$ together and regress these on a group-level design matrix, $\boldsymbol{X}$. In this example, $\boldsymbol{X}$ represents an $(N \times P)$ matrix, where N is the number of participants, each measured on P corresponding regression predictors. Often, $\boldsymbol{X}$ may simply contain a group-level intercept term. The goal of this straightforward two stage strategy is to recapitulate inference from voxel-wise mixed-effects models in a computationally efficient way. In general, assuming the errors in the first stage analyses are conditionally Gaussian, then the second stage analysis can exactly recover the desired mixed effects estimates by using weighted regression [e.g., 93]. In practice, this assumption can be reasonable, and the $\gamma_i(\mathbf{v})$ can, if desired, be weighted by the inverse of their variance estimates in the second stage analysis. It can also be perfectly reasonable, however, to omit inverse variance weighting during the second stage. Use of simplified assumptions in second stage modeling has been carefully considered and validated by Mumford and Nichols [115].

The brilliance of the SPM framework lies in its computational efficiency and straightforward implementation. Throughout this dissertation, we work entirely within the two stage analysis paradigm and specifically attempt to improve upon the second, group-level stage of the analysis. Among the most immediate statistical dissatsifactions with the classical method is the loss of power that can result from (i) lack of explicit spatial regularization, and (ii) the voxel-wise multiple testing formulation. One of the central themes of this dissertation is that it can be beneficial to construct an explicit model for spatial dependence in effects of interest rather than rely on a somewhat *ad hoc* smoothing step. In the subsequent chapters, we build different Bayesian hierarchical models to address distinct clinical and research questions; underlying each proposed method is the idea to specify a probability law for governing spatial regression coefficient processes. We accomplish modeling of spatial dependence through use of Gaussian process priors, and explore several different methods to render posterior computation tractable in this setting. Chapter 4 shows to what extent our fully spatial modeling strategy can support enhanced statistical power relative to the classical method, and eliminate the need for additional multiple comparisons corrections.

1.2 Gaussian processes

In this section, we discuss notation for general Gaussian process models, and outline the typical computational strategy used to evaluate them.

1.2.1 Definition and notation

Gaussian processes represent an extension of the Gaussian distribution to an infinite dimensional setting. As such, they can be used to specify a probability distribution on a functional space.

Throughout this dissertation, we will borrow the notational convention in [134], and write for example,

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}')), \tag{1.1}$$

to mean that we model the distribution of a function $f(\cdot)$ with a Gaussian process. In (1.1) we take $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ to represent generic inputs to $f(\cdot)$, $m(\cdot)$, and $K(\cdot, \cdot)$. Although generalizations to multivariate processes exist, we will take $f(\cdot)$ to denote a univariate process that maps domain $\mathcal{X}$ onto the real line ($f : \mathcal{X} \to \mathbb{R}$).

Gaussian processes are characterized by their mean and covariance functions: $m(\cdot)$ and $K(\cdot, \cdot)$ in (1.1) above, respectively. More formally, we can write that for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ we have $\mathbb{E}\, f(\boldsymbol{x}) = m(\boldsymbol{x})$, and $\mathbb{E}\{f(\boldsymbol{x}) - m(\boldsymbol{x})\}\{f(\boldsymbol{x}') - m(\boldsymbol{x}')\} = K(\boldsymbol{x}, \boldsymbol{x}')$, where $\mathbb{E}(\cdot)$ denotes the expectation under (1.1). By this definition, for any finite set of unique $\{\boldsymbol{x}_i\}_{i=1}^n$, writing (1.1) as above implies that $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^\mathsf{T}$ is multivariate Gaussian distributed,

$$\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}),$$

where $\boldsymbol{m} = [m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_n)]^\mathsf{T}$, and $\boldsymbol{K} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n$. This notation suppresses potential dependence on any hyperparameters in $m(\cdot)$ and $K(\cdot, \cdot)$. Throughout this dissertation, we will as a rule formulate our problems so that it makes sense to take $m(\cdot)$ to be the zero function ($m(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$). It is also quite common to use covariance functions of the form $K(\boldsymbol{x}, \boldsymbol{x}') = \tau^2 \rho(\|\boldsymbol{x} - \boldsymbol{x}'\|; \boldsymbol{\theta})$, where τ^2 is the marginal variance of $f(\cdot)$, and $\rho(\cdot; \boldsymbol{\theta})$ is a stationary, positive definite correlation function that depends on additional parameters $\boldsymbol{\theta}$.[1] Due to the substantial history of Gaussian smoothing in applied neuroimaging research, we will take the radial basis covariance function,

$$K(\boldsymbol{x}, \boldsymbol{x}') = \tau^2 \exp(-\psi \|\boldsymbol{x} - \boldsymbol{x}'\|^\nu), \qquad \tau^2, \psi > 0, \quad \nu \in (0, 2], \tag{1.2}$$

as a canonical example throughout, although many alternative choices exist (see for example the small compendium of spatial covariance functions listed in [6, pp. 25–26]). In addition, we work exclusively with Gaussian processes defined over a spatial input domain, where in this example $\mathcal{X}$ will be a stand-in for a subset of $\mathbb{R}^2$ or $\mathbb{R}^3$ where brain signals are measured

1.2.2 Computation

Here, we outline the typical approach to posterior computation with Gaussian process models using a toy example. In Fig. 1.1, we suppose that we have observed noisy data generated from some unknown, nonlinear regression function with independent stationary Gaussian errors. We write our

[1] Note that in Chapters 2 and 3 we parameterize our problems such that $\boldsymbol{\theta}$ includes τ^2, and thus denotes the hyperparameters of the covariance, not the correlation.

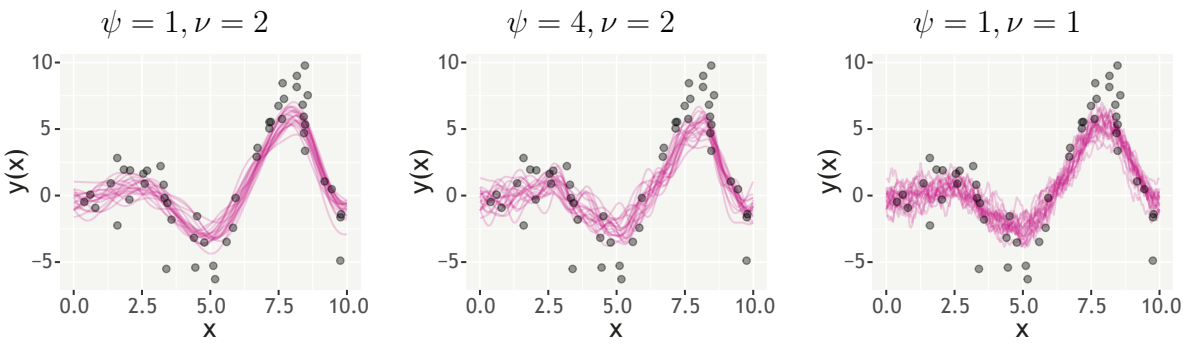


Figure 1.1: Gaussian process regression: toy example. In each subplot, observed data ($y(x_i)$; gray dots) are shown overlaid with 20 random draws from the posterior distribution of the regression function (magenta lines). In each case, the data are identical but posterior inference has been conditioned on distinct choices of the Gaussian process covariance hyperparameters.

corresponding functional regression model,

$$y(x_i) = f(x_i) + \epsilon(x_i), \qquad \epsilon(x_i) \sim \mathcal{N}(0, \sigma^2),$$

where $x_i \in \mathbb{R}$ for $i = 1, \ldots, n$. To complete a Bayesian hierarchical regression model for these data, we assume

$$f(x) \sim \mathcal{GP}(0, K(x, x')), \qquad \sigma^{-2} \sim \text{Gamma}(1, 0),$$

where $K(\cdot, \cdot)$ is the radial basis covariance as in (1.2) with $\tau^2 = 1$, and the error precision σ^{-2} is assigned an improper Gamma prior. In Fig. 1.1, we plot our toy data and overlay 20 draws (in magenta) from the posterior distribution of $f(\cdot)$ for a coarse set of choices for the bandwidth and exponent parameters, ψ and ν. In this limited example, although the individual samples of $f(\cdot)$ look quite different across the panels in the figure, posterior estimation of quantities like the pointwise mean and variance of $f(\cdot)$ at a given input are relatively insensitive to the covariance hyperparameters.

For any new point $\tilde{x} \in \mathbb{R}$, we can sample the corresponding $f(\tilde{x})$ from its full conditional distribution. Let $\boldsymbol{y} = [y(x_1), \ldots, y(x_n)]^\mathsf{T}$ denote the vector of n observed responses, and similarly let $\boldsymbol{x} = [x_1, \ldots, x_n]^\mathsf{T}$ denote the locations where we have observed $\boldsymbol{y}$. Also, as above let $\boldsymbol{K} = [K(x_i, x_j)]_{i,j=1}^n$ denote the $n \times n$ covariance matrix of $\boldsymbol{f} = [f(x_1), \ldots, f(x_n)]^\mathsf{T}$, and let $K(\boldsymbol{x}, \tilde{x}) = [K(x_i, \tilde{x})]_{i=1}^n$ denote the n-dimensional vector of cross-covariances between $\boldsymbol{f}$ and $f(\tilde{x})$. In this example, we have constructed a model with full posterior conjugacy so that the

conditional posterior of $f(\tilde{x})$ is trivially Gaussian with,

$$\begin{aligned}\mathbb{E}\{f(\tilde{x}) \mid \boldsymbol{y}, \cdot\} &= K^{\mathsf{T}}(\boldsymbol{x}, \tilde{x})(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{y}, \quad \text{and} \\ \operatorname{var}\{f(\tilde{x}) \mid \boldsymbol{y}, \cdot\} &= K(\tilde{x}, \tilde{x}) - K^{\mathsf{T}}(\boldsymbol{x}, \tilde{x})(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1}K(\boldsymbol{x}, \tilde{x}),\end{aligned} \tag{1.3}$$

using shorthand to express conditioning on σ^2, ψ, and ν.

Notice how both the mean and variance in (1.3) require decomposition or inversion of the $n \times n$ matrix $\boldsymbol{K} + \sigma^2 \boldsymbol{I}$. For n up to a few thousand, this computation can be handled efficiently using Cholesky decomposition and fast routines to solve triangular linear systems. For very large n, however, it can be impractical to even construct a dense $n \times n$ matrix, let alone compute its Cholesky decomposition (an $\mathcal{O}(n^3)$ operation). In the context of our applications in Chapters 2–4, n will equivalently be the number of spatial locations in an MR image where we observe brain data. This number will range from over 200,000 in Chapter 2 to around 30,000 in Chapter 4. In each chapter we explore a different computational scheme for dealing with data of this size. Chapter 2 uses a data augmentation approach based on the work of Wood and Chan [178]. In Chapter 3, we work with a low-rank projection of the spatial process [after e.g., 7], and in Chapter 4 we apply a sparse approximation of $\boldsymbol{K}^{-1}$ that retains a full-rank spatial process in the prior [following e.g., 35].

1.3 Dissertation outline

The rest of this dissertation is organized as follows. In Chapter 2, we describe a model for presurgical fMRI data designed to integrate information from scans collected at different spatial resolutions. This chapter is based on work published in [175]. We model z-statistic outcome images that summarize task related activation patterns and use Gaussian process regression to infer the mean spatial activation pattern at the highest available spatial resolution. The resulting information can be incorporated by neuroradiologists and neurosurgeons into presurgical planning to help navigate patients' individual functional neuroanatomy. In Chapter 3, we consider the classical second stage or group-level analysis paradigm in MRI analysis on the basis of the question: can we identify latent subgroups of participants across which imaging outcomes may differ in some systematic, meaningful way? We develop a semiparametric mixture of spatial regression models with mixture weights governed by a logistic stick-breaking process [136]. Spatial regression coefficient functions for each mixture component are modeled with Gaussian process priors. We apply this model to neurotyping Autism spectrum patients, and find subgroups related to component scores of an Autism Diagnostic Interview questionnaire [104]. In Chapter 4, we again consider group-level regression analyses for MRI data with the specific goal of precise estimation of spatially vary-

ing regression coefficient functions. At the time of writing, the companion paper for this chapter has been submitted for publication [176]. With this project, we also give special attention to the geometry of the cortical surface, and meaningful measures of distance in this context. We apply this method to analyze fMRI working memory task contrast data collected from over 3,000 children enrolled in the Adolescent Brain and Cognitive Development study. Finally, in Chapter 5 we conclude with general discussions of our proposed methods and considerations for future work.

CHAPTER 2

Bayesian Inference for Brain Activity from Functional Magnetic Resonance Imaging Collected at Two Spatial Resolutions

Neuroradiologists and neurosurgeons increasingly opt to use functional magnetic resonance imaging (fMRI) to map functionally relevant brain regions for noninvasive presurgical planning and intraoperative neuronavigation. This application requires a high degree of spatial accuracy, but the fMRI signal-to-noise ratio (SNR) decreases as spatial resolution increases. In practice, fMRI scans can be collected at multiple spatial resolutions, and it is of interest to make more accurate inference on brain activity by combining data with different resolutions. To this end, we develop a new Bayesian model to leverage both better anatomical precision in high resolution fMRI and higher SNR in standard resolution fMRI. We assign a Gaussian process prior to the mean intensity function and develop an efficient, scalable posterior computation algorithm to integrate both sources of data. We draw posterior samples using an algorithm analogous to Riemann manifold Hamiltonian Monte Carlo in an expanded parameter space. We illustrate our method in analysis of presurgical fMRI data, and show in simulation that it infers the mean intensity more accurately than alternatives that use either the high or standard resolution fMRI data alone.

2.1 Introduction

Neurosurgery presents a unique set of challenges to the operating surgeon. Treatment of brain tumors, for example, is handled primarily by surgical resection when possible. Gliomas are often infiltrative, however, and as a result may be impossible to remove entirely [88, 159]. Requiring precise structural and functional information, the neurosurgeon's goal is typically to resect as much of the tumor as possible while avoiding damage to surrounding healthy areas of brain tissue. Although the structure of the human brain shares a gross organization common across individuals, functional neuroanatomy may vary between patients and within regions [e.g. 94], highlighting the

need for within-patient precision. Here we propose a model that leverages the massive amount of spatial data available in individual functional magnetic resonance imaging (fMRI) scans to help guide presurgical planning by identifying functionally relevant brain regions in a patient-specific manner.

Traditionally, electrocortical interference is used to map brain functional organization during surgery [e.g. 28], but this procedure is highly invasive, lengthens surgery duration, and cannot be incorporated into presurgical planning [159]. Clinicians can also opt to use imaging methods to help inform patient-specific presurgical planning and intraoperative neuronavigation [e.g. 4, 118, 39, 150]. FMRI may be used, for example, to map patient-specific functional areas, but the data come with an inherent trade off. Surgeons would like to collect information that is spatially precise, but the fMRI signal-to-noise ratio (SNR) decreases as spatial resolution increases, potentially making functional mapping more difficult [16]. In practice, modern scanners are equipped to handle a variety of image resolutions by modifying magnetic pulse sequences, so radiologists are in principle able to collect any combination of scans advantageous for presurgical planning.

Our motivating datasets come from two separate fMRI experiments in which preoperative patients performed cognitive tasks chosen to localize brain regions involved in language processing (see sections 2.2 and 2.4 for details). Each individual patient was administered their task over two separate scanning runs, collected at different spatial resolutions. Details vary by patient, but in both instances one run was collected at "standard" spatial resolution with voxel (volumetric pixel) dimensions measuring approximately $3 \times 3 \times 3$ mm^3, and the other was collected at "high" spatial resolution with approximately $2 \times 2 \times 2$ mm^3 voxels. Raw image time series data were preprocessed using standard software [84, 179] to yield statistical parametric maps for each spatial resolution that summarized patients' fMRI activation over time. In this paper, we propose a new Bayesian model to integrate both sources of data, leveraging the anatomical/spatial precision of high resolution fMRI and the SNR of standard resolution fMRI for enhanced within-patient precision. The primary goal of our model is to reduce spatial noise while making inferential statements identifying functional regions at the highest resolution available. Conceptually, we accomplish this goal by modeling the mean intensity function of both data sources as a Gaussian process. Gaussian processes induce a probability measure on a functional space with distribution characterized by a mean and covariance function [134]. Conditional on the covariance function hyperparameters, which we estimate from data, we conduct fully Bayesian inference on the mean function measured at voxel locations in the high spatial resolution image.

In addition to spatial precision, computational complexity is also a major concern since excessive latency between preoperative scanning and a patient's actual surgery is undesirable. Computation with spatial Gaussian process models typically involve decomposition of an $n \times n$ matrix, where n is the number of spatial locations. Between the two image types there are over

200,000 unique spatial locations in each of our motivating datasets, rendering usual computational approaches to inference intractable in most computing environments. Here, we outline a modification of the typical Hamiltonian Monte Carlo (HMC) algorithm that makes this inference not only feasible but computationally efficient. To do so, we propose a dual resolution mapping prior that generalizes the existing Gaussian predictive process framework [e.g. 146, 7] to our setting with multiple data sources. Our algorithm further harnesses a parameter expansion idea from [178] to sample from the posterior using Riemann manifold Hamiltonian dynamics [62] in an ultrahigh dimensional parameter space.

Our model is related to existing literature from the field of spatial satistics that consider the "change of support problem" [e.g. 57, 53, 11]. Such models have been used, for example, to combine data from air pollution monitoring sites with simulations from physical models for prediction at unobserved locations and model validation. Studies such as these commonly model conditional relationships between data sources, for example by regressing measured air pollution onto physical model output. Our multi-resolution imaging paradigm is related in the sense that we would like to use standard resolution data to improve inference in high resolution space. This goal, however, is complicated by the fact that high and standard spatial resolution voxels in general only partially overlap with their neighbors in their complementary image (see Fig. 2.1). We will, however, take a different approach by modeling both sources of data as joint outcomes. Not only does this approach perhaps make more conceptual sense for modeling multiple image types, it permits flexible and natural reconfiguration in response to real world challenges. For example, if only one fMRI resolution or session is available presurgically, the missing data can be removed from the joint outcome. Though we discuss our method exclusively in a functional neuroimaging context, the method can easily generalize to other imaging modalities or indeed to spatial data with mixed supports more broadly.

Whereas the inferential goal of most neuroimaging studies is to identify activated or deactivated brain regions while controlling the family-wise error rate, we take a somewhat different approach given specific presurgical needs. In a neurosurgical context, clinicians are typically more concerned with inaccurate labeling of functionally important tissue as unimportant. To this end, we adopt a decision theoretic rule from previous work to control the ratio of false negative to false positive errors [101, 100, 114]. We show in simulation that our dual resolution method achieves good accuracy for realistic effect sizes. Specifically, our method outperformed single spatial resolution alternatives in terms of both false negative and false positive error rates when the number of discoveries was fixed across methods. Software to fit the dual and single resolution models discussed in this paper to data stored in the NIfTI data standard [29] is available online at https://github.com/asw221/dualres.

The body of this paper contains descriptions of our motivating clinical datasets in section 2.2,

and a summary of the method we propose to handle the unique challenges of those data in section 2.2.1. In sections 2.2.2 and 2.2.3, we elaborate on our approach to enable precise estimation and computation in such a large parameter space. We discuss a strategy to conduct inference based on weighted trade offs between false negative and false positive errors in section 2.2.4. We quantify our method's performance against single resolution alternative methods in section 2.3. Section 2.4 reports on analyses of real patient data using our proposed method for dual resolution fMRI. Finally, we present an overall evaluation of our contributions in section 2.5.

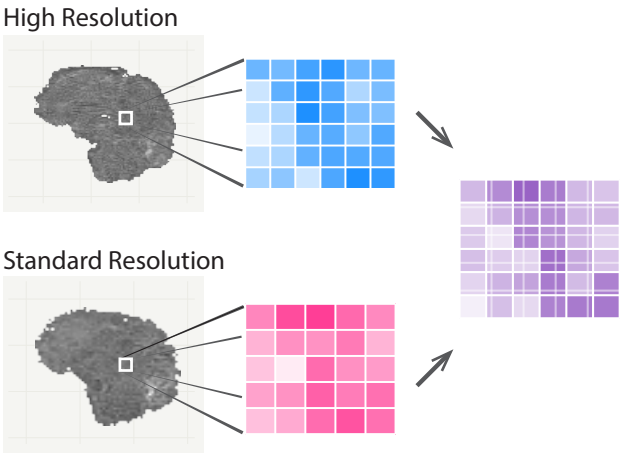


Figure 2.1: Schematic of aims and difficulties with integration of fMRI data collected at multiple spatial resolutions. Images collected at different resolutions exhibit inherently different levels of noise. We would like to reduce spatial noise while making inferential statements at the highest resolution available, but voxel locations may not align in general.

2.2 Data and methods

We developed the method presented here to analyze single-patient presurgical fMRI data collected at two spatial resolutions. Our first motivating dataset comes from a 62 year old right handed woman—"patient 1"—who presented difficulties with reading, finding, and comprehending words. This patient was subsequently found to have a tumor in her left middle and inferior temporal gyrus. Prior to surgery, the patient was scanned while performing a reading task to map brain areas associated with reading non-final embedded clause sentences and language processing. Scans were collected in two separate runs: once at standard $3 \times 3 \times 3.45$ mm^3 resolution ($64 \times 64 \times 48$ grid), and once at high $1.8 \times 1.8 \times 2.3$ mm^3 resolution ($120 \times 120 \times 62$ grid).

Our second motivating dataset comes from an 18 year old right handed woman—"patient 2"—who presented after a general seizure and was subsequently found to have a cavernoma in her left temporal lobe (see Appendix G for more detail). For cavernomas in critical areas, presurgi-

cal fMRI is considered one option—as with brain tumors—to map brain function noninvasively for presurgical planning and intraoperative neuronavigation. Patient 2 was also scanned prior to surgery while performing a language processing task. Her standard resolution data were collected with slightly smaller $3 \times 3 \times 3.3$ mm^3 voxels ($64 \times 64 \times 48$ grid), and her high resolution data with $1.8 \times 1.8 \times 2.2$ mm^3 voxels ($120 \times 120 \times 62$ grid). As in this patient, cavernomas typically cause profound T2*-weighted MR signal loss, with blooming into surrounding brain tissue. Signal loss is caused by abrupt differences in magnetic susceptibility in apposed tissues and is a common occurrence in clinical fMRI (e.g. intratumoral hemorrhages can cause similar dropout). We use this patient's data to illustrate our model's capacity to recover an estimate of activation in areas of such fMRI signal loss.

FMRI time series preprocessing without spatial smoothing was performed prior to our analysis using FSL software [84] and the FEAT tool [179]. As will become clear in section 2.2.1, our model imposes smoothness on the image mean function, and so we avoided smoothing the data during preprocessing (beyond the small amount of unavoidable smoothing that can occur when time series images from the two spatial resolutions are motion corrected and co-registered with one another). Smoothing is an otherwise ubiquitous step in typical fMRI pipelines, but over smoothing is not desirable for presurgical planning applications as it may reduce spatial precision by, for example, smearing activation into adjacent areas when the smoothing kernel is too wide. Data were corrected for motion and temporally high pass filtered, and marginal linear models were fit to the time series data at each voxel to create summary statistic maps of task-related activation.

Preprocessing resulted in one unsmoothed z-statistic contrast image for each fMRI resolution that summarized task-related activation over the course of each respective scan. We went on to use the generated test statistic maps as outcome data in our subsequent analysis, treating the images as noisy measures of true activation. Although we may find it beneficial to include both spatial and temporal data in our modeling framework in future work, the present model only explicitly represents a spatial process. As such, throughout the rest of this paper we will use "high resolution," for example, as a stand in for "high *spatial* resolution" etc. In the greater imaging community, however, "resolution" could in general relate to frequency of either spatial or temporal data collection, or both. We give additional details regarding patient data collection and image preprocessing in Appendix F.

2.2.1 Bayesian dual resolution mapping

Let $\mathcal{B}$ denote a generic brain image space, and let $B_h \subset \mathcal{B}$ and $B_s \subset \mathcal{B}$ denote the sets of spatial locations in the brain where high and standard resolution functional MRI data are collected, respectively. For reference, the number of voxels $|B_h| \approx 200,000$ in the high resolution image, and

$|B_s| \approx 50,000$ in the standard resolution image. Each atom $\mathbf{v} \in \mathcal{B}$ is a three dimensional vector of spatial coordinates relative to some origin point $\mathbf{v}_0 \in \mathcal{B}$; the Euclidean distance between any two points, $\mathbf{v}, \mathbf{v}' \in \mathcal{B}$ can be represented $\|\mathbf{v} - \mathbf{v}'\|_2$, and is typically measured in millimeters. Although data at a given voxel is associated with a small volume, we follow common practice and essentially treat that data as observed on location $\mathbf{v}$ exactly. In general, even voxels that overlap between the two image types may not have the same centers, so that the set of points in the intersection $B_h \cap B_s$ may be empty.

Conceptually, we motivate our proposed model as follows. Let $Y_h(\mathbf{v}_h)$ denote the high resolution imaging outcome at voxel $\mathbf{v}_h$, and let $Y_s(\mathbf{v}_s)$ denote the standard resolution imaging outcome at voxel $\mathbf{v}_s$. For the same patient performing the same cognitive task in the same scanner, we make the assumption that $Y_h(\mathbf{v}_h)$ and $Y_s(\mathbf{v}_s)$ are realizations from a unifying generative process. Let $\mathcal{N}(\mu, \sigma^2)$ denote a Gaussian distribution with mean μ and variance σ^2. We model the data as jointly Gaussian,

$$\begin{aligned} Y_h(\mathbf{v}_h) &\sim \mathcal{N}(\mu(\mathbf{v}_h), \sigma_h^2), \qquad \mathbf{v}_h \in B_h \\ Y_s(\mathbf{v}_s) &\sim \mathcal{N}(\mu(\mathbf{v}_s), \sigma_s^2), \qquad \mathbf{v}_s \in B_s \end{aligned} \tag{2.1}$$

where $\mu(\mathbf{v})$ represents the expected intensity of brain activity in voxel $\mathbf{v} \in \mathcal{B}$, and σ_h^2 and σ_s^2 are noise variances in the high and standard resolution images, respectively. Because our data were not smoothed, we modeled noise as a spatially independent and additive process. Given the known phenomenon that SNR increases with voxel volume [e.g. 16], we expect standard resolution images to be less noisy than high resolution images. We therefore adopted a weakly informative prior for the noise variances with the restriction $\sigma_h^2 > \sigma_s^2$:

$$\pi(\sigma_h^2, \sigma_s^2) \propto \sigma_h^{-2} \sigma_s^{-2} \, \mathbb{1}(0 < \sigma_s^2 < \sigma_h^2), \tag{2.2}$$

where $\mathbb{1}(\cdot) \in \{0, 1\}$ is the event indicator function ($\mathbb{1}(\mathcal{A}) = 1$ if $\mathcal{A}$ occurs, and 0 otherwise).

For functional maps, we are primarily interested in making inferences about the mean intensity function, $\mu(\cdot)$, to which we assign a mean zero Gaussian process prior,

$$\mu(\mathbf{v}) \sim \mathcal{GP}(0, K(\mathbf{v}, \mathbf{v}')). \tag{2.3}$$

In our formulation, the function $\mu(\cdot)$ captures all of the correlation between voxels and between the two images; conditional on $\mu(\cdot)$, $Y_h(\mathbf{v}_h)$ and $Y_s(\mathbf{v}_s)$ are mutually independent across all $\mathbf{v}_h \in B_h$ and $\mathbf{v}_s \in B_s$. We implicitly assume that the two brain images share a real-world coordinate system, and that $\mu(\mathbf{v})$ is correlated with $\mu(\mathbf{v}')$ if the distance $\|\mathbf{v} - \mathbf{v}'\|_2$ is small. A variety of preprocessing techniques have been developed to align 3D images and ensure the former assumption holds with

minimal error [e.g. 137, 84].

Since anatomical precision is paramount in our application, we would like to conduct inference on $\mu(\cdot)$ for all locations in B_h. To facilitate this goal while simultaneously modeling the cross correlation between $\mu(\cdot)$ evaluated on locations in B_h and B_s, we introduce a nonstationary covariance function to map between data sets. For any $\mathbf{v}, \mathbf{v}' \in \mathcal{B}$, let,

$$K(\mathbf{v}, \mathbf{v}') = \begin{cases} k(\mathbf{v}, \mathbf{v}') & \text{if } \mathbf{v}' \in B_h \\ w^{\mathsf{T}}(\mathbf{v}) k(B_h, \mathbf{v}') & \text{otherwise,} \end{cases} \tag{2.4}$$

where $w(\cdot)$ is a vector of weights in a finite basis (defined below; chosen so that the covariance function is symmetric for all $\mathbf{v}, \mathbf{v}' \in \mathcal{B}$), and $k(\cdot, \cdot)$ is some positive definite function with range $\mathbb{R}_{>0}$. In our application, we take $k(\cdot, \cdot)$ to be the isotropic radial basis function,

$$k(\mathbf{v}, \mathbf{v}') = \tau^2 \exp(-\psi \|\mathbf{v} - \mathbf{v}'\|_2^{\nu}), \qquad \tau^2, \psi > 0, \quad \nu \in (0, 2], \tag{2.5}$$

and extend the notation to apply to sets of locations so that $k(B_h, \mathbf{v}') = [k(\mathbf{v}_h, \mathbf{v}')]_{\mathbf{v}_h \in B_h}$ is a vector in $\mathbb{R}^{|B_h|}$. In (2.5), $\tau^2 > 0$ is the "partial sill" or marginal prior variance of $\mu(\cdot)$, the decay parameter $\psi > 0$ defines the correlation bandwidth, and $\nu \in (0, 2]$ is the kernel exponent or smoothness parameter. We define the covariance parameters $\boldsymbol{\theta} = (\tau^2, \psi, \nu)^{\mathsf{T}}$; ψ and ν, are commonly fixed prior to analyses, but because of the abundance of spatial data in even a single brain image, in practice we recommend estimating these parameters from data (see section 2.4.1 for details). The custom kernel in (2.4) was designed to approximate a "Gaussian parent process" [7] with isotropic radial basis covariance (2.5) everywhere in $\mathcal{B}$. We arrange our presentation here to make clear that we use (2.4) directly, and obtain exact inference with the prior (2.3) specified in this way. A careful choice of the weight function $w(\cdot)$, moreover, can render the problem more computationally tractable.

2.2.2 Construction of the covariance weights

By the definition of a Gaussian process, $\mu(\mathbf{v})$ and $\mu(\mathbf{v}')$ are jointly multivariate Gaussian distributed for any distinct locations $\mathbf{v}$ and $\mathbf{v}'$. As a result, Gaussian process models promote natural and flexible predictions of values of $\mu(\cdot)$ at unobserved locations. For arbitrary collections of locations $U = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathcal{B}$ and $V = \{\mathbf{v}'_1, \ldots, \mathbf{v}'_m\} \subset \mathcal{B}$, we define $\mu(U) = [\mu(\mathbf{v}_i)]_{i=1}^n$ as a vector in $\mathbb{R}^n$; $K(U, \mathbf{v}) = [K(\mathbf{v}_i, \mathbf{v})]_{i=1}^n$ as a vector in $\mathbb{R}^n$; and $K(U, V) = [K(\mathbf{v}_i, \mathbf{v}'_j)]_{i,j=1}^{n,m}$ as a matrix in $\mathbb{R}^{n \times m}$. If, for example, V is a set of observed locations, and U are unobserved locations, then conditional on $\mu(V)$, $\mu(U)$ is multivariate Gaussian distributed with mean $K(U, V)K(V, V)^{-1}\mu(V)$ and variance $K(U, U) - K(U, V)K(V, V)^{-1}K(V, U)$.

Since imaging data is collected on a dense grid we often have no need to predict outcomes at unobserved or non-brain locations, except in cases of signal loss or other artifact. We made use of the kriging or conditional distribution relationships above primarily to define the basis weight function $w(\cdot)$ to integrate information from both high and standard resolution images. We constructed the basis weights in (2.4) so that $w(\mathbf{v}) \approx K(B_h, B_h)^{-1}k(B_h, \mathbf{v})$, with the "approximate" relation explained below. This formulation allowed us to leverage the relationship that $w^{\mathsf{T}}(\mathbf{v})\boldsymbol{\mu}_h$ approximates the prior conditional expectation of $\mu(\mathbf{v})$ given $\boldsymbol{\mu}_h = \mu(B_h)$. As such, our construction in (2.4) generalizes a Gaussian predictive process framework [e.g. 146, 7] to our setting with multiple data sources by using B_h as a high-dimensional reference set. In general, within this framework we could have defined the weights $w(\cdot)$ based on any arbitrary set of knot locations $B_* \subset \mathcal{B}$. Since inference at a fine spatial scale typically requires a dense set of knot locations [e.g. 157], we preferred to define $w(\cdot)$ based on all of B_h.

Our covariance function in (2.4) effectively employs kriging methods to map $\mu(B_h)$ onto the locations in B_s so that the standard resolution data can still inform $\mu(B_h)$ in the posterior. Switching to vector notation, let $\boldsymbol{\mu}_s = \mu(B_s)$. We express the prior in (2.3),

$$\pi(\boldsymbol{\mu}_h, \boldsymbol{\mu}_s) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{K}_h & \boldsymbol{K}_{h,s} \\ \boldsymbol{K}_{s,h} & \boldsymbol{K}_{s,h}\boldsymbol{K}_h^{-1}\boldsymbol{K}_{h,s} \end{bmatrix}\right). \tag{2.6}$$

where $\boldsymbol{\mu}_h$ and $\boldsymbol{\mu}_s$ are the means of the high, and standard resolution images, respectively; we denote the marginal prior variance of $\boldsymbol{\mu}_h$ by $\boldsymbol{K}_h = K(B_h, B_h)$, the prior covariance of $\boldsymbol{\mu}_h$ and $\boldsymbol{\mu}_s$ by $\boldsymbol{K}_{h,s} = K(B_h, B_s)$, etc. The obvious difficulty working with (2.6) directly is that the covariance matrix is large and dense and we need to be able to compute its inverse in order to evaluate the prior. We would like to make inferential statements about $\boldsymbol{\mu}_h$, but the dimension of the submatrix $\boldsymbol{K}_h$ ($n_h \approx$ 200,000) alone is prohibitive on most hardware architectures—such a matrix would require over $(1.8 \times 10^5)^2 \times 32 = 129.6$ Gb of memory just to store in a single precision floating point format. Though the memory requirement could be reduced by storing just the upper or lower triangle, to sample $\boldsymbol{\mu}_h$ Cholesky decomposition of $\boldsymbol{K}_h$ would still require $\approx 1.9 \times 10^{15}$ floating point operations (FLOPs) to compute.

In (2.6), the covariance matrix has rank of at most n_h, and the implied conditional density $\pi(\boldsymbol{\mu}_s \mid \boldsymbol{\mu}_h)$ is degenerate on $\boldsymbol{K}_{s,h}\boldsymbol{K}_h^{-1}\boldsymbol{\mu}_h$. Additionally, we represent the product $\boldsymbol{K}_h^{-1}\boldsymbol{K}_{h,s}$ by matrix $\boldsymbol{W}^{\mathsf{T}} = [w(\mathbf{v}_s)]_{\mathbf{v}_s \in B_s}$. To induce sparsity and save computational resources, we defined $\boldsymbol{W}$ in terms of neighborhoods of voxels in B_h. For any $\mathbf{v} \in \mathcal{B}$, let $N_h(\mathbf{v})$ denote a set of locations in B_h in an r-neighborhood of location $\mathbf{v}$, $N_h(\mathbf{v}) = \{\mathbf{v}_h \in B_h : \|\mathbf{v}_h - \mathbf{v}\|_2 \leq r\}$. If $N_h(\mathbf{v})$ is empty, then we defined $w(\mathbf{v}) = \mathbf{0}$; otherwise let $\boldsymbol{K}_{N_h(\mathbf{v})} = [k(\mathbf{v}_i, \mathbf{v}_j)]_{\mathbf{v}_i, \mathbf{v}_j \in N_h(\mathbf{v})}$, let $\boldsymbol{k}_{N_h(\mathbf{v})} = [k(\mathbf{v}_i, \mathbf{v})]_{\mathbf{v}_i \in N_h(\mathbf{v})}$, and let $\tilde{\boldsymbol{w}} = \boldsymbol{K}_{N_h(\mathbf{v})}^{-1}\boldsymbol{k}_{N_h(\mathbf{v})}$ denote a vector with implicit dependence on $N_h(\mathbf{v})$ where each element corresponds with one location in $N_h(\mathbf{v})$. For non-empty $N_h(\mathbf{v})$, each element

of $w(\mathbf{v})$ similarly corresponds with one location in B_h. We defined those elements to be,

$$w_i(\mathbf{v}) = \begin{cases} \tilde{w}_j & \text{if the } j^{th} \text{ location in } N_h(\mathbf{v}) \text{ corresponds to the } i^{th} \text{ location in } B_h \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

With $\boldsymbol{W}^\mathsf{T} = [w(\mathbf{v}_s)]_{\mathbf{v}_s \in B_s}$, the product $\boldsymbol{W}\boldsymbol{\mu}_h$ can be interpreted as a local kriging approximation of $\boldsymbol{\mu}_s$ conditional on $\boldsymbol{\mu}_h$. Our definition of $\boldsymbol{W}$ is conceptually somewhat inspired by work on Nearest Neighbor Gaussian Processes by [35, 44]. A sensitivity analysis over choice of r is available in the Appendices.

The matrix $\boldsymbol{W}$ can be entirely precomputed given the kernel parameters, ψ, ν, and a neighborhood radius, r. Equipped with the matrix $\boldsymbol{W}$, samples from (2.6) can be drawn by first sampling $\boldsymbol{\mu}_h \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}_h)$, and then computing $\boldsymbol{\mu}_s = \boldsymbol{W}\boldsymbol{\mu}_h$. In practice we treat r as a hyperparameter and condition analyses on it. In our data example (see section 2.4) we took the radius r to be roughly one FWHM length based on estimated prior covariance and hyperparameters $\boldsymbol{\theta}$ (section 2.4.1). This choice was motivated by the desire to keep r roughly in line with the width of (2.5) while keeping $\boldsymbol{W}$ only modestly expensive to compute: for this choice of r, typical neighborhood sizes $|N_h(\mathbf{v})|$ were on the order of 300–700 voxels in patient data. We next outline an efficient posterior computation algorithm for $\boldsymbol{\mu}_h$.

2.2.3 Posterior computation

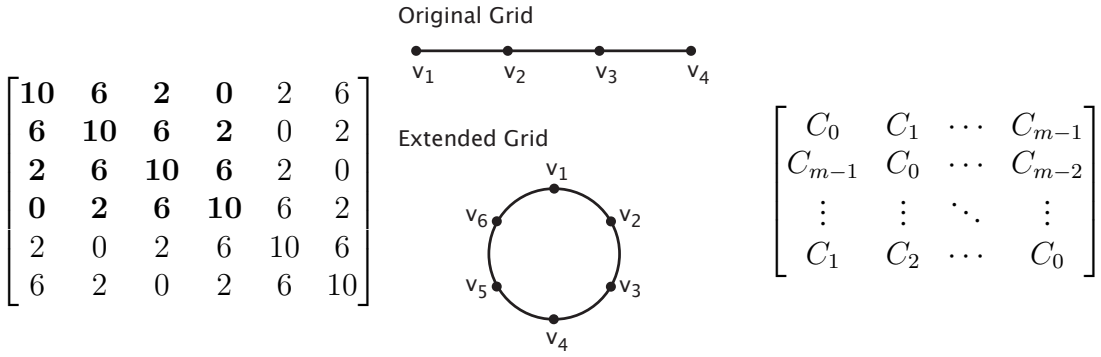


Figure 2.2: Example circulant matrix embedding. The left-most panel shows an example 4×4 Toeplitz matrix (bold) embedded within a 6×6 circulant matrix. In this simple example, the inner Toeplitz matrix might correspond with locations on a 1D grid (center panel). Conceptually, the outer circulant matrix can be taken to correspond with an extended grid, where an extended set of vertices have been "wrapped around" a circle. In the more general case (right-most panel), blocks C_i of a circulant-family matrix have symmetry such that $C_{m-i} \equiv C_i$.

To facilitate computation we embedded the random field $\boldsymbol{\mu}_h$ within a much larger random field,

which we will call $\boldsymbol{u}$. Our goal in doing so was to be able to replace expensive matrix operations with computations involving discrete Fourier transformations (DFTs) as we show below. Conceptually, the augmented parameter space we chose can be viewed to correspond with an extended grid of locations with toroidal geometry. In our data application, the resulting extended grid has about 8.4×10^6 elements (grid dimensions $256 \times 256 \times 128$). We treat this extended grid as if it were a part of B_h in the prior, with the result that the covariance of the field $\boldsymbol{u}$ has a nested block-circulant structure. In the discussion to follow, we will use $\boldsymbol{C}$ to denote the prior variance of $\boldsymbol{u}$. With this construction, we have added a large number of auxiliary parameters, but have not changed the effective prior on $\boldsymbol{\mu}_h$: the matrix $\boldsymbol{K}_h$ is a principal submatrix of $\boldsymbol{C}$. Any Toeplitz-family matrix can be embedded in a larger circulant-family matrix in this way. For additional exposition, Fig. 2.2 shows a simple example of this type of circulant embedding. In the figure, the left and center panels illustrate circulant embedding for a 1D grid. If working on a 2D grid, then schematically each block C_i in the right-most panel of Fig. 2.2 will be a circulant matrix; on a 3D grid each block will itself be block-circulant, etc. This construction can be used to enable efficient simulation of random Gaussian fields over dense grids as others have shown [e.g. 178, 140] and as we summarize below.

Circulant matrix–vector products can be computed efficiently with DFT software. Given the first row or column $\boldsymbol{a}$—the so called base—of a circulant matrix $\boldsymbol{A}$ the product $\boldsymbol{Ax}$ can be expressed as the discrete convolution $\boldsymbol{a} * \boldsymbol{x}$. Equivalently, by the discrete convolution theorem, $\text{DFT}(\boldsymbol{Ax}) = \text{DFT}(\boldsymbol{a}) \odot \text{DFT}(\boldsymbol{x})$, where $\odot$ denotes an elementwise or Hadamard product. The principle is the same with a nested block-circulant matrix like $\boldsymbol{C}$: the matrix has a base, $\boldsymbol{c}$, that is efficient to work with using 3D DFTs. In the present case, $\boldsymbol{c}$ can be precomputed (see Appendix C.2) from the original grid dimensions and covariance function $K(\cdot, \cdot)$. As with any circulant matrix, $\boldsymbol{C}$ can be diagonalized by two Fourier matrices. If $\boldsymbol{F}$ denotes a scaled 3D DFT matrix, and $\boldsymbol{F}^{\mathsf{H}}$ its adjugate, $\boldsymbol{F}^{\mathsf{H}}\boldsymbol{C}\boldsymbol{F} = \text{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ are the (complex) eigenvalues of $\boldsymbol{C}$. With only the base $\boldsymbol{c}$ in memory, $\boldsymbol{\lambda} = \mathcal{F}(\boldsymbol{c})/N$ can be computed directly, where N is the number of elements in $\boldsymbol{c}$, and $\mathcal{F}(\cdot)$ denotes the 3D discrete Fourier transform. We provide a simple algorithm to construct $\boldsymbol{c}$ for any dense 3D grid in Appendix C.2.

Wood and Chan [178] took advantage of this relationship to propose an efficient algorithm for simulation of random Gaussian fields when the covariance of the field can be embedded within a circulant matrix. For example, in our setting, we could sample from the prior $\pi(\boldsymbol{\mu}_h \mid \boldsymbol{\theta})$ by first drawing $\boldsymbol{z} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, where $\mathcal{CN}(0, v)$ denotes the circularly symmetric complex normal distribution with variance $2v$. Let $\mathcal{F}^{-1}(\cdot)$ denote the 3D inverse DFT, let $\boldsymbol{a}^{\circ b} = [a_i^b]$ denote elementwise or Hadamard exponentiation, and let $\mathcal{R}e(\boldsymbol{a})$ extract the real part of a complex vector $\boldsymbol{a}$. With $\boldsymbol{\lambda}$ computed as above, we could then set $\boldsymbol{u} \leftarrow \mathcal{R}e[\mathcal{F}\{\boldsymbol{\lambda}^{\circ 1/2} \odot \mathcal{F}^{-1}(\boldsymbol{z})\}]$, and obtain a prior sample of $\boldsymbol{\mu}_h$ by simply discarding extraneous elements of $\boldsymbol{u}$.

There is no direct extension of the above [178] algorithm for posterior simulation in our setting. In part, this is because we use different noise variance terms for our two data sources (2.2). Unless the diagonal noise terms are exactly equal, the joint posterior variance of $(\boldsymbol{\mu}_h^\mathsf{T}, \boldsymbol{\mu}_s^\mathsf{T})^\mathsf{T}$ will not be Toeplitz in general. We still draw inspiration from the work of [178], however, and use the circulant matrix relationships above to write an efficient Hamiltonian Monte Carlo algorithm for posterior inference. Details of this algorithm are presented in Appendix C.1, but the key components are: (i) as discussed, we embed $\boldsymbol{\mu}_h$ in a higher dimensional random Gaussian field with a circulant covariance matrix; and (ii) we construct a circulant "mass matrix" for our HMC. Modification (i) allows us to be able to evaluate the log prior and compute its gradient, and modification (ii) dramatically improves mixing of the HMC chains. As a result, our algorithm reduces the computational requirement to evaluate the log prior on $\boldsymbol{\mu}_h$ roughly to < 0.01 Gb and $\approx 2 \times 10^9$ FLOPs. We now turn to remark on how we summarize inference from our model in practice.

2.2.4 Functional region detection

FMRI detects functionally relevant brain regions by recording changes in oxygenated blood flow (BOLD signal). In a typical study, practitioners identify these regions by thresholding voxelwise statistical summaries in a manner that controls the false discovery rate [e.g. 61]. For presurgical applications, it is at least as important to limit false negative reports, since errors of this kind may potentially lead to damage of healthy tissue. To this end, we adapted a decision theoretic approach following previous work [114, 101, 100]. We consider the loss function,

$$L(\boldsymbol{m}, \boldsymbol{\delta}) = \sum_i -f(m_i)\delta_i - \{1 - f(m_i)\}(1 - \delta_i) \\ + k_1 f(m_i)(1 - \delta_i) + k_2\{1 - f(m_i)\}\delta_i + t\delta_i, \quad (2.8)$$

where (k_1, k_2, t) are tunable constants, the $m_i = |\mu_{h,i}|/\sqrt{\text{var}(\mu_{h,i})}$ are posterior t-statistic analogs measuring pointwise signal strength in $\boldsymbol{\mu}_h$, and the $\delta_i \in \{0, 1\}$ are pointwise statistical decisions (i.e. $\delta_i = 1$ reports a finding at voxel i, and $\delta_i = 0$ otherwise). The function $f(\cdot)$ can be any monotonically increasing function restricted to $[0, 1]$, and is intended to act as a proxy for $\pi(\delta_i = 1 \mid \boldsymbol{Y}_h, \boldsymbol{Y}_s, \boldsymbol{\theta}, r)$. Again, following previous work [101, 100], we take $f(m) = m/M$, where $M = \max_i\{m_i\}$.

The loss function (2.8) is composed of five terms, each with a distinct importance: $-\sum_i f(m_i)\delta_i$ and $-\sum_i\{1 - f(m_i)\}(1 - \delta_i)$ induce gains for correct discoveries and correct non discoveries, respectively; $k_1 \sum_i f(m_i)(1-\delta_i)$ penalizes false negative errors; $k_2 \sum_i\{1-f(m_i)\}\delta_i$ penalizes false positive errors; and $t \sum_i \delta_i$ penalizes the total number of discoveries. Optimal

decisions δ_i^* minimize the posterior risk and follow,

$$\delta_i^* = \mathbb{1}\{\bar{f}_i \geq (1 + k_2 + t)/(2 + k_1 + k_2)\}, \tag{2.9}$$

where $\bar{f}_i$ is the posterior expectation $\mathbb{E}\{f(m_i) \mid \boldsymbol{Y}_h, \boldsymbol{Y}_s, \boldsymbol{\theta}, r\}$, and the parameters (k_1, k_2, t) suggest a threshold based on a trade off between false negative and false positive errors.

Thresholds can be tuned with domain expert guidance and/or varied dynamically, as a single static threshold may not be sufficient for a surgeon's needs [e.g. 159]. As a practical note, setting $k_2 = t = 1$ and varying k_1 over the range $[5, 12]$ can provide good guidance, with $k_1 = 7$ a reasonable default. In one of our patient data analyses (below), we set $t = 1$, $k_1 = 12$, and $k_2 = 1$ for inference. As per our coauthor and collaborating neuroradiologist's advice, this tuning parameter choice penalizes false negative errors 12 times more heavily than false positive errors. The other patient in our data was somewhat younger and less ill, with no interictal speech or language impairments. Consequentially her z-statistic images appeared to have a better signal to noise ratio. For this patient, the suggestion was to set $k_1 = 7$ and use a seven fold penalty ratio (not shown). In both cases, the corresponding activation thresholds were confirmed visually by comparison with results from intraoperative electrocortical interference mapping.

2.3 Simulation studies

We quantified the advantages of our proposed method with easier to visualize simulations in two dimensions. Our goal in simulation was to evaluate how well the proposed model and alternative methods recovered activation patterns in data. Typical fMRI studies use significance testing as a means to identify functionally relevant brain regions. To mimic this setting, our simulation designs considered active regions embedded within low variance signal (see Fig. 2.3). As we discuss below, we further tried to mimic the patient data by roughly matching simulated spatial signal smoothness and signal-to-noise ratios to the real data.

2.3.1 Simulations on 2D grids

Figure 2.3 illustrates our general approach to data simulation. In the figure, active regions were drawn in a midsagittal plane including a T-shaped region, a circular region, and a four voxel square. These signals were created by smoothing binary images with a six millimeter full width at half maximum (FWHM) Gaussian kernel, scaling by a factor of two, and thresholding the result at 0.4. We then embedded the active regions within random draws from a 2D random field with mean zero, marginal variance 0.2, and 6 mm FWHM Exponential or Gaussian correlation functions. We treated the resultant images as true nonzero mean intensity images, with "active" voxels given only

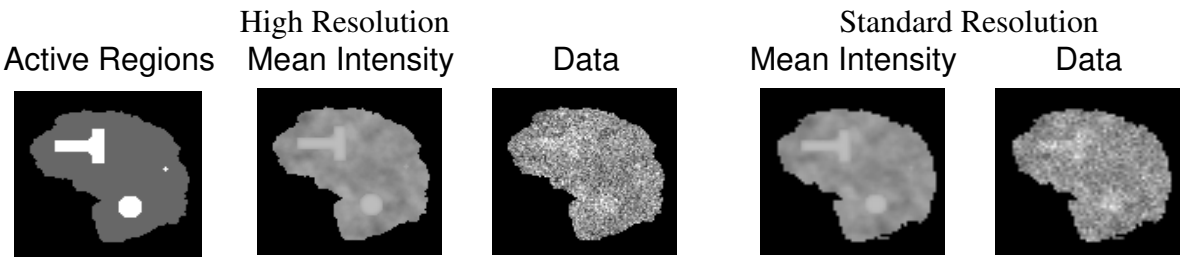


Figure 2.3: Simulation design example with SNR_h = 0.1 and SNR_s = 0.2. Non-activation smooth signal has marginal variance 0.2 and 6 mm FWHM Exponential correlation; activation signal has mean 2.

by the smoothed T, circle, and square shapes; on average about 11% of activation-adjacent voxels would have had signal strength within ± 0.25 standard deviations of their active neighbors. As per the patient data, we treated voxels in this plane as (1.8×1.8) mm for high resolution (4,722 voxels total), and as (3×3) mm for standard resolution (1,853 voxels). With our design, there were exactly 450 active voxels in the high resolution slice (9.5%; see Fig. 2.3).

We adopted this method to generate "high resolution" mean images, or $\boldsymbol{\mu}_h$ as in section 2.2.3, and projected $\boldsymbol{\mu}_h$ into "standard resolution" space by multiplying by $\boldsymbol{W}$ as in section 2.2.2 to generate corresponding standard resolution mean images. In all simulation settings, $\boldsymbol{W}$ was constructed using the true 6 mm FWHM Exponential or Gaussian background signal correlation functions, and an extent radius r defined as the distance after which the correlation would drop below 0.05. To simulate observed outcome data, we added independent Gaussian noise to the mean intensity images, modulated the noise variances to control SNRs of the simulated high and standard resolution images, and ran 100 replicates per parameter combination. We took the SNR to be the ratio of the second moment of the mean to the variance of the noise, and set this to be one of $\{0.1, 0.2\}$ for high resolution images (SNR_h). We parameterized standard resolution noise in terms of the ratio of standard to high resolution SNR (SNR_s:SNR_h), and set this ratio to one of $\{1, 2, 4\}$. In the first case, the standard resolution image would not provide additional signal-to-noise support as it typically would in real data. We considered this a worst case scenario. The latter two settings were chosen so that the standard resolution image provided increasingly large signal-to-noise support, where we expected our dual resolution method to dominate. In our analysis of the patient 1 data, based on the fits of our high and standard single resolution alternative models, we estimated $\text{SNR}_h \approx 0.18$ and $\text{SNR}_s \approx 0.44$ based on their posterior means (ratio SNR_s:$\text{SNR}_h \approx 2.4$).

Model	*Kernel*	SNR_s:SNR_h	SNR_h	*MSE*	*False –*
Dual	Exponential	1	0.1	**0.20**	**31.8% (0.4)**
High	Exponential	1	0.1	0.23	34.0% (0.5)
Naive	Exponential	1	0.1	0.30	43.6% (0.4)
Std	Exponential	1	0.1	0.47	43.1% (0.6)
Dual	Exponential	2	0.1	**0.18**	**30.6% (0.4)**
High	Exponential	2	0.1	0.23	34.0% (0.5)
Naive	Exponential	2	0.1	0.29	42.7% (0.4)
Std	Exponential	2	0.1	0.43	40.6% (0.4)

Table 2.1: Selected results for estimation and inference quality in 2D simulations. Results for the *High* resolution method do not change across the different SNR ratios, but are repeated to facilitate comparison. *Model* denotes the image combination used in the analysis, and *Kernel* gives the correlation pattern of low variance background signal. *MSE* refers to mean squared error computed over the entire high resolution mean parameter vector; the simulation standard error of this metric was on the order of 10^{-3} for all simulation settings and so was omitted for brevity. *False –* reports the mean (SE) false negative error rate when the number of discoveries was fixed at 450. One hundred replicates per parameter combination; additional results with different kernel and SNR_h parameter settings are summarized in Appendix D.

2.3.2 Recovery of simulated activation regions in 2D images

In each simulation, models were conditioned on the true $\boldsymbol{\theta} = (\tau^2, \psi, \nu)^\mathsf{T}$ used to generate the low variance mean fields. We chose to condition on the true $\boldsymbol{\theta}$ so as to explicitly focus our simulation results on estimation of and inference on the image mean intensities. We compared performance of our dual resolution model (2.1) against single resolution alternative methods: (i) a related Gaussian process model that only considered the high resolution data, (ii) the same model but considering only standard resolution data (kriging the posterior mean of $\boldsymbol{\mu}_s$ to the locations in B_h), and (iii) a method that we term naive data averaging. For the alternative high and standard resolution models, we used a Gaussian process to model the the mean of the data as in (2.3). For the naive alternative, we estimated the matrix $\boldsymbol{W}$ (defined in section 2.2.2) from the data and used it to interpolate standard resolution data into the high resolution space. We then treated a simple pointwise average of the high and interpolated standard resolution images—i.e. $\bar{\boldsymbol{Y}}_{hs} = (\boldsymbol{Y}_h + \boldsymbol{W}^\mathsf{T}\boldsymbol{Y}_s)/2$—as data in the alternative high resolution model (i). This approach is conceptually similar to previous work in this area [100]. The high resolution method (i) served as our primary comparison point both because of its inherent spatial resolution and because it tended to be the best competing method in our simulations (see section 2.3).

Table 2.1 presents selected results for estimation and inference quality in our 2D simulation settings. Results are presented predominantly for the setting with SNR_h = 0.1, the SNR_s:SNR_h ratio set to two, and an Exponential correlation function to roughly approximate our patient data (also

reflected in Fig. 2.3). We provide a comparison point with the SNR ratio equal to one for additional interest. More extensive results are available in Appendix D. In the table, *MSE* denotes the mean squared error of the estimated $\boldsymbol{\mu}_h$, computed over pixels in our simulated high resolution slices. We treat MSE as a measure of estimation quality, and report that in all simulation settings considered, MSE was lowest for the dual resolution models. This result indicates that when the same mean intensity function underlies both high and standard resolution images and the kernel function is estimated accurately, the model that used joint information from both imaging modalities outperformed possible single resolution alternatives. Interestingly, when the background intensity was generated with an Exponential kernel, as in Table 2.1, the model that used only high resolution data was the second best performer, underscoring the importance of spatial precision in estimation.

We also report false negative rates—the measure of inference we are most concerned with in our framework—for each model in Table 2.1. In the table, we set parameters k_1, k_2, and t in our decision rule (2.9) independently for each model type so as to control the total number of discoveries to exactly 450 (the same as the number of pixels we considered truly active in the simulations). The actual decisions corresponding to these thresholds are shown in Fig. 2.4 (*right*) for a single representative simulation iteration. We emphasize that thresholds here were chosen as an objective point of comparison across the alternative methods, not by optimizing any kind of inferential criteria. In Fig. 2.4 (*left*), we show that the dual resolution model would give superior inference for any set of decision rule thresholds that fix the false negative rate at a single value across all methods.

2.4 Patient data analysis

As noted, our first motivating dataset—"patient 1"—comes from a right handed 62 year old woman who presented primarily with difficulties reading (pure alexia). This patient was found to have a large tumor in her left middle and inferior temporal gyrus; following partial surgical resection, the tumor was classified as a glioblastoma multiforme. Our second motivating dataset—"patient 2"—comes from an 18 year old right handed woman who presented after a general seizure. This patient was found to have a relatively large cavernoma adjacent to insular cortex and the transverse temporal gyrus. Both patients were scanned prior to surgery while performing a reading task with a 30 second on/off block design to map brain areas associated with reading and subsequent language processing. The task consisted of silent reading in interleaved blocks of non-final embedded clause sentences (on; eight blocks) and strings of consonants (control; eight blocks).

Details of our fMRI acquisition protocol and preprocessing are given in Appendix F. Preprocessing resulted in one unsmoothed z-statistic image for each fMRI resolution that summarized task-related activation over the course of the functional scans. We fit our model to the patient 1

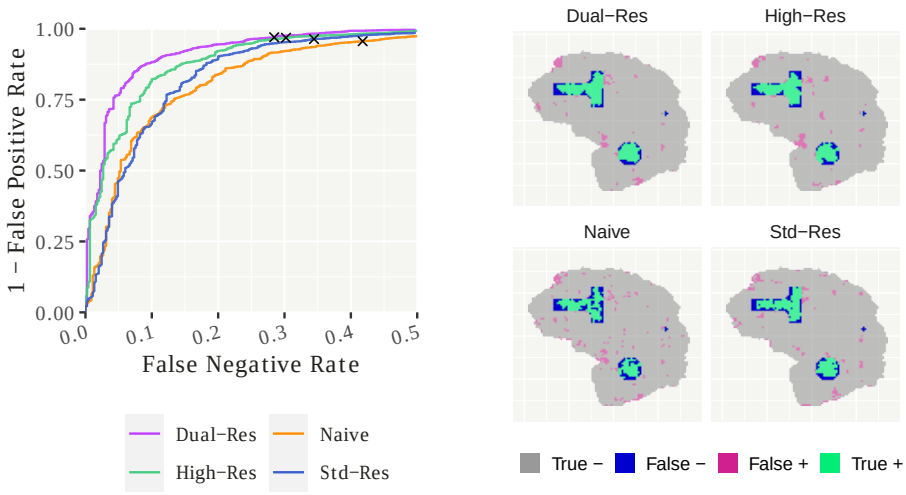


Figure 2.4: Inference quality in 2D simulations. (*Left*) Receiver operating characteristic (ROC) curves comparing dual and single resolution methods to a naive data averaging approach in a setting that matches the data in Fig. 2.3. The curves show that for almost any given false negative rate, the dual resolution method can have a uniformly lower false positive rate than alternative single resolution methods. The $\times$'s mark the thresholds used to generate the inferential summary on the (*right*). These thresholds limit the total number of discoveries to 450 across all four methods.

z-statistic image data to compare relative performance against a set of similar single-resolution alternative methods. With this analysis, our goal was to show how our method can be applied to identify peritumoral activations in patient data and to illustrate potential benefits to inference using combined spatial resolutions. In addition, we fit our model to the z-statistic images from patient 2 to illustrate the method's capacity to recover an estimate of activation in regions with signal loss. Signal loss in fMRI data can occur where tissue types with different magnetic field susceptibilities neighbor one another. This is a common problem encountered in presurgical applications, and can potentially lead to exclusion of areas of interest from the analysis [e.g. 68, 159].

2.4.1 Covariance estimation

We chose to estimate the Gaussian process covariance hyperparameters $\boldsymbol{\theta}$ in the spirit of empirical Bayes using the method of minimum contrast. Minimum contrast estimation (MCE) originates from [37] as a moment estimation approach to spatial modeling. The method seeks to estimate parameters of a function with a known form by minimizing some discrepancy criterion given data. In our case we extracted empirical covariances between voxels at different distances ("empirical covariogram"). We then selected $\boldsymbol{\theta}$ to minimize a nonlinear least squares objective over (2.5), treating the empirical covariogram as pseudo data. Appendix C gives a detailed overview of this

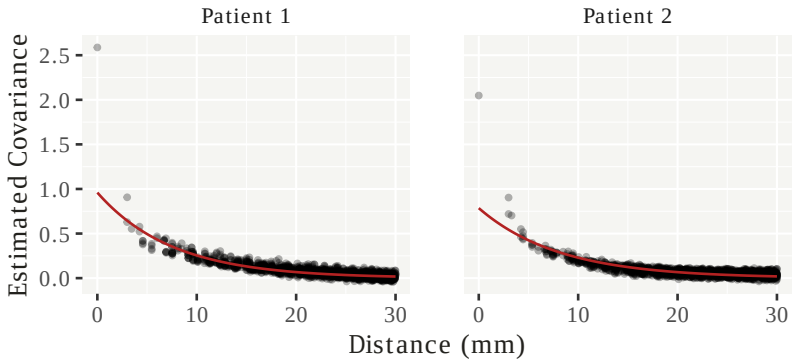


Figure 2.5: Covariograms show empirical covariances between neighboring standard resolution voxels as a function of distance overlaid with a parametric estimate of the covariance function.

procedure for interested readers, as well as a brief sensitivity analysis over our choice of covariance function in (2.5).

This method is not without difficulty. For example, asymptotic theory suggests that empirical covariogram estimation is biased [e.g. 34]. Although this bias does not decrease with increased sampling density ("infill asymptotics"), it can be decreased by sampling data over increasing domains [110, 158, 184]. This point is worth acknowledging because the kriging identity encoded in our prior (2.6) makes estimation of the correlation relatively important. On the other hand, MCE is computationally efficient and scalable to large datasets, and we found that it produced reasonable estimates of the true covariance function in simulation (see Appendix C.3). One reason for this may be that with fMRI data we have a tremendous amount of spatial information collected on a dense grid. Although we can only estimate empirical covariances at a fixed set of distances, we typically have tens of thousands of unique pairs of voxels separated by those distances. In Fig. 2.5, we used the standard resolution images to estimate $\boldsymbol{\theta}$ as we expect these data to have better SNR and there is no theoretical benefit to adding infill locations as with the high resolution images. In doing so, we make an appeal to the notion of a parent process [7] for $\pi\{\mu(\cdot)\}$, which could be defined such that in the prior $\mathrm{cov}\{\mu(\mathbf{v}), \mu(\mathbf{v}')\} = k(\mathbf{v}, \mathbf{v}')$ for all $\mathbf{v}, \mathbf{v}' \in \mathcal{B}$.

For patient 1, an initial unrestricted estimate of $\boldsymbol{\theta}$ yielded an estimated kernel exponent of $\nu \approx 1.25$; for improved interpretability we reran our MCE procedure fixing $\nu = 1$ to yield $\boldsymbol{\theta} = (0.887, 0.135, 1)^{\mathsf{T}}$. Optimization was performed using the COBYLA algorithm [127] as implemented by [86] in the popular NLopt library. The resulting covariance function is shown in the left panel of Fig. 2.5, and corresponds to a 10.47 mm full width at half maximum (FWHM) exponential correlation function. This estimate of $\boldsymbol{\theta}$ was used for all of our analyses of patient 1's data; correspondingly, we set the neighborhood radius r to 10.35 mm in analyses of this patient's data.

Similarly, we estimated $\boldsymbol{\theta} = (0.785, 0.132, 1)^{\mathsf{T}}$ for patient 2. The resulting covariance function (also shown in Fig. 2.5, right) corresponds to a 11.28 mm FWHM exponential correlation function; we set r to 11 mm for analysis of this patient's data. In Fig. 2.5, the exponential kernels appear to fit the empirical covariograms quite well. The points at distances of 0 mm are not outliers but estimates of the "sill" or marginal variance of the $\boldsymbol{Y}_s$ which in our model is $\tau^2 + \sigma_s^2$. Consequentially, our algorithm constrains τ^2 to be strictly less than the empirical variance of $\boldsymbol{Y}_s$, or whichever image is used to construct the covariogram. In addition, the exponential models in Fig. 2.5 tend to mildly but systematically underestimate the empirical covariances at displacements around 3 mm. We discuss how these data points can be modeled more accurately, and elaborate on why it may or may not be optimal to do so in the Appendix B.2.

2.4.2 Patient 1: Inference on the functional signals

We fit our model to the data from patient 1 described in section 2.2 with custom software written in C++ that uses the Eigen [66] and FFTW [48] libraries for linear algebra and DFT operations, respectively. For these analysis, we set the number of leapfrog steps $L = 25$ and ran three independent HMC chains of 4,000 iterations each, discarding the first 1,000 as burnin, and thinning the output to every third iteration thereafter. Univariate Gelman–Rubin statistics [60] were used to evaluate voxelwise convergence of $\boldsymbol{\mu}_h$. This statistic was ≤ 1.03 for every voxel, suggesting approximate convergence. Additionally, trace plots of means from six randomly selected voxels are shown in in Appendix B.1 and show good mixing of the Markov chains.

Fig. 2.6 (*left*) shows posterior mean activation maps for a series of sagittal slices through the patient's tumor in left temporal lobe. Activations are overlaid on a high resolution, gadolinium enhanced T1-weighted anatomical scan. In the figure, we have circled a peritumoral region that was deemed to determine the surgical access considered. The patient's tumor can be seen within this circled region in all slices. As in our simulation studies, we compared performance of our dual resolution model against single resolution alternatives: models considering only the high or standard resolution data, and an additional setting using a naive average as data.

In Fig. 2.6 (*right*), we show that no matter the threshold applied to whole brain posterior activation maps, our dual resolution method identified at least as many active voxels in the peritumoral region than if we had ignored the standard resolution data. A visual comparison of the posterior mean of $\mu(\cdot)$ for all four methods is shown in Fig. 2.7. We chose a single sagittal slice to represent this comparison although the analysis was over the whole brain. Qualitatively, posterior means from the dual and high resolution analyses appear substantially sharper than for the standard resolution analysis. At the same time, differences are apparent in the dual and high resolution posterior means, particularly around the edges of areas with high magnitude signal. We also plot voxelwise

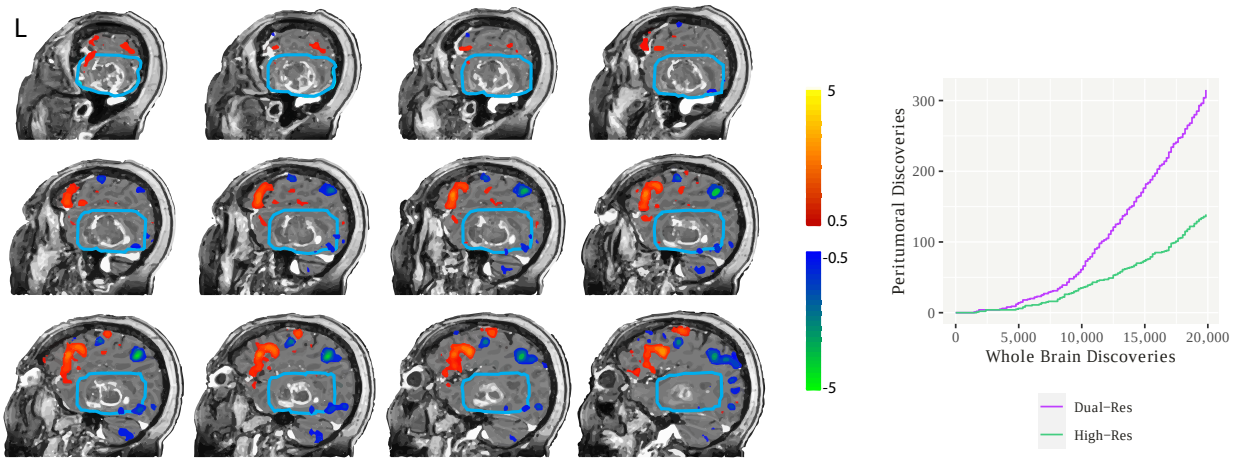


Figure 2.6: Patient 1: (*Left*) Thresholded posterior mean image shows peritumoral activation identified using our dual resolution method. The tumor is the region of mixed hypo- and hyperintensity in the temporal lobe across slices; the peritumoral region is outlined in each panel (in cyan). Functional activations are shown in warm colors, and functional deactivations are shown in cool colors, with units on the z-statistic scale. Activation regions are shown setting k_1 in our decision rule (2.9) to 17 to enhance the visualization. Slices are shown proceeding lateral-to-medial through the left hemisphere in left-to-right, top-to-bottom order. (*Right*) Cumulative counts of discoveries at varying decision thresholds. Voxelwise discoveries in the peritumoral region plotted against whole brain discoveries for both dual and high resolution methods.

comparisons of dual and high resolution posterior means and variances of $\boldsymbol{\mu}_h$ in Fig. 2.7. In the figure, voxels with high signal strength typically had higher magnitude posterior means estimated with the dual resolution model; marginal variances of the $\mu_{h,i}$, moreover, were lower with the dual resolution model in about 72.4% of voxels. With respect to mean image smoothness, we estimated (using our MCE procedure; see section 2.4.1) the standard resolution posterior mean image had a kernel FWHM of about 17.3 mm, and the high resolution posterior mean image had a kernel FWHM of about 13 mm. Appropriately, the dual resolution posterior mean image had a kernel FWHM between these two, at about 14.4 mm. Relating back to Fig. 2.1, our initial goal in modeling joint data sources was to reduce noise inherent in the high resolution signal and leverage signal strength from the standard resolution data. Taken all together, these results demonstrate that we have met that goal.

Additional patient 1 model fit and diagnostic evaluations are given in Appendix B.1. In particular, we evaluated the residual independence approximation present in our model likelihood by running our kernel estimation procedure (see section 2.4.1) on the model residual images. These analyses suggested that residual correlation decayed to near zero within the smallest voxel dimension widths, leading us to conclude that residual independence was a reasonable approximation in our data. Full results are available in Appendix B.1.

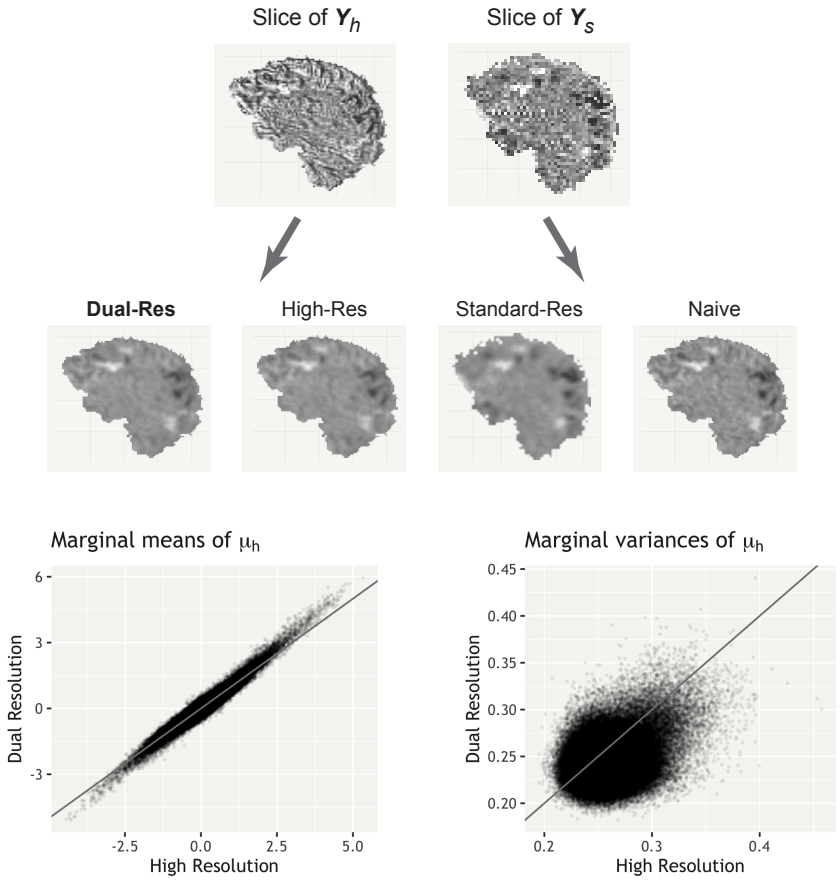


Figure 2.7: Patient 1: visual comparison of posterior means in a single sagittal slice from four models fit to different combinations of whole brain patient data (*middle*). The (*top*) row of the figure shows the raw data from the same slice at both high and standard resolution. Grayscale intensity is shared across all subfigures. The (*bottom*) row shows a comparison of voxelwise posterior means (*bottom, left*) and variances (*bottom, right*) of the elements of $\boldsymbol{\mu}_h$ estimated using the proposed model and a single (high) resolution alternative. The gray lines show identity relationships for comparison; variances were lower using the dual resolution model in about 72.4% of voxels.

2.4.3 Patient 2: Recovery of lost signal

Similar to our analysis of patient 1, we fit our dual and single resolution models to the data from patient 2. In this case, we ran five independent HMC chains for each model and set the chain length, burnin, thinning rate, and number of leapfrog steps identically as above. Univariate Gelman–Rubin statistics were ≤ 1.05, again suggesting approximate voxelwise convergence of $\boldsymbol{\mu}_h$. Our primary goal with this analysis was to illustrate our method's ability to recover estimates of activation from regions of signal loss.

In this particular data set, the patient's cavernoma caused a region of GE-EPI signal dropout, with blooming along the left insular and upper temporal lobe (see Fig. 2.8, *left*). This is a com-

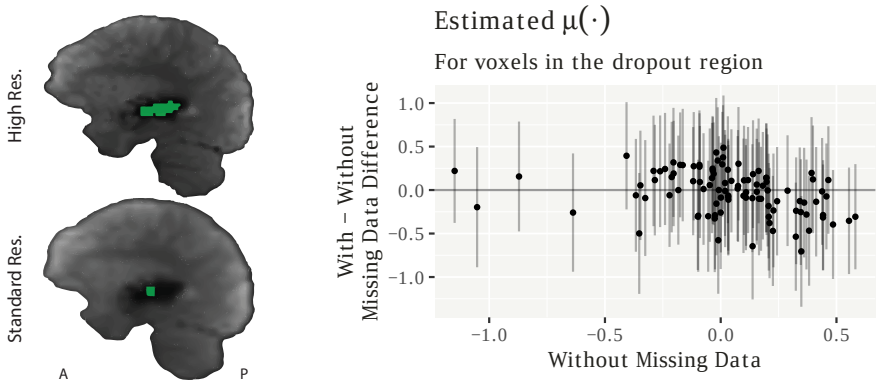


Figure 2.8: Patient 2: (*Left*) Core regions of fMRI signal loss across the left temporal and insular cortex are highlighted on high and standard resolution T2*-weighted slices. (*Right*) Comparison of the mean parameter for voxels in the core high resolution dropout region. We fit our dual resolution model to parallel versions of the data with and without missingness. The posterior mean estimate of $\mu(\cdot)$ without missing data is shown on x-axis, with the difference in the estimates shown on y-axis. Error bars give $\pm$ one standard error of the difference estimated across five HMC chains.

mon occurrence in clinical fMRI: brain lesions can induce signal loss and magnetic susceptibility artifacts in gradient echo imaging. The functional signal is not "missing," per se, but local signal hypointensities can cause image preprocessing software to exclude affected areas from analysis [68]. Such was the case here, and we leveraged this data structure to highlight our model's predictive ability. By adjusting brain/background thresholding in FSL, we created two versions of this patient's task contrast data: one where data in the core dropout region was completely masked out of all analyses (with missing data), and one where voxels in this same region were included in all analyses (without missing data). In this case, we were able to create parallel versions of the data for this patient, though in more general practice it may sometimes be difficult to engineer the data without missingness [68].

We fit our models to both versions of the data to compare resulting functional estimates. In the case of missing data, the Gaussian process formulation of our model enables natural prediction of values of $\mu(\cdot)$ for voxels with missing data. In the high resolution image, the region masked out due to signal loss encompassed exactly 100 voxels, had a maximum length (measured anterior to posterior) of around 25–26 mm, and was on average about 9.4 mm wide (lateral to medial) and 4.5 mm tall (dorsal to ventral). In the low resolution image, masked dropout was limited to only two voxels given default brain/background thresholding in FSL. With typical preprocessing pipelines we would generally expect more signal loss voxels to be excluded from high resolution images. In Fig. 2.8 (*right*) we show, for our dual resolution method, the correspondence of predicted/estimated voxelwise means in the dropout region for the two data variants. The figure shows excellent correspondence: the Pearson correlation between the predictions and the estimates is 0.673, suggesting

our method has a strong capacity for signal recovery in dropout regions of this size. Similar predictions/estimations using only the high resolution data also show good correspondence, but were made with higher variance relative to our dual resolution approach in 81 of the 100 dropout region voxels.

Similarly, across the whole brain, the marginal variances of $\boldsymbol{\mu}_h$ for patient 2 were lower with our dual resolution approach in about 62.4% of voxels (compared to the high resolution model). Moreover, we estimated that the dual resolution posterior mean image smoothness had a kernel FWHM of about 7.5 mm, while the high and standard resolution posterior mean image smoothness FWHMs were about 7.6 mm and 13.5 mm, respectively. Overall, estimation and inference about $\boldsymbol{\mu}_h$ was more similar between our dual and high resolution models here than for the previous patient. Altogether, results from our analysis of the patient 2 data again suggested our inference benefited from combining information from both spatial resolutions, though the benefit may be less pronounced compared to in patient 1. Based on our single data source high and standard resolution alternative methods, we estimated that for patient 2, the standard resolution data provided only a modest 5.4% improvement in SNR compared to the high resolution data. By contrast, for patient 1, the standard resolution data provided approximately a 139% improvement in SNR, lending context to the above result.

2.5 Discussion

Preoperative fMRI presents many interesting and unique statistical challenges from an applied perspective. Presurgical planning requires spatially precise localization of patient specific functional neuroanatomy, but the current physical limitations of MR imaging technology lead to reductions in the signal to noise ratio (SNR) with increases in spatial resolution. This inherent limitation has led to the hypothesis that collecting fMRI data at multiple spatial resolutions may result in improved functional region detection; our simulations in the present paper suggest that this may indeed be the case. We have also shown how a simple decision rule can be applied by practitioners to infer about functional regions given some desired trade off between false positive and false negative errors. This is important because neuroradiologists and neurosurgeons may be more concerned with false negative errors, which could lead to resection of functionally relevant tissue in practice.

With our present work, we propose to base inferences about functional regions on a joint model for images collected at each spatial resolution. Modeling high dimensional correlated outcomes can be quite challenging computationally, and the dataset presented the additional burden of integrating two data sources with different spatial support sets. We circumvented this problem with a Gaussian parent process approximation using only the highest collected resolution image's voxel locations as a primary support set, and embedding these locations within a larger, toroidal space,

leading to computational gains. As a consequence, our Gaussian process model and related algorithm has very natural extensions to cases with different numbers of data sources. For example, we recognize that not every preoperative plan will rely on collecting both high and standard resolution fMRI data. Our model can easily accommodate the situation where only one spatial resolution is collected by simply dropping unobserved data from the joint outcome.

Just as easily, our model could accommodate data collected at additional spatial resolutions with minimal added computational cost. In fact, in a different setting, we imagine our method could be used for image based meta-analysis to synthesize results from multiple experimental studies. In such a setting, posterior credible sets—Bayesian analogs of spatial confidence sets from [18]—could be used to shift inferential focus back to limiting a family wise false positive error rate. With the recent proliferation of large, multi-center imaging collectives [e.g. 166], we feel this may be a promising area for further applied research.

One important limitation of our present model is that we treat both the prior mean model and the errors within each image as stationary processes. In general, stationarity may not be a realistic assumption for imaging data [e.g. 180]. In the case of our mean model, stationarity is only a limitation of the prior: given the data the posterior may still reflect a non-stationary process. In our analysis of the residual images from patient data we found that while there was some residual spatial autocorrelation, this autocorrelation in general decayed to near zero within one to two voxel widths (see Appendix B.1 for figures). Thus we concluded that prior model mean-field stationarity can lead to reasonable posterior approximations for these data. Our model on the error structure, however, is a bit more restrictive. We further note that residuals tended to show modestly higher dispersion in gray matter than in white (see Appendix B.1). While we do not believe this difference is so pronounced as to negatively impact our analyses, it may be worthwhile to explore non-stationary error models. In general, this is a difficult issue. Allowing too much flexibility in the error process may, for example, lead to model non-identifiability or similar complications.

In our simulations and analysis of patient data we showed that our dual resolution method borrows strength from both data sources to improve inference, especially around the edges of active regions, without sacrificing the spatial resolution of the high resolution data (confer from Figs. 2.7 and 2.4). To accomplish this task with patient images, we started from the output of typical single subject fMRI analyses, treating summary statistics from voxelwise marginal time series models as data. [17] similarly used summary statistics from voxelwise marginal models as data in a group analysis in an experimental setting. Although their approach and setting was different from ours, the authors also chose not to smooth their data during preprocessing and made a similar independent noise approximation in their model likelihood as we do here [17]. We provide additional evaluation of our independent and homogeneous residual noise approximations in Appendix B.1 and conclude that the approximations are reasonable in our patient data. While we might even-

tually like to incorporate available time series information into our model, doing so would only add to computational complexity, and it is unclear to what extent spatial inference would improve as a result. At present, a handful of integrated spatiotemporal models have been developed for fMRI studies, but nearly all of these are intended to be fit to single slice data, not whole brain [e.g. 124, 65, 99]. Only more recently have variational approximations been leveraged to enable whole brain spatiotemporal inference at a reasonable computational cost [149]. In its current form, our work uses summaries of temporal data to enable whole brain inference at a very fine spatial scale, but there may be room to incorporate richer temporal information into our model as part of future study.

Finally, in our work we estimated the Gaussian process hyperparameters $\boldsymbol{\theta} = (\tau^2, \psi, \nu)^\mathsf{T}$ from the data in the spirit of empirical Bayes. We accomplished this goal by minimizing a least squares contrast function over an empirical covariogram estimated from the data. Other approaches to learning these parameters include maximizing the data marginal likelihood [e.g. 110], and fully Bayesian estimation [e.g. 7]. We chose our minimum contrast estimation (MCE) type approach as it is generally more extensible to the size of our dataset. Computing the marginal likelihood would involve inversion of an $(n \times n)$ matrix where n is the number of voxels or spatial locations. Our posterior computation algorithm specifically avoids even constructing such a matrix, which is impossible to store on most computer systems (see section 2.2.2). Fully Bayesian estimation of $\boldsymbol{\theta}$ on the other hand is possible, though still computationally demanding. The kernel bandwidth and exponent parameters, ψ and ν, respectively, can be quite slow to update with multiple data sources, and computation time is a concern in a preoperative setting. In contrast, the partial sill variance τ^2 is straightforward to update in our framework, and an abundance of spatial data make this parameter strongly identifiable. We considered updating τ^2 by default in our algorithm, but found that it did not dramatically affect spatial inference in our data and sometimes led to slower Markov chain mixing. As a result, we decided to condition inference on fixed $\boldsymbol{\theta}$ by default in our analyses and consider alternative estimation methods a possibility for future extension.

Conditional on $\boldsymbol{\theta}$, our method enables spatially precise inference on whole brain fMRI data collected at multiple spatial resolutions. Despite the very high dimensional nature of our data, our method is computationally efficient enough to be viable for application in presurgical planning. In addition, we have shown through simulation that inference drawn from a joint model using both available data sources can lead to substantial improvement over inference with single resolution alternatives. We hope that this body of work will benefit the presurgical fMRI community, and may find extension in experimental fields by supporting image based meta analysis and results synthesis.

CHAPTER 3

A Semiparametric Mixture of Spatial Regression Models for Subgroup Effect Estimation in Group-Level Imaging Studies

Unsupervised learning of class labels is an area of broad applied interest. Here, we propose a semiparametric hierarchical model for image-on-scalar regression in the presence of potential unknown subgroup heterogeneity. We model the mean intensity of imaging outcomes as a mixture of regression models with spatially varying coefficient functions. In turn we use Gaussian process priors to model spatial correlation in the regression coefficients. Additional participant-level covariates are used to inform the mixing distribution by way of a logistic stick-breaking process prior. This general class of prior allows construction of individual-specific mixture weights, inducing correlation in mixture component assignments between individuals with similar covariate features. We solve the problem of computing with high dimensional mixture components by projecting the original data onto a lower dimensional subspace. An appealing consequence of this construction is that subgroups are identified based on low-rank features in the data. We show in a simulated toy example how our proposed method can lead to clustering and estimation performance superior to common unsupervised methods. Finally, we illustrate use of our stick-breaking model in the context of neurotyping in patients with Autism spectrum disorders using preprocessed single-site data from the Autism Brain Imaging Data Exchange.

3.1 Introduction

Detecting meaningful subgroups in the absence of explicit cluster labels (unsupervised learning) is an area of interest in many applied settings. For instance, one of the central goals of the precision medicine initiative is ultimately to improve patient prognoses through use of individually targeted treatment and prevention regimes [27]. Disease subtyping is widely recognized as an important component of this general goal [see e.g. 77, 145, for reviews], and consequentially is an area

of active research. Researchers, for example, have long sought meaningful subtyping within the Autism spectrum disorders (ASD) [see 76, 167, for reviews]. Classically, subtyping of ASD has been accomplished through analysis of patient profiles of one or more clinical severity indicators [e.g. 64]. Only more recently have researchers begun to study possible ASD subtyping through imaging biomarkers (neurotyping) [e.g. 40, 162].

In cases like these, the prevailing methodology is to apply a clustering algorithm or factor analysis model to part of the data—e.g. the imaging outcomes—and then test for cluster-based differences among remaining covariates—e.g. the clinical indicators. We posit that a fully generative model may lead to more accurate and meaningful estimation of subgroups in data with this type of structure. With the present paper, we propose a semiparametric hierarchical model for image-on-scalar regression in the presence of potential subgroup effect heterogeneity. Our proposed method will be useful when group structure is not known but can be reliably inferred given observed covariates. We accomplish this by modeling the mean intensity function of imaging outcomes as a mixture of regression models with spatially varying coefficient functions. In turn, we model the uncertainty in each spatially varying coefficient function using Gaussian process priors. Individual-level covariates are used to inform the mixing distribution by way of a logistic stick-breaking process (LSBP) prior on individual-specific mixture weights [136]. This general class of prior induces correlation in mixture component assignments between individuals with similar covariate features. Given data, the posterior distribution of the mixture weights can be used to help impute and make inferences about latent subgroups and participant membership.

Here, we solve the problem of computing with high-dimensional mixture components by conditioning on the hyperparameters of the Gaussian process covariance function and projecting the original data onto a lower dimensional subspace. In practice, the covariance hyperparameters can first be estimated from the data, and an optimal projection can be defined following the ideas of [146, 134, 7, 45]. Though this form of low-rank projection entails some loss of information, in practice it has been shown to perform reasonably well in the mean squared prediction error sense [71]. Moreover, this construction implies that subgroups are defined based on low-rank spatial features in imaging data: an idea we find intuitively appealing. We write our prior hierarchy so that, using existing data augmentation schemes, full conditional updates are available for all of our model parameters and reasonably efficient posterior inference can be achieved via Gibbs sampling. We show in a simulated toy example how our proposed method can lead to clustering and estimation performance superior to common unsupervised methods. Finally, we illustrate use of our method in the context of ASD neurotyping using preprocessed single-site data from the Autism Brain Imaging Data Exchange (ABIDE). Software for our method is available online.

The rest of this paper is organized as follows. We briefly introduce the LSBP prior in section 3.2, and present our use thereof within the rest of our model hierarchy in section 3.3.1. We continue by

describing our spatial regression approach in section 3.3.2, and outline our posterior computation scheme in section 3.3.4. In section 3.4 we demonstrate the feasibility of our method in a simulated toy example and compare against a modified k-means procedure. We illustrate use of our method neurotyping ASD patients from the New York University cohort of the ABIDE I preprocessed data in section 3.5. To end, in section 3.6 we discuss issues of computational scalability and consider potential data-specific extensions of our method.

3.2 Logistic stick-breaking process models

Dirichlet process models [42, 14] have garnered wide use as a backbone of nonparametric Bayesian statistics. The typical goal of such models is to approximate any general density function $f(\cdot)$ by a weighted sum of simpler densities. The stick-breaking constructive definition of the Dirichlet process [147] explicitly represents,

$$f(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k}, \quad \omega_k = p_k \prod_{m<k} (1 - p_m), \quad p_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), \quad \theta_k \sim P_0, \tag{3.1}$$

where δ_θ is the Dirac measure with mass on θ, and P_0 is some base distribution which generates atoms $(\theta_k)_{k=1}^{\infty}$. With this construction, the ω_k can be thought of as a countably infinite sequence of random probability measures on each corresponding atom θ_k, such that $\lim_{k\to\infty} \omega_k = 0$ and $\sum_k \omega_k = 1$ almost surely. The "concentration" parameter $\alpha > 0$ controls how quickly the ω_k converge to zero. In general, P_0 need not be a distribution over a subset of the real line, and generalizations exist with P_0 a distribution over some arbitrary multivariate or functional space. For computational tractability, common solutions to estimate the posterior distribution of $f(\cdot)$ given data $(y_i)_{i=1}^{N}$ typically rely on imputing auxiliary cluster labels $\mathcal{C}_i$ for each datum y_i. Cluster labels can then be integrated out of the posterior to recover the nonparametric density.

Dependent Dirichlet process-type models were introduced by [108] to induce correlation in the latent clustering process for data indexed by similar additional features, $\boldsymbol{z}$ [131]. For example, to formulate the logistic stick-breaking process (LSBP), [136] began from the stick-breaking construction above and redefined,

$$\omega_k(\boldsymbol{z}) = p_k(\boldsymbol{z}) \prod_{m<k} \{1 - p_m(\boldsymbol{z})\}, \quad \ln\left\{\frac{p_k(\boldsymbol{z})}{1 - p_k(\boldsymbol{z})}\right\} = \boldsymbol{z}^{\mathsf{T}} \boldsymbol{\gamma}_k, \quad \boldsymbol{\gamma}_k \sim \pi_\gamma, \tag{3.2}$$

where π_γ is the prior placed on elements of the series of logistic regression coefficients, $\boldsymbol{\gamma}_k$. This construction, which we abbreviate by $\boldsymbol{\omega}(\boldsymbol{z}) \sim \mathcal{LSBP}(\pi_\gamma)$, allows $p_k(\boldsymbol{z})$ and $p_k(\boldsymbol{z}')$ to be close to one another for "similar" inputs $\boldsymbol{z}$ and $\boldsymbol{z}'$, leading to correlated cluster assignments. In our present

work, we leverage this idea to cluster outcome images given participant-level covariates $\boldsymbol{z}$.

3.3 Methods

3.3.1 Proposed model

Let $\mathcal{B}$ denote a generic brain image space, and let $B \subset \mathcal{B}$ denote the sets of spatial locations in the brain where fMRI data are collected. Each atom $\mathbf{v} \in \mathcal{B}$ is a three dimensional vector of spatial coordinates measuring distance from some origin point, $\mathbf{v}_0 \in \mathcal{B}$; the Euclidean distance between any two points, $\mathbf{v}, \mathbf{v}' \in \mathcal{B}$ can be represented $\|\mathbf{v} - \mathbf{v}'\|$, and is typically measured in millimeters. Although data at a given voxel is associated with a small volume, we follow common practice and treat that data as observed on location $\mathbf{v}$ exactly.

We denote imaging outcomes at specific voxels by $y_i(\mathbf{v})$, $\mathbf{v} \in B$, where $i = 1, \ldots, N$ indexes participants or images. Additionally, we define notation $\boldsymbol{y}_i = [y_i(\mathbf{v})]_{\mathbf{v} \in B}$ to refer to the entire brain image outcome for participant i. We take each participant's data to be referenced by sets of scalar of covariates $\boldsymbol{x}_i \in \mathbb{R}^P$ and $\boldsymbol{z}_i \in \mathbb{R}^Q$, that can describe participant or measurement characteristics such as age or gender. We will link the $\boldsymbol{x}_i$ to mean model structure, and the $\boldsymbol{z}_i$ to the mixture weight structure below. For brevity we also allow $\boldsymbol{z}_i$ to include covariates that may be linked to random effects in the study population. Let $\mathcal{N}(\mu, \Sigma)$ denote the normal distribution with mean μ and variance Σ. For many problems in applied imaging research it can be reasonable to model the data as jointly Gaussian. We write,

$$y_i(\mathbf{v}) = \mu_i(\mathbf{v}, \boldsymbol{x}_i, \boldsymbol{z}_i) + \epsilon_i(\mathbf{v}) \quad \text{where} \quad \epsilon_i(\mathbf{v}) \sim \mathcal{N}(0, \sigma^2(\mathbf{v})), \tag{3.3}$$

for $i = 1, \ldots, N$, and $\mathbf{v} \in B$. In equation (3.3), we express imaging outcome $y_i(\cdot)$ as the sum of a mean image intensity function $\mu_i(\cdot)$ and a non-stationary white noise process described by variance function $\sigma^2(\cdot)$. Mean image intensity functions $\mu_i(\cdot)$ are taken to depend on spatial location as well as participant-level covariates. In writing the model this way, we make the (potentially approximate) assumption that $\epsilon_i(\mathbf{v})$ is conditionally independent of $\epsilon_i(\mathbf{v}')$ for any two $\mathbf{v}, \mathbf{v}' \in \mathcal{B}$, $\mathbf{v} \neq \mathbf{v}'$. Our prior on the $\mu_i(\cdot)$, however, will induce spatial correlation in the corresponding $y_i(\mathbf{v})$ and $y_i(\mathbf{v}')$ in section 3.3.2. Using the shape-rate parameterization of the Gamma distribution, we assign the error process precision function $\sigma^{-2}(\cdot)$ a weakly-informative hierarchical prior such that,

$$\sigma^{-2}(\mathbf{v}) \overset{\text{iid}}{\sim} \text{Gamma}(1, \xi), \qquad \pi(\xi) \propto \mathbb{1}(\xi > 0), \tag{3.4}$$

taking $\mathbb{1}(\mathcal{A}) \in \{0, 1\}$ to denote the event indicator function ($\mathbb{1}(\mathcal{A}) = 1$ if event $\mathcal{A}$ occurs, and 0 otherwise). Given the large number of spatial locations in a typical imaging analysis, the error

process rate parameter, ξ, will usually be strongly identified even with minimal prior information.

The likelihood in (3.3) is quite general and could be adapted to a variety of methods for imaging. As written, we consider this model best for use with unimodal imaging outcomes, such as structural images or task-based contrast data. Although the model could be extended to handle multimodal images or time series data by incorporating additional structure in the $\mu_i(\cdot)$, this is not our present focus. The defining feature of our method lies in the construction of the prior for the regression function $\mu_i(\cdot)$. To induce a subgrouping effect construct, we model the mean intensity function for each image using a mixture prior with,

$$\mu_i(\mathbf{v}, \boldsymbol{x}, \boldsymbol{z}) \sim \sum_{k=1}^{\infty} \omega_k^\dagger(\boldsymbol{z}) \delta_{\mu_k^\dagger(\mathbf{v}, \boldsymbol{x})}, \tag{3.5}$$

where the $\omega_k^\dagger(\cdot)$ are mixture weights, and $\delta_{\mu^\dagger}$ is the Dirac measure as in section 3.2. Throughout, we will use the dagger superscript to help us distinguish mixture component-specific parameters. We use an LSBP prior (discussed in Section 3.2) to inform the mixture weights for each individual given known subject-level covariates $\boldsymbol{z}$,

$$\boldsymbol{\omega}^\dagger(\boldsymbol{z}) \sim \mathcal{LSBP}(\pi_\gamma), \tag{3.6}$$

where each $\omega_k^\dagger(\cdot)$ maps $\mathbb{R}^Q$ onto the unit interval. Notice from (3.2) that this choice of prior can lead to different mixture weights for participants i and i' when $\boldsymbol{z}_i \neq \boldsymbol{z}_{i'}$. When individuals have similar covariate profiles (e.g. in the supremum norm sense), however, their mixture weight functions will also be similar *a priori*. This feature of the prior represents the belief that individuals with similar covariate profiles are more likely to share patterns in their imaging outcomes. For the sake of exposition, we will reserve discussion of our choice of prior π_γ for the moment. It is enough to say that it will be convenient to work within the Gaussian scale mixture family for conjugacy, and that we center π_γ on zero to encourage approximate sparsity in the $\boldsymbol{\gamma}_k^\dagger$. These choices roughly follow those in [136].

3.3.2 Spatial regression model mixture components

In equation (3.5), the mixture components are random functions $\mu_k^\dagger(\cdot)$, that we in turn express as a linear model with spatially varying coefficient functions. To induce spatial correlation in the mean intensity across imaging outcomes, we use Gaussian processes to model the probability law

governing each spatially varying coefficient function,

$$\mu_k^\dagger(\mathbf{v}, \boldsymbol{x}) = \sum_{j=0}^{P-1} x_j \beta_{k,j}^\dagger(\mathbf{v}), \quad \text{where} \quad \beta_{k,j}^\dagger(\mathbf{v}) \sim \mathcal{GP}(0, K(\mathbf{v}, \mathbf{v}')), \tag{3.7}$$

and where covariance function $K(\cdot)$ is taken to parameterize the covariance of the spatial processes $\beta_{k,j}^\dagger(\cdot)$. Here, we center the spatially varying coefficient functions on zero without loss of generality: this choice can typically be made reasonable in practice for example by centering the outcome images $\boldsymbol{y}_i$ on their global mean. In our application, we define $K(\cdot)$ based on a the notion of a Gaussian parent process [7] with a stationary, isotropic correlation structure. We define our $K(\cdot)$ so that the full-rank parent processes are projected onto low-rank bases. One appealing feature of this construction is that it implies the components of our mixture prior can be distinguished based on relatively low-dimensional features. We also make this choice for computational convenience since for a typical imaging data set the number of voxels in B can make it difficult to impossible to work with the full-rank Gaussian parent process.

To formalize our choice of $K(\cdot)$, let $\mathcal{V}_* \subset \mathcal{B}$ denote a set of $M_* = |\mathcal{V}_*|$ knot locations in the brain, and let $\rho(\mathbf{v}, \mathbf{v}')$ denote a positive definite correlation function such that $\rho(\mathbf{v}, \mathbf{v}') = 1$ when $\mathbf{v} = \mathbf{v}'$ and $\rho(\mathbf{v}, \mathbf{v}') \leq 1$ otherwise. Though many options are available for the choice of correlation function, here we take $\rho(\mathbf{v}, \mathbf{v}') = \exp(-\psi \|\mathbf{v} - \mathbf{v}'\|^\nu)$ to represent the two parameter radial basis correlation function with decay hyperparameter $\psi > 0$ and exponent parameter $\nu \in (0, 2]$. We chose the radial basis function family here per the long history of Gaussian smoothing in applied imaging analyses. This function is synonymous with the Gaussian kernel when the exponent $\nu = 2$; additional literature suggests that the exponential kernel (with $\nu = 1$) may be more appropriate for analysis of fMRI task data [65]. Let $\boldsymbol{c}_*(\mathbf{v}) = [\rho(\mathbf{v}, \mathbf{v}_*)]_{\mathbf{v}_* \in \mathcal{V}_*}$ denote a vector in $\mathbb{R}^{M_*}$, and let $\boldsymbol{C}_* = [\rho(\mathbf{v}_*, \mathbf{v}'_*)]_{\mathbf{v}_*, \mathbf{v}'_* \in \mathcal{V}_*}$ denote the $(M_* \times M_*)$ dimensional correlation matrix evaluated at locations in $\mathcal{V}_*$. For any two $\mathbf{v}, \mathbf{v}' \in \mathcal{B}$, we take,

$$K(\mathbf{v}, \mathbf{v}') = \tau^2 \boldsymbol{c}_*^\mathsf{T}(\mathbf{v}) \boldsymbol{C}_*^{-1} \boldsymbol{c}_*(\mathbf{v}'), \tag{3.8}$$

where $\tau^2 > 0$ is the marginal variance of the $\beta_{k,j}^\dagger(\cdot)$. Expressing the Gaussian process covariance function as in (3.8) has been shown to be "optimal" in the sense of minimizing the Kullback-Leibler divergence from the full-rank Gaussian parent process [146, 7]. That is to say that if Π is the full-rank parent process distribution defined with covariance function $K_\Pi(\mathbf{v}, \mathbf{v}') = \tau^2 \rho(\mathbf{v}, \mathbf{v}')$ everywhere in $\mathcal{B}$, and Π_* is a distribution from the class of distributions of projected processes conditioned on knots $\mathcal{V}_*$, then defining the projected process covariance as in (3.8) minimizes $D(\Pi_* \,\|\, \Pi)$, where $D(\cdot)$ denotes Kullback-Leibler divergence. With this construction, τ^2 can be treated as a known hyperparameter or assigned a weakly-informative prior and estimated from the

data. In the present paper, we take a hybrid approach and estimate $\boldsymbol{\theta} = (\tau^2, \psi, \nu)^{\mathsf{T}}$ from the data in an empirically Bayesian fashion. We will discuss this procedure in greater detail in section 3.3.5.

3.3.3 Prior on the logistic model sequence coefficients

Importantly, the mixture prior expressed in (3.5) and (3.7) assigns individual-specific weights to the series of mixture components. The logistic stick-breaking prior in (3.6) uses individual-level covariates to help inform the mixture weights. In this section, we complete construction of our LSBP by describing the prior π_γ on the sequence of logistic model coefficients $(\boldsymbol{\gamma}_k^\dagger)_{k=1}^\infty$.

Above, we consolidated notation by lumping fixed and random clustering effects within each vector $\boldsymbol{\gamma}_k^\dagger$. In our ensuing discussion, we assume the covariates $\boldsymbol{z}_i$ contain at minimum a global intercept term, and possibly additional terms linked to fixed and random clustering effects. We use the Gaussian scale-mixture family to express our prior model for the corresponding $\boldsymbol{\gamma}_k^\dagger$. Let $J_f \subset \{1, \ldots, Q-1\}$ denote the index set of the fixed clustering effect coefficients for our model, and let $J_r \subset \{1, \ldots, Q-1\}$ with $J_f \cap J_r = \emptyset$, denote the index set of one random clustering effect component. For all k, we let,

$$\begin{aligned}
\gamma_{k,0}^\dagger &\sim \mathcal{N}(m_0, \eta_0^2), \\
\gamma_{k,j}^\dagger &\sim \mathcal{N}(0, \zeta_{k,j}^{\dagger 2}\eta_k^{\dagger 2}), \quad j \in J_f, \\
\gamma_{k,j}^\dagger &\sim \mathcal{N}(0, \zeta_{k,r}^{\dagger 2}), \quad j \in J_r,
\end{aligned} \tag{3.9}$$

where m_0 and $\eta_0^2 > 0$ are considered known hyperparameters, the product $\zeta_{k,j}^{\dagger 2}\eta_k^{\dagger 2} > 0$ is the prior variance of the fixed clustering effect coefficients, and $\zeta_{k,r}^{\dagger 2} > 0$ is the prior variance of the random clustering effect coefficients indexed by J_r. More complex models can be written, for example, by adding extra random clustering effect and associated variance terms. In the present case, we introduce approximate selection of the fixed clustering effects by assigning those coefficients the so called horseshoe prior [21]. Other choices of prior could be made. For example, [136] explore the spike-and-slab and Laplace priors, which are also both special cases of the Gaussian scale-mixture. We express the prior on the on the fixed clustering effect hierarchical variance components,

$$\eta_k^\dagger \sim \text{Cauchy}^+(0, 1), \qquad \zeta_{k,j}^\dagger \overset{\text{iid}}{\sim} \text{Cauchy}^+(0, 1), \quad \text{for } j \in J_f, \tag{3.10}$$

where $\text{Cauchy}^+(a, b)$ denotes the half-Cauchy distribution with location a, shape b, and support on non-negative real line $\mathbb{R}_{\geq 0}$. This class of prior aggressively shrinks "small" coefficients towards zero, while leaving larger magnitude coefficients relatively unpenalized. For random clustering

effect coefficient blocks, we assign the associated variance terms relatively non-informative priors,

$$\pi(\zeta_{k,r}^{\dagger -2}) \propto \mathbb{1}(\zeta_{k,r}^{\dagger -2} > 0). \tag{3.11}$$

This construction mirrors the typical prior hierarchy for describing random effects in generalized linear models.

We partition the component logistic model coefficients in this way to highlight the importance of the component intercept parameters $\gamma_{k,0}^{\dagger}$. The general Dirichlet process model (3.1) contains a concentration parameter α which controls how quickly the sequence of mixture weights converges to zero. In the posterior, the data inform the number of occupied mixture components or clusters. A well known result for this family of models is that the expected number of clusters grows in proportion to the concentration parameter α times the log of the sample size N. Unlike in the Dirichlet process, the LSBP does not contain a single parameter that controls the mixture weight convergence rate. Rather, the $\boldsymbol{z}_i$ and the $\boldsymbol{\gamma}_k^{\dagger}$ control the component mixture weights in a complex way. In the absence of additional covariate information, however, the $\gamma_{k,0}^{\dagger}$ play a role analogous to the Dirichlet process concentration parameter.

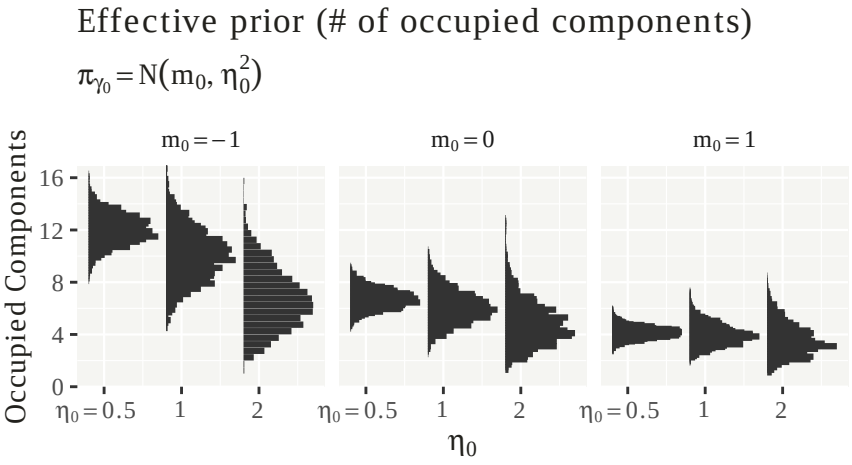


Figure 3.1: "Effective prior" on the number of active mixture components for $N = 100$ samples as a function of the prior on the LSBP intercept parameters.

By assuming m_0 and η_0^2 fixed and constant across mixture components $k = 1, 2, \ldots$ we can control the effective prior on the number of clusters we can expect to see occupied by N data points. We study the nature of this "effective prior" in a small simulation, the results of which are summarized in Fig. 3.1. In the figure, we sampled mixture weights for $N = 100$ individuals in the absence of other covariate information, assigned individuals to specific mixture components using multinomial sampling (truncating the infinite measure at N components), and counted the number of assigned or active clusters. We repeated this procedure 2,000 times for handful of different m_0

and η_0 combinations. Fig. 3.1 shows sideways histograms (counts unlabeled) of the number of active clusters generated under varying m_0 and η_0. In general, as might be expected intuitively, the hyperparameter m_0 influences the expectation and η_0^2 controls the spread of the number of occupied clusters, given N (though this is not strictly true as can be seen in the $m_0 = -1$ panel). Further, we note that for a given choice of m_0 and η_0^2, the expected number of active clusters grows roughly linearly with $\log(N)$ (not shown).

3.3.4 Posterior computation

In this section we briefly outline a blocked Gibbs sampling algorithm for posterior computation with our model. Derivations of the full conditional distributions for each block of parameters is given in Appendix J. This approach requires us chiefly to truncate the infinite sequence of random measures $\omega_k^\dagger(\cdot)$ at some manageable upper bound, which we will call T. Choice of T will in general depend on the problem at hand, the sample size N, and, as we have seen in section 3.3.3, the prior on the $\boldsymbol{\gamma}_k^\dagger$. In the greater Dirichlet process mixture model family literature, T is commonly chosen between 10–50 to limit the total number of mixture components. When following this strategy for computing with infinite mixture models, it can be important to assess sampling during run time to ensure that the number of active mixture components is somewhat less than T.

Our Gibbs sampler also relies heavily on the introduction of latent data components. Most importantly, we will introduce latent cluster labels $\mathcal{C}_i \in \{1, \ldots, T\}$ to resolve the mixture model for each individual $i = 1, \ldots, N$. Given the cluster labels, the other model parameters can be updated with relative ease. To update the $\boldsymbol{\gamma}_k^\dagger$, we will additionally have to introduce latent data to facilitate sampling of logistic models. Published latent data formulations for logistic models are available in [74] and [126].

Critically, by saving samples of $\mathcal{C}$, we can also obtain inferential summaries about cluster assignments, leading to latent group identification. For statements about cluster assignment to make sense, we must ensure that cluster labels retain their meaning over Markov chain Monte Carlo (MCMC) iterations. It is a well known problem when working with mixture models that the likelihood is frequently invariant to permutations of the mixture component indices [e.g. 83]. This can ultimately lead to the so called "label switching problem" where quantities associated with cluster k at iteration t become associated with cluster k' at iteration t', and the notion of coherent subgroup membership disintegrates in MCMC averages. Papastamoulis and Iliopoulos [122] proposed an elegant solution to this problem that we employ here. Briefly, given a reference set of cluster labels, $(\mathcal{C}_i^*)_{i=1}^N$, the labels at iteration t, $(\mathcal{C}_i^{(t)})_{i=1}^N$, can be realigned by minimizing some function that measures the cost associated with changing label k into label k'. We follow [122],

and use,

$$L(\mathcal{C}^*, \mathcal{C}) = -\sum_{i=1}^{N} \mathbb{1}\left(\mathcal{C}_i^* = \mathcal{C}_i\right)$$

as our cost function. Finding a minimum cost relabeling solution is known as the "assignment problem" [91], and can be solved quickly using a number of existing algorithms. For our purposes, we have implemented the "primal-dual" algorithm of [20] to solve our cluster label assignment problem efficiently.

Though this procedure can be fooled if, for example, the estimate of $\mathcal{C}^*$ is poor, we have observed it to work well practically. With a reasonable estimate of $\mathcal{C}^*$, the [122] algorithm can mitigate interpretive issues that arise due to label switching. Importantly, the per-iteration computational cost of doing so grows with N but not with model complexity, and in many cases can be negligible compared to other parameter updates. In practice, we allow MCMC chains free exploration of the posterior during burnin including label switching moves. We keep track of the overall log posterior during warm up, and estimate $\mathcal{C}^*$ as the clustering associated with the highest value of the log posterior over burnin iterations. We can then permute the cluster labels in $\mathcal{C}^*$ so that indices $k = \{1, \ldots, T\}$ are sorted in descending order of their associated cluster sizes. Then during sampling, we relabel and reorder the clusters based on $\mathcal{C}^*$ to maintain consistent subgroup meaning over saved iterations.

As a final note, it can be useful to think about reasonable soft starting points for our algorithm, which may help accelerate convergence. In the present case, we have chosen to initialize our posterior computation algorithm with a single k-means scan of the outcome images without reference to any of the clustering covariates $\boldsymbol{z}_i$. With this scheme, we require only that the initial number of clusters be less than or equal to T.

3.3.5 Estimation of the Gaussian process hyperparameters

In typical 2D spatial regression settings it is common to use Gaussian processes to model a spatially correlated variance component [e.g. 33]. In cases like these, practitioners can estimate the Gaussian process hyperparameters by integrating the Gaussian random field out of the likelihood and maximizing with respect to $\boldsymbol{\theta}$ [e.g. 110]. We can estimate $\boldsymbol{\theta}$ prior to MCMC in the same manner by first marginalizing the random function $\mu_i(\cdot)$ out of the likelihood. This turns out to be relatively easy since (i) our prior on the $\beta_{k,j}^{\dagger}(\cdot)$ is conjugate to the likelihood; (ii) the mixture weights $\omega_k^{\dagger}(\cdot)$ sum to one for any input; and (iii) we write our model with $\boldsymbol{\theta}$ homogeneous across all of the mixture components.

To develop the marginal likelihood for our problem, let $s = 1, \ldots, M$ index the set of voxel locations B, with $M = |B|$. Let $\boldsymbol{y}_i$ denote the vectorized set of image outcomes for individual

i so that $y_{i,s} = y_i(\mathbf{v}_s)$, and let $\boldsymbol{\Sigma} = \text{diag}\{\sigma^2(\mathbf{v}_1), \ldots, \sigma^2(\mathbf{v}_M)\}$. Also, let $\boldsymbol{K}$ denote the prior covariance of one of the $\beta^{\dagger}_{k,j}(\cdot)$ over all locations in B so that $K_{s,s'} = \tau^2\rho(\mathbf{v}_s, \mathbf{v}_{s'})$. Then, ignoring the integration constant, we can write the log marginal likelihood of individual i's outcome image,

$$f(\boldsymbol{y}_i \mid \boldsymbol{\Sigma}, \boldsymbol{\theta}) = -\frac{1}{2}\ln\det\boldsymbol{\Lambda}_i - \frac{1}{2}\boldsymbol{y}_i^{\mathsf{T}}\boldsymbol{\Omega}_i^{-1}\boldsymbol{y}_i,$$

where $\boldsymbol{\Lambda}_i = \boldsymbol{\Sigma} + \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_i\boldsymbol{K}$. In practice, evaluating this function is not computationally tractable if the number of locations in B is large enough: decomposition of the $\boldsymbol{\Lambda}_i$ is in general an $\mathcal{O}(M^3)$ operation, and can be very slow when M is more than several thousand. We work around this issue programmatically by simply evaluating $f(\cdot)$, etc. over some fixed subset of the locations in B. In practice we have handled this by taking a random subset of B of a given size. For this optimization step, we make the further simplifying approximation that $\sigma^2(\mathbf{v}) \equiv \sigma^2$ for locations $\mathbf{v}$ in the spatial subset. If for example the covariance function is taken of radial basis family as in (3.8), then these simplifications mean that we only have to optimize $f(\cdot)$ over four parameters, $(\sigma^2, \tau^2, \psi, \nu)$. In the present work, we have accomplished this with off-the-shelf software for constrained optimization [127, 129]

3.4 Simulation study

3.4.1 Small 3D simulation design

We created a small scale simulation to study the performance of our method when subgroup membership is known, and to illustrate what we can expect to gain compared to simpler methods of subgroup discovery. Fig. 3.2 gives a schematic example of our simulation design in this setting. We simulated data on a $32 \times 32 \times 16$ grid as follows. Each simulated individual i was assigned to a discrete grouping covariate with ten levels. We treated this grouping factor as the $\boldsymbol{z}_i$; individuals were evenly divided amongst the levels of this factor. As shown in the bottom of Fig. 3.2, the levels of this grouping factor were differentially associated with each of nine true subgroups indicated by the letters "A" through "I." Conditional subgroup probabilities (shown in grayscale in the bottom of Fig. 3.2) were sampled from an LSBP once and then fixed across our simulations. Letter subgroup assignments, however, were resampled for each simulation iteration. Here we refer to the subgroup assignments as the true cluster labels, which we denote by $\mathcal{C}^{\star}$, where $\mathcal{C}^{\star}_i = 1$ denotes letter "A" group membership, etc.

For simplicity, we generated individuals' outcome images using a simple model with just two spatially varying terms: a global intercept and a coefficient linked to a scalar covariate. Let $\beta_{i,j}(\cdot)$ be shorthand to refer to the true jth covariate function for individual i—i.e. $\beta^{\dagger}_{\mathcal{C}^{\star}_i,j}(\cdot)$ more pre-

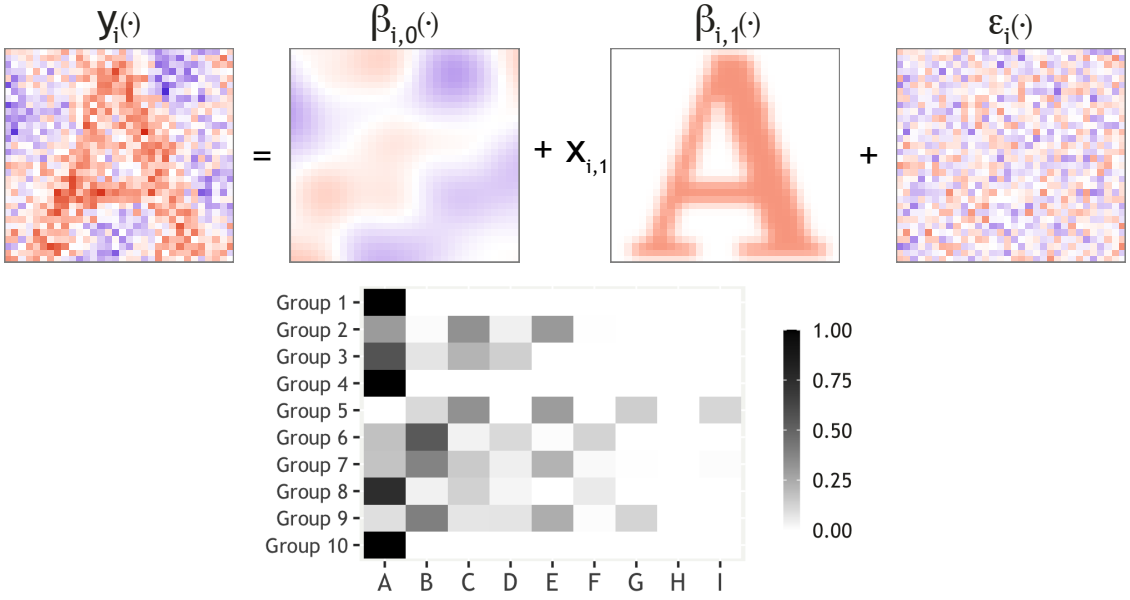


Figure 3.2: Simulation design with up to nine true subgroup effects corresponding to $\beta_{i,1}(\cdot)$ equal to a smoothed letter image (one of "A" through "I") for each simulated individual $i = 1, \ldots, N$. The top row of the figure shows a schematic example of our data generating mechanism, while the bottom row shows the true probability of subgroup assignment for simulated individuals in one of ten discrete bins (Groups 1–10).

cisely. We sampled the intercept coefficients, $\beta_{i,0}(\cdot)$, from a mean zero Gaussian process with a six voxel full-width-at-half-maximum (FWHM) Gaussian covariance function (unit marginal variance). These parameters were resampled for each simulation iteration, but fixed within iteration to be the same across all individuals so that $\beta_{i,0}(\cdot) \equiv \beta_0(\cdot)$. Covariate coefficient images, $\beta_{i,1}(\cdot)$, were created by embedding binary images of capital English alphabet letters within the central six (32×32) slices of otherwise null 3D images. We then smoothed the letter boundaries by convolving the entire image with a two voxel FWHM Gaussian kernel. Letters were chosen as convenient examples of partially overlapping non-convex shapes. Certain pairs of letters exhibit a high degree of overlap as well. In the B-D pair, for example, coefficient images have correlation close to 0.9. The resultant $\beta_{i,1}(\cdot)$ images were linked with scalar covariates $x_{i,1}$, which we simulated as standard normal variates. Finally, error images $\epsilon_i(\cdot)$ were drawn from a stationary Gaussian white-noise process with $\sigma^2(\mathbf{v}) \equiv \sigma^2$ set to control the overall spatial signal-to-noise-ratio (SNR) to be either 0.1 or 1. Although stationarity of the error process represents a simplification of typical real data settings, the spatial SNRs are somewhat more realistic. Our setting with SNR = 1 was designed to mimic the (rather large) SNRs we have observed throughout different outcome image types in the ABIDE data (see section 3.5.2), while in our experience SNR = 0.1 may be more realistic for analysis of task-based fMRI contrast images. The top row of Fig. 3.2 shows the central 32×32

slice of data for an example individual.

3.4.2 Simulation results

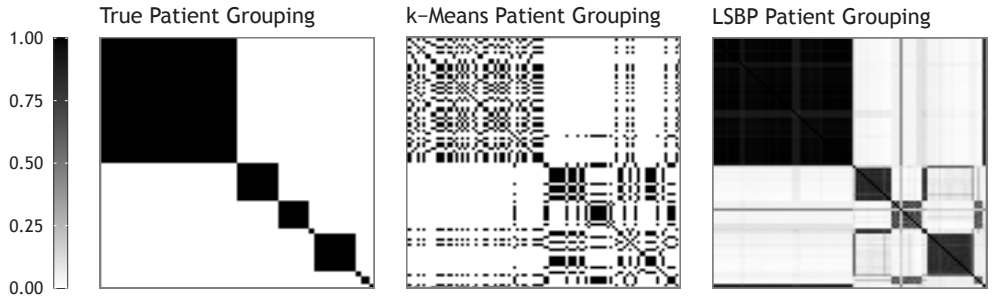


Figure 3.3: Individual-level co-clustering incidence matrix. Each sub figure represents an $N \times N$ grid of pairwise probabilities that individual i belongs to the same group as individual j (dark—high, light—low). (*Left-most*) In the true individual grouping matrix, blocks of dark tone correspond, in this particular example, to data simulated with one of eight letter images for the covariate coefficient ($\{$A–G, I$\}$). Compare against estimated individual co-clustering matrices discovered by k-means regression (*center*) and our method (*right*).

In this section we first give an illustration of the gain that can result from using a covariate-aware clustering method like ours versus a more typical approach to subgroup discovery. A number of published papers have attempted to identify latent subgroups in neuroimage data using algorithmic methods like k-means [40, 26, 82], or hierarchical clustering [78, 75, 38]. This procedure is known as "neurotyping" [see e.g. 76]. Practitioners then try to associate identified clusters with individual-level covariates for some *post hoc* measure of cluster validation. While this method of analysis may work well for a given data set, we have noticed in simulation that clustering performance can be unsatisfactory.

Here, we take k-means as an archetype for comparison. In the presence of mean model covariates, the general k-means algorithm can be improved upon by clustering regression models instead of outcomes. Typically, k-means starts from an initial partition of the data and iteratively reassigns each observational unit to the cluster with the nearest centroid [69]. If a regression-based mean model is useful we can generalize this procedure to treat within-cluster regression models as centroids rather than within-cluster means [e.g. discussed in 183]. In the case of neuroimaging data, it is convenient and computationally fast to use the typical voxel-wise analysis to form regression-based centroids for whole images. We have implemented such an algorithm here for comparison, and illustrate its relative clustering performance in Fig. 3.3. In the figure, we chose a representative example from our simulated data with the relatively easy setting spatial SNR = 1, and $N = 100$. There were exactly eight true clusters in this particular iteration (see the *left* panel

of Fig. 3.3), and we thus asked k-means to identify eight clusters given this oracle-like knowledge of the truth. Anecdotally, applying *some* spatial smoothing to the outcome data prior to analysis with our k-means regression procedure was critical to improve clustering performance, but only up to a point. In this particular example, application of a six (or greater) voxel FWHM Gaussian smoothing kernel led to degraded clustering performance; a two voxel FWHM kernel led to best performance; not smoothing the data led to clustering with the worst concordance with the truth. In contrast, our hierarchical mixture of spatial regressions model does not require the data to be pre-smoothed: spatial smoothing is handled flexibly in (3.7) with Gaussian process priors on the spatially varying coefficient functions.

In Fig. 3.3, we present the best case scenario for k-means regression (algorithm initialized to eight clusters, outcome smoothed using a two voxel FWHM Gaussian kernel), and compare against our covariate-informed LSBP model. The figure shows a set of $N \times N$ individual co-clustering matrices where individuals have been sorted by their true subgroup membership. The leftmost panel of the figure shows the true blocks of {A–G, I} individuals on the diagonal, the center panel shows the co-clustering resulting from k-means regression, and the rightmost panel summarizes the co-clustering probabilities estimated by our proposed LSBP method. The performance of our LSBP method is far from perfect, but concordance with the truth is high nonetheless. Here, our model only identifies the four major clusters ({A, B, C, E}), and generally confuses the other four minority groups as likely belonging to their nearest majority relative. For example, most of the errors made by our method result from combining subgroups D with B, F with E, and G with C. Within the majority blocks, however, performance is quite good. For example, all of the individuals in the true A block have over 90% posterior probability of membership to the first and largest cluster. In contrast, k-means has split individuals in the true A block across four different clusters here (this granularity is not quite apparent in the figure), and overall concordance with the truth is somewhat limited.

With this example in mind, we went on to quantify both the clustering and estimation performance of our proposed method in repeated simulation. We simulated data as in section 3.4.1 for $N = 100$ and $N = 300$ individuals, repeating the simulation 50 times for each combination of the spatial SNR and N. We fit our LSBP model to the data from each iteration using Gibbs sampling and truncating the stick-breaking process at $T = 15$ components. In all cases, we set the hyperparameters for the LSBP intercepts $m_0 = 0$ and $\eta_0^2 = 0.5$ to encourage the number of occupied mixture components to be less than 15, and set the initial number of clusters to five. For the spatial process components of our model, we conditioned on a set of 300 random knots sampled uniformly throughout our $32 \times 32 \times 16$ image grid. We also conditioned as well as on the true Gaussian process covariance hyperparameters used to generate our spatial intercept $\beta_0(\cdot)$. For each simulation, we ran 7,000 total Gibbs sampling iterations, discarding the first 4,000 as burnin, and

SNR	N	Method	Mutual Info.	RMSE: β_0	RMSE: β_1
0.1	100	LSBP	35.2% (0.6)	81.5 (3.6)	149.5 (2.3)
0.1	300	LSBP	52.5% (0.6)	139.2 (2.4)	181.1 (3.1)
1.0	100	LSBP	51.2% (0.8)	137.7 (3.0)	180.8 (2.5)
1.0	300	LSBP	66.7% (0.4)	153.7 (2.7)	188.0 (1.8)
0.1	100	k-means	36.6% (0.8)	1674.9 (101.4)	1365.1 (41.0)
0.1	300	k-means	31.9% (0.6)	808.3 (18.4)	755.6 (12.1)
1.0	100	k-means	36.3% (0.8)	592.0 (34.0)	492.6 (13.2)
1.0	300	k-means	32.4% (0.6)	280.2 (5.6)	294.0 (3.8)

Table 3.1: Inference quality for our proposed method when data exhibit different levels of noise. Rows marked "LSBP" correspond to our proposed method; we have also included results from k-means regression for reference. The spatial signal-to-noise-ratio (SNR; averaged over simulated individuals) is given in the first column and reflects high (SNR = 0.1) and low (SNR = 1) noise settings. The column "Mutual Info." gives the mutual information between the true group labels and the posterior distribution of cluster labels $\mathcal{C}$. We express mutual information as a percentage of the maximum possible value (perfect, noiseless concordance between the true and estimated group labels). The "RMSE" columns give the root mean squared error for each spatial coefficient, averaged over simulated subjects. Values in the RMSE columns have been scaled by multiplying by 10^3 to facilitate comparison. Results are presented as mean (standard error).

saving 1,000 samples over the remaining 3,000 iterates.

Table 3.1 summarizes our LSBP method's clustering accuracy over simulations using the mutual information (MI) [148] between the true subgroup labels and posterior distribution of the mixture component weights for each simulated individual. When used as a clustering metric, MI can help us distinguish between two posterior distributions of cluster allocations that have the same mode, say, but different levels of noise. In the table, we have expressed MI as a percentage of the MI between the truth and itself (i.e. perfect, noise-free clustering); MI of 0%, moreover, would correspond to independence between the truth and posterior distribution of the component weights. For a frame of reference, in the example in Fig. 3.3, k-means regression has 35.1% MI with the truth, while our LSBP approach has 57.9% MI. The root mean squared error (RMSE) columns of Table 3.1, meanwhile, suggest good accuracy estimating the cluster-specific spatially varying coefficient functions. RMSE is generally a little higher for the "β_1" parameters than for the spatial intercept parameters due to errors in the clustering. To help give a sense of scale, Table 3.1 also provides results for k-means regression (bottom rows). With the exception of the lowest SNR/lowest N setting, our LSBP method performs dramatically better in terms of both clustering and estimation accuracy.

3.5 Analysis of data from the Autism Brain Imaging Data Exchange (ABIDE)

In this section we illustrate use of our proposed method using data from the Autism Brain Imaging Data Exchange (ABIDE; Release I). ABIDE is the product of an international research consortium whereby previously collected imaging and demographic data have been aggregated for broad scientific use [36]. The ABIDE repository contains data from hundreds of patients with Autism-spectrum behavioral disorders (ASD) and age-matched neurotypical control participants. Data aggregated by the exchange focus on measures derived from resting state fMRI, an imaging modality that has had broad utility in both clinical and neuropsychological research settings [e.g. see 96, 154, for reviews]. Resting state fMRI measures low frequency fluctuations in the blood oxygen level dependent signal in the absence of any explicit experimental paradigm, and has been widely used to derive patterns of correlation or functional connectivity between brain regions [e.g. 13].

3.5.1 Description of the outcome image data

In total, the ABIDE contains data collected at 17 research institutions from across the continental United States and Europe. We downloaded imaging data from the Processed Connectomes Project [32] that were fully preprocessed using the Configural Pipeline for the Analysis of Connectomes (C-PAC[1]). A complete description of preprocessing methods for this pipeline is available online.[2] Briefly, standard fMRI time series correction steps were applied using AFNI software [30, 31], and the data were registered to anatomical space using tools from FSL [152, 84]. Here, we consider weighted degree centrality images as our outcome summary of resting state connectivity. Weighted degree centrality is a graph theory summary statistic and simply reflects, for each node in a weighted graph, a summation over all the connecting edge weights. In this case, each voxel is treated as a node in a large graph, and edge weights are taken to be the correlations between the time series data measured at each node. Additional details on centrality computation for the C-PAC pipeline are available online.[3] The version of the data we consider was derived from time series that were temporally filtered using a 0.01–0.1 Hz band pass.

[1] https://fcp-indi.github.io
[2] http://preprocessed-connectomes-project.org/abide/cpac.html
[3] https://fcp-indi.github.io/docs/latest/user/centrality.html

3.5.2 New York University subsample

In initial exploratory analyses of these data, we noticed that some of the largest differences between individual images could be explained by imaging site effects. Around four natural clusters emerged dominantly on the bases of groups of collection sites. This result is not especially surprising: although the weighted degree centrality images we consider were derived from intensity-normalized time series, site effects due to scanner, fMRI protocol, or even demographic differences can still persist in practice. Preprocessing for multi-site imaging studies is still an area of active research [e.g. 25]. Exploring this issue further, we estimated the spatial signal-to-noise-ratios for the data from each site. We constructed these estimates in the context of our proposed model using maximum marginal likelihood to optimize the Gaussian process hyperparameters and spatial error variance. We optimized over a set of 4,000 voxels sampled randomly from within a gray matter mask. The range of spatial SNRs we estimated this way was roughly 0.46–0.84.

With the general goal in mind of estimating latent clusters related to potential Autism neurotypes we simply subset the data to the cohort of ASD individuals imaged at New York University. The NYU cohort represents the largest single-site group in the ABIDE repository. We further subset to the group of ASD patients with a complete set of Autism Diagnostic Interview-Revised (ADI-R) scores available, resulting in a final sample of 64 ASD individuals. The ADI-R [143] is an interview-based instrument with four component scores designed to differentiate Autism from other developmental verbal or intellectual impairments. ADI-R scores have been validated against clinical diagnoses [e.g. 104, 95], and have been used in research to help characterize and understand Autism phenotypes [e.g. 79, 155]. A demographic summary of our subset of 64 ASD patients is available in Table 3.2. In subsetting the data in this way we have chosen to prefer simplicity and model structure generality over complete use of all of the information in the data. A more data-specific method could of course be developed based on our model to more appropriately incorporate the complex site heterogeneity we observe into the analysis.

3.5.3 Structure of our mean and clustering models

In general, given the somewhat elaborate hierarchy of our model, we find it most helpful to work with relatively simple mean models and build up complexity in the clustering covariates. This strategy can also be advantageous computationally. In the present context, we take our mean model to be a spatially varying intercept so that the $\boldsymbol{x}_i$ are simply scalars equal to one for all i. In our clustering model, however, we consider a number of covariates including: diagnosis category (Primary Autism, Asperger syndrome, Pervasive developmental disorder–not otherwise specified); patient gender; patient age at scan; and the ADI-R component scores (A–Reciprocal Social Interaction; B–Communication; C–Restricted, Repetitive, and Stereotyped Patterns of Behavior; and

Diagnosis	Percentage (SD)
Primary Autism	71.88% (42.36)
Asperger Syndrome	23.43% (44.96)
PDD-NOS	4.69% (21.14)

Measure	Mean (SD)
Age (yrs)	13.87 (5.61)
ADI-R Social	19.00 (5.73)
ADI-R Verbal	15.59 (4.49)
ADI-R RRB	5.41 (2.57)
ADI-R Onset	3.33 (1.36)
ADI-R Total	43.33 (10.92)
Fluid IQ	107.59 (15.72)
Performance IQ	108.89 (17.24)
Verbal IQ	105.34 (15.14)

	Percentage (SD)
Male	90.62% (29.15)
Comorbidities	51.56% (49.98)
Medication	21.88% (41.34)

Table 3.2: Demographic information for the 64 Autism-spectrum patients scanned as part of the NYU cohort. Rows "Comorbidities" and "Medication" respectively denote the proportion of patients experiencing one or more comorbidity factors, and patients prescribed some medication to treat their behavioral disorder. ADI-R—Autism diagnostic interview, revised; RRB—Restricted, repetitive, and stereotyped patterns of behavior; PDD-NOS—Pervasive developmental disorder, not otherwise specified.

D–Abnormality of Development Evident at or before 36 Months). Each of the four ADI-R component scores can take multiple ordinal values where higher numbers indicate increasing presence of ASD symptoms. Linear terms for each of these covariates were treated as fixed clustering effects in our prior (3.9). Non-binary covariates were first mean-centered and scaled by dividing by two standard deviations. In addition to the linear terms, we created five sets of Gaussian bases to model potential non-linear effects of age and the four ADI-R component scores. We associated each non-linear basis with a separate random effect-like variance component in our prior (3.9). Bases were constructed identically for each covariate, and followed the general form,

$$g(z; z^*, b) = \exp\big(-b\|z - z^*\|^2\big),$$

given knot location z^* and bandwidth parameter b. For each covariate, we took the ten internal $5^{\text{th}}, \ldots, 95^{\text{th}}$ quantiles as knot locations, and set the corresponding bandwidth b to be two divided

by the maximum absolute distance between the knots—i.e. if z_l^*, $l = 1, \ldots, 10$, represent knot locations, we fixed $b = 2/\max_{l,l'} |z_l^* - z_{l'}^*|$. This choice resulted in smooth bases for the age and ADI-R score covariates. More work could be done to estimate or validate the knots and bandwidths; our present focus is simply to illustrate how generalized additive-like constructs can be incorporated into the logistic stick-breaking framework.

3.5.4 MCMC and model evaluation

Prior to fitting our model, we further preprocessed the degree centrality images by centering and rescaling them (with scalar factors) to have marginally zero mean and unit variance. We computed the scalar recentering and rescaling factors collapsing the data over all voxels and patients. We estimated the Gaussian process hyperparameters using maximum marginal likelihood (see section 3.3.5). The resulting estimate, $\boldsymbol{\theta} = (0.27, 0.03, 1.67)^{\mathsf{T}}$ corresponds to a sub-Gaussian, 12.6 mm FWHM radial spatial basis with marginal variance 0.27. Spatial knot locations $\mathcal{V}_*$ were sampled uniformly over voxel locations in a gray matter mask. The number of knot locations $|\mathcal{V}_*| = 5,600$ was set to comprise about 11.3% of the total number of locations in the mask; with this configuration, the minimax distance between spatial knots was 9.5 mm. To set the hyperparameters of our stick-breaking process (m_0 and η_0^2; see section 3.3.3), we looked to experimental literature. Previous work on neurotyping ASD generally suggests between two and four neural subtypes [see 76, for a comprehensive review]. Given this prior knowledge, we set $m_0 = 0$, and $\eta_0^2 = 0.5$ to place high "effective prior" mass around or slightly above this range. Algorithmically, we similarly truncated stick-breaking after $T = 12$ components.

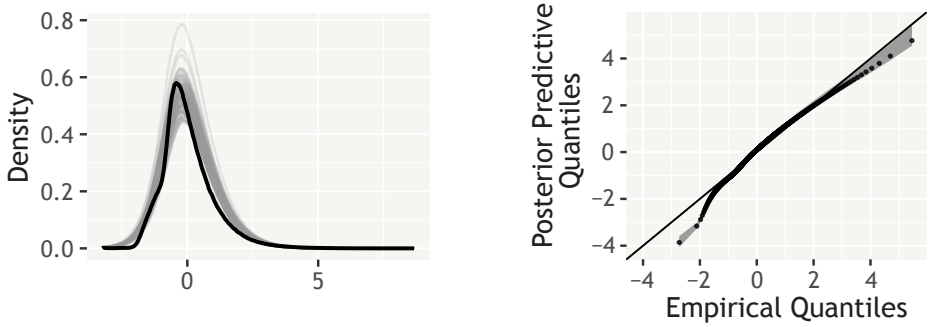


Figure 3.4: Comparison of the empirical and posterior predictive distributions. The black line in the left panel depicts a kernel density estimate of the empirical distribution of weighted degree centrality across all voxels and individuals in our subsample. Gray lines reflect the uncertainty in the kernel density estimate over the posterior predictive distribution. Similarly, the right panel provides a Q-Q plot for the empirical and posterior predictive distributions of degree centrality. Dots show the mean, and the dark gray ribbon shows 95% credible intervals for the predictive quantiles.

We fit our proposed model using Gibbs sampling as outlined in section 3.3.4. We ran a total of 4,500 Markov chain Monte Carlo iterations, discarding the first 2,500 as burnin and saving sampled parameters from every other iteration thereafter. As discussed in [70] it can be difficult to assess Markov chain mixing and convergence in models that have a stick-breaking representation. In part, this issue arises due to multi-modality in the posterior. Hastie and colleagues suggest monitoring the log marginal partition posterior ($\ln \pi \left[\mathcal{C} \mid \{y_i(\cdot), \boldsymbol{x}_i, \boldsymbol{z}_i\}_{i=1}^N\right]$ in our notation) as a convergence criterion [70], though this quantity is not analytically tractable in our case. Here, we somewhat rudimentarily use additional Markov chains to assess stability of the between-chain modal cluster assignments for each patient. We further appeal to the posterior predictive distribution for an overall representation of model fit.

For the present analysis, we ran three total MCMC chains in parallel, and present inference from the chain with the lowest deviance information criterion (DIC) [24, 156]. We found that the modal cluster assignments for the other chains were 95.3% and 75.0%, respectively, in agreement with that of the lowest DIC chain, suggesting reasonably stable cluster identification. In Fig. 3.4, we present an overall assessment of the posterior predictive distribution. In the left panel of the figure, we have overlaid a kernel density estimate (KDE) derived from the empirical distribution of degree centrality on KDEs from derived from posterior predictive simulations. This sub-figure shows how our mixture model is able to capture most features of the skewed degree centrality distribution. Confer with the right panel of Fig. 3.4, where we translate this same information into a Q-Q plot. Between the two panels of the figure, it is evident that our model fits the data reasonably well, with the exception of the unusual flat left tail in the empirical distribution.

3.5.5 Posterior inference

In our model, the mixture components reflect expected imaging outcomes for hypothetical neurotypes related to degree centrality in ASD patients. Fig. 3.5 summarizes covariate relationships with modal cluster assignment as well as the posterior uncertainty in the clustering. We found that there were three occupied clusters at the posterior mode (Fig. 3.5, *right*). The largest, cluster one, was occupied by 50% of the sample; cluster two was occupied by 17.2% of the sample; and the third cluster was occupied by 32.8% of the sample. Further, we found that identified clusters were to some extent related to certain ADI-R component scores. In the left panel of Fig. 3.5, we show box plots of ADI-R Communication and Repetitive, and Stereotyped Patterns of Behavior (RRB) component scores for each cluster. As a convenient summary of this information we fit proportional odds logistic regression models to the ordinal ADI-R component scores using modal cluster membership indicators as predictors and controlling for age, sex, and diagnostic category. Cluster two was associated in particular with lower ADI-R RRB scores: the odds of higher ADI-R

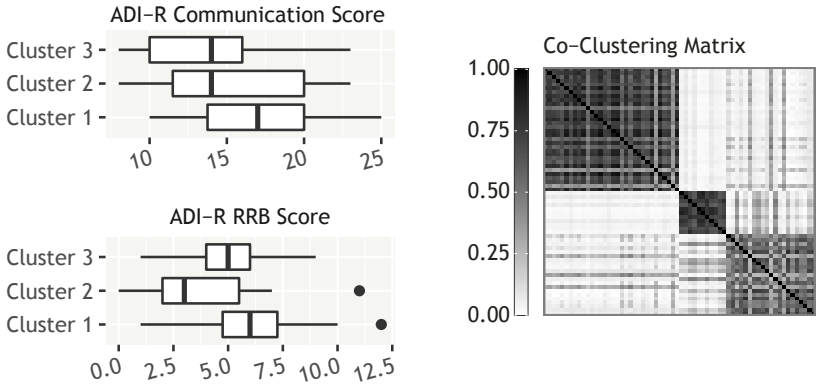


Figure 3.5: Relationship between ADI-R components and hypothetical neurotypes (*left*). Box plots show medians and interquartile ranges of ADI-R component scores for each cluster. Clusters themselves are derived from the posterior mode summary. (*Right*) Patient co-clustering matrix. Rows and columns in the matrix correspond to ASD patients from the NYU cohort: each (i, j) cell in the grid reflects the pairwise posterior probability that patient i belongs to the same cluster as patient j.

RRB scores were roughly 83% lower for cluster two patients than for cluster one ($z = 2.49$; confer with Fig. 3.5). Similarly, we observed a suggestive association between Cluster three and ADI-R Communication scores, with the odds of higher communication scores roughly 58% lower for cluster three than for cluster one ($z = 1.61$). To a lesser extent, we also observed similar patterns of results for the other two ADI-R component scores (not shown).

In Fig. 3.6, we show posterior mean z-statistic images for the differences between hypothetical neurotypes. In the figure, we have thresholded the z-statistic contrasts based on posterior credible bands. Posterior credible bands summarize the joint behavior of the contrasts over all voxels simultaneously, and can be easily estimated from MCMC samples [see e.g. 142]. In Fig. 3.6, the highlighted areas mark a set of voxels over which our model posterior suggests an 80% probability the cluster differences are simultaneously less than zero. As the figure suggests, we observed marked differences between cluster one and cluster two, with cluster two exhibiting lower average degree centrality throughout most of the brain. This result can be interpreted to mean that the average voxel-wise resting state functional connectivity graph is more dense overall for cluster one patients than for cluster two patients. Concordantly, cluster one was associated with higher median ADI-R subscores, and cluster two with lower ADI-R subscores (Fig. 3.5). This result may partially help resolve differential reports of hypo- [e.g. 113] and hyperconnectivity [e.g. 161] related to ASD. Addressing these differential findings is not a goal of the present analysis, however, and it is difficult to say more without a formal comparison to neurotypical individuals.

Cluster three largely represented a middle ground between clusters one and two. In fact, as can be seen in Fig. 3.5 (*right*), a small handful of patients tended to switch between clusters three and

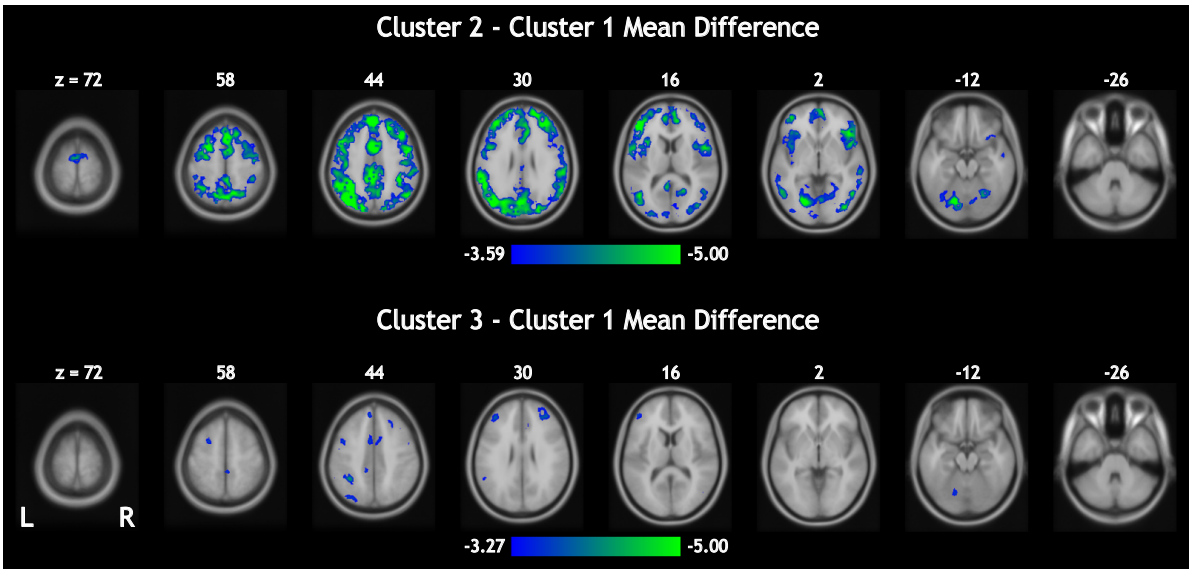


Figure 3.6: Differences in the expected degree centrality images between hypothetical neurotypes. In the figure, color bars measure differences on a z-statistic scale, while "z $= 72, \ldots$" corresponds to the axial slice in anatomical MNI152 coordinates. Images were thresholded using a simultaneous 80% posterior credible band.

one or three and two. As such, in the present analysis no brain regions appeared significantly different between clusters two and three at the 80% simultaneous posterior credible level. Differences between clusters one and three, however, are shown in Fig. 3.6. In this case, our credible band highlighted a more select set of brain areas with significant contrast between the clusters. Selected voxels appeared in the anterior cingulate cortex, the right frontal pole, and the frontal gyrus including the superior, inferior, and middle divisions. This result is consistent with applied work that has implicated the frontal pole and middle frontal gyrus in ASD neurotyping [82]. This study used a k-means approach to cluster regional functional connectivity patterns derived from resting state fMRI. Further, within their ASD clusters, the authors reported correlations between connectivity (using network edge weights) and Communication and RRB subscores for the Autism Diagnostic Observation Schedule (ADOS) [103], another clinical tool for characterizing ASD severity [82]. This overall pattern of results seems quite consistent with what we observe here.

Other similar findings have been reported for neurotypes estimated from structural MRI data measures (regional gray matter volume or cortical thickness). In particular, two studies estimated neurotypes from structural images using either k-means or hierarchical clustering [26, 76]. In both cases the reports found relationships between their neurotypes and ADOS scores, though it should be noted that the studies used overlapping samples.

3.5.6 Comparison with other methods

We compared our analysis against results from unsupervised clustering methods: k-means; hierarchical clustering using Euclidean distance and average, complete, or single linkage; and hierarchical clustering with feature selection [177] using Euclidean distance and average, complete, or single linkage. In all cases, we fixed the number of desired clusters to three to be able to compare with the results from our method. None of these methods were able to produce clustering with greater than 48% similarity to ours, with the closest coming from standard hierarchical clustering with average linkage. In addition, none of these comparison methods produced clusters that recovered any kind of association with ADI-R subscores with the exception of k-means. Interestingly, k-means identified clusters that were associated with ADI-R Communication scores controlling for patient age, sex, and diagnostic category as in section 3.5.5. We assessed the stability of this finding by running the k-means algorithm from 100 different starting points. In this small follow up analysis, we found a significant relationship between clusters and the ADI-R Communication scores in 54 out of 100 replicates.

3.6 Discussion

Models based on Dirichlet process mixtures of Gaussian processes have been proposed for nonparametric regression-prediction with univariate outcomes [e.g. 132], and for modeling nonstationarity in spatial processes in a single sample setting [e.g. 56]. Our framework, however, is defined by high dimensional outcomes and potentially many subjects, and necessitates a somewhat different approach. In our case, we utilize the logistic stick-breaking process [136] to model effect heterogeneity across participants. We have used this mixture model formulation to induce a subgrouping effect construct where the mixture components are different regression models with spatially varying coefficient functions. In a toy example, we have shown how our covariate-informed approach to clustering and subgroup identification can yield superior results to a commonly used unsupervised method. Finally, we have illustrated use of our method in application to neurotyping Autism spectrum patient data from the ABIDE I database.

In our analysis of the ABIDE I data we selected patients from a single imaging site (New York University) in order to avoid undesirable clustering dominated by site effects. In principle, our model could be modified in a data-specific manner to be able to take full advantage of the rich information in the ABIDE repository. These particular data seem to imply the need for some form of a site-specific variance component external to the main clustering process. How best to model this information in an identifiable way, however, requires careful thought. Assume for the sake of argument that random effect-like, site-specific, spatial variance components were to be

included as explicit parameters in the model hierarchy. Estimation may then possibly become numerically unstable if, for example, the LSBP returns a cluster allocation similar to the inherent patient-within-site grouping. In addition, in our analysis of the ABIDE I data, we observed that the signal-to-noise-ratios may differ quite substantially across the different data collection facilities, potentially complicating the issue further. The need to address difficulties like these represents, in miniature, the generally abundant possibilities for novel methods development in this area.

The method we present here may be limited in practical use due to the computational cost associated with clustering spatial effects in a high-dimensional regime. To that end, it may be of interest to adapt one of the fast, approximate posterior inference algorithms that have been developed for Dirichlet process-family models [e.g. 174, 72] into the present setting. Use of either the [72] stochastic variational or [174] local mode-finding algorithms would additionally simplify the difficult task of assessing algorithmic convergence. As we have seen, and as has been studied elsewhere [e.g. 70], diagnostic checking of MCMC output for infinite mixture models can be far from trivial in large data settings. Nonetheless, our simulation results and data analysis demonstrate the potential advantage that our proposed model (or indeed perhaps covariate-informed clustering methods in general) may provide in certain settings. That we were able, albeit with a different image modality and generally smaller sample size, to closely mirror existing findings from Autism spectrum disorder neurotyping studies [82, 26, 76] is an encouraging result.

CHAPTER 4

Bayesian Inference for Group-Level Cortical Surface Image-on-Scalar-Regression with Gaussian Process Priors

In regression-based analyses of group-level neuroimage data researchers typically fit a series of marginal general linear models to image outcomes at each location in the images. Spatial regularization of effects of interest is usually induced indirectly by applying spatial smoothing to the imaging data prior to analyses. While this procedure often works well, resulting inference can be poorly calibrated, particularly in smaller samples. Full spatial models for effects of interest should lead to more powerful analyses, however the number of locations in a typical neuroimage can preclude standard computation with explicit spatial models. Here we contribute a Bayesian spatial regression model for group-level analyses, and study the utility of such a model in the context of neuroimage data referenced by locations on the cortical surface. We induce regularization of spatially varying regression coefficient functions through Gaussian process priors. When combined with a simple nonstationary model for the error process, our prior hierarchy can lead to more data-adaptive smoothing than standard methods. We achieve computational tractability through Vecchia approximation of our prior which, critically, permits estimation of spatially varying coefficient processes that retain full spatial rank. We outline several ways of working with our model in practice and compare performance against standard vertex-wise analyses. Finally we illustrate use of our method in an analysis of fMRI task contrast data (n-back task) from a large cohort of children enrolled in the Adolescent Brain Cognitive Development (ABCD) study.

4.1 Introduction

Modern large-scale neuroimaging studies collect massive amounts of data, often across thousands of patients, sometimes across several years [e.g. 2, 153, 166, 171]. Typically these studies collect multiple structural and/or functional scans, with the aim to probe relationships between the images

and patient-level characteristics. We focus here on an image-on-scalar regression treatment for this general framework, where patients' images are taken to be the response, and sets of individual-level scalars are considered covariates.

Since neuroimages are spatially referenced data, we can cast the image-on-scalar problem as a functional regression of the form,

$$y_i(\mathbf{s}) = \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}(\mathbf{s}) + \omega_i(\mathbf{s}) + \epsilon_i(\mathbf{s}). \tag{4.1}$$

In (4.1) we take $y_i(\mathbf{s})$ to be the imaging outcome for patient i ($i = 1, \ldots, N$) at location $\mathbf{s} \in \mathcal{S}$, and coefficients of interest, $\boldsymbol{\beta} : \mathcal{S} \to \mathbb{R}^P$, are treated as spatially varying. Further, we decompose the error into a sum of $\omega_i(\cdot)$ and $\epsilon_i(\cdot)$ terms, where $\omega_i(\cdot)$ reflects individual-level deviations from the mean with an assumed spatial structure, and $\epsilon_i(\cdot)$ is taken to be a white noise process. Many classical analysis methods in imaging can be cast within this framework. For example, in the typical group-level functional magnetic resonance imaging (fMRI) analysis, the $y_i(\cdot)$ might represent contrasts of parameter estimates from within-participant first level time series analyses, and $\boldsymbol{x}_i \in \mathbb{R}^P$ might include an intercept term along with any relevant covariate information. Often, univariate models are fit marginally to the data from each location $\mathbf{s}$ in practice [e.g. 115]. This procedure tremendously simplifies estimation by avoiding modeling spatial correlations in $\boldsymbol{\beta}(\cdot)$ and $\omega_i(\cdot)$, but can lead to poorly calibrated inference (for example, see attempts to improve the power of tests derived from marginal ordinary least squares models by spatially pooling variance estimates in [117, 160, 173]).

For model 4.1 to make sense practically, the images must have reasonably comparable support in the spatial domain $\mathcal{S}$. Though it is still an area of active research, a tremendous amount of study has focused on methods to preprocess raw neuroimage data to help coregister the images across patients and data collection sites [e.g. 46, 47, 84, 137]. In particular, certain neuroimage preprocessing tools compute state-of-the art cross-subject alignment of cortical features by first mapping each hemisphere of the cortex onto the surface of a sphere with minimal distortion [46, 47]. Fig. 4.1 gives an example of such a mapping. This procedure standardizes the spatial support for each hemisphere of cortex, and has already been shown to lead to reduced spatial signal contamination and result in more sensitive analyses [e.g. 19]. Part of the gain from this methodology is due to the natural construction of a gray matter surface-based coordinate system which more accurately reflects the topology of primate cortex versus simple Euclidean distance in 3D space [46]. Recently, within the statistical community, Mejia and colleagues [111] highlighted this preprocessing pipeline by developing a cortical-surface-on-scalar regression model for task-based fMRI data. In their paper [111], the authors propose a joint multi-subject spatio-temporal regression model, model their spatial regression coefficients with Gaussian random fields, and derive an integrated nested Laplace

approximation routine for approximate Bayesian inference. Per their data application, Mejia *et al.* develop their model primarily for analysis of multi-subject fMRI data where the number of subjects is not large [111].

Such joint multi-subject spatio-temporal methods are not easily extensible to large-scale imaging studies. The number of spatial locations in a conventional neuroimage typically precludes Bayesian computation in most computing environments except by methods that either approximate (a) the spatial process by low-rank projection or downsampling, or that approximate (b) the posterior distribution with variational or Laplace family approximations [see e.g. 124, 149, 111]. In general, low-rank projection methods can tend to miss or over smooth local features in data [e.g. 157], and both low-rank projection and variational approximation can commonly underestimate posterior variance [e.g. 172, 133]. Integrated nested Laplace approximation, moreover, is thought to give accurate and scalable approximations within a wide class of posterior distributions [e.g. 141], but its accuracy can sometimes suffer when model structure is complex [see e.g. 163]. Here, we expand on this body of work and show how a Bayesian model with a prior hierarchy related to that in [111] can permit estimation of coefficient functions that are realizations of numerically full-rank spatial processes. To be able to extend our method to large-scale imaging studies we contribute a spatial regression model intended primarily for group-level analyses of data indexed by locations on the cortical surface. In the context of group-level fMRI studies, for example, our method could simply be "plugged in" at the classical second-stage analysis, with individual-level task contrast images taken to be the response. Our method can also be flexibly applied to analysis of cortical thickness outcomes, or other structural indicators. We model the probability law governing prior uncertainty in the functions $\boldsymbol{\beta}(\cdot)$ and $\omega_i(\cdot)$ with Gaussian processes. Posterior computation is enabled by Vecchia approximation of the spatial process [169, 35, 90] and empirical Bayesian estimation of the spatial process hyperparameters. As a result of this computational innovation, we are able to estimate spatially varying regression coefficient functions that are numerically full spatial rank and thus suffer minimal approximation error.

Our model can be reasonably fit to the data from whole hemispheres of cortex using fast optimization or scalable Markov chain Monte Carlo (MCMC) routines without the need to downsample the original data. Additionally, we elaborate on an approximate working model and related Bayesian sampling scheme with computational complexity that scales almost independently of N, further allowing our method to be viable for application to large-scale neuroimaging studies. Model computation with MCMC permits natural posterior inference on the spatial extent of activation regions with simultaneous credible bands, which can facilitate spatial and multiple comparisons consistent-inference. We show our method's accuracy and sensitivity estimating the spatial coefficient functions in simulation. Finally, we use our method to analyze n-back task contrast data (z-statistic images) from the second annual release of the Adolescent Brain Cognitive

Development (ABCD) imaging collective data. Software for our methods is available online at https://github.com/asw221/gourd.

The body of this paper contains an elaboration of our spatial regression model hierarchy at the beginning of section 4.2. We continue to discuss ways to consider working with this model in sections 4.2.1 and 4.2.2. In section 4.2.3, we introduce an approximate working model that results in comparable inference on the regression function of interest $\boldsymbol{\beta}(\cdot)$ when N is moderate to large. We elaborate on our general strategy for computation in the context of this working model in section 4.2.4, and compare the performance of our approaches with standard vertex-wise univariate regression methods in section 4.3. We apply our working model method in an example analysis of n-back task contrast data from a large sample of children enrolled in the ABCD study in section 4.4. Finally, we discuss limitations and possible extensions of our methodology in section 4.5.

4.2 Methods

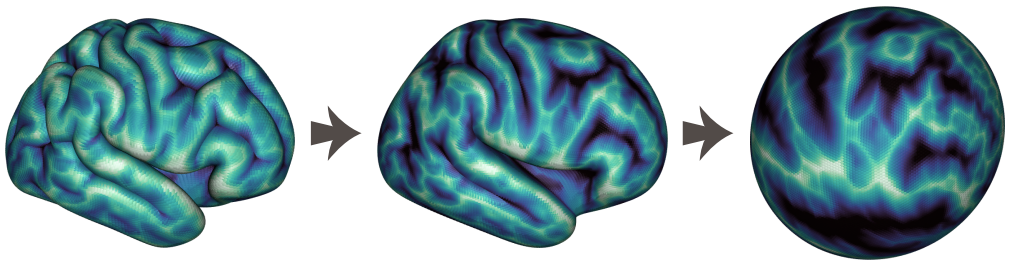


Figure 4.1: Example mapping of cortical surface coordinates onto a sphere. Left to right, the figure shows progressive inflation and warping of the right hemisphere of cortex. Gross anatomical features are highlighted to help visualize the mapping. This procedure was introduced to facilitate state-of-the-art cross-subject alignment of cortical features, but can also be leveraged into a mathematically convenient measure of geodesic distance along the cortical surface.

Throughout this work, we assume the single hemisphere, cortical surface-based, spherical coordinate system of [46]. By isolating data from the cortical sheet we gain anatomical specificity and a better connection to the underlying neurobiology. We simplify notation etc. by considering the left and right hemispheres of cortex as separate outcomes in separate analyses. Let $\mathcal{S}$ denote the set of coordinates on a sphere with a known radius R, and let $S \subset \mathcal{S}$ denote the set of vertices for a single hemisphere of cortex at which we have observed MRI data. For reference, the data in our application have all been mapped to a normalized template brain space with approximately 30,000 vertices in S. In native patient brain space, S may contain on the order of 150,000 vertices.

For any two $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$, let $d(\mathbf{s}, \mathbf{s}')$ measure the great-circle distance between $\mathbf{s}$ and $\mathbf{s}'$. Great-circle distance is sufficient for our purpose; more generally, however, $\mathcal{S}$ might represent any topo-

logical surface, etc., and $d(\cdot, \cdot)$ any appropriate metric. As has been discussed by [46, 47, 111], geodesic distances along the cortical surface are more meaningful than, say, simple Euclidean distances in the compact 3D volume. This is due to the fact that primate cortex is thought to be organized by function topographically [e.g. 151], and exhibits a folded structure in higher mammals to accommodate a larger cell body area [e.g. 23, 87]. Beginning from (4.1) above, we model the data likelihood as multivariate Gaussian with a particular error structure. We write the data likelihood:

$$y_i(\mathbf{s}) \sim \mathcal{N}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}(\mathbf{s}) + \omega_i(\mathbf{s}), \sigma^2(\mathbf{s})), \qquad i = 1, \ldots, N, \text{ and } \mathbf{s} \in S, \tag{4.2}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes the Normal distribution with mean μ and variance Σ; $\boldsymbol{x}_i \in \mathbb{R}^P$ are covariates; $\boldsymbol{\beta} : \mathcal{S} \to \mathbb{R}^P$ are the primary effects of interest; and $\omega_i : \mathcal{S} \to \mathbb{R}$ reflect individual-level deviations from $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}(\cdot)$. Conditional on $\boldsymbol{x}_i$, $\boldsymbol{\beta}(\cdot)$, and $\omega_i(\cdot)$, we model the errors as a non-stationary white noise process with spatial variances denoted by $\sigma^2 : \mathcal{S} \to \mathbb{R}_{>0}$. Given the nature of typical data in group-level functional or structural MR image analyses, this data-level model may be sufficient for a variety of studies.

Spatial dependence in our model arises entirely through our prior hierarchy on the effects $\boldsymbol{\beta}(\cdot)$ and $\omega_i(\cdot)$. Let $C_{\boldsymbol{\theta}}\{d(\mathbf{s}, \mathbf{s}')\}$ denote a positive definite stationary spatial correlation function defined on $\mathcal{S}$ with parameter $\boldsymbol{\theta}$. For notational simplicity, we will drop the subscript $\boldsymbol{\theta}$ throughout and use $C(\cdot)$ to represent a correlation function with implicit dependence on $\boldsymbol{\theta}$. We model the distributions of each spatially varying coefficient function,

$$\beta_j(\mathbf{s}) \sim \mathcal{GP}(0, \zeta_j^2\tau^2 C\{d(\mathbf{s}, \mathbf{s}')\}), \qquad j = 0, \ldots, P-1, \tag{4.3}$$

with mean zero Gaussian processes with marginal variances given by the product $\zeta_j^2\tau^2$. This class of prior for functional regression coefficients was originally proposed by [55] for general spatial regression problems. We write the coefficient processes this way without loss of generality: while zero mean processes are reasonable in our application, data from other imaging modalities may, for example, require centering at the global mean for this formulation to make the most sense. Many common constructions in linear regression can be extended to our spatial setting within this framework. For example, by creating special groups of covariates in $\boldsymbol{x}_i$ and placing additional constraints on the corresponding ζ_j^2, we can easily extend our method to accommodate spatially varying random effects and/or spatially varying penalized splines.

We similarly treat the individual-level deviations $\omega_i(\cdot)$ as spatially varying random effects with prior mean zero and marginal variance τ^2,

$$\omega_i(\mathbf{s}) \sim \mathcal{GP}(0, \tau^2 C\{d(\mathbf{s}, \mathbf{s}')\}). \tag{4.4}$$

Typically, we would not think of the $\omega_i(\cdot)$ as parameters of interest, but they are not without utility. We write the model this way to make clear how we decompose sources of spatial and non-spatial signal. Though the model can be estimated by first integrating out the $\omega_i(\cdot)$, it can be helpful to think about the $\omega_i(\cdot)$ as distinct components. For example, one could use a summary statistic of each $\omega_i(\cdot)$—such as their posterior variance, or ℓ_∞-norm, say—to help diagnose outliers in the data. With this in place, we specify a relatively simple nonstationary process for the error precisions,

$$\sigma^{-2}(\mathbf{s}) \mid \xi \overset{\text{iid}}{\sim} \text{Gamma}(1/2, \xi), \qquad \xi \sim \text{Gamma}(1/2, 1), \tag{4.5}$$

using the shape-rate parameterization of the Gamma distribution. With this formulation, the the prior on each $\sigma(\mathbf{s})$ is marginally half Cauchy with location zero and scale one. To round out our model hierarchy, we place weakly informative priors on the remaining spatial variance components,

$$\tau^{-2} \sim \text{Gamma}(1, 1/2), \qquad \zeta_j^{-2} \overset{\text{iid}}{\sim} \text{Gamma}(1, 1/2). \tag{4.6}$$

In practice, we have so much spatial data in imaging studies that the Gamma hyperparameters in (4.6) above can be set to any "small" value without undue influence on the posterior.

As noted above, the correlation function $C(\cdot)$ can in general be any positive definite kernel function defined so that $C(0) = 1$ and $C(\alpha) \leq 1$ for all $\alpha > 0$. Given the substantial history of Gaussian smoothing in applied MRI analysis, we will work chiefly with the two parameter exponential radial basis function,

$$C(\alpha) = \exp(-\psi|\alpha|^\nu), \qquad \boldsymbol{\theta} = (\psi, \nu)^\mathsf{T}, \quad \psi > 0, \nu \in (0, 2], \tag{4.7}$$

which is stationary, isotropic, and synonymous with the Gaussian kernel when $\nu = 2$. In (4.7), ψ is sometimes called the bandwidth or inverse length-scale parameter and controls how rapidly the correlations decay, and ν is the kernel exponent or smoothness parameter. Alternative correlation functions could be used instead. For example, Mejia and colleagues use the Matérn correlation function [111] which is synonymous with the Gaussian kernel in a limiting case. Correlation functions with polynomial tails [e.g. 107] may be even more relevant for situations in imaging where the practitioner could expect parameters to have potential dependence at long-range. Myriad options exist: the choice of correlation function can in some cases be more art than science, and results may vary slightly depending on the selected kernel. We will discuss one data-driven way the correlation function might be selected in practice in section 4.2.5. The same method can also be used to estimate the correlation parameters $\boldsymbol{\theta}$ for a given functional family.

4.2.1 Conditional model

We outline two ways of working with model (4.1) in our setting, and also study the relative behavior of an approximate working model with connections to the standard vertex-wise analysis framework. The regression model that we have outlined is difficult to work with without simplification for several reasons. The first and perhaps most obvious reason is the dimension of the parameter space. Computational strategies for spatial modeling typically involve decomposition of a dense spatial covariance matrix. In our case, a somewhat naive decomposition of the joint covariance of the $\beta_j(\cdot)$ and the $\omega_i(\cdot)$ would be an $\mathcal{O}(M^3(N+P)^3)$ operation, where M is number of vertices in S, and N and P are the sample size and number of regression predictors, respectively. In Bayesian sampling algorithms, this decomposition often needs to be recomputed for each sample, which would be prohibitively expensive in our setting. For large enough data, even simply evaluating the joint covariance of all of the $\beta_j(\cdot)$ and $\omega_i(\cdot)$ is infeasible. The other difficulty working with the model as written is that decomposing the error structure into the sum of two spatially varying terms (i.e., the $\omega_i(\cdot)$ and the $\epsilon_i(\cdot)$) renders the whole model at best weakly identifiable.

As we lay out in greater detail below in Section 4.2.4, we overcome the first difficulty by using a conditional independence approximation to the model parameters' spatial covariance, inducing sparsity in the parameters' spatial precision. This type of approximation can greatly reduce the computational burden while retaining a covariance structure with full spatial rank, leading to high accuracy and scalability [e.g. 35, 44]. We overcome the second difficulty in several different ways, and we first introduce what we term the "conditional" approach to working with our model. We base this approach off of the observation that if we knew the correct $\omega_i(\cdot)$ the remaining terms in the model would be relatively easy to estimate. For this approach, our strategy will be first to obtain a pseudo maximum a posteriori estimate of the $\omega_i(\cdot)$, and second to condition on those estimates, sampling the other model parameters in an Empirically Bayesian way. To obtain these estimates, we work with an approximate model that considers $\sigma^2(\mathbf{s}) \equiv \sigma^2$ constant over all vertices in S, and alternate conditional maximization of $\boldsymbol{\beta}(\cdot)$ and the $\omega_i(\cdot)$ until convergence. Full details are available in Appendix M. Once we have obtained our estimate of the $\omega_i(\cdot)$ in this way we simply subtract the $\omega_i(\cdot)$ from the $y_i(\cdot)$, and switch to an efficient Bayesian sampling algorithm for the remaining parameters in the model (see section 4.2.4 for an overview). As we will see in simulation below, this approach tends to work quite well, but carries the potential downside that, by conditioning on a point estimate of the $\omega_i(\cdot)$, we may underestimate uncertainty in the $\beta_j(\cdot)$ for example.

4.2.2 Marginal model

Alternatively, since the individual deviations $\omega_i(\cdot)$ are not typically of direct interest, we can first integrate them out, leading to a marginal model with respect to the $\beta_j(\cdot)$, $\sigma^2(\cdot)$, etc. Marginalizing out the $\omega_i(\cdot)$ is relatively straightforward given the conjugacy in our model hierarchy, and leads to a rewritten likelihood with,

$$y_i(\mathbf{s}) = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}(\mathbf{s}) + \epsilon_i^*(\mathbf{s}), \qquad \epsilon_i^*(\mathbf{s}) \sim \mathcal{GP}(0, H\{d(\mathbf{s}, \mathbf{s}')\}), \tag{4.8}$$

where $H\{d(\mathbf{s}, \mathbf{s}')\} = \tau^2 C\{d(\mathbf{s}, \mathbf{s}')\} + \sigma^2(\mathbf{s})\, \mathbb{1}\{d(\mathbf{s}, \mathbf{s}') = 0\}$, and $\mathbb{1}(\mathcal{A})$ the event indicator function ($\mathbb{1}(\mathcal{A}) = 1$ if event $\mathcal{A}$ occurs, and 0 otherwise). A computational approach to (4.8) can then follow by additional application of a conditional independence approximation [35, 44] to the covariance of the $\epsilon_i^*(\cdot)$. In Appendix M, we outline a means of computing with model (4.8) based on estimating $\boldsymbol{\theta}$, τ^2, and $\sigma^2(\cdot)$ in an Empirically Bayesian way. Briefly, we take a two stage approach to computation, first obtaining approximate (up to optimization tolerance) maximum a posteriori estimates of $\boldsymbol{\beta}(\cdot)$, $\boldsymbol{\theta}$, τ^2, and $\sigma^2(\cdot)$. Second, we fix the covariance parameters $\boldsymbol{\theta}$, τ^2, and $\sigma^2(\cdot)$ at their approximate posterior modes and switch to an efficient MCMC routine to sample from the conditional posterior of $\boldsymbol{\beta}(\cdot)$. A sketch of our sampling algorithm is presented for a related model in section 4.2.4. As will be seen in simulation, the marginal approach works quite well for estimation of the $\beta_j(\cdot)$, but has the disadvantage that since the $\omega_i(\cdot)$ have been integrated away, they are not immediately available for outlier diagnosis, etc.

4.2.3 Working model

We also introduce a third, working model as a way to obtain approximate inference on the $\beta_j(\cdot)$. In general, including the $\omega_i(\cdot)$ as a separate correlated error component will not influence standard estimators of the center of the posterior of the $\beta_j(\cdot)$, such as the posterior mean. If out of sample prediction of imaging outcomes is not a goal of the analysis, then the primary reason to include a spatially correlated error component is to reduce the posterior variance of the $\beta_j(\cdot)$. In a large data setting, the gain in efficiency from including a correlated error component can be minimal to negligible. A natural question, then is how well the resulting model performs when we replace the likelihood with the approximation,

$$y_i(\mathbf{s}) = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}^w(\mathbf{s}) + \epsilon_i^w(\mathbf{s}), \qquad \epsilon_i^w(\mathbf{s}) \sim \mathcal{N}(0, \sigma^2(\mathbf{s})), \tag{4.9}$$

where the prior structure on the $\beta_j^w(\cdot)$ and $\sigma^2(\cdot)$ is the same as in (4.3) and (4.5) above. We term this approximation our "working" model here. Our working model can be viewed as a generalization of the standard vertex-wise GLM analysis paradigm in a spatial Bayesian context. The model

implied by fitting vertex-wise marginal GLMs is a limiting case of our working model as $\tau^2 \to \infty$ for select choices of the correlation function, $C(\alpha) = \mathbb{1}(\alpha = 0)$, and (improper) prior on $\sigma^{-2}(\cdot) \sim \text{Gamma}(1, 0)$. Considering comparisons among our suite of methods, we will show in simulation that for moderate to large sample sizes, the posterior of the $\beta_j^w(\cdot)$ for our working model is quite similar to that of the $\beta_j(\cdot)$ from either our conditional or marginal models.

4.2.4 Posterior computation

Since computation with the working model (4.9) is relatively simpler than for either the conditional (4.2) or marginal (4.8) models, we will outline our general approach to computation in the working model context. Posterior computation with the conditional and marginal models can be accomplished in very similar fashion, and we reserve explicit discussion thereof for Appendix M.

Since we typically work on a fixed spatial domain S, let $\boldsymbol{\beta}_j$ (dropping the superscript w for simplicity) denote the random field $[\beta_j^w(\mathbf{s})]_{\mathbf{s}\in S}$ for $j = 0, \ldots, P-1$, and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^\mathsf{T}, \ldots, \boldsymbol{\beta}_{P-1}^\mathsf{T})^\mathsf{T}$. Let $\boldsymbol{C} = [C\{d(\mathbf{s}, \mathbf{s}')\}]_{\mathbf{s},\mathbf{s}'\in S}$ represent the $(M \times M)$ spatial correlation matrix such that the prior on each $\boldsymbol{\beta}_j$ is equivalently $\mathcal{N}(\boldsymbol{\beta}_j \mid \mathbf{0}, \zeta_j^2\tau^2\boldsymbol{C})$. Similarly, let $\boldsymbol{\Sigma}$ represent the variance of $\epsilon_i^w(\cdot)$, here an $(M \times M)$ diagonal matrix with the $\sigma^2(\mathbf{s})$, $\mathbf{s} \in S$ on the diagonal; let $\boldsymbol{X}$ denote the $(N \times P)$ matrix of participant-level covariates; let $\boldsymbol{y}_i = [y_i(\mathbf{s})]_{\mathbf{s}\in S}$ denote the vectorized outcome image for participant i; and let $\boldsymbol{y} = (\boldsymbol{y}_1^\mathsf{T}, \ldots, \boldsymbol{y}_N^\mathsf{T})^\mathsf{T}$ represent the $(NM \times 1)$ vector of concatenated subject outcomes.

With the data in this "long" format, the model can be conveniently expressed in terms of Kronecker products. With $\boldsymbol{Z} = \text{diag}(\zeta_0^2, \ldots, \zeta_{P-1}^2)$, the conditional posterior variance of $\boldsymbol{\beta}$ can be written,

$$\text{var}(\boldsymbol{\beta} \mid \boldsymbol{y}, \cdot) = \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} \otimes \boldsymbol{\Sigma}^{-1} + \boldsymbol{Z}^{-1} \otimes \tau^{-2}\boldsymbol{C}^{-1}\right)^{-1}, \tag{4.10}$$

using shorthand to express conditioning on $\boldsymbol{\Sigma}$, $\boldsymbol{Z}$, $\boldsymbol{\theta}$, and τ^2. Since the dimension of $\boldsymbol{\beta}$ grows rapidly with P, it can be difficult or even impossible to work with (4.10) directly. Instead, we outline two strategies to enable efficient posterior computation at this scale. The first strategy, as alluded to above, is to replace $\boldsymbol{C}^{-1}$ with a sparse approximation $\tilde{\boldsymbol{C}}^{-1}$ such that $\tilde{\boldsymbol{C}} \approx \boldsymbol{C}$. In doing so, we follow work on the so called "Nearest Neighbor Gaussian Process" [35, 44], replacing the idea of k-nearest neighbors with small neighborhoods of fixed physical radius r. Briefly, we replace $\boldsymbol{C}^{-1}$ with a conditional independence approximation, enforcing that $\tilde{C}_{ij}^{-1} = 0$ if $d(\mathbf{s}_i, \mathbf{s}_j) > r$ for $\mathbf{s}_i, \mathbf{s}_j \in S$. Similar ideas have been alternately called Vecchia approximation [169, 90], composite likelihood [168], or Markov random field approximation [140], but in general can lead to highly accurate and scalable approximations of full rank spatial models [see e.g. 163, 35, 71]. Working with such an approximation of course introduces a hyperparameter, r, for the neighborhood radius size. In practice we found that in a large data setting choice of r had very little effect on our

analysis (see Appendix L.2 for a sensitivity analysis). In a small N setting, however, when the prior has more influence on the posterior, r must generally be chosen large enough to obtain a good approximation of the log prior. Anecdotally, we found that taking $r \geq 6$ mm worked well in simulation.

Although replacing $\boldsymbol{C}^{-1}$ with $\tilde{\boldsymbol{C}}^{-1}$ in (4.10) above lends sparsity and efficiency to computation in our setting, it can still be burdensome to evaluate or decompose (4.10) even for moderate P. To overcome this issue we propose an approximate quasi-Newton Hamiltonian Monte Carlo (HMC) algorithm for sampling from the posterior of $\boldsymbol{\beta}$, conditional on the other model parameters. HMC is a hybrid, gradient-based MCMC method that is often more efficient in high dimensions relative to other MCMC algorithms [116]. HMC can be used, here, to help avoid direct computation with the very high dimensional matrix in (4.10). In the general HMC algorithm, sampling can be improved by scaling the gradients by a "mass matrix," $\boldsymbol{M}$. In their highly influential paper, Girolami and Calderhead showed that the most efficient version of this algorithm updates $\boldsymbol{M}$ to be proportional to the posterior Fisher information matrix of the updated parameter [62]. A later extension of HMC estimates the information matrix using first-order gradient information akin to quasi-Newton optimization algorithms [52]. Obviously, since we are trying to avoid direct computation with (4.10), neither working with the information matrix $[\mathrm{var}(\boldsymbol{\beta} \mid \boldsymbol{y}, \cdot)]^{-1}$, nor some first-order estimate of this inverse is a practical solution. Instead, we can choose to use the prior information matrix to "estimate" the posterior information in the spirit of these algorithms. Doing so results in an efficient alternative. Taking $\boldsymbol{M} \propto (\boldsymbol{Z}^{-1} \otimes \tau^{-2}\boldsymbol{C}^{-1})$ and plugging in a sparse approximation of $\boldsymbol{C}^{-1}$ as above can result in dramatic improvement in Markov chain mixing with minimal increase in computation time. In practice, we found that we need not use the same $\tilde{\boldsymbol{C}}^{-1}$ in $\boldsymbol{M}$ as in our approximation of the log prior. In fact, we found it better to use smaller neighborhood radii in our construction of $\boldsymbol{M}$, and that keeping the neighborhood radius within the 2–4 mm range here resulted in the best Markov chain mixing.

The algorithm we have outlined above can be reasonably used for efficient posterior computation even in very large data sets. In fact, given a set of sufficient statistics that can be computed with one pass through the images, all of the parameter updates in our working model can be performed without reference to the original data. This leads to computational time complexity that, save for an initial data streaming step, is entirely independent of N. This has obvious advantages in large data regimes. In applied fMRI analysis, for example, a common use case when working with task contrast images is to use a simple set of predictors: practitioners often require an intercept-only model, or perhaps additionally desire to control for covariates age and sex, etc. We benchmarked our working model software for these use cases, analyzing task contrast data from the right hemisphere ($\approx$ 30,000 vertices) for close to 4,000 participants (a more detailed analysis is presented in section 4.4). We found that streaming the images typically took around 100 ms or less per image (for images

stored in the CIFTI/NIFTI-2 file format: https://www.nitrc.org/projects/cifti/). Once the images were streamed and sufficient statistics computed, analysis with HMC took around 3.3 min per 1,000 iterations for the intercept-only model, or around 18.8 min per 1,000 iterations for the three predictor model (intercept, age, and sex). Each analysis required less than 300 Mb of free RAM to run, demonstrating the scalability of our approach. We ran this comparison on a Dell PowerEdge R440 server with Intel® Xeon® Gold 6230 processors (2.1 GHz), limiting our processes to use eight cores each.

4.2.5 Estimation of $\boldsymbol{\theta}$ and $C(\cdot)$

There are a number of ways to estimate $\boldsymbol{\theta}$ in practice for a given correlation function $C(\cdot)$. Experience can guide practitioners to some extent: in applied imaging it is common to apply a Gaussian smoothing kernel to data prior to analyses. In part, the goal of this practice is to approximate a full spatial model for effects of interest [e.g. 165]. Commonly applied smoothing kernels are specified by their full-widths-at-half-maxima (FWHMs), which are often chosen to be within a 4–12 mm range [e.g. 112]. A 6 mm FWHM Gaussian kernel, for example, is nominally equivalent to a radial basis correlation function (4.7) with bandwidth parameter $\psi = 0.077$ and exponent parameter $\nu = 2$. These parameters can make for perfectly reasonable choices in practice: in our setting, the posterior will typically not be overly sensitive to the choice of $\boldsymbol{\theta}$, especially for problems with moderate to large N.

For general spatial modeling with Gaussian processes, other commonly used methods to estimate the spatial correlation include variogram or covariogram estimation [e.g. 5, 34], and maximum marginal likelihood methods [e.g. 110]. These methods can also be used to select the correlation function itself by taking, for example, the correlation function resulting in the best fit to the variogram or the highest marginal likelihood. Here, we have used a maximum marginal likelihood-based approach for a surrogate model to estimate the correlation function and corresponding parameters in the spirit of Empirical Bayes. Appendix M.2 provides a full description of this selection method for interested readers. In our analysis of the ABCD study data (section 4.4), we estimated $\boldsymbol{\theta} = (0.17, 1.38)^{\mathsf{T}}$, which corresponds to a sub-Gaussian correlation function with 5.57 mm full-width-at-half-maximum.

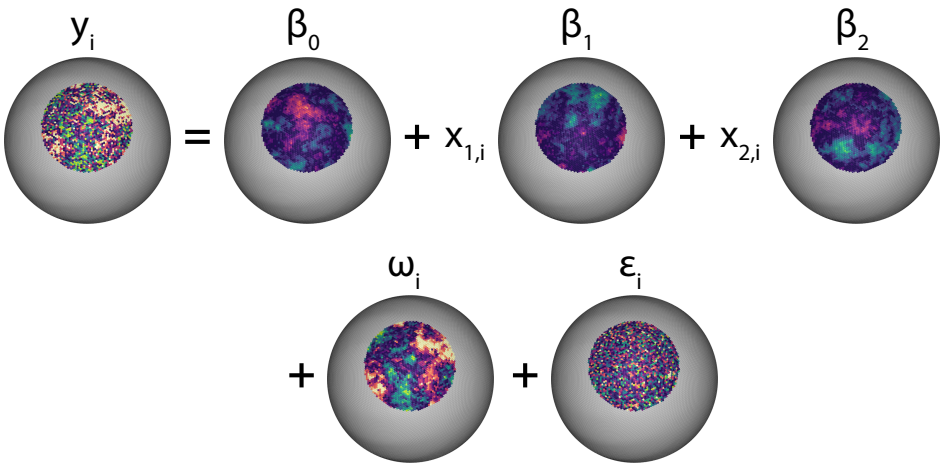


Figure 4.2: Simulation design. Data were simulated over a disc of 2,000 vertices on a spherical surface. Effects of interest $\boldsymbol{\beta}_j$, $j = 0, 1, 2$ were simulated as hard-thresholded Gaussian fields each with approximate 30% sparsity. Error terms $\boldsymbol{\omega}_i$ and $\boldsymbol{\epsilon}_i$ were drawn from larger variance spatial processes and dominate the spatial signals of interest such that the spatial signal-to-noise ratio was controlled to be approximately 0.04. We have enhanced the contrast of the $\boldsymbol{\beta}_j$ images for visual clarity.

4.3 Simulation study

4.3.1 Simulation design

Our goal in simulation was to compare the performance of our two methods for estimating our model against our "working model" and the standard vertex-wise marginal linear model approach. In all cases, data were simulated from model (4.2) on a disc of 2,000 vertices on the cortical surface. We designed our simulation to mimic the spatial smoothness and signal-to-noise ratio we estimated from real data. Fig. 4.2 illustrates our approach to simulation. For each simulation iteration, we generated spatially correlated and sparse $\boldsymbol{\beta}_j = [\beta_j(\mathbf{s})]_{\mathbf{s}\in S}$, $j = 0, 1, 2$, by hard thresholding draws from independent Gaussian processes with 6 mm FWHM exponential correlation functions. We set the marginal variance parameter of each Gaussian field drawn this way to 0.04 and thresholded the result at 0.08 so that each $\boldsymbol{\beta}_j$ would be approximately 30% sparse on average. This level of sparsity roughly matches the pseudo-sparsity we estimated from the real data: applying standard vertex-wise GLM methods with a Bonferroni correction-based p-value threshold resulted in significant findings over about 70% of the cortical surface. Since our prior model in (4.3) is non-sparse, simulating the $\boldsymbol{\beta}_j$ in this way actually reflects a setting with slight model misspecification. It is advantageous to consider such a setting, however, since our criteria for model evaluation includes measures of inferential accuracy. While we treated the field $\boldsymbol{\beta}_0$ as a spatially varying intercept parameter, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ were each paired with covariates. For subject $i = 1, \ldots, N$, covariates

$(x_{1,i}, x_{2,i})^{\mathsf{T}}$ were drawn jointly from a multivariate Gaussian distribution with mean zero, unit marginal variance, and correlation parameter 0.5.

In each simulation, we also generated the subject-level deviations $\boldsymbol{\omega}_i = [\omega_i(\mathbf{s})]_{\mathbf{s}\in S}$ as draws from independent Gaussian processes with 6 mm FWHM exponential correlation functions. Again to mirror estimates from the real data, we set the marginal variance of each $\boldsymbol{\omega}_i$ to 1.75. In a similar vein, we drew each $\boldsymbol{\epsilon}_i = [\epsilon_i(\mathbf{s})]_{\mathbf{s}\in S}$ following a white-noise process with spatially constant variance 1.25. Under the above parameter settings, we controlled the spatial signal-to-noise ratio to be approximately 0.04 (or equivalently, the spatial R^2 was controlled to be approximately 3.8%). As can be seen from Fig. 4.2, the error terms $\boldsymbol{\omega}_i$ and $\boldsymbol{\epsilon}_i$ largely dominate the spatial signal. Within this regime, we studied the behavior of our various comparison methods for increasing sample size, replicating the simulation 50 times per sample size.

In all cases, we then fit our suite of methods to compared against the standard vertex-wise GLM conditioning on the true correlation parameters (or smoothing the outcome images with exactly a 6 mm FWHM exponential kernel in the case of the standard method). For our suite of methods, we used Vecchia approximations with 8 mm neighborhood radii to approximate the Gaussian process priors on the $\beta_j(\cdot)$ as discussed in section 4.2.4. Here, we give the standard vertex-wise analysis paradigm a Bayesian treatment by replacing our priors on the $\beta_j(\cdot)$ and $\sigma^2(\cdot)$ with independent Jeffreys priors as alluded to in section 4.2.3. Since the full conditional posterior distributions of the resulting model parameters are quite easy to sample from we fit the vertex-wise model using Gibbs sampling. Working with the model in this fashion allowed us to compare the standard vertex-wise analysis to our proposed models in terms of full posterior inference. Namely, we used posterior credible bands as a way to summarize the joint uncertainty in the $\beta_j(\cdot)$ over all vertices simultaneously. In a spatial modeling context, posterior credible bands are very natural and fully Bayesian approach to inference, and can be easily estimated from MCMC samples [see e.g. 142]. Since credible bands reflect posterior probability statements about the joint behavior of the $\beta_j(\cdot)$ for all spatial locations, inference derived from them is fully multiple comparisons consistent.

4.3.2 Results of simulation comparisons

Table 4.1 summarizes the results of our simulation for increasing sample size. For each method in the table, we report scaled absolute bias and variance as well as sensitivity and specificity rates (*True* $+$ and *True* $-$, respectively; expressed as percentages. Since the scale of each $\beta_j(\cdot)$ is the same for all j in simulation we report absolute bias and variance as averages over the entire parameter vector $\boldsymbol{\beta}$, so that the values in Table 4.1 can be interpreted as the scaled expected point-wise bias or marginal variance of each $\beta_j(\mathbf{s})$. For example, the absolute bias column reports $10^3/(3 \times 2,000) \sum_{j,\mathbf{s}} |\hat{\beta}_j(\mathbf{s}) - \beta_j(\mathbf{s})|$, where $\hat{\beta}_j(\cdot)$ is the posterior mean estimate for a given

Method	*N*	$\lvert Bias \rvert$	*Variance*	*True* $+$	*True* $-$
Conditional	20	249.8 (3.9)	19.4 (0.4)	14.7 (0.7)	91.5 (0.4)
Marginal	20	178.1 (2.4)	17.5 (0.2)	4.0 (0.5)	98.5 (0.3)
Working Model	20	264.9 (4.4)	59.3 (1.8)	1.6 (0.1)	99.5 (0.1)
Vertex-wise GLM	20	228.1 (4.1)	77.7 (2.6)	0.0 (–)	100.0 (–)
Conditional	100	121.2 (1.5)	5.9 (<0.1)	26.2 (0.6)	95.1 (0.3)
Marginal	100	111.0 (1.3)	10.0 (<0.1)	7.2 (0.5)	99.6 (0.1)
Working Model	100	122.9 (1.5)	14.6 (0.2)	6.8 (0.3)	99.8 (<0.1)
Vertex-wise GLM	100	123.0 (1.3)	13.7 (0.2)	2.8 (0.4)	99.8 (<0.1)
Conditional	500	61.1 (0.4)	2.1 (<0.1)	53.6 (0.6)	98.0 (0.1)
Marginal	500	61.3 (0.4)	4.2 (<0.1)	27.5 (0.7)	99.9 (<0.1)
Working Model	500	61.2 (0.4)	4.4 (<0.1)	29.7 (0.5)	99.9 (<0.1)
Vertex-wise GLM	500	89.8 (0.3)	2.7 (<0.1)	29.3 (1.0)	96.2 (0.2)

Table 4.1: Simulation results focusing on parameter estimation (absolute bias and variance) and inferential accuracy (true positive and true negative rates). Results are reported as mean (standard error). Absolute bias and variance have been scaled by a factor of 10^3 to facilitate comparison; true positive and negative rates (sensitivity and specificity, respectively) are expressed as percentages.

method (three predictors; 2,000 spatial locations; scaled by a factor of 10^3 to enhance the clarity of the table). We constructed example inferential decisions based on 80% simultaneous credible bands. The 80% credibility threshold was chosen to represent a selection that might reasonably be applied in practice rather than by optimizing any kind of inferential criterion. In Table 4.1, the *True* $+$ column corresponds to the average percentage of cases where the true $\beta_j(\mathbf{s}) \neq 0$ and the corresponding credible band does not include zero. Similarly, the *True* $-$ column reports the average percentage of cases $\beta_j(\mathbf{s}) = 0$ and the corresponding credible band covers zero.

The most immediate result of our simulations is that the marginal method of estimating our model typically leads to the most accurate posterior mean estimate of $\boldsymbol{\beta}$ in the mean squared error sense. For all methods under consideration, the point-wise bias dominates the point-wise variance across all simulation settings. At large sample size, we note that both the marginal and conditional methods of estimating our model as well as our working model tend to produce very similar estimates and results. We explore this pattern further in Fig. 4.3, which summarizes the similarity of the full posterior distribution of $\boldsymbol{\beta}$, as estimated with our suite of methods. It is clear from the figure that differences in estimation between methods decay as the sample size increases. Interestingly, at small sample sizes, our conditional and working model methods have higher absolute bias than the vertex-wise GLM, but quickly overtake the standard method as the sample size increases. Bias does not decrease with increasing N as rapidly for the vertex-wise GLM as for our suite of methods. The other major result indicated by our simulations is that, using simultaneous credible bands for inference, the conditional method of estimating our model is the

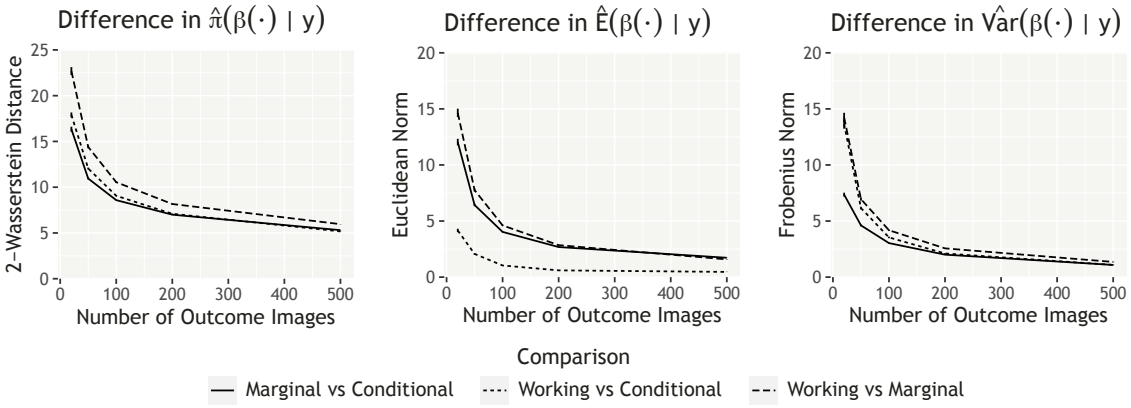


Figure 4.3: Comparison of the posterior distributions of $\boldsymbol{\beta}$ across our suite of methods. In the left panel, 2-Wasserstein distance was computed using a Gaussian approximation to the posterior, derived from MCMC samples. The center and right panels of the figure show the similarity in the posterior mean and variance of $\boldsymbol{\beta}$, summarized as the Euclidean and Frobenius norms of the differences, respectively. Error bars are minimally visible, but show $\pm$ 1 simulation standard error.

most powerful or sensitive among our methods under comparison. The sensitivity of this method, however, is modestly lacking compared to our marginal and working model methods, which make virtually no false positive errors even at smaller sample sizes.

This pattern of results can be somewhat difficult to summarize. Based on this simulation, we cannot uniformly recommend any one method without knowledge of the practical goals of an intended analysis. If estimation of $\boldsymbol{\beta}$ is of primary concern, then we would generally recommend our marginal method if the sample size is small, or any of our suite of methods at larger sample sizes. If inference is of primary concern, we might recommend our conditional method, for example, for its high sensitivity. Alternatively, we might also recommend either our marginal or working methods for their rather low false positive to true positive ratio.

4.4 Illustrative analysis of fMRI task contrast data

4.4.1 Description of the data and model setup

To illustrate use of our methodology, we applied our working model to analyze n-back task contrast data from the ABCD study, release 2.0.1 [85]. A brief comparison of estimation differences between our working, conditional, and marginal model variants is available for these data in Appendix L.1. The ABCD study is the product of a large collaborative effort to study longitudinal changes in the developing brain through childhood and adolescence, and to track biological and environmental correlates of development [41]. Data collection and processing has been harmo-

nized across 21 research sites in the continental United States. At the time of writing, the study has collected baseline environmental, behavioral, genetic, and neuroimage data from over 11,800 children between the ages of 9–10 years. Longitudinal data has been collected at six month intervals for a subset of children in the study; already over 3,600 children have been enrolled for over 2.5 years. Details regarding study design and recruitment [54], neurocognitive assessment [105], neuroimage acquisition [22], and image preprocessing [67] are available in published literature.

For this illustration we will work exclusively with data from a subset of 3,267 children in the baseline cohort that were scanned while performing an n-back task [8, 22] with pictures of human faces expressing emotion as stimuli. The n-back task has enjoyed wide use in the neuropsychological and imaging community for its relationship with executive function and as a correlate of working memory processes [e.g. 81, 121, 80]. Our subsample of children is limited to those who scored at or above 60% correct on both 0-back and 2-back task conditions.

We focus our analysis on the relationships between task-related activation and individual-level task accuracy, which has been studied previously [see e.g. 97, for a recent article]. In concert, our analysis controls for various child-level characteristics and family-level demographic information. We took 2- vs 0-back task contrast data (z-statistic scale) as our primary outcome and modeled it as a function of 2-back task accuracy; child fluid intelligence; child age (months); child gender (binary); parental education (five levels); parental marital status (binary); and family income (three levels). We included first-order interactions between child gender and parental education; child age and parental education; child age and child gender; child age and 2-back accuracy; and child gender and 2-back accuracy. Table 4.2 gives a summary of the demographic information for this sample. For interpretive purposes, we centered continuous covariates in the analysis on their respective in-sample means, and we treated the in-sample modal demographic categories as baseline (female child from a married household, at least one parent with a post graduate degree, and household income greater than $100,000 USD/year). Given this coding scheme, the intercept parameters in our spatial regression can be interpreted as the expected task contrast image for a typical in-sample female child of average fluid intelligence that scored 80% correct on the 2-back task condition. For visualization purposes, we scaled each continuous covariate by two standard deviations [58] so that resultant coefficient images are more directly comparable with coefficient images for categorical covariates. Although we will not have room here to give a full account of all of the effects we estimate for our demographic and socio-economic predictors, interested readers can find a more comprehensive report in Appendix L.1.

We chose covariates largely on the basis of known associations with general n-back task accuracy [123]. In addition, we performed sets of exploratory analyses in the classic vertex-wise framework without any spatial smoothing (not shown). These analyses served to help us visualize and understand several important aspects of the data. First, we observed modest but present nonlin-

ear patterns in the relationship between the contrast data and 2-back accuracy. Preferring simplicity here, we found that these trends were reasonably well characterized by a quadratic model for 2-back accuracy. Including this term in the analysis resulted in a total of $P = 24$ predictors including the global intercept.

Additionally, since the ABCD data are naturally grouped by the study's 21 data collection sites, we explored the utility of including random site effects. For these data, the random site effects explained less than 1% of the total variance in over 97% of vertices, and less than 0.1% of the variance in nearly half of vertices. We ultimately concluded that site-specific random effects do not critically influence results here. Again preferring simplicity, the results we show in the main text do not include site effects as a variance component. Interested readers can find further analysis to support this claim in Appendix L.1. As a final note before presenting results, the ABCD study more broadly contains imaging data acquired from siblings. Around 20% of families in the ABCD release 2.0 baseline data have two or more children enrolled in the study. This might additionally suggest the need for a random family effects analysis. We avoid this issue entirely here: the cohort that we analyze contains data from only one child per family in our subset. While our method is capable of estimating effects like this in general, it would be very slow computationally to give a fully Bayesian treatment to a large number of random spatial effects. A more specific tool could be built on top of the methods we present here to include such variance components and/or treat them as nuisance parameters.

We fit our model with Hamiltonian Monte Carlo (HMC) as noted in section 4.2.4. For this analysis, we ran eight chains of 7,000 iterations each, discarding the first 5,000 as adaptation and burnin, and saving 200 samples from the final 2,000 iterations of each chain. Convergence was assessed via univariate folded and non-folded rank-normalized split $\hat{R}$ [170] for each parameter $\beta_j(\cdot)$, and by visual examination of trace plots for subsets of these parameters. The folded split $\hat{R}$ statistic was below the recommended threshold of 1.01 for over 99.9% of the $\beta_j(\cdot)$ (the worst case scenario was 1.02), indicating reasonable convergence in the posterior spread and tail behavior for these parameters. Similarly, the worst-case non-folded split $\hat{R}$ statistic was 1.04 across all $\beta_j(\cdot)$, indicating reasonable convergence of the center of the posterior distribution for these parameters. We set the neighborhood radius of the Vecchia approximation of our prior precision to 8 mm, and the neighborhood radius of our HMC mass matrix to 3 mm. While the algorithm can be quite sensitive to the choice of mass matrix neighborhood radius, values in the range 2–4 mm led to efficient and well-mixing chains both here and in simulation. For readers familiar with Hamiltonian Monte Carlo: Metropolis-Hastings rates were tuned during burnin to be approximately 65%; automatic tuning was achieved using the dual-averaging method presented in [73]. Additionally, we fixed the number of numerical integration steps in our HMC to 35, which we noted produced well-mixing chains.

4.4.2 Summary of primary results

The primary results of our analysis are presented in Figs. 4.4 and 4.6. We consolidate the output by focusing on results in the right hemisphere, and we note that in general results in the left hemisphere are highly symmetric. Figs. 4.4 and 4.6 follow the same general format for different terms in our model. In particular, Fig. 4.4 shows the posterior mean estimate of our model intercept parameters ($\boldsymbol{\beta}_0$) and also gives a region of interest level summary of this term. Regions of interest were taken from the Gordon 2016 cortical surface parcellation, which created was in part from resting state functional connectivity maps and naturally groups brain regions within a network community structure [63]. The atlas delimits 172 brain regions in the right hemisphere (161 in the left), each grouped within one of 13 functional network communities.

To summarize our model intercept by brain region, we fit a series of mixed effect models to MCMC samples of $\boldsymbol{\beta}_0$, taking advantage of the Gordon atlas's grouped structure. We used this modeling strategy to obtain region-level averages of our spatial intercept parameters, where each region-level average is shrunk towards its network community mean. By repeatedly fitting this model to each sample of $\boldsymbol{\beta}_0$, we obtain fully Bayesian point and interval estimates for the region-level averages. The bottom panel of Fig. 4.4 displays point and multiple comparisons consistent 95% interval estimates for a subset of regions in the Gordon 2016 atlas. Although we model and adjust the intervals based on all 172 regions in the atlas, we only show the region-level estimates for regions belonging to communities: Cingulo-Opercular network, Default mode network, Dorsal attention network, Fronto-Parietal network, and "None" (see Fig. 4.4). Results of this analysis show that the largest activations occur in regions associated with the Dorsal attention, Fronto-Parietal, and Cingulo-Opercular networks, with a handful of regions associated with the Default mode network also showing significant activations. Similar conclusions were reached by [22] in a smaller, preliminary subset of these data ($N = 517$). Data from another large collective imaging study also support similar results in adults aged 22–37 years ($N = 949$) [97].

Fig. 4.5 depicts example spatial inference on the intercept parameters in the right hemisphere, thresholding at what we might consider a small to medium effect size. In the figure, colored regions denote areas where the posterior mean estimate of $|\beta_0(\mathbf{s})|$ is greater than 0.4. Since we are modeling z-statistic outcomes, $\beta_0(\mathbf{s}) > 0.4$ can be interpreted to mean that roughly 2 out of every 3 "average" children in our sample would show task-related activation at location $\mathbf{s}$ (versus 1 in 3 showing deactivation; the statement can be reversed for $\beta_0(\mathbf{s}) < 0.4$). Regions of darker color in Fig. 4.5 mark areas where our analysis suggests the probability that $|\beta_0(\mathbf{s})| > 0.4$ is greater than or equal to 80% simultaneously for all vertices $\mathbf{s}$ within those areas. This interpretation is similar to the notion of "upper confidence sets" from [18]. Here, as in in section 4.3, we use posterior credible bands [see e.g. 142] to create these inferential summaries. In principle, this type of summary can be generated for other standardized or real-world measures of effect size, such as

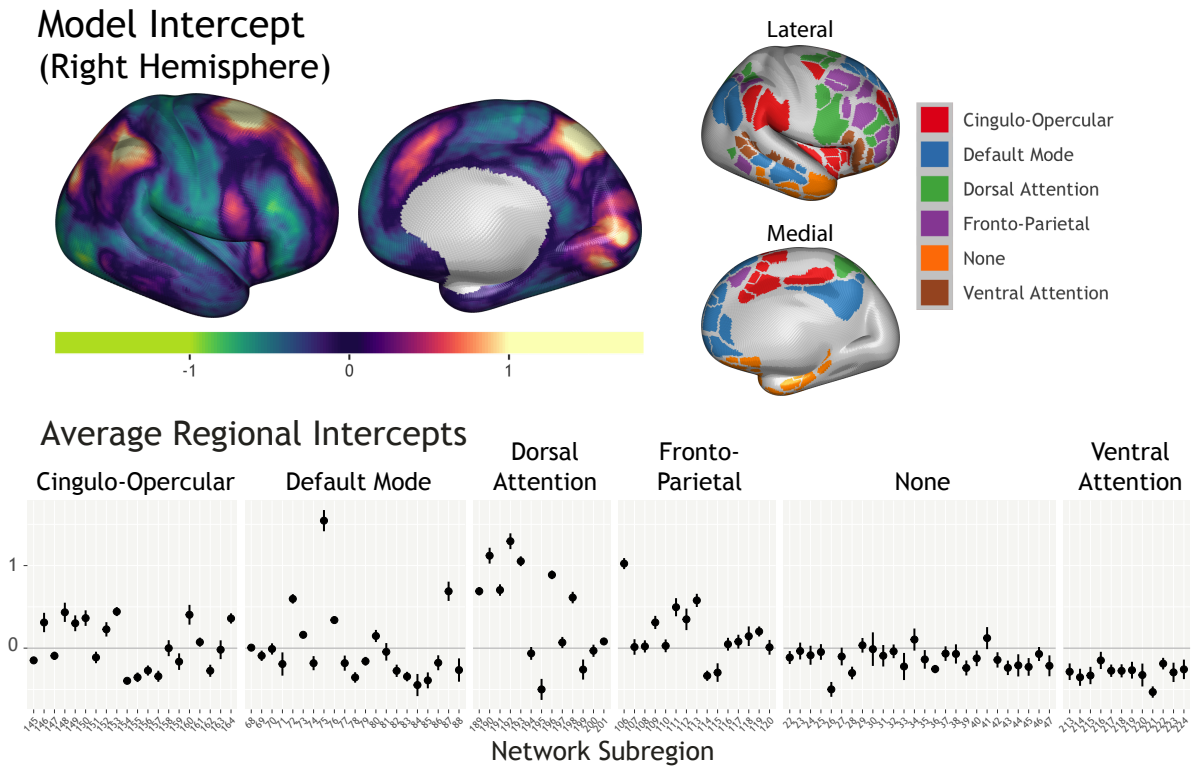


Figure 4.4: Model intercept coefficients summary. The upper left corner of the figure shows the posterior mean estimate of the intercept, which can be interpreted as a one-sample t-test for the 2- vs 0-back contrast, controlling for demographic information (see the main text for details). Forest plots in the bottom row of the figure summarize the intercept parameters in terms of region-level averages, with regions taken from the Gordon 2016 cortical surface parcellation [63]. Error bars in the forest plots correspond to fully Bayesian 95% intervals that have been widened to be multiple-comparisons consistent (Bonferroni adjustment). The upper right panel of the figure shows the brain regions represented on the x-axis in the bottom row forest plots. Region numbers correspond to the Freesurfer (`https://surfer.nmr.mgh.harvard.edu/fswiki`) labels for the Gordon parcellation. Left to right the region labels read, Cingulo–Opercular: 145–164; Default Mode: 68–88; Dorsal Attention: 189–201; Fronto-Parietal: 106–120; None: 22–47; Ventral Attention: 213–224.

Cohen's d or percent signal change, etc.

Similarly to Fig. 4.4, Fig. 4.6 summarizes results for the effect of 2-back accuracy rate on the 2- vs 0-back contrast. In the figure, coefficients for the linear and quadratic accuracy terms reflect the expected change in activation between ten year old female children scoring 96% and 80% correct on the 2-back condition, respectively, holding all other demographic covariates constant. Our analysis suggests high spatial overlap between the intercept and areas where average activation increased linearly with increasing 2-back accuracy (confer from Figs. 4.4 and 4.6). Interestingly, however, the quadratic accuracy term largely seems to reflect areas where average activation in-

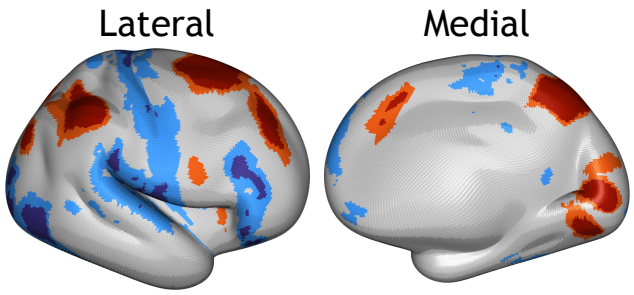


Figure 4.5: Model intercept for the right hemisphere: example signed discoveries using an 80% posterior simultaneous credible band to infer locations where $|\beta_0(\cdot)| > 0.4$. Red regions correspond to functional activations and blue regions correspond to deactivations. Darker colors indicate regions of simultaneous posterior confidence that $|\beta_0(\mathbf{s})|$ is greater than 0.4 for all vertices $\mathbf{s}$ in those regions. Lighter colors can be thought of as reflecting the spatial uncertainty in that claim of posterior credibility.

creased supra-linearly with increasing 2-back accuracy. Based on our analysis, these areas are more constrained to regions associated with the Dorsal-Attention and Fronto-Parietal networks (Fig. 4.6). Model residual standard deviations for both hemispheres are shown in Fig. 4.7. In general, areas with the highest residual variance overlap with areas activated in the 2- vs 0-back contrast (confer from Figs. 4.4 and 4.7). This result indicates substantial variability in individual responses in these regions. Overall, our fitted model explained about 6.2% of the total variance in the task contrast images.

4.4.3 Goodness-of-fit evaluation

Finally, we assess the fit of our model using posterior predictive simulation [139, 59] and analysis of model residuals. Selected results of these comparisons are presented in Fig. 4.8. In the figure, we summarize the extent of discrepancy between the observed data and posterior predictions the model would make for replicated data. To do this, we again leveraged the Gordon 2016 cortical surface parcellation [63] and computed test descriptive statistics across subjects within each brain region, comparing against the same statistics computed over synthetic data of the same size simulated from our model. We explored the discrepancies in the predictive and empirical data distributions based on measures of central tendency, spread, and several quantiles. Absolute differences in the empirical values and the posterior predictive mean value are shown in Fig. 4.8 for each brain region and for three such test statistics. To give a sense of scale, in the figure the largest regional difference is < 0.2 (10^{th} Quantile panel), whereas the range of the data is approximately -13.7 to 17.1. In general, we found that discrepancies between the empirical and predictive distributions were extremely low for test statistics that summarize features of the central bulk of the distributions. Features in the very tails of the empirical distribution were less well captured in

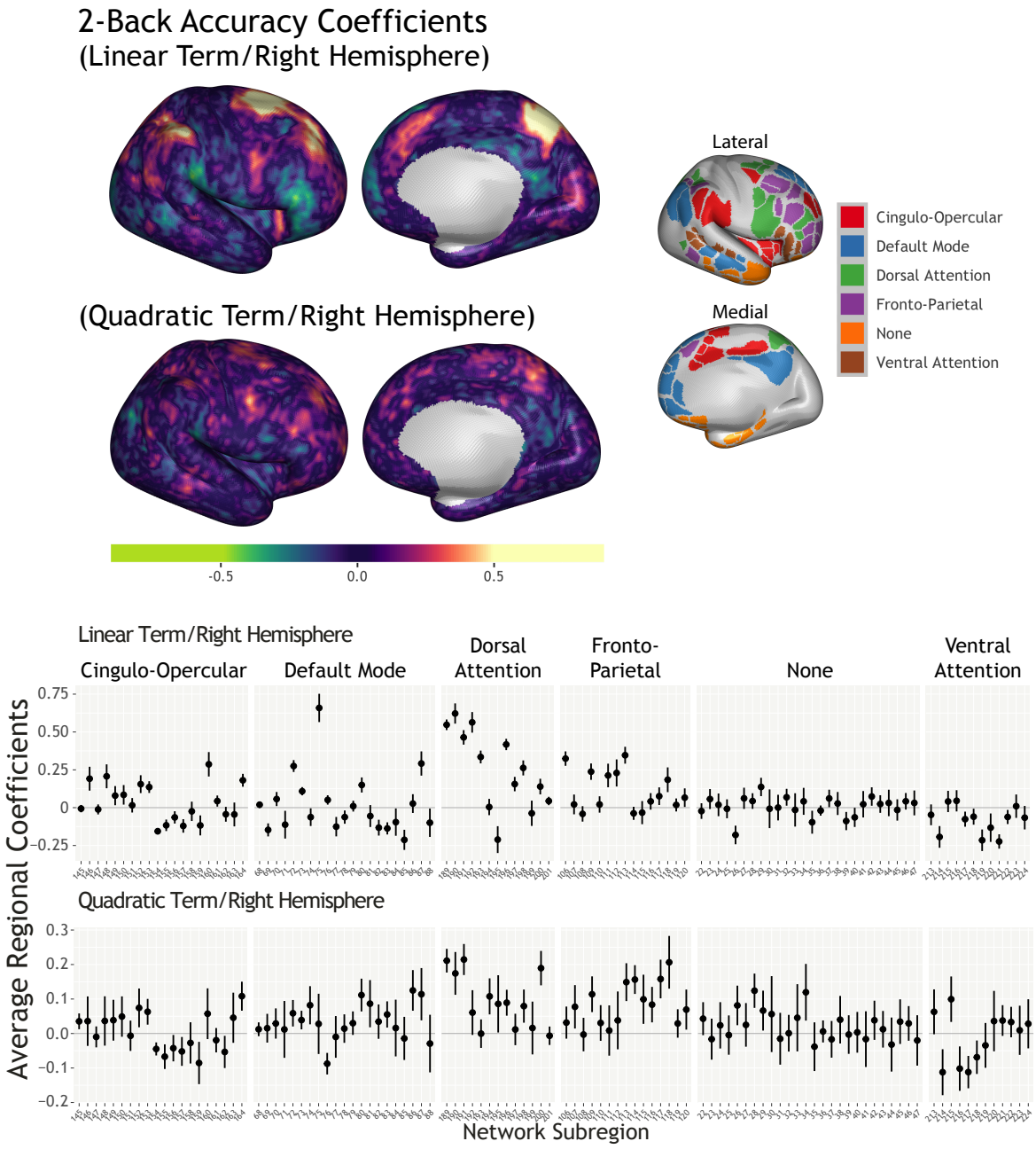


Figure 4.6: Coefficient summary for 2-back condition accuracy rate (linear and quadratic terms). The overall format of the figure is the same as in Fig. 4.4 above.

the predictive distribution, as might be expected for a normal model (not shown). In Fig. 4.8, we also summarize goodness-of-fit by comparing standardized residual histograms for each brain region. In the lower panel of Fig. 4.8, we ranked each brain region by their discrepancy with a normal model and show residual histograms for the best, median, and worst-case regions. In general, we again see evidence of excellent model fit throughout the central bulk of the data. Inter-

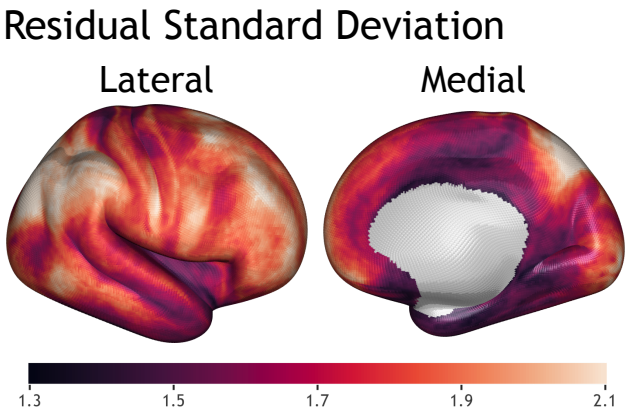


Figure 4.7: Residual standard deviation for the right hemisphere. Areas of high residual variation generally overlap with activation areas in the 2- vs 0-back contrast (confer with Fig. 4.4).

estingly, region 192 (worst-case fit) contained the highest overall mean parameter estimate within the Dorsal-Attention network community for both the intercept and linear 2-back accuracy term (Figs. 4.4 and 4.6; as in Fig. 4.4, "192" corresponds to the Freesurfer label for the Gordon atlas region). This result may indicate, for example, that while the (simple) quadratic model we have used for 2-back accuracy provides a reasonable fit to the task contrast data across most of the right hemisphere, it may fail to perfectly encapsulate the complex task-related activation patterns in this sample.

4.5 Discussion

This chapter proposes a Bayesian spatial model for group-level image-on-scalar regression analyses, and illustrates several ways to consider working with the model in practice. We also show how the spatial Gaussian process prior formulation and related approximation through conditional independence methods can enable flexible and reasonably efficient computation with MCMC. Critically, our approach allows us to work with full-rank spatial processes, and does not rely on lossy compression schemes like down-sampling or low-rank projection, which can be dissatisfying in practice [e.g. 157]. We have shown in simulation that our strategy can improve on the standard analysis stream in terms of finite sample bias, sensitivity, and specificity. We have also shown that an approximate working model produces similar inference on the spatially varying coefficients $\boldsymbol{\beta}(\cdot)$ in settings with moderate to large N. Our working model is relatively easy to compute with, and can be thought of as a generalization of the standard analysis stream. Finally, we illustrate use of our method on task (n-back) contrast data from the Adolescent Brain Cognitive Development study.

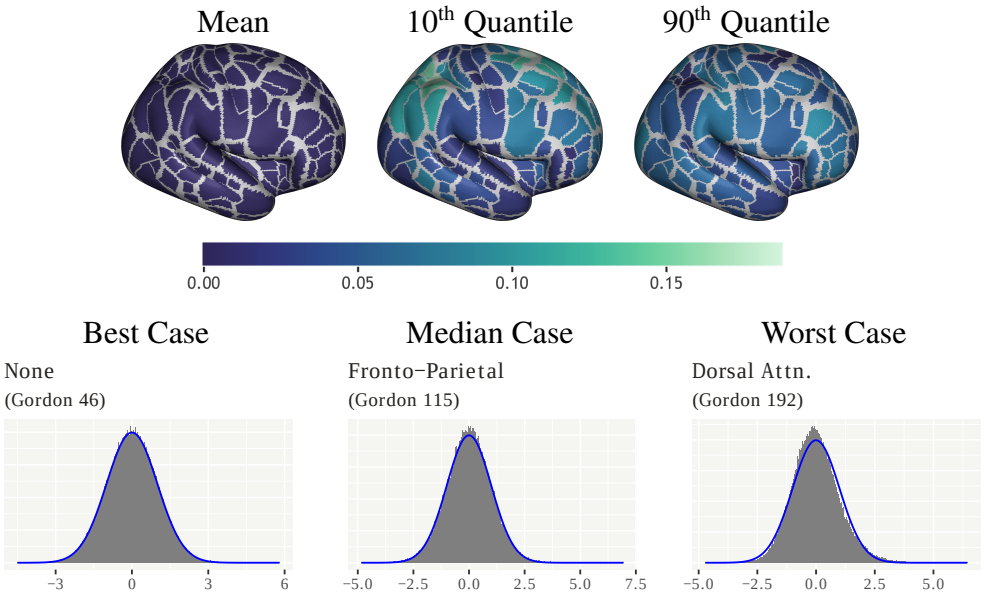


Figure 4.8: Goodness of fit checking. (*Top*) Absolute differences in the observed and mean posterior predictive value for three different test statistics computed across all subjects and vertices for each brain region in the Gordon 2016 parcellation [63]. Test statistics shown are the regional mean and 10^{th} and 90^{th} quantiles. In the figure, predictive checking is homogeneous within each brain region; the gray shows a boundary area not assigned to any particular region. (*Bottom*) Histograms of standardized residuals from three different brain regions. Blue lines show the fit of model at the posterior mean. The three different regions chosen show the best, the median, and the worst-case scenarios for the model's goodness-of-fit in these areas.

With the exception of a white-noise component our model error process, we express our model as a sum of terms assigned stationary spatial priors in section 4.2. In general, spatial stationarity is not considered a realistic assumption for imaging data [see e.g. 180, 1]. Although we use stationary priors throughout for simplicity, in practice, given the data the posterior distribution of our model parameters can still reflect non-stationary processes. In fact Figs. 4.4 and 4.6 illustrate clear posterior mean field non-stationarity. In particular, the posterior mean of our model intercept and linear 2-back accuracy rate coefficients suggest obvious mean-field non-stationarity. Moreover, since our model on the white-noise process is inherently non-stationary, our prior hierarchy can lead to more data-adaptive smoothing in the regression coefficients compared to standard analysis streams where stationary spatial smoothing is applied, at some level, to the data.

As we alluded to in section 4.4.1, it may be of interest to build extensions to our method to incorporate additional variance components for more complex or specific study designs. Such an extension might correspond, for example, to the addition of random family effects in analyses of the greater ABCD study sample. While our present model is technically capable of estimating such effects by pooling the corresponding ζ_j^2 across related terms, including a large number of random

spatial effects in the analysis can be extremely demanding computationally. One workaround might be to omit modeling spatial correlation structures for these terms, and treat them as pure nuisance parameters. At the time of writing, we have not yet studied the practical consequences of doing so.

Other possible extensions of our method include modeling of fMRI time series data at the individual or group level. Our present method does not require more than a minor modification of the likelihood to be appropriate for individual-level fMRI time series analysis. In this single-patient setting we might write a new data-level model,

$$y_t(\mathbf{s}) = \boldsymbol{x}_t^{\mathsf{T}}\boldsymbol{\beta}(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad t = 1, \ldots, T,$$

where T is the number of time points in the series. Here, we could retain the prior on $\boldsymbol{\beta}(\cdot)$ from (4.3) and assume the spatio-temporal errors $\epsilon_t(\cdot)$ can be modeled approximately as an auto-regressive process [e.g., 51, 182]. Additional flexibility could be incorporated into the model for the $\epsilon_t(\cdot)$ for example by allowing the temporal auto-regressive order also to vary with spatial location [similar to 164].

As discussed in section 4.1, a Bayesian spatio-temporal model for analyzing multi-participant time series data has already been proposed by Mejia *et al.* [111]. Although such an integrative model is very attractive conceptually, the number of parameters required to be estimated grows rapidly with increasing sample size N, creating the potential for a significant computational bottleneck. Mejia *et al.* [111] approach this very difficult problem by limiting their method to small sample studies, substantially down sampling the available spatial data, and constructing a numerical approximation of the posterior. Though this procedure appears to work quite well in small sample settings, cases where N is moderate to large may demand an alternative approach. With the present chapter, we have proposed one such alternative. Our method can be reasonably used in a large-sample group-level fMRI analysis setting by simply taking first-stage contrast images to be the response. Within this framework, our proposed method gains the advantage of being able to generate scalable, fully Bayesian inference for group-level image-on-scalar regression models. Our choice of prior, moreover, additionally allows us to construct this inference in a way that retains a full spatial rank posterior over the varying coefficient function $\boldsymbol{\beta}(\cdot)$. Compared to relatively more common low-rank posterior approximation schemes, the ability to work with full-rank spatial processes may be highly desirable in applied settings.

Descriptor	Mean	SD	IQR
0-Back Accuracy	0.87	0.09	0.11
2-Back Accuracy	0.80	0.08	0.12
{0 - 2}-Back Difference	0.07	0.09	0.12
Age (yrs)	9.99	0.62	1.08
Fluid IQ	0.29	0.75	0.97
	Percentage		
Child Gender			
Female	50.8%		
Male	49.2%		
Child Race/Ethnicity			
Asian	2.4%		
Black	8.8%		
Hispanic	17.1%		
Other	9.5%		
White	62.2%		
Household Income (US$/yr)			
$<$ 50K	22.4%		
50K–100K	30.7%		
$\geq$ 100K	46.9%		
Parental Education			
$<$ HS Diploma	2.1%		
HS Diploma/GED	5.3%		
Some College	23.0%		
Bachelor	28.6%		
Post Graduate Degree	41.0%		
Parental Marital Status			
Married Household	76.0%		
Unmarried Household	24.0%		

Table 4.2: Demographic information for children in our sample. Continuous covariates are summarized by their mean, standard deviation and interquartile range; categorical covariates are summarized by percentage of the sample in the respective category.

CHAPTER 5

Discussion and Future Work

In this dissertation we have considered three different applications of Gaussian process priors to model spatial structure in neuroimage data. Additionally, in each chapter we explored a different method of computing with high-dimensional Gaussian processes [178, 7, 35]. We further deployed the methods we proposed in each chapter to address three very distinct clinical and research questions each motivated by different data sets.

In Chapter 2, we were motivated by single patient fMRI task contrast data collected at two different spatial resolutions. Single patient data was collected from a surgical candidate prior to their operation. The objective was to map individual patient functional neuroanatomy and identify eloquent brain tissue. Such a procedure can be used to inform preoperative planning of the surgical access route. Presurgical planning requires spatially precise localization of functionally healthy brain regions, but current physical limitations of MR imaging technology lead to a reduction in the signal-to-noise ratio (SNR) with increasing spatial resolution. In a worst-case scenario, lower SNR might lead to reduced sensitivity and resection of healthy brain tissue. To try to work with this inverse SNR-voxel size limitation rather than against it, our collaborator collected scans at both high and low spatial resolution. The resulting data, however, inherently exhibit different levels of noise and lack a common spatial support, rendering them difficult to combine in a straightforward manner.

We achieved integration of data over different spatial supports by modeling the mean intensity function of both images with a unifying generative Gaussian process. To handle the massive amount of spatial information in the data we leveraged a parameter expansion idea from [178] into a relatively efficient posterior sampling algorithm using Riemann manifold Hamiltonian Monte Carlo methods. Although our sampling algorithm is quite efficient considering the ultra high-dimensional parameter space it operates in (the Wood and Chan parameter expansion trick increases the dimension of our sampled parameter from about $300,000$ to approximately 8.4×10^6), it can still take over 2 hours per $1,000$ iterations to run. While this is fast enough to be run overnight, in general excess latency between presurgical scanning and inference is not desirable. One of the most obvious next steps for this project is to consider faster posterior approximation schemes, and

study the cost to inference they entail. In Chapter 2, our loss function-based approach to inference relies only on pointwise estimates of the mean and variance of the function $\mu(\cdot)$. Since this is the case full posterior inference with MCMC may not be entirely necessary. I am also optimistic about applying Vecchia approximation [169, 90] of the prior (similar to in Chapter 4) to our model in this setting. Here, Vecchia approximation could for example be substituted for the joint prior precision of $(\boldsymbol{\mu}_h^\mathsf{T}, \boldsymbol{\mu}_s^\mathsf{T})^\mathsf{T}$ and (potentially) eliminate the need to expand the parameter space or rely on Hamiltonian Monte Carlo.

In Chapter 3 we proposed a model for covariate-informed unsupervised clustering of imaging outcomes based on a mixture of spatially varying regression models. In this framework, the mixture components can be thought of as the mean varying coefficient regression models for latent subgroups. Given data, the model then learns patient-specific distributions for subgroup or cluster assignment. We used a logistic stick-breaking process [136] to model patient-specific component weights, and projected or "predictive" Gaussian processes [146, 7] to model the spatially varying regression coefficient functions for each mixture component. Using existing data augmentation schemes [74, 126], full conditional posterior updates are readily available for each parameter, leading to relatively straightforward posterior inference using Gibbs sampling.

With this project, we were inspired by disease subtyping efforts in precision medicine. In particular, there has been a recent push to identify Autism spectrum disorder neurotypes from imaging-derived data [for reviews, see 76, 167]. The prevailing method in these types of studies is to apply some unsupervised clustering method like k-means or hierarchical clustering to imaging outcomes. In follow up *post hoc* analyses, researchers then partially validate identified clusters using correlations with patient covariate information like clinical measures of Autism severity. We have shown in simulation that this two-stage approach may produce very noisy clustering patterns and thus be of limited use. Instead of this procedure, we propose our covariate-informed clustering model for imaging outcomes, and show how such a model can lead to dramatically better clustering and estimation. We have applied this method to single-site, resting state fMRI-derived data from Autism spectrum patients and have obtained results that closely mirror existing findings from different imaging modalities [82, 26, 76].

As we have seen, working with stick-breaking type models can be quite technically challenging. This is especially true in high-dimensional outcome settings; doubly so if inference conditional on cluster assignment is a requirement of the analysis. Both of these conditions are met by our model in Chapter 3. For example, [70] study Markov chain mixing properties with Dirichlet process models in large data settings. In order to improve Markov chain mixing, [70] specifically induce label switching moves in their sampling algorithm. In contrast, in order to obtain inference for cluster-specific parameters, we have gone to some length to ensure that label-switching moves happen at most infrequently after burnin. In doing so, we are essentially trying to trap our Markov chains

"close" a single local clustering mode. From this perspective, it might make more sense to adopt an alternative computational approach for our problem. Wang and Dunson [174], for example, propose a Dirichlet process model approximation scheme using a greedy search algorithm to find a local clustering mode. The authors then propose to fix that estimate of the clustering allocation and condition any remaining analysis on it. This scheme, though not fully Bayesian, might ultimately be more in line with the goal of estimating neurotypes. Given that there is so much applied interest in this area, there may be a number of model extensions, reformulations, or estimation strategies that could conceivably be useful for such problems.

In one sense, we took a step backward with Chapter 4. Rather than add complexity to or continue to explore aspects of the model in Chapter 3, I wanted to simplify and focus in on the spatially varying coefficient (SVC) regression part of the model. In Chapter 3, we used a reduced-rank projection method to compute with spatially varying regression coefficient processes within each mixture component. While this design encodes an intuitively appealing assumption that latent mixture components in imaging should be determined by low-rank features in the data, in practice results can be sensitive to the number of low-rank bases and model performance can suffer. In the transition to Chapter 4, my goal was to study what it would take to compose a highly accurate spatial regression model for group-level neuroimaging data. To this end, we developed several models that treat regression coefficient functions $\beta_j(\cdot)$ as spatially varying. In turn, we used priors based on Gaussian processes and a measure of geodesic distance along the cortical surface to model correlations in the $\beta_j(\cdot)$ functions at nearby locations. Further, we achieved all of this using a computational technique that is both scalable to number of locations on the cortical surface *and* retains full spatial rank in the posterior distribution of the $\beta_j(\cdot)$.

Typical practice in applied imaging studies is to induce indirect spatial regularization of regression coefficients by smoothing the image data with a fixed-width kernel, and ignore spatial information otherwise. In the statistical community, current approaches to SVC-type models for group neuroimage data largely rely on lossy compression schemes like down sampling the data or other form of low-rank projection (as we did in Chapter 3). In practice, however, low-rank methods can be dissatisfying to work with (from a goodness-of-fit perspective [e.g. 157]; from an estimation perspective [e.g. discussed in 35]; and to some extent from a predictive error perspective [e.g. see methods comparison in 71]).

With Chapter 4, we propose an SVC regression model for group-level imaging data. We model the probability law governing our SVC functions with Gaussian processes, and show how this construction can naturally extend to data indexed by locations on the cortical surface. Critically, we propose Vecchia approximation [169, 90] to enable computation of the log prior on our spatial regression coefficient functions. As a result, we are able to estimate SVC functions that are numerically full spatial rank and suffer minimal approximation error. Our method, moreover, permits

relatively scalable, fully Bayesian estimation with Hamiltonian Monte Carlo.

One general issue with this framework is that by "scalable" we mostly mean in terms of the computer memory required to run our associated software, which can be minimal even for massive data sets. Our method is less scalable in terms of wall-clock time for models with a large number of covariates. Since our entire goal is to be able to estimate full spatial rank SVC functions, the number of raw parameters we are required to estimate increases dramatically with the number of covariates in the model. This can pose a practical issue if, say, the analyst would like to compare models over a set of candidates, or if study design entails a complex mean structure. An extension of our current method is most immediately needed for just these scenarios: (i) efficient model comparison and (ii) fast approximation in high-dimensional covariate settings. On the other hand, since our work in Chapter 4 enables gold-standard full-rank, full posterior inference for SVC regression models, we can more fully explore the practical consequences of using various other posterior approximation strategies such as mean-field variational inference [15], etc.

As discussed at the end of Chapter 4, another possible future direction for our cortical surface regression modeling framework is an extension to spatio-temporal data. Such an extension may be conceptually quite desirable for application to general fMRI studies. In the case of single-patient data, our present method would not require more than a minor modification of the likelihood to be applicable. We might write, for example,

$$y_t(\mathbf{s}) = \boldsymbol{x}_t^{\mathsf{T}}\boldsymbol{\beta}(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad t = 1, \ldots, T,$$

where T is the number of time points in the series, and $\epsilon_t(\cdot)$ reflects a spatio-temporal error process. In a typical applied fMRI analysis, working spatial independence approximations are used to compute $\boldsymbol{\beta}(\cdot)$ and any temporal autocorrelation parameters related to $\epsilon_t(\cdot)$ in a tractable manner. We might instead approach this problem by retaining our spatial Gaussian process prior on $\boldsymbol{\beta}(\cdot)$ from (4.3). Additional flexibility could then be built into the model by treating temporal autoregressive order in the error process as spatially varying. Inspiration could be drawn from Teng *et al.* [164], who similarly allowed spatial variation in the autoregressive order of temporal error processes.

In the single-patient spatio-temporal model proposed above, evaluation of the model log-likelihood could be achieved in a manner very similar to that we used for our marginal model (4.8) in Chapter 4. One of the defining features of autoregressive processes is conditional independence between observations outside of some neighborhood. For example, modeling the temporal autocorrelation pattern in the $\epsilon_t(\mathbf{s})$ with an AR(1) process would imply that, given $\epsilon_{t-1}(\mathbf{s})$ and $\epsilon_{t+1}(\mathbf{s})$, $\epsilon_t(\mathbf{s})$ is independent of $\epsilon_{t+k}(\mathbf{s})$ for any $|k| > 1$. As a direct consequence of this specific structure, all of the sparse precision techniques we develop and use throughout Chapter 4 are immediately applicable to computation with the spatio-temporal model outlined above.

APPENDIX A

Chapter 2: Software

A.1 Software

In this section we give a brief sketch of the command line tools we have written to implement the methods discussed in Chapter 2. Software is available for download at `https://github.com/asw221/dualres`, and should be compatible with any Unix-based system.

A.1.1 Dependencies

We require a `C`/`C++` compiler compatible with the `C++17` standard (e.g. `gcc >= 8.3.0` should suffice). At the time of writing, external dependencies include:

- The boost `filesystem` and `math` libraries
- The Eigen (3) linear algebra library
- The fftw Fourier transform library
- The nlopt library for non-linear optimization
- OpenMP
- zlib - (Likely already on your system)

A.1.2 Installation

We have used the `cmake` build system. Installation instructions assume dependencies have been preinstalled and our source code downloaded from GitHub. We first require compilation of an included NIFTI library:

```
cd /path/to/dualres/lib/nifti && make all
```

Then from `*/dualres/lib/nifti`, run:

```
mkdir ../../build && cd ../../build
cmake .. -DCMAKE_BUILD_TYPE=Release
make
```

A.1.3 Analysis of single-patient presurgical fMRI data

Dual- or single-resolution models can be fit to data stored using the NIFTI file standard with the `dualgpmf` command.

Basic syntax might look like the following:

```
./dualres/build/bin/dualgpmf \
  --highres /path/to/highres.nii \  # REQUIRED. Defines inference space
  --stdres /path/to/stdres.nii \    # Auxiliary data
  --covariance 0.806 0.131966 1 \   # [marg. var., bandwidth, exponent]
  --neighborhood 6.9 \              # Kriging approximation extent (mm)
  --output output_basename \        # Output file base name
  --hmask /path/to/hresmask.nii \   # Mask for highres image input
  --omask /path/to/outmask.nii \    # (Optional) Output image mask
  --smask /path/to/sresmask.nii \   # Mask for auxiliary image input
  --burnin 1000 \                   # MCMC burnin iterations
  --nsave 1000 \                    # MCMC iterations to save
  --thin 3 \                        # MCMC post-burnin thinning factor
  --leapfrog 25 \                   # HMC number of integrator steps
  --mhtarget 0.65 \                 # HMC target acceptance rate
  --threads 6 \                     # Number of cores to use
  --seed 48109                      # URNG seed
```

The assumed Gaussian parent process covariance function for this project is the three parameter radial basis function in equation (2.5). Above, the arguments to `--covariance` reflect the marginal variance, bandwidth, and exponent parameter for that function. See the main text for further details (Chapter 2).

A.1.4 Estimation of mean process covariance parameters

The `dualgpmf` program will estimate the covariance parameters using a minimum contrast method if they are not specified by the user, but the user control over this feature is sparse. For an enhanced interface and control over the estimation we provide `estimate_rbf`, which exposes more user options.

For example:

```
./dualres/build/bin/estimate_rbf \
  /path/to/input.nii \        # REQUIRED. Input image/data
  --mask /path/to/mask.nii \  # Mask for image input
  --xtol 1e-5 \               # Set numerical tolerance
  --bandwidth 1.0 \           # } \
  --exponent 1.5 \            # }  - Fix given RBF parameters
  --variance 1.0 \            # } /
  --constraint                # } - Constrain b/width <= expon
```

Covariance parameters estimated using `estimate_rbf` can then be passed to `dualgpmf` using the `--covariance` flag as above. For further details about our minimum contrast estimation method, see section 2.4.1 and Appendix C.

APPENDIX B

Chapter 2: Additional Patient Data Analysis Results and MCMC Diagnostics

B.1 Model diagnostics for Patient 1

In this section we include several of the general attempts we have made to probe Markov chain convergence and model fit in our analysis of patient data.

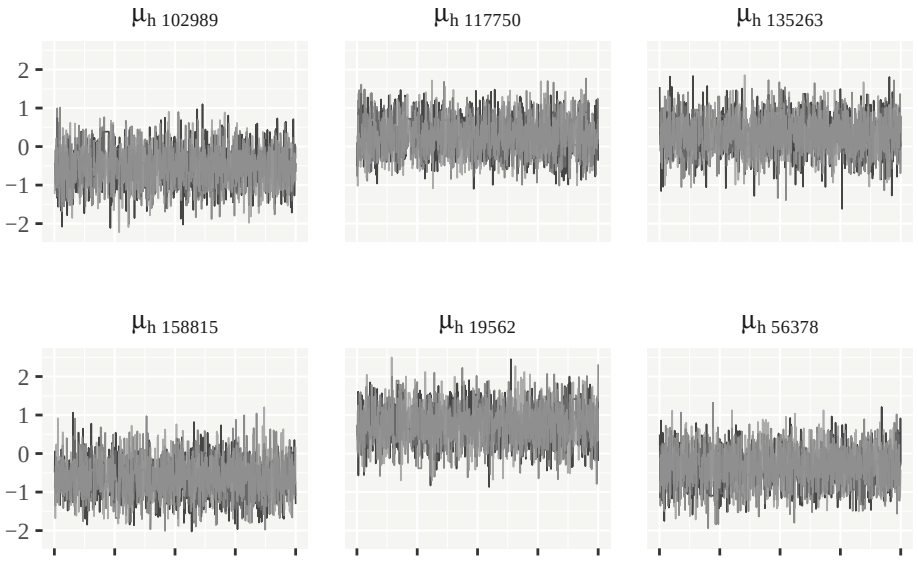


Figure B.1: Trace plots for the mean parameter of six random voxels from analysis of patient 1's data with our dual resolution model. Three different HMC chains are overlaid on one another in each subfigure.

As discussed in the main text, Fig. B.1 shows trace plots for three chains of Hamiltonian Monte Carlo (HMC) draws of the mean parameter for six random voxels. In all cases we examined, chains appear to show good convergence and mixing. Fig. B.2 (*left*) shows empirical covariograms and

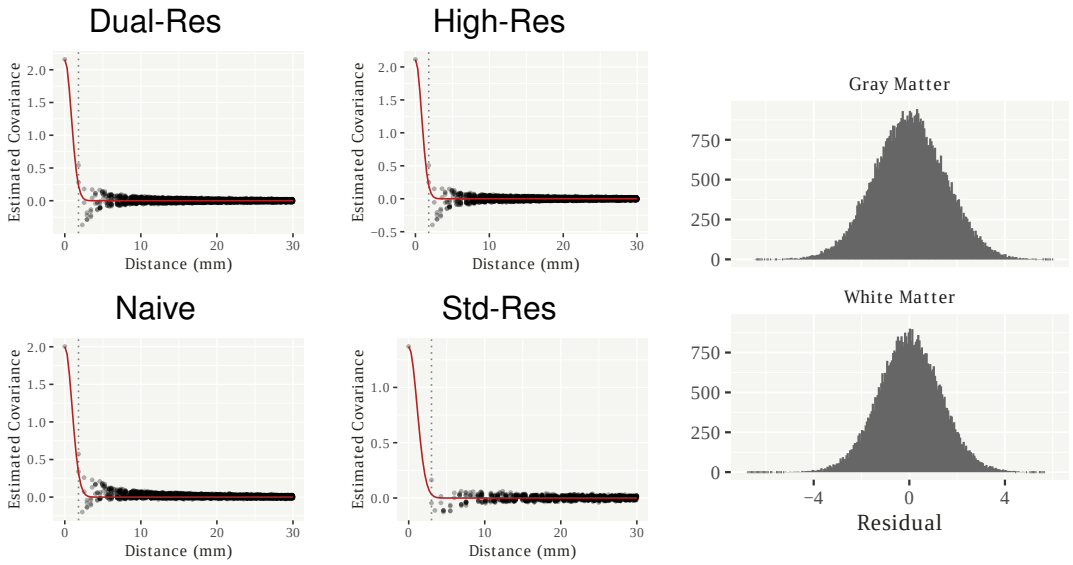


Figure B.2: (*Left*) Residual covariograms for each method. The dotted lines show minimum voxel dimensions for each resolution, suggesting that the residual independence approximation is reasonable in these data. (*Right*) Dual resolution method residual histograms roughly separated by gross tissue type. Residuals have modestly higher dispersion in gray matter than in white.

estimated covariance functions for residual images from each method. We found that the estimated residual correlation functions' full widths at half maxima were on the order of the minimum voxel dimensions in all cases. These analyses suggested that residual correlation decayed to near zero within the smallest voxel dimension widths, leading us to conclude that residual independence was a reasonable approximation in our data.

The right panel of Fig. B.2 shows histograms of the residuals from our dual resolution method roughly separated by gross tissue type. We chose to parse the residuals in this way due to some concern that a homogeneous residual variance approximation may not be fully justifiable across the whole brain. To construct this figure, we created non-overlapping gray and white matter tissue labels using the FAST program from the FSL software suite [185], though the presence of the tumor complicates this procedure. The figure suggests that residuals had modestly higher dispersion in gray matter (standard deviation = 1.49) than in white (standard deviation = 1.31). If it were not for the tumor, we might ideally only want to analyze gray matter voxels for signs of task-related activation. Given the present context, however, this strategy is not completely possible. As it stands, although it appears homogeneous residual variance may not strictly hold across different tissue types, we do not believe the approximation is so poor as to grossly impact our analyses in a negative way.

We further examined posterior predictive distributions for the data from each voxel in the high resolution image, and compared the distributions against the observed data (analysis not shown).

Dual resolution model posterior predictive inverse quantiles for the observed data were roughly uniform, suggesting that data outliers occurred no more or less frequently than would be expected given the model.

B.2 Patient 2: Sensitivity analyses

In this section we include a brief sensitivity analyses related to the choice of neighborhood size and covariance function in our dual resolution method.

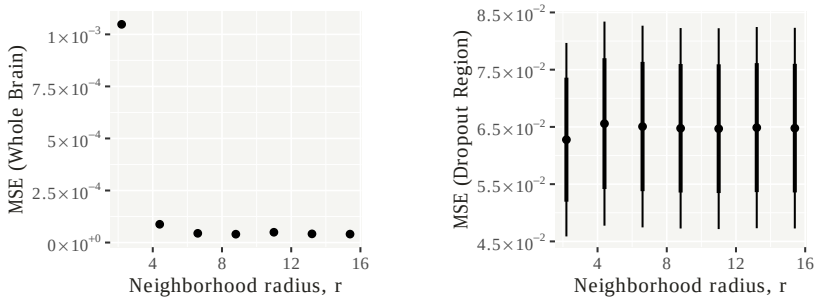


Figure B.3: Mean squared error (MSE) of the posterior expectation of $\mu(\cdot)$ given fixed $\boldsymbol{\theta}$ but different values of r. The (*left*) panel shows MSE of $\mu(\cdot)$ evaluated across the whole brain, while the (*right*) panel shows the predictive MSE for voxels in patient 2's dropout region. Thick and thin lines give approximate 80% and 95% confidence intervals.

Our dual resolution mapping method relies on a neighborhood radius parameter r to construct locally kriged samples of $\boldsymbol{\mu}_s$ given $\boldsymbol{\mu}_h$ (see section 2.2.2 in the main text). Conceptually, this construction is somewhat inspired by the so called nearest neighbor Gaussian process [35, 44]. In practice, we treat r as a hyperparameter and condition analyses on it, though it is of interest to understand how the choice of r affects inference about $\boldsymbol{\mu}_h$. In our patient data analyses (sections 2.4.2 and 2.4.3 in the main text), we set r to be approximately equal to the estimated full kernel width at half maximum. This resulted in neighborhood sizes on the order of 300–700 voxels for our patient data.

To explore the influence of r on estimation and prediction, we fit our dual resolution model to the patient 2 data under several different settings, all for fixed $\boldsymbol{\theta}$. As a comparison point, we took the posterior mean of $\mu(\cdot)$ fit to the data without missingness and conditioned on $r = 11$ mm. We then compared against the posterior mean of $\mu(\cdot)$ from repeat analyses of the with-missingness data and varying values of r (see section 2.4.3 in the main text for an explanation of the two data sets). For these repeat analyses, we chose values of r based on multiples of the largest high resolution image voxel dimension (2.2 mm). Fig. B.3 summarizes this experiment in terms of the squared error of $\mu(\cdot)$ averaged over the whole brain (*left*) and voxels in the dropout region (*right*). From

these results we conclude that as long as r is sufficiently large ($\geq$ 6.6 mm or so; corresponding to neighborhood sizes of at least 100–200 voxels), it does not appear to have much influence on posterior estimates.

In our analysis of Patient 2's covariogram, the exponential model we used tends to underestimate the proximal empirical covariances. We chose to use the radial basis covariance function-family largely because of the substantial history of gaussian smoothing in applied MRI analysis. Additional literature suggests exponential smoothing kernels are perhaps more appropriate for fMRI data [65]. We considered alternative covariance functions and their impact on our analysis, and we summarize one such alternative here.

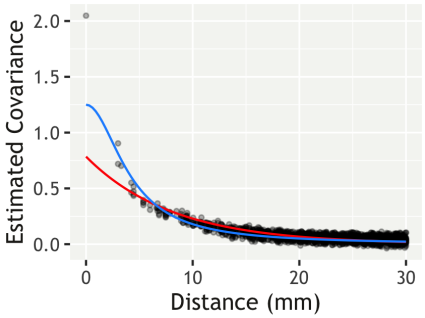


Figure B.4: Reanalysis of Patient 2's covariogram. The red line reproduces the exponential covariance model from the main text; the blue line shows a rational quadratic covariance model.

In Fig. B.4, we compare the exponential covariance model from the main text against a rational quadratic model, and find that the rational quadratic model fits the proximal empirical covariances quite well. The specific rational quadratic model in the above figure is,

$$k_{R.Q.}(\mathbf{v}, \mathbf{v}') = 1.25\left(1 + \frac{\|\mathbf{v} - \mathbf{v}'\|^2}{16.67 \times 0.99}\right)^{-0.99}.$$

It is impossible to tell visually, but the rational quadratic model in Fig. B.4 is *sub*-optimal in the sense that it has a very slightly higher residual weighted sum of squares than the exponential model. Better than either might be some weighted linear or piecewise combination of the two.

Choice of the covariance function is more art than science. An interesting feature of this problem is that the empirical covariances will tend to overestimate the true mean field covariance if the noise is positively correlated spatially. Assuming as we do in the main text that $Y(\mathbf{v}) = \mu(\mathbf{v}) + \epsilon(\mathbf{v})$ and that $\mu(\mathbf{v}) \perp \epsilon(\mathbf{v}')$ for all $\mathbf{v}, \mathbf{v}'$, the empirical covariances will be overestimates since,

$$\mathrm{cov}\{Y(\mathbf{v}), Y(\mathbf{v}')\} = \mathrm{cov}\{\mu(\mathbf{v}), \mu(\mathbf{v}')\} + \mathrm{cov}\{\epsilon(\mathbf{v}), \epsilon(\mathbf{v}')\}.$$

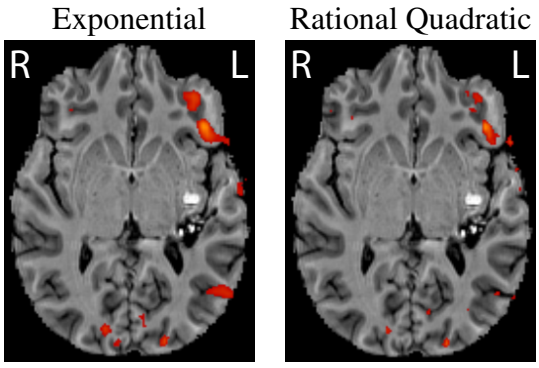


Figure B.5: The figure shows thresholded posterior inference of activation regions for Patient 2 in an example horizontal slice. The color scale is shared between sub figures and reflects an approximate posterior probability of activation (range 0.3–1.0).

Although in our work we modeled the error structure as a white noise-type process for simplicity, it is perhaps more realistic to assume the errors may be positively correlated over short distances. If the errors are in fact correlated spatially, then it may be preferable to use a covariance model that underestimates the proximal empirical covariances.

In practice we observed mild, intermittent spatial autocorrelation patterns in our residual images (see Section B.1 for example). We switched from the exponential to the rational quadratic covariance function and did not find that this change ameliorated residual autocorrelation patterns. Because the rational quadratic covariance model places higher correlation between proximal elements and decays more rapidly, we found that it yielded posterior inference that was both noisier and less sensitive than our primary analysis (see Fig. B.5). Clearly results are sensitive to the choice of covariance function to some degree, underscoring the importance of these issues.

APPENDIX C

Chapter 2: Computational Details

C.1 Detailed explanation of posterior computation scheme

In Section 2.2.3, we outlined a posterior computation algorithm for our model that relies on embedding the covariance of $\boldsymbol{\mu}_h$ in a higher dimensional nested block-circulant matrix. We present the details of this algorithm here. Broadly, our posterior computation algorithm has a Hamiltonian Monte Carlo (HMC) -within-Gibbs sampling structure. Full conditional updates are available for all of our model parameters, but it is numerically challenging to evaluate or sample from the full conditional distribution of $\boldsymbol{\mu}_h$.

In the main text, we discussed how we drew inspiration from the work of [178] to design an efficient HMC algorithm to facilitate sampling of $\boldsymbol{\mu}_h$. We elaborate on that algorithm in detail here. First, we embed $\boldsymbol{\mu}_h$ in a higher dimensional random field $\boldsymbol{u}$, which is constructed so that the prior variance of $\boldsymbol{u}$ is a nested block-circulant matrix $\boldsymbol{C}$. The prior variance of $\boldsymbol{\mu}_h$—$\boldsymbol{K}_h$—is a principal submatrix of $\boldsymbol{C}$ (see Fig. 2.2 in the main text for a schematic picture). We never actually construct or store the full matrix $\boldsymbol{C}$: its base $\boldsymbol{c}$ can be computed following Algorithm 2 below. With only the base $\boldsymbol{c}$ in memory, the complex eigenvalues of $\boldsymbol{C}$ can be computed using discrete Fourier transform (DFT) software:

$$\boldsymbol{\lambda} \leftarrow \mathcal{F}(\boldsymbol{c})/N,$$

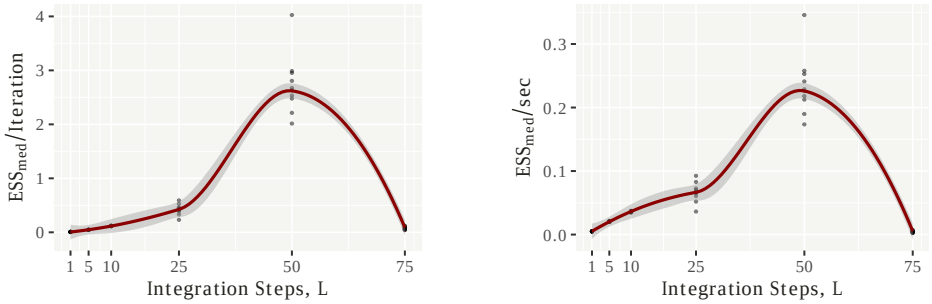
where N is the number of elements in $\boldsymbol{c}$.

Then, let $\boldsymbol{\xi} = \boldsymbol{u} + \boldsymbol{v}i$ represent a complex Gaussian random field with real part $\boldsymbol{u}$, imaginary part $\boldsymbol{v}$, and with the prior properties that $\boldsymbol{u} \perp \boldsymbol{v}$ and $\mathrm{var}(\boldsymbol{u}) \equiv \mathrm{var}(\boldsymbol{v}) \equiv \boldsymbol{C}$. Writing out the prior in terms of $\boldsymbol{\xi}$,

$$\boldsymbol{\xi} = \boldsymbol{u} + \boldsymbol{v}i, \qquad \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}), \quad \boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}),$$

does not change our model, moreover: the imaginary and non-brain parts of $\boldsymbol{\xi}$ can simply be integrated away to recover our original prior on $\boldsymbol{\mu}_h$. Similarly, our plan will be to obtain posterior samples of $\boldsymbol{\xi}$ through HMC, and then simply discard extraneous elements to be left with a posterior

sample of $\boldsymbol{\mu}_h$.



Method	Run time (hrs)	Total RAM (Gb)
Dual	2.76	2.32
High	1.88	2.05
Std	0.44	1.50

Figure C.1: Dual resolution algorithm efficiency (median ESS per iteration and per second) as a function of integration steps L in analysis of whole brain patient data. *ESS* denotes the effective sample size of elements of $\boldsymbol{\mu}_h$. Peak efficiency was estimated around $L = 50$. Analyses were replicated 10 times for each value of L, and were timed on a Thelio System76 desktop with 62 Gb of free RAM and 20 logical cores (3.3 GHz Intel® Core™ i9 processors). Below the figure, we summarize the overall computational burden for real patient data on this hardware and at $L = 25$ steps. Run time is given in hours per 1,000 iterations; our naive method has the same cost as the high-resolution only method.

HMC relies on several tuning parameters, including the choice of momentum distribution, mass matrix, step size, and number of numerical integration steps [116]. While a review of HMC-flavored algorithms and tuning parameter selection is beyond the scope of this paper, we will detail our approach to tuning parameter selection for model (2.1). Given the other tuning parameters and a target Metropolis-Hastings rate (which we fixed at 65%), we tuned the step size ϵ during warm up following the dual averaging method of [73]. We then fixed ϵ_0 at the value of ϵ on the last burnin iteration, and drew $\epsilon \sim \text{Uniform}(0.9\,\epsilon_0, 1.1\,\epsilon_0)$ to induce random integration path lengths (the product ϵL) during sampling, potentially helping the algorithm escape local modes [116]. To inform selection of the number of leapfrog integration steps L, we performed repeated analyses of patient data. Results of this experiment suggest $L = 25$ or $L = 50$ as practical starting points for best algorithmic efficiency (see Fig. C.1).

Let,

$$\mathcal{L}(\boldsymbol{\xi}) = \ln \pi(\boldsymbol{\xi} \mid \boldsymbol{Y}_h, \boldsymbol{Y}_s, \boldsymbol{\mu}_s, \boldsymbol{\theta}, \sigma_h^2, \sigma_s^2, r)$$

represent the full conditional log posterior of $\boldsymbol{\xi}$. Since $\exp\{\mathcal{L}(\boldsymbol{\xi})\}$ is complex Gaussian, we in turn chose a complex Gaussian distribution for HMC momenta. [62] suggest exploiting Riemannian geometry in HMC by adapting the algorithm's mass matrix, $\boldsymbol{M}$, to the local curvature of the

log posterior. The authors suggest that taking $\boldsymbol{M}$ proportional to the negative Hessian of the log posterior leads to improved algorithmic efficiency in high dimensions, though this approach is not typically feasible when the dimension of $\boldsymbol{M}$ is more than a few thousand. Up to a permutation of $\boldsymbol{\xi}$, in our model, we have that,

$$-\nabla^2\mathcal{L}(\boldsymbol{\xi}) = \begin{pmatrix} \sigma_h^{-2}\boldsymbol{I} + \sigma_s^{-2}\boldsymbol{W}^\mathsf{T}\boldsymbol{W} \\ \mathbf{0} \end{pmatrix} + \boldsymbol{F}\boldsymbol{\Lambda}^{-1}\boldsymbol{F}^\mathsf{H}, \tag{C.1}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$, and $\boldsymbol{F}$ is the 3D DFT matrix as in the main text. In the present case, (C.1) is ultrahigh dimensional and impossible to work with directly, but by dropping the term involving $\boldsymbol{W}^\mathsf{T}\boldsymbol{W}$, which is dense, and extending $\sigma_h^{-2}\boldsymbol{I}$ we can arrive at an alternative choice of mass matrix. Let $\boldsymbol{M}(\sigma_h^2)$ denote the matrix-valued function,

$$\boldsymbol{M}(\sigma_h^2) = \boldsymbol{F}[\boldsymbol{\Lambda}^{-1} + \sigma_h^{-2}\boldsymbol{I}]\boldsymbol{F}^\mathsf{H}, \tag{C.2}$$

which, like $\boldsymbol{C}$, is nested block-circulant, and easy to compute with. Circulant matrices have been used successfully as preconditioners in other gradient-based optimization schemes for imaging problems [e.g. 43]. If each element in $\mathcal{R}e(\boldsymbol{\lambda})$ is strictly greater than zero, $\boldsymbol{M}(\sigma_h^2)$ is positive definite and so can be used to define a metric tensor on a Riemannian manifold as in [62]. Some additional intuition can be gained by considering how (C.1) relates to a missing data problem. If we were only modeling high resolution data, and those data were observed on the entire extended grid with variance σ_h^2, then (C.2) would be exactly the negative Hessian of $\mathcal{L}(\boldsymbol{\xi})$.

Algorithm 1 Riemann manifold HMC for dual resolution mapping models. $\boldsymbol{F}$ denotes the scaled 3D DFT matrix; products of the form, $\boldsymbol{F}^{\mathsf{H}}\boldsymbol{p} = \mathcal{F}^{-1}(\boldsymbol{p})$, for example, can be computed efficiently using DFT software.

procedure UPDATEMEAN($\boldsymbol{\xi}$; $\boldsymbol{\lambda}$, σ_h^2, ϵ, L)
  Compute eigenvalues of $\boldsymbol{M}(\sigma_h^2)$:
    $\lambda_i^M \leftarrow \sigma_h^{-2} + \lambda_i^{-1}$
    Set $\boldsymbol{\Lambda}_M \leftarrow \text{diag}\{\lambda_i^M\}$, $i = 1, \ldots, \dim(\boldsymbol{\xi})$
  Sample momentum, $\boldsymbol{p} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{F}\boldsymbol{\Lambda}_M\boldsymbol{F}^{\mathsf{H}})$
  Compute total energy, $H \leftarrow -\mathcal{L}(\boldsymbol{\xi}) + \frac{1}{2}\boldsymbol{p}^{\mathsf{H}}\boldsymbol{F}\boldsymbol{\Lambda}_M^{-1}\boldsymbol{F}^{\mathsf{H}}\boldsymbol{p}$
  Set $\boldsymbol{\xi}^{\text{new}} \leftarrow \boldsymbol{\xi}$
  for l in $1, \ldots, L$ **do** ▷ Leapfrog integrator
    $\boldsymbol{p} \leftarrow \boldsymbol{p} + \frac{\epsilon}{2}\nabla\mathcal{L}(\boldsymbol{\xi}^{\text{new}})$
    $\boldsymbol{\xi}^{\text{new}} \leftarrow \boldsymbol{\xi}^{\text{new}} + \epsilon\boldsymbol{F}\boldsymbol{\Lambda}_M^{-1}\boldsymbol{F}^{\mathsf{H}}\boldsymbol{p}$
    $\boldsymbol{p} \leftarrow \boldsymbol{p} + \frac{\epsilon}{2}\nabla\mathcal{L}(\boldsymbol{\xi}^{\text{new}})$
  Compute $H^{\text{new}} \leftarrow -\mathcal{L}(\boldsymbol{\xi}^{\text{new}}) + \frac{1}{2}\boldsymbol{p}^{\mathsf{H}}\boldsymbol{F}\boldsymbol{\Lambda}_M^{-1}\boldsymbol{F}^{\mathsf{H}}\boldsymbol{p}$
  Set $\boldsymbol{\xi} \leftarrow \boldsymbol{\xi}^{\text{new}}$ with probability $\alpha = \min\{1, \exp(H - H^{\text{new}})\}$
  Discard all elements of $\boldsymbol{\xi}$ that do not correspond to $\boldsymbol{\mu}_h$
  Return posterior sample of $\boldsymbol{\mu}_h$

With all this in hand, samples of $\boldsymbol{\mu}_h$ can be drawn following Algorithm 1. In particular, note how all products involving $\boldsymbol{F}$ can be computed with DFT software. In addition, the quadratic forms in Algorithm 1 represent computations over ultrahigh dimensional components. The quadratic form $\boldsymbol{p}^{\mathsf{H}}\boldsymbol{F}\boldsymbol{\Lambda}_M^{-1}\boldsymbol{F}^{\mathsf{H}}\boldsymbol{p}$, for example, can be evaluated by computing,

$$\boldsymbol{\phi} \leftarrow \mathcal{F}^{-1}(\boldsymbol{p}),$$

into a temporary product, $\boldsymbol{\phi}$, and then summing over terms $\sum_i \bar{\phi}_i \cdot \phi_i/\lambda_i^M$, where $\bar{a}$ denotes the complex conjugate of a. When working in single precision, we found it necessary to use the Kahan summation algorithm [89], or similar correction, to evaluate these long sums accurately.

Finally, our other parameters, $\boldsymbol{\mu}_s$, σ_h^2, and σ_s^2 can easily be sampled with full conditional Gibbs updates. We particularly note that our prior places the restriction $\sigma_h^2 > \sigma_s^2$ so that the conditional posteriors of both nugget variance parameters are truncated inverse Gamma. We sometimes encountered numerical difficulty sampling these parameters during warm up. As a result, we chose to ignore the restriction on σ_h^2 and σ_s^2 programmatically, and simply discard posterior samples where the restriction was not satisfied. After warm up, however, we found that even when working with patient data the posterior probability that $\sigma_h^2 > \sigma_s^2$ was effectively unity, and that we never had to discard or post-process MCMC samples in this way.

C.2 Circulant base construction

This section presents a simple algorithm to illustrate circulant matrix base computation for our applications.

Algorithm 2 Compute the base of a circulant matrix associated with a 3D grid

procedure COMPUTECIRCULANTBASE($\boldsymbol{d}$, $K(\cdot,\cdot;\boldsymbol{\theta})$)
  Inputs: $\boldsymbol{d}$, original 3D grid dimensions; $K(\cdot,\cdot;\boldsymbol{\theta})$ covariance function parameterized by $\boldsymbol{\theta}$
  Compute extended grid dimensions, $d_i^\star \leftarrow 2^{\log_2\lceil 2(d_i-1)\rceil}$ for $i = 1, 2, 3$
  $k \leftarrow 0$, $h \leftarrow 1$
  Find location $\mathbf{v}_1$ associated with grid position $(1, 1, 1)$
  for l in $1, \ldots, d_3^\star$ **do** ▷ Column-major order
    $j \leftarrow 0$
    if $l \leq d_3$ **then** $k \leftarrow k + 1$ **else** $k \leftarrow k - 1$
    for m in $1, \ldots, d_2^\star$ **do**
      $i \leftarrow 0$
      if $m \leq d_2$ **then** $j \leftarrow j + 1$ **else** $j \leftarrow j - 1$
      for n in $1, \ldots, d_1^\star$ **do**
        if $n \leq d_1$ **then** $i \leftarrow i + 1$ **else** $i \leftarrow i - 1$
        Find location $\mathbf{v}$ associated with grid position (i, j, k)
        Compute $c_h \leftarrow K(\mathbf{v}_1, \mathbf{v}; \boldsymbol{\theta})$
        $h \leftarrow h + 1$
  Return circulant matrix base, $\boldsymbol{c}$

C.3 Covariance function estimation

In this section we detail our procedure to estimate isotropic covariance functions from 3D data; in practice the method could be extended to arbitrary n dimensional data sources. The methods considered herein are not new but are included for completeness. We also report simulation results using this method to estimate the covariance from small three dimensional images and show that the method has relatively small bias in most simulation settings.

C.3.1 Minimum contrast estimation procedure

Algorithm 3 Minimum contrast estimation of $\boldsymbol{\theta}$: high level overview

procedure ESTIMATEIMAGECOVARIANCE($\boldsymbol{Y}$, $\mathrm{k}(\cdot;\boldsymbol{\theta})$, $\boldsymbol{\Theta}$) ▷ With the argument to $\mathrm{k}(\cdot;\boldsymbol{\theta})$ the Euclidean distance between any two points, $\|\mathbf{v}-\mathbf{v}'\|$

Inputs: Image $\boldsymbol{Y}$; covariance function $\mathrm{k}(\cdot;\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$ with feasible region $\boldsymbol{\Theta}$

Construct $\mathcal{D} \leftarrow$ EXTRACTCOVARIANCESUMMARY($\boldsymbol{Y}$) ▷ With $\mathcal{D} = (\boldsymbol{d}, \hat{\boldsymbol{c}}, \boldsymbol{\omega})$

Return $\arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{\dim(\hat{\boldsymbol{c}})} \omega_i[\hat{c}_i - \mathrm{k}(d_i;\boldsymbol{\theta})]^2$

Algorithm 3 outlines our minimum contrast estimation (MCE) procedure at a high level. The algorithm first extracts summary data $\mathcal{D} = (\boldsymbol{d}, \hat{\boldsymbol{c}}, \boldsymbol{\omega})$ from the input data source $\boldsymbol{Y}$, where $\hat{\boldsymbol{c}}$ are empirical covariances between elements of $\boldsymbol{Y}$ offset by corresponding distances $\boldsymbol{d}$, and $\boldsymbol{\omega}$ is a set of corresponding weights (defined below in algorithm 4). The algorithm then finds $\boldsymbol{\theta}$ from within constraint region $\boldsymbol{\Theta}$ to minimize a weighted least squares contrast between the $\hat{c}_i$ and $\mathrm{k}(d_i;\boldsymbol{\theta})$.

With $\mathrm{k}(\cdot;\boldsymbol{\theta})$ taken to be the radial basis function as in (2.5), for example, the parameters $\boldsymbol{\theta}$ correspond to the marginal variance τ^2, correlation bandwidth ψ, and exponent ν. For this problem, we took the feasible region $\boldsymbol{\Theta}$ to constrain $0 < \tau^2 < \hat{c}_0$, $0 < \psi$, and $0 < \nu \leq 2$, where $\hat{c}_0$ is the empirical variance of $\boldsymbol{Y}$. For problems we consider, we found that the additional constraint $\psi \leq \nu$ frequently helped improve estimation.

Construction of $\mathcal{D}$ using a modified 3D raster scan is outlined in algorithm 4. In the algorithm, empirical covariances between voxels and their neighbors are computed by shifting the (i, j, k) index of each voxel by the rows of the matrix $\boldsymbol{P}$ (which is constructed with the procedure outlined in Algorithm 5). The rows of $\boldsymbol{P}$ define a series of perturbations in a dense 3D raster scan. In one dimension, a raster scan might only look ahead one pixel at a time so as to visit each pair of adjacent pixels only once. In two dimensions, the procedure might be defined to look ahead one pixel and look down one pixel for the same reason. In three dimensions, a simple raster might look ahead, down, and to the right by one or more voxels. We designed our procedure to sample local pairs of voxels more densely than this while still only visiting each unique pair once. Briefly, our algorithm "looks ahead" by visiting pairs of voxels within an $(n_0 \times n_0 \times n_0)$ voxel cube such that the polar and azimuthal angles of the search are between $[0°, 180°)$. We further extended this search by adding simple raster scan perturbations out to an n_1 voxel distance. In algorithm 5, we defined $n_0 = 18$ voxels and $n_1 = 25$ voxels by default. Our default values encompass a large number of perturbations while limiting the total computation time to a few seconds for full scale brain images.

Algorithm 4 Compute empirical covariance summary data

procedure EXTRACTCOVARIANCESUMMARY($\boldsymbol{Y}$, n_0, n_1)
  Inputs: Image $\boldsymbol{Y}$ with dimensions $\boldsymbol{q} \in \mathbb{R}^3$. We set $n_0 = 18$, and $n_1 = 25$ by default
  Set $N \leftarrow q_1 \cdot q_2 \cdot q_3$ ▷ N is the total number of voxels in image $\boldsymbol{Y}$
  Store $\boldsymbol{P} \leftarrow$ IMAGESCANPERTURBATIONS(n_0, n_1) ▷ $\boldsymbol{P}$ is an $(M \times 3)$ matrix of integers
  Allocate $\boldsymbol{d} \in \mathbb{R}^M$, $\hat{\boldsymbol{c}} \in \mathbb{R}^M$ ▷ $\boldsymbol{d}$—perturbation distances; $\hat{\boldsymbol{c}}$—empirical covariances
  $\boldsymbol{s}^{ab} \leftarrow \mathbf{0}_M$, $\boldsymbol{s}^{a} \leftarrow \mathbf{0}_M$, $\boldsymbol{s}^{b} \leftarrow \mathbf{0}_M$ ▷ Accumulators for sufficient statistics
  $\boldsymbol{r} \leftarrow \mathbf{0}_M$ ▷ Accumulators for counts of voxel pairs
  Compute sufficient statistics for pairs of voxels separated by perturbation distances:
  for h in $1, \ldots, N$ **do** ▷ Outer loop over voxels
    Locate grid position (i, j, k) such that corresponds to voxel $\mathbf{v}_h$
    if Y_{ijk} corresponds to brain data **then**
      for m in $1, \ldots, M$ **do** ▷ Inner loop over perturbations
        $(i', j', k') \leftarrow (i, j, k) + \boldsymbol{P}_m^{\mathsf{T}}$
        if $Y_{i'j'k'}$ corresponds to brain data **then** ▷ Update sufficient statistics
          $s_m^{ab} \leftarrow s_m^{ab} + Y_{ijk} \cdot Y_{i'j'k'}$
          $s_m^{a} \leftarrow s_m^{a} + Y_{ijk}$; $s_m^{b} \leftarrow s_m^{b} + Y_{i'j'k'}$
          $r_m \leftarrow r_m + 1$
  Compute distances and empirical covariances associated with grid perturbations:
  "Locate" voxel $\mathbf{v}_0$ associated with grid position $(i, j, k) = \mathbf{0}_3$
  for m in $1, \ldots, M$ **do**
    "Locate" voxel $\mathbf{v}'$ associated with grid position $\boldsymbol{P}_m$
    $d_m \leftarrow \|\mathbf{v}_0 - \mathbf{v}'\|$
    if $r_m > 1$ **then**
      $\hat{c}_m \leftarrow (s_m^{ab} - s_m^{a} s_m^{b} / r_m)/(r_m - 1)$
  Set $\omega_m \leftarrow \#(\boldsymbol{d} = d_m)$ **for** m in $1, \ldots, M$ ▷ Count of instances of unique elements in $\boldsymbol{d}$
  Set $\omega_m \leftarrow 1/\omega_m$ **if** $\omega_m > 0$ **and** $\omega_m \leftarrow 0$ otherwise **for** m in $1, \ldots, M$
  Return $\mathcal{D} = (\boldsymbol{d}, \hat{\boldsymbol{c}}, \boldsymbol{\omega})$

Algorithm 5 Construct matrix $\boldsymbol{P}$ of grid index perturbations for minimum contrast estimation procedure.

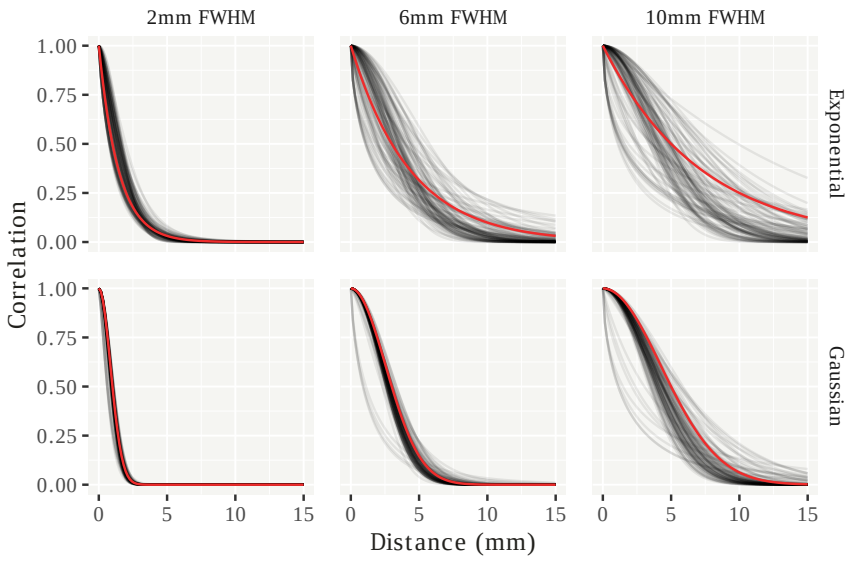
procedure IMAGESCANPERTURBATIONS(n_0, n_1)
    Inputs: positive integers n_0, n_1, $n_0 < n_1$
    Construct principal direction matrix $\boldsymbol{U} \in \mathbb{R}^{(14\times 3)}$ such that each element $U_{ij} \in \{-1, 0, 1\}$; the polar angle of each row of $\boldsymbol{U}$ is between $[0°, 180°)$; and the azimuthal angle of each row of $\boldsymbol{U}$ is between $[0°, 180°)$. In our construction, $\boldsymbol{U}$ includes a row of all 0's
    Construct $\boldsymbol{Q} \in \mathbb{R}^{(n_0^3\times 3)}$ with rows consisting of all 3-element permutations of elements of $(1, \ldots, n_0)$
    Compute $\boldsymbol{P} \leftarrow \boldsymbol{Q} * \boldsymbol{U}$, where $*$ denotes the column-wise Khatri-Rao product
    for k in $n_0 + 1, \ldots, n_1$ **do**
        $\boldsymbol{P} \leftarrow [\boldsymbol{P}^\mathsf{T} \;\; k\,\boldsymbol{I}_3]^\mathsf{T}$
    Remove duplicate rows from $\boldsymbol{P}$
    Return $\boldsymbol{P}$

C.3.2 Simulations with minimum contrast estimation

Fig. C.2 presents the results of a simulation assessing the performance of our MCE procedure. We simulated small 3D images on a $(32 \times 32 \times 16)$ grid, treating voxels as isotropic 1 mm^3. In our simulation, we drew mean images from Gaussian processes with either Exponential or Gaussian correlation functions; unit marginal variance; and either two, six, or ten mm full widths at half maxima. Mean images were corrupted with independent Gaussian noise with the signal to noise ratio set to 0.2 roughly to match our observed patient data and our 2D simulations in section 2.3.

Since the Gaussian predictive process basis described in section 2.2.2 relies only on the correlation bandwidth and exponent parameters ψ and ν, the most important measure of estimation success in our setting is recovery of the correlation function, not necessarily estimation of $\boldsymbol{\theta}$ itself. For any given dataset, the nonlinear least squares objective in algorithm 3 might be multimodal in $\boldsymbol{\theta}$, but this is relatively unimportant if the resulting correlation functions at different modes behave similarly.

In Fig. C.2, the red line in each panel shows the true correlation function used to generate underlying mean images in simulation, and the corresponding gray lines show estimated correlation functions from 100 repeated simulations. The table below the panels summarizes pointwise bias and variance averaged over a grid of 1,000 equally spaced points from [0, 15] (mm). In the worst case scenario (10 mm FWHM Exponential correlation function), pointwise mean squared error was on average only about 3.8×10^{-2}, and was between $[1.1, 7.5] \times 10^{-2}$ for 95% of points on the grid. Even with relatively small 3D images, these results suggest our MCE procedure can recover



Kernel	*FWHM*	*Bias*	*Variance*
Exponential	2	-6.73×10^{-3}	8.46×10^{-4}
Exponential	6	-6.04×10^{-2}	7.22×10^{-3}
Exponential	10	-1.47×10^{-1}	1.56×10^{-2}
Gaussian	2	6.70×10^{-3}	1.17×10^{-3}
Gaussian	6	-3.86×10^{-3}	3.45×10^{-3}
Gaussian	10	-3.07×10^{-2}	1.73×10^{-2}

Figure C.2: Recovery of the correlation function in small 3D images. Each gray line shows a correlation function estimated in repeated simulation (true correlation functions for each panel shown in red). In the table, *Bias* and *Variance* were computed pointwise and averaged over a dense grid from $[0, 15]$ (mm).

the true correlation functions reasonably well.

APPENDIX D

Chapter 2: Additional 2D Simulation Results

D.1 Detailed Simulation Results with 2D Images

In this section we give results designed augment those reported in section 2.3.2 with additional simulation settings. Tables follow the exact format of Table 2.1 in the main text. In all cases considered, our dual resolution method had the lowest mean squared error (MSE) and false negative rate. Of potential interest, however, is that the high resolution-only method was the second best performer when data were simulated with a marginal exponential correlation structure (Table D.1), and the naive data averaging method was the second best performer when data were simulated with a marginal Gaussian correlation structure (Table D.2).

Model	*Kernel*	SNR_s:SNR_h	SNR_h	*MSE*	*False –*
Dual	Exponential	1	0.1	**0.20**	**31.8% (0.4)**
High	Exponential	1	0.1	0.23	34.0% (0.5)
Naive	Exponential	1	0.1	0.30	43.6% (0.4)
Std	Exponential	1	0.1	0.47	43.1% (0.6)
Dual	Exponential	1	0.2	**0.17**	**29.3% (0.4)**
High	Exponential	1	0.2	0.20	31.0% (0.4)
Naive	Exponential	1	0.2	0.29	43.0% (0.3)
Std	Exponential	1	0.2	0.43	40.6% (0.4)
Dual	Exponential	2	0.1	**0.18**	**30.6% (0.4)**
High	Exponential	2	0.1	0.23	34.0% (0.5)
Naive	Exponential	2	0.1	0.29	42.7% (0.4)
Std	Exponential	2	0.1	0.43	40.6% (0.4)
Dual	Exponential	2	0.2	**0.15**	**28.5% (0.3)**
High	Exponential	2	0.2	0.20	31.0% (0.4)
Naive	Exponential	2	0.2	0.29	42.4% (0.3)
Std	Exponential	2	0.2	0.41	40.5% (0.3)
Dual	Exponential	4	0.1	**0.16**	**29.5% (0.3)**
High	Exponential	4	0.1	0.23	34.0% (0.5)
Naive	Exponential	4	0.1	0.29	42.3% (0.3)
Std	Exponential	4	0.1	0.41	40.5% (0.3)
Dual	Exponential	4	0.2	**0.14**	**27.9% (0.3)**
High	Exponential	4	0.2	0.20	31.0% (0.4)
Naive	Exponential	4	0.2	0.28	42.1% (0.3)
Std	Exponential	4	0.2	0.40	40.8% (0.3)

Table D.1: Results for estimation and inference quality in 2D simulations when background signal has an Exponential correlation structure. As in Table 2.1, results for the *High* resolution method do not change across the different SNR ratios, but are repeated to facilitate comparison. *MSE* refers to mean squared error computed over the entire high resolution mean parameter vector. *False –* reports the mean (SE) false negative error rate when the number of discoveries was fixed at 450. One hundred replicates per parameter combination.

Model	*Kernel*	SNR_s:SNR_h	SNR_h	*MSE*	*False –*
Dual	Gaussian	1	0.1	**0.24**	**29.8% (0.3)**
High	Gaussian	1	0.1	0.28	34.1% (0.3)
Naive	Gaussian	1	0.1	0.25	34.5% (0.3)
Std	Gaussian	1	0.1	0.59	50.0% (0.4)
Dual	Gaussian	1	0.2	**0.17**	**24.2% (0.2)**
High	Gaussian	1	0.2	0.21	27.1% (0.2)
Naive	Gaussian	1	0.2	0.19	27.1% (0.2)
Std	Gaussian	1	0.2	0.58	43.5% (0.3)
Dual	Gaussian	2	0.1	**0.21**	**27.5% (0.2)**
High	Gaussian	2	0.1	0.28	34.1% (0.3)
Naive	Gaussian	2	0.1	0.24	33.3% (0.3)
Std	Gaussian	2	0.1	0.58	43.5% (0.3)
Dual	Gaussian	2	0.2	**0.15**	**22.8% (0.2)**
High	Gaussian	2	0.2	0.21	27.1% (0.2)
Naive	Gaussian	2	0.2	0.18	26.4% (0.2)
Std	Gaussian	2	0.2	0.57	38.6% (0.2)
Dual	Gaussian	4	0.1	**0.18**	**25.0% (0.2)**
High	Gaussian	4	0.1	0.28	34.1% (0.3)
Naive	Gaussian	4	0.1	0.24	33.0% (0.3)
Std	Gaussian	4	0.1	0.57	38.6% (0.2)
Dual	Gaussian	4	0.2	**0.12**	**21.1% (0.2)**
High	Gaussian	4	0.2	0.21	27.1% (0.2)
Naive	Gaussian	4	0.2	0.18	25.9% (0.2)
Std	Gaussian	4	0.2	0.54	34.8% (0.2)

Table D.2: Results for estimation and inference quality in 2D simulations when background signal has a Gaussian correlation structure. As in Tables 2.1 and D.1, results for the *High* resolution method do not change across the different SNR ratios, but are repeated to facilitate comparison. *MSE* refers to mean squared error computed over the entire high resolution mean parameter vector. *False –* reports the mean (SE) false negative error rate when the number of discoveries was fixed at 450. One hundred replicates per parameter combination.

APPENDIX E

Chapter 2: Symmetry of the Custom Covariance Function

E.1 A Brief remark on the cross covariance between $\mu(B_h)$ and $\mu(B_s)$ in our dual resolution mapping prior.

In the main body text, we defined a custom covariance function to help map between high and standard spatial resolution images. We reproduce that covariance function here for convenience:

$$K(\mathbf{v}, \mathbf{v}') = \begin{cases} k(\mathbf{v}, \mathbf{v}') & \text{if } \mathbf{v}' \in B_h \\ w^{\mathsf{T}}(\mathbf{v}) k(B_h, \mathbf{v}') & \text{otherwise,} \end{cases}$$

where $w(\cdot) \approx K(B_h, B_h)^{-1} k(B_h, \cdot)$ (equations (2.4) and (2.5) in the main text). In our application, we take $k(\cdot, \cdot)$ to be the isotropic radial basis function,

$$k(\mathbf{v}, \mathbf{v}') = \tau^2 \exp(-\psi \|\mathbf{v} - \mathbf{v}'\|_2^{\nu}), \qquad \tau^2, \psi > 0, \quad \nu \in (0, 2].$$

Remark 1. *Under our prior,* $\operatorname{cov}\{\mu(\mathbf{v}_h), \mu(\mathbf{v}_s)\} = k(\mathbf{v}_h, \mathbf{v}_s)$ *for any pair of* $\mathbf{v}_h \in B_h$ *and* $\mathbf{v}_s \in B_s$.

Proof. Notationally, it is most convenient to show this relationship when $w(\mathbf{v}) = K(B_h, B_h)^{-1} k(B_h, \mathbf{v})$ exactly, though the method is still valid given our approximation in section 2.2.2, equation (2.7). Per the definition of $K(\cdot, \cdot)$,

$$\begin{aligned} \operatorname{cov}\{\mu(\mathbf{v}_h), \mu(\mathbf{v}_s)\} &= w^{\mathsf{T}}(\mathbf{v}_h) k(B_h, \mathbf{v}_s) \\ &= k^{\mathsf{T}}(B_h, \mathbf{v}_h) K(B_h, B_h)^{-1} k(B_h, \mathbf{v}_s). \end{aligned}$$

Let $\boldsymbol{d} = [\mathbb{1}(\mathbf{v}_i = \mathbf{v}_h)]_{\mathbf{v}_i \in B_h}$. Since $K(B_h, B_h)\boldsymbol{d} = k(B_h, \mathbf{v}_h)$ by definition, it follows that,

$$\begin{aligned} k^{\mathsf{T}}(B_h, \mathbf{v}_h) K(B_h, B_h)^{-1} k(B_h, \mathbf{v}_s) &= \boldsymbol{d}^{\mathsf{T}} k(B_h, \mathbf{v}_s) \\ &= k(\mathbf{v}_h, \mathbf{v}_s). \end{aligned}$$

□

APPENDIX F

Chapter 2: Technical Details Regarding fMRI Data Acquisition

F.1 Details of fMRI data collection and preprocessing

FMRI data collection and methods have been described previously [101]. Briefly, the patients were scanned using a 3 Tesla TrioTim scanner (TQ engine, 32 channel head coil; Siemens Medical Solutions, Erlangen) using gradient-echo echo-planar imaging (GE-EPI; 3000 ms repetition time; 30 ms echo time; 0.69 ms echo spacing; GRAPPA acceleration factor 2). High resolution structural T1 weighted MPRAGE and T2 weighted FLAIR scans were also acquired to aid intraoperative neuronavigation and fMRI data preprocessing. The high and standard spatial resolution scans largely followed the same protocols, except that multi-band acceleration was used to increase the spatial resolution of high resolution acquisitions while keeping the temporal resolution the same between protocols (160 volumes were collected for each run).

FMRI time series preprocessing without spatial smoothing was performed prior to our analysis using FSL software [version 6.0.4; 84] and the FEAT tool [version 6.00; 179]. Standard resolution fMRI data were padded by 8 voxels in x and y (resulting in a $72 \times 72 \times 48$ grid), and high resolution data by 10 voxels in z (resulting in $120 \times 120 \times 72$ grid). Given standard resolution voxel sizes of $3 \times 3 \times 3.45$ mm^3 (patient 1) and $3 \times 3 \times 3.3$ mm^3 (patient 2), and high resolution voxel sizes of $1.8 \times 1.8 \times 2.3$ mm^3 (patient 1) and $1.8 \times 1.8 \times 2.2$ mm^3 (patient 2) this padding ensured that standard and high resolution data spanned the same field-of-view (FoV) within subject prior to further processing. The difference in effective resolution between the two patients resulted only from different interslice gaps (15% for patient 1 vs. 10% for patient 2; interslice gap was lowered for patient 2 because of a smaller head size). Optimal within-subject alignment of the two runs was then achieved by downsampling the volume used as the target reference for motion correction in the high resolution run and supplying this downsampled image as an alternative reference image for motion correction of the standard resolution time series. Per FSL default, we used the middle volume of the recorded frames (the 80$^{\text{th}}$ of our 160 volume time series) as the target reference.

This volume was downsampled to the gridding of the standard resolution run using FSLeyes (part of FSL) using nearest-neighbor interpolation and no additional smoothing.

Data were temporally filtered using a 0.011 Hz high pass filter to remove low frequency drifts, and marginal linear models were fit to the time series data at each voxel to create summary statistic maps of task-related activation. In this last step, task related regressors were convolved with the canonical hemodynamic response function; temporal derivatives of resulting functions were also used as covariates of no interest. Preprocessing resulted in one unsmoothed z-statistic image for each fMRI resolution that summarized task-related activation over the course of the scans. We went on to use the generated test statistic maps as outcome data in our subsequent analysis, treating the contrast images as noisy measures of true activation.

APPENDIX G

Chapter 2: Cavernomas

G.1 Cavernomas and additional details about Patient 2

Cavernomas are a specific type of arteriovenous malformation without shunting. They contain closely apposed, angiogenetically immature blood vessels, typically with intralesional bleeding residuals. Cavernomas can be treated via microsurgical removal [e.g. 12]; if left untreated, they may lead to seizures or progressive neurological deficits upon symptomatic micro- or macrohemorrhages. Our patient 2 was found to have a cavernous malformation (cavernoma) with chronic and subacute hemorrhage (Zabramski type I) in her left temporal lobe close to the transverse temporal gyrus and insular cortex.

APPENDIX H

Chapter 3: Software

H.1 Software

In this section we give a brief sketch of the command line tools we have written to implement the methods discussed in Chapter 3. Software is available for download at https://github.com/asw221/stickygpm, and should be compatible with any Unix-based system.

H.1.1 Dependencies

We require a C/C++ compiler compatible with the C++17 standard (e.g. gcc >= 8.3.0 should suffice). At the time of writing, external dependencies include:

- The boost filesystem and math libraries
- The Eigen (3) linear algebra library
- The nlopt library for non-linear optimization
- OpenMP
- zlib - (Likely already on your system)

H.1.2 Installation

We have used the cmake build system. Installation instructions assume dependencies have been preinstalled and our source code downloaded from GitHub. We first require compilation of an included NIFTI library:

```
cd /path/to/stickygpm/lib/nifti && make all
```

Then from */stickygpm/lib/nifti, run:

```
mkdir ../../build && cd ../../build
cmake .. -DCMAKE_BUILD_TYPE=Release
make
```

H.1.3 Analysis of group-level MRI data with latent subgroup detection

The clustered spatially varying coefficient regression model we propose in Chapter 3 can be fit to data stored in the NIFTI file standard with our `sgpreg` command.

Basic syntax might look like the following:

```
./sgpreg path/to/data*.nii.gz \
  -x /path/to/x.csv             # (R) Mean regression covars \
  --mask /path/to/mask.nii      # (R) Analysis mask (*.nii) \
  --covariance 1 0.08 1         # Spatial cov. func. parameters \
  --kmeans 5                    # Initial clustering K \
  --knots 1000                  # Number of GP bases \
  --lsbp-mu 0                   # LSBP intercept hyper (mean) \
  --lsbp-sigma 0.5              # LSBP intercept hyper (sd) \
  -z /path/to/z.csv             # LSBP fixed effects (*.csv) \
  -re /path/to/re.csv           # LSBP random effects (*.csv) \
  --truncate 10                 # LSBP max clusters \
  --burnin 1000                 # MCMC burnin iterations \
  --nsave 1000                  # MCMC samples to save \
  --samples                     # MCMC/Flag: output _all_ samples \
  --thin 5                      # MCMC thinning factor \
  --monitor                     # Flag: verbose messaging \
  --output /output/path/prefix  # Basename for output files \
  --subset /path/to/subs.csv    # Filename tokens: subset data*.nii \
  --seed 48109                  # URNG seed \
  --threads 6                   # Threads to use (OpenMP)
```

The assumed Gaussian parent process covariance function for this project is the three parameter radial basis function in equation (see section 3.3.2). Above, the arguments to `--covariance` reflect the marginal variance, bandwidth, and exponent parameter for that function. All of the name/value paired arguments have reasonable default values except for those marked with an '`(R)`,' which are required to be specified. Random clustering effects can be included in the model using the `-re` argument followed by a corresponding file name (with the file in *.csv format). All of the covariates in the referenced file will be associated with their own series of variance components (the $\zeta^{\dagger 2}_{k,r}$ in section 3.3.3). To include multiple random clustering effects in a generalized additive way, simply include multiple `-re re_1.csv ... -re re_n.csv` name/value pairs. Random clustering effect parameters will be ordered in the $\boldsymbol{\gamma}^{\dagger}_{k}$ in the same sequence they are input in the call to `sgpreg`. They will always appear after any fixed clustering effects, mirroring the layout of the prior in equation (3.9).

H.1.4 Estimation of mean process covariance parameters

Estimation of the spatial mean process covariance parameters $\boldsymbol{\theta}$ can be accomplished using the `covest` program. This function requires, at minimum, a list of data files stored (in the NIFTI file format), a comma delimited file containing the mean regression model design matrix, and an explicit analysis mask (also in the NIFTI format). Paths to multiple data files can be typed one after the other, or wildcard completions can be used for convenience. Spatial covariance function estimation is accomplished here by maximizing the marginal likelihood over a subset of locations as noted in section 3.3.5.

Basic syntax might look like the following:

```
./covest path/to/data*.nii.gz \
  --covariates /path/to/x.csv  # REQUIRED. Mean model covars \
  --mask /path/to/mask.nii.gz  # REQUIRED. Analysis mask \
  --huge 100                   # Large parameter upper bound \
  --seed 48109                 # URNG seed \
  --subsample 4000             # Subsample size (locations) \
  --xtol 1e-8                  # Convergence tolerance
```

Covariance parameters estimated using `covest` can then be passed to `sgpreg` using the `--covariance` flag as above. The argument to `--huge` puts an upper bound constraint on the covariance function marginal variance and bandwidth parameters. This can be useful for faster convergence of the algorithm, and more stable results.

We provide the option to toggle between three different gradient-free optimizers. The default option is to use the NEWUOA algorithm [128]; from experience BOBYQA [129] can give more stable results, but is typically a little slower.

```
  --bobyqa        # Use BOBYQA optimizer
  --cobyla        # Use COBYLA optimizer
  --newuoa        # (D) Use NEWUOA optimizer
```

Three parameter radial basis and Matérn covariance functions are available for use with `covest` and `sgpreg`. These options can be toggled using the flags below (without argument). The default option is to use a radial basis covariance.

```
  --radial-basis  # (D) Radial basis covariance
  --matern        # Matern covariance
```

APPENDIX I

Chapter 3: Site Effects in ABIDE I

I.1 Site effects in ABIDE I

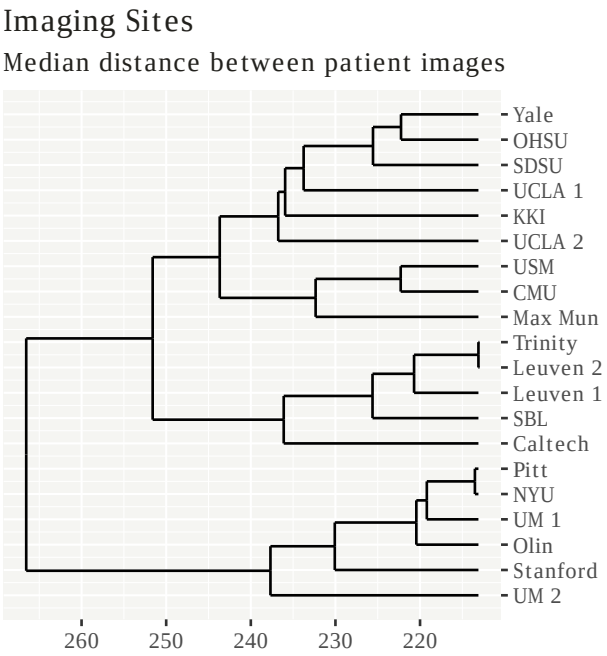


Figure I.1: Illustration of ABIDE I imaging site differences via hierarchical clustering of the median Euclidean distance between patient images across sites.

Here we give a brief illustration of how site effects can dominate the process of cluster identification. In Fig. I.1, we have used hierarchical clustering (Ward linkage) to show proximity between ABIDE I research sites. To make this figure, we computed the median Euclidean distance over all pairs of participant images between each site and clustered the sites based on the median distances. Median distances are shown on the x-axis in the figure. For reference, in the main text, the largest Euclidean distance between the posterior means for clusters derived from single site patient data was approximately 159.3 units (between clusters one and two).

APPENDIX J

Chapter 3: Model Parameter Full Conditional Distributions

In this section we present the full conditional distributions for all of our model parameters in block-wise fashion. We will need to introduce several latent data constructs—the most central of these being the cluster labels $\mathcal{C}$—in order to express the full conditional posteriors of several model parameters.

Nota bene: throughout this section, we always express the Gamma distribution using its shape-rate parameterization.

J.1 Full conditional distributions for model parameters

J.1.1 Cluster labels

Let $\boldsymbol{\beta}_k^\dagger : \mathcal{B} \to \mathbb{R}^P$ denote the regression coefficient function for cluster k so that $\boldsymbol{\beta}_k^\dagger(\cdot) = [\beta_{k,0}^\dagger(\cdot), \ldots, \beta_{k,P-1}^\dagger(\cdot)]^\mathsf{T}$. Similarly, let $\boldsymbol{\beta}_i : \mathcal{B} \to \mathbb{R}^P$ be shorthand to denote the spatially varying regression coefficient function for the ith individual. Given cluster assignment $\mathcal{C}_i = k$, each $\boldsymbol{\beta}_i(\cdot) \equiv \boldsymbol{\beta}_{\mathcal{C}_i}^\dagger(\cdot)$.

Truncating the logistic stick-breaking process at T components, the full conditional distribution of each $\mathcal{C}_i$ is categorical with,

$$\pi(\mathcal{C}_i = k \mid -) \propto \omega_k^\dagger(\boldsymbol{z}_i) \exp\left(-\frac{1}{2} \sum_{\mathbf{v} \in B} \sigma^{-2}(\mathbf{v}) \big\{ y_i(\mathbf{v}) - \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}_k^\dagger(\mathbf{v}) \big\}^2 \right).$$

This expression has an immediate "prior $\times$ likelihood" flavor. In our case, since the outcome images are high dimensional objects, the exponential "likelihood part" will usually dominate in the posterior. Proto-probabilities can be evaluated first on the log scale and then adjusted so that, when exponentiated, the largest is numerically finite and greater than zero. These values can then be rescaled to sum to one to recover $\pi(\mathcal{C}_i = k \mid -)$ for $k = 1, \ldots, T$.

J.1.2 White noise process parameters

The full conditional distribution of each $\sigma^{-2}(\mathbf{v})$ is,

$$\pi\{\sigma^{-2}(\mathbf{v}) \mid -\} \equiv \text{Gamma}\left(1 + \frac{N}{2}, \xi + \frac{1}{2}\sum_i \{y_i(\mathbf{v}) - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}_i(\mathbf{v})\}^2\right).$$

The full conditional distribution of ξ is,

$$\pi(\xi \mid -) \equiv \text{Gamma}\left(1 + M, \sum_{\mathbf{v}\in B} \sigma^{-2}(\mathbf{v})\right),$$

where M is the number of voxels in B.

J.1.3 LSBP parameters

Updating the truncated LSBP parameters $(\boldsymbol{\gamma}_k^\dagger)_{k=1}^T$, requires the conceptual inclusion of auxiliary data. The LSBP parameters are associated with the observed data only through the cluster labels $\mathcal{C}$. Working from the ground up, recall the constructive definition of the LSBP in equation (3.2). Let Υ_{ik} denote auxiliary latent data that takes the value 1 with probability $p_k(\boldsymbol{z}_i)$ and 0 with probability $1 - p_k(\boldsymbol{z}_i)$ for individual $i = 1, \ldots, N$ and component $k = 1, \ldots, T$. The $\boldsymbol{\gamma}_k^\dagger$ can then be directly thought of as logistic regression coefficients, where the $[\Upsilon_{ik}]_{i=1}^N$ act as binary outcome data. Using the Υ_{ik}, we construct,

$$\Omega_{ik} = \Upsilon_{ik} \prod_{k'<k} (1 - \Upsilon_{ik'}),$$

so that the Ω_{ik} can be thought of as binary auxiliary latent data that take the value 1 with probability $\omega_k^\dagger(\boldsymbol{z}_i)$ and 0 with probability $1 - \omega_k^\dagger(\boldsymbol{z}_i)$. Notice also that the Ω_{ik} are an equivalent way of expressing the cluster labels $\mathcal{C}_i$. If $\Omega_{ik} = 1$ and $\Omega_{ik'} = 0$ for all $k' \neq k$, then $\mathcal{C}_i$ must equal k. Similarly, reversing the construction, if $\mathcal{C}_i = k$, then it must be that $\Upsilon_{ik} = 1$ and $\Upsilon_{ik'} = 0$ for all $k' < k$. The remaining $\Upsilon_{ik'}$ can be conceptually imputed conditional on the $\boldsymbol{\gamma}_{k'}^\dagger$ for all other $k' > k$.

Given the Υ_{ik}, and the LSBP component hyperparameters, m_0, η_0^2, $\eta_k^{\dagger 2}$, $(\zeta_{k,j}^{\dagger 2})_{j\in J_f}$, and $\zeta_{k,r}^{\dagger 2}$, the LSBP coefficients $\boldsymbol{\gamma}_k^\dagger$ can be updated according to one of the latent data formulations for logistic regression models [i.e. 74, 126]. We have used the method due to Holmes and Held [74] in our current implementation. This procedure requires the inclusion of further latent data so that the conditional posterior of the $\boldsymbol{\gamma}_k^\dagger$ can be expressed as a Gaussian scale-mixture.

In brief, to use the construction given in [74] we introduce a set of Gaussian scale parameters Φ_{ik} such that the $\sqrt{\Phi_{ik}/4}$ are marginally Kolmogorov-Smirnov distributed. Let $\boldsymbol{\Phi}_k = \text{diag}(\Phi_{1k}, \ldots, \Phi_{Nk})$ be the diagonal matrix of the auxiliary Gaussian scales, and let $\boldsymbol{Z} =$

$[\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]^\mathsf{T}$ denote the $N \times Q$ matrix of clustering model terms. Now introduce latent data $\varkappa_{ik}$ into the prior hierarchy so that,

$$\Upsilon_{ik} = \begin{cases} 1 \text{ if } \varkappa_{ik} > 0 \\ 0 \text{ otherwise} \end{cases} \quad \text{with,}$$
$$\varkappa_{ik} \sim \mathcal{N}(\boldsymbol{z}_i^\mathsf{T} \boldsymbol{\gamma}_k^\dagger, \Phi_{ik}).$$

Let $\boldsymbol{\varkappa}_k$ collect the latent data into an N-vector with $\boldsymbol{\varkappa}_k = (\varkappa_{1k}, \ldots, \varkappa_{Nk})^\mathsf{T}$. Finally, let $\boldsymbol{m}_k$ represent the prior mean, and $\boldsymbol{V}_k$ represent the prior variance of $\boldsymbol{\gamma}_k^\dagger$. In this case, $\boldsymbol{V}_k$ is simply a diagonal matrix with the various η_0^2, etc., terms on the diagonal. The full conditional posterior of the $\boldsymbol{\gamma}_k^\dagger$ is then,

$$\pi(\boldsymbol{\gamma}_k^\dagger \mid -) \equiv \mathcal{N}\left((\boldsymbol{V}_k^{-1} + \boldsymbol{Z}^\mathsf{T} \boldsymbol{\Phi}_k^{-1} \boldsymbol{Z})^{-1} (\boldsymbol{V}_k^{-1} \boldsymbol{m}_k + \boldsymbol{Z}^\mathsf{T} \boldsymbol{\Phi}_k^{-1} \boldsymbol{\varkappa}_k), (\boldsymbol{V}_k^{-1} + \boldsymbol{Z}^\mathsf{T} \boldsymbol{\Phi}_k^{-1} \boldsymbol{Z})^{-1} \right).$$

The algorithm(s) given in [74] are well documented, and include pseudocode for sampling what in our notation are the Φ_{ik}. Lastly, we note that neither the Ω_{ik} nor the Υ_{ik} need to be evaluated or kept in memory for this update scheme to work. Rather, we can use the cluster labels $\mathcal{C}_i$ directly to infer the signs of the $\varkappa_{ik}$, thus skipping the step of constructing the Υ_{ik}.

J.1.4 LSBP hyperparameters

The global shrinkage parameters on the fixed clustering effects coefficients $(\eta_k^\dagger)_{k=1}^T$, can be sampled from their full conditional distributions by reexpressing the induced prior on the $\eta_k^{\dagger -2}$ as a Gamma scale mixture [following 109]. Our prior in equation (3.10) can be equivalently expressed,

$$\eta_k^{\dagger -2} \sim \text{Gamma}(1/2, \xi_{\eta,k}), \qquad \xi_{\eta,k} \sim \text{Gamma}(1/2, 1).$$

Let $Q_f = |J_f|$ denote the number of fixed clustering effect parameters in each $\boldsymbol{\gamma}_k^\dagger$. The above parameters then have full conditional distributions,

$$\pi(\eta_k^{\dagger -2} \mid -) \equiv \text{Gamma}\left(\frac{1}{2} + \frac{Q_f}{2}, \xi_{\eta,k} + \frac{1}{2} \sum_{j \in J_f} \frac{\gamma_{k,j}^{\dagger 2}}{\zeta_{k,j}^{\dagger 2}} \right),$$
$$\pi(\xi_{\eta,k} \mid -) \equiv \text{Gamma}\big(1, 1 + \eta_k^{\dagger -2}\big).$$

The prior on the local shrinkage parameters for the fixed clustering effects $(\zeta_{k,j}^{\dagger 2})_{j \in J_f}$ can be expressed similarly with auxiliary parameters $\xi_{\zeta,kj}$. Their full conditional posterior distributions can

be written,

$$\pi(\zeta_{k,j}^{\dagger -2} \mid -) \equiv \text{Gamma}\left(1, \xi_{\zeta,kj} + \frac{\gamma_{k,j}^{\dagger 2}}{2\eta_k^{\dagger 2}}\right),$$
$$\pi(\xi_{\zeta,kj} \mid -) \equiv \text{Gamma}\big(1, 1 + \zeta_{k,j}^{\dagger -2}\big).$$

Distributions for the pooled shrinkage parameters for any random clustering effects can also be written similarly. Let $Q_r = |J_r|$ denote the number of random clustering effect terms in the $\boldsymbol{\gamma}_k^\dagger$ indexed by J_r. The full conditional distribution of the corresponding $\zeta_{k,r}^{\dagger -2}$ is,

$$\pi(\zeta_{k,r}^{\dagger -2} \mid -) \equiv \text{Gamma}\left(1 + \frac{Q_r}{2}, \frac{1}{2}\sum_{j \in J_r} \gamma_{k,j}^{\dagger 2}\right).$$

J.1.5 Cluster component regression parameters

To detail the full conditional distribution of the cluster-specific spatially varying regression coefficient functions, we must again define some extra notation. It will be convenient here to work in a vector-based format. Given our specific Gaussian process covariance structure in equation (3.8), the regression coefficient functions $\beta_{k,j}^\dagger(\cdot)$ can be thought of as projections of low-rank spatial processes. For example, let

$$\tilde{\boldsymbol{\beta}}_{k,j}^\dagger \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{C}_*^{-1}\right)$$

denote an M_*-vector of random basis weights associated with knot locations $\mathcal{V}_*$ for cluster k and covariate j. The corresponding regression coefficient function evaluated at location $\mathbf{v} \in \mathcal{B}$ is equivalently, $\beta_{k,j}^\dagger(\mathbf{v}) = \tau \boldsymbol{c}_*^\mathsf{T}(\mathbf{v})\tilde{\boldsymbol{\beta}}_{k,j}^\dagger$. A little algebra and standard multivariate normal theory is enough to show this equivalence.

Let $\tilde{\boldsymbol{C}} = [\tau \boldsymbol{c}_*(\mathbf{v})]_{\mathbf{v} \in B}^\mathsf{T}$ denote the $M \times M_*$ basis matrix such that for any $\mathbf{v} \in B$ there is a corresponding row of $\tilde{\boldsymbol{C}}$—say row s—such that $\tilde{\boldsymbol{C}}_s^\mathsf{T}\tilde{\boldsymbol{\beta}}_{k,j}^\dagger = \beta_{k,j}^\dagger(\mathbf{v})$. Let us also gather the $\tilde{\boldsymbol{\beta}}_{k,j}^\dagger$ together for $j = 0, \ldots, P-1$ and write the $(PM_* \times 1)$ vector $\tilde{\boldsymbol{\beta}}_k^\dagger = (\boldsymbol{\beta}_{k,0}^{\dagger\mathsf{T}}, \ldots, \boldsymbol{\beta}_{k,P-1}^{\dagger\mathsf{T}})^\mathsf{T}$. Given that $\mathcal{C}_i = k$, the likelihood for the ith individual is,

$$\pi(\boldsymbol{y}_i \mid -) \equiv \mathcal{N}\{(\boldsymbol{x}_i^\mathsf{T} \otimes \tilde{\boldsymbol{C}})\tilde{\boldsymbol{\beta}}_k^\dagger, \boldsymbol{\Sigma}\},$$

reusing our notation $\boldsymbol{y}_i$ and $\boldsymbol{\Sigma}$ from section 3.3.5, and where $\otimes$ denotes the standard Kronecker product.

Let $\mathcal{I}_k = \{i : \mathcal{C}_i = k\}$ denote the index set for the kth cluster. The full conditional posterior

distribution of the $\tilde{\boldsymbol{\beta}}_k^\dagger$ can now be expressed as Gaussian with,

$$\mathbb{E}(\tilde{\boldsymbol{\beta}}_k^\dagger \mid -) = \left(\boldsymbol{I}_P \otimes \boldsymbol{C}_* + \sum_{i \in \mathcal{I}_k} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \otimes \tilde{\boldsymbol{C}}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{C}} \right)^{-1} \operatorname{vec} \left(\tilde{\boldsymbol{C}}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \sum_{i \in \mathcal{I}_k} \boldsymbol{y}_i \boldsymbol{x}_i^\mathsf{T} \right), \text{ and}$$

$$\operatorname{var}(\tilde{\boldsymbol{\beta}}_k^\dagger \mid -) = \left(\boldsymbol{I}_P \otimes \boldsymbol{C}_* + \sum_{i \in \mathcal{I}_k} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \otimes \tilde{\boldsymbol{C}}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{C}} \right)^{-1},$$

where $\operatorname{vec}(\boldsymbol{A})$ denotes the vectorization of matrix $\boldsymbol{A}$. For clusters k such that $\mathcal{I}_k = \emptyset$, the $\tilde{\boldsymbol{\beta}}_k^\dagger$ can be updated by sampling from the prior.

APPENDIX K

Chapter 4: Software

K.1 Software

In this section we give a brief sketch of the command line tools we have written to implement the methods discussed in Chapter 4. Software is available for download at `https://github.com/asw221/gourd`, and should be compatible with any Unix-based system.

K.1.1 Dependencies

We require a `C`/`C++` compiler compatible with the `C++17` standard (e.g. `gcc >= 8.3.0` should suffice). External dependencies have been kept to a near minimum. At the time of writing, they include:

- The Eigen (3) linear algebra library
- Expat - (Likely already on your system)
- zlib - (Likely already on your system)

Additionally, our software will make use of mathematical functions from boost if it is available.

K.1.2 Installation

We have used the `cmake` build system. With the dependencies preinstalled and our source code downloaded from GitHub, compilation can be as simple as:

```
mkdir path/to/gourd/build && cd path/to/gourd/build
cmake .. -DCMAKE_BUILD_TYPE=Release
make
```

K.1.3 Estimation of mean process covariance parameters

Estimation of the spatial mean process covariance parameters $\boldsymbol{\theta}$ can be accomplished using the `gourd_covest` program. This function requires, at minimum, a list of data files stored in the CIFTI/NIFTI-2 file format, and a GIFTI shape file. Paths to multiple data files can be typed one after the other, or wildcard completions can be used for convenience. The `gourd_covest` program estimates the mean process spatial correlation hyperparameters by maximizing a surrogate marginal likelihood over the full spatial data. We achieve this numerically by using a Vecchia approximation of the marginal likelihood. Fur additional details, please refer to section 4.2.5 and Appendix M.

Basic syntax might look like the following:

```
./gourd_covest path/to/data*.nii --surface path/to/surf.gii \
  --radius 6.0    # Vecchia approximation radius (mm) \
  --tol 1e-8      # Optimization tolerance (Default = 1e-6) \
  --radial-basis
```

Implemented covariance function options include:

```
Covariance Functions:
  --radial-basis       (Default)
  --rational-quadratic
  --matern
```

All of these are treated as three parameter covariance functions, with the parameters corresponding roughly to mean process (i) marginal variance, (ii) correlation bandwidth, and (iii) smoothness. The `--matern` option is implemented directly using modified cylindrical Bessel functions, and may be somewhat slow.

Options for the distance metric include:

```
Distance Metrics:
  --great-circle       (Default)
  --euclidean
```

K.1.4 Bayesian estimation of spatially varying coefficient (SVC) regression models

Our software contains several different programs that can be used to estimate group-level cortical surface spatially varying coefficient regression models. The `gourd_gplm` function, for example, fits our working regression model (see section 4.2.3) and is suitable for moderate to very large data sets. As above, we require input in the form of CIFTI/NIFTI-2 outcome images and a GIFTI shape file.

Basic syntax might look like the following:

```
./gourd_gplm path/to/data*.nii --surface path/to/surf.gii \
  --covariates path/to/x.csv  # Mean model design matrix \
  --radial-basis              # GP Covariance function selection \
  --theta 1 0.08 1            # GP Covariance function hyperparams \
  --neighborhood 8            # NNGP radius (for SVCs) \
  --subset path/to/subs.csv   # Filename tokens to subset data*.nii \
  -o path/to/output/prefix    # Output file location and prefix \
  --burnin 4000               # MCMC burnin iterations \
  --samples 1000              # MCMC samples to save \
  --thin 5                    # MCMC post-burnin thinning rate \
  --steps 12                  # HMC numerical integration steps \
  --neighborhood-mass 2       # HMC mass matrix radius \
  --seed 48109                # URNG seed
```

This program will output posterior mean and standard deviation images for the spatially varying regression coefficients in CIFTI format. It will also produce a posterior mean residual standard deviation image and a set of MCMC log files (tab delimited) containing saved samples of the model parameters. Rows in the log files correspond to MCMC iterations. Posterior credible bands can be constructed from the output MCMC log files using the `gourd_credband` program:

```
./gourd_credband path/to/logfile.dat \
  --surface path/to/surf.gii \
  -ref path/to/reference.dtseries.nii \
  -p 0.8 0.9 0.95
```

All arguments are required. The above syntax will read MCMC samples from '`logfile.dat`' and pair them with a reference CIFTI file and GIFTI shape file to format the output. The program will then compute 80%, 90%, and 95% simultaneous posterior credible bands for the associated spatial parameter, and output the bands in separate CIFTI-format files.

Several other executables are packaged with this software suite. The `gourd_gplmce` function fits the conditional model variant described in section 4.2.1. The `gourd_gplmme` function fits the marginal model variant described in section 4.2.2. The `gourd_vwise_glm` function fits the Bayesian equivalent of the vertex-wise general linear model we refer to throughout Chapter 4.

APPENDIX L

Chapter 4: Additional Data Results

L.1 Additional ABCD study results

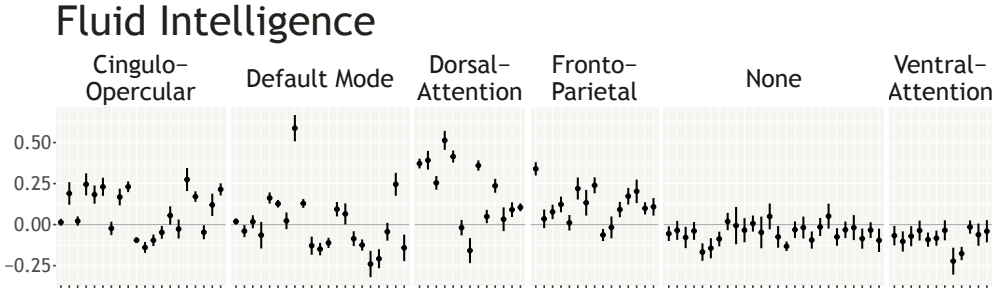


Figure L.1: Regional average coefficients: fluid intelligence, linear term. Consistent with previous studies, fluid intelligence is positively correlated with task-related activation in functionally relevant cingulo-opercular, dorsal-attention, and fronto-parietal network regions.

Here we complete our report of demographic effects on the 2- vs 0-back task contrast data from the ABCD study. We again summarize results from the right hemisphere by averaging over all vertices within brain regions from the Gordon 2016 cortical surface parcellation [63]. Figures follow the same format as the primary model intercept and 2-back accuracy rate results figures from the main text. Results in the left hemisphere were generally highly symmetric. Note that the fluid intelligence results in Fig. L.1 are consistent with evidence from experimental studies [e.g. 130, 97] regarding cingulo-opercular, dorsal-attention, and fronto-parietal network region recruitment.

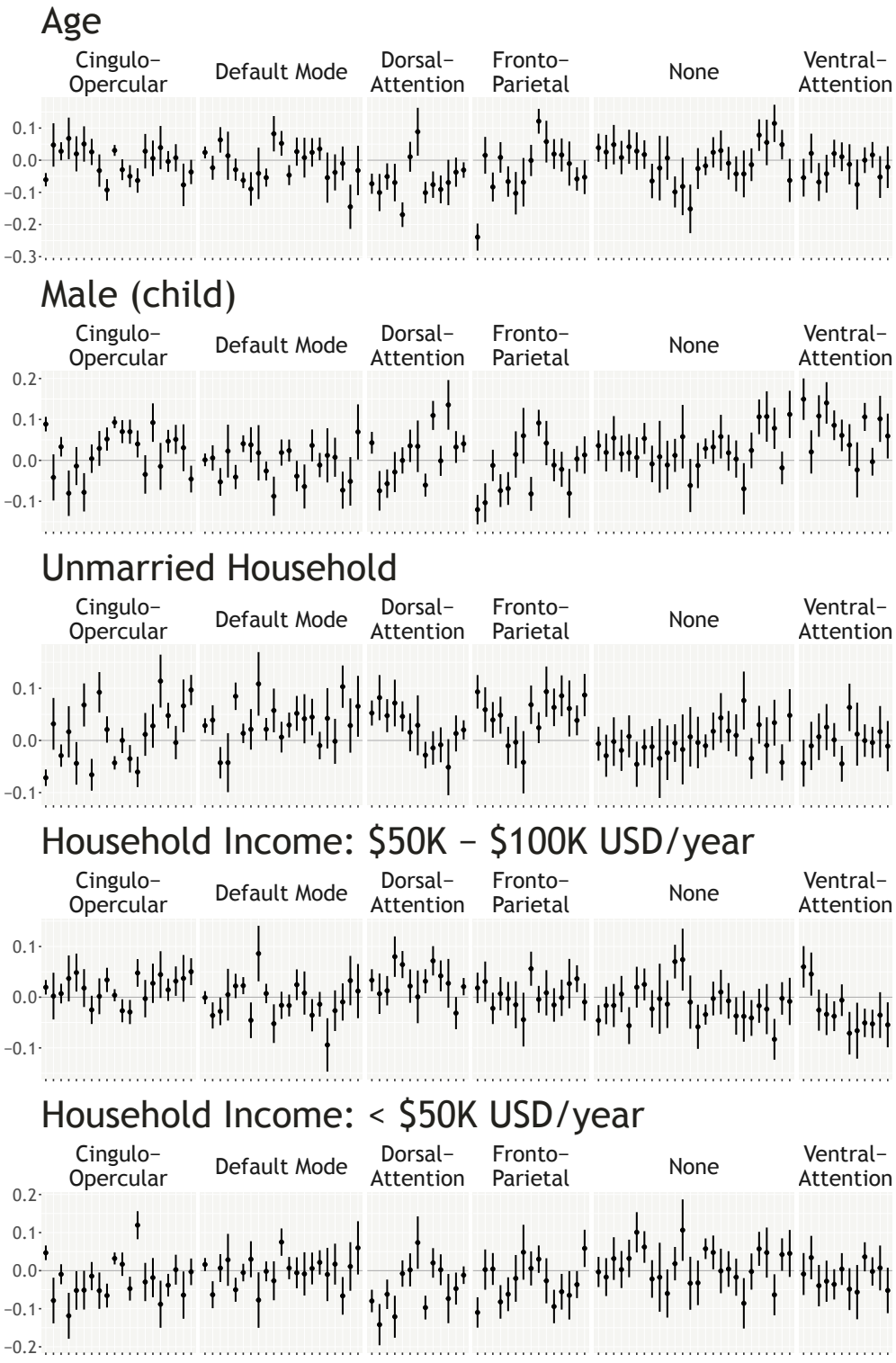


Figure L.2: Regional average coefficients: additional demographic covariates. The majority of these effects are relatively small in magnitude with the notable exception of a negative association between child age and task-related activation in a functionally relevant fronto-parietal network region (Freesurfer label: 106).

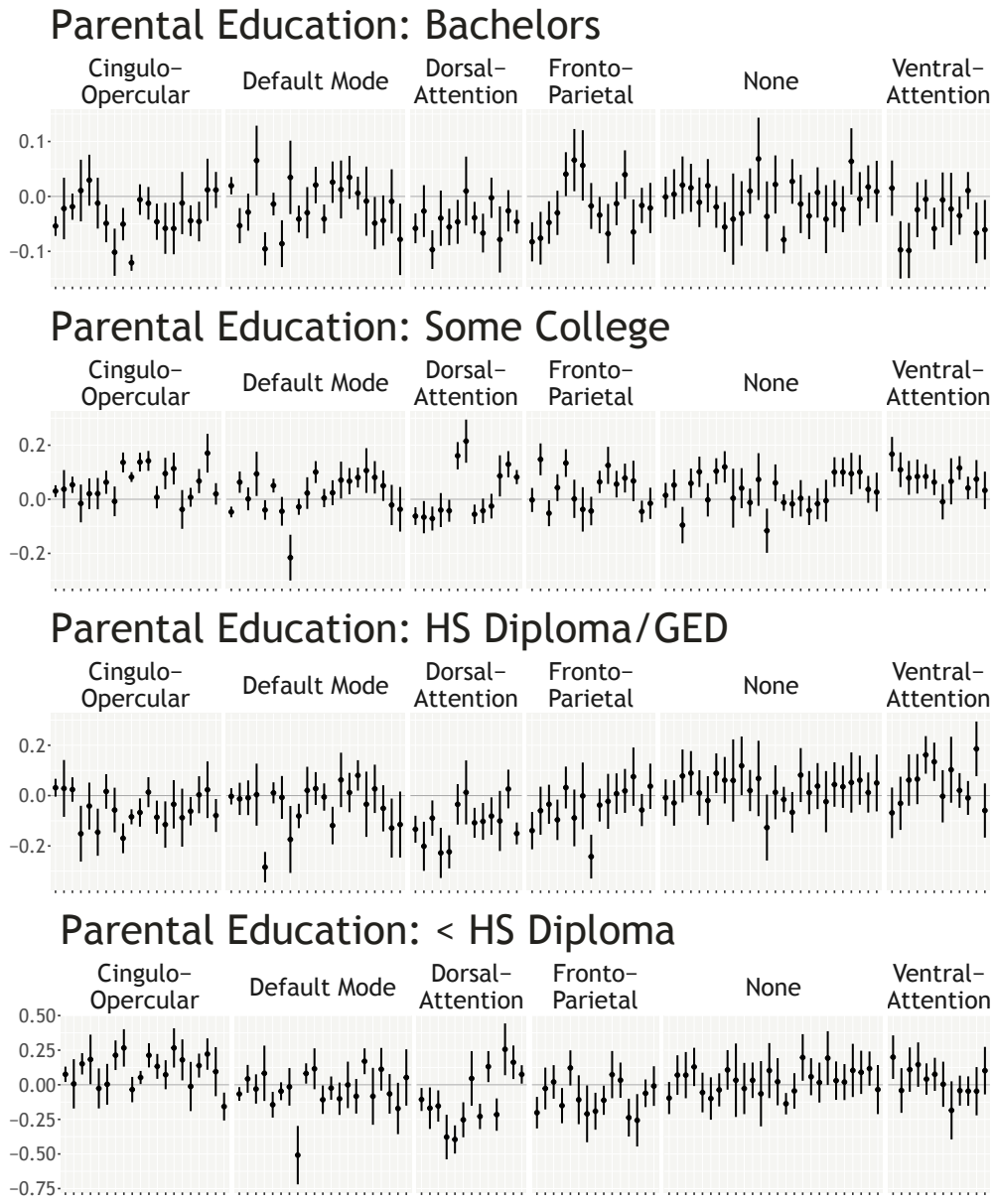


Figure L.3: Regional average coefficients: parental education (compared to a "post-graduate degree" reference group). The largest magnitude effects may suggest a pattern of decreased activation in functionally relevant dorsal-attention and fronto-parietal network regions in children of parents with less than "some college" education.

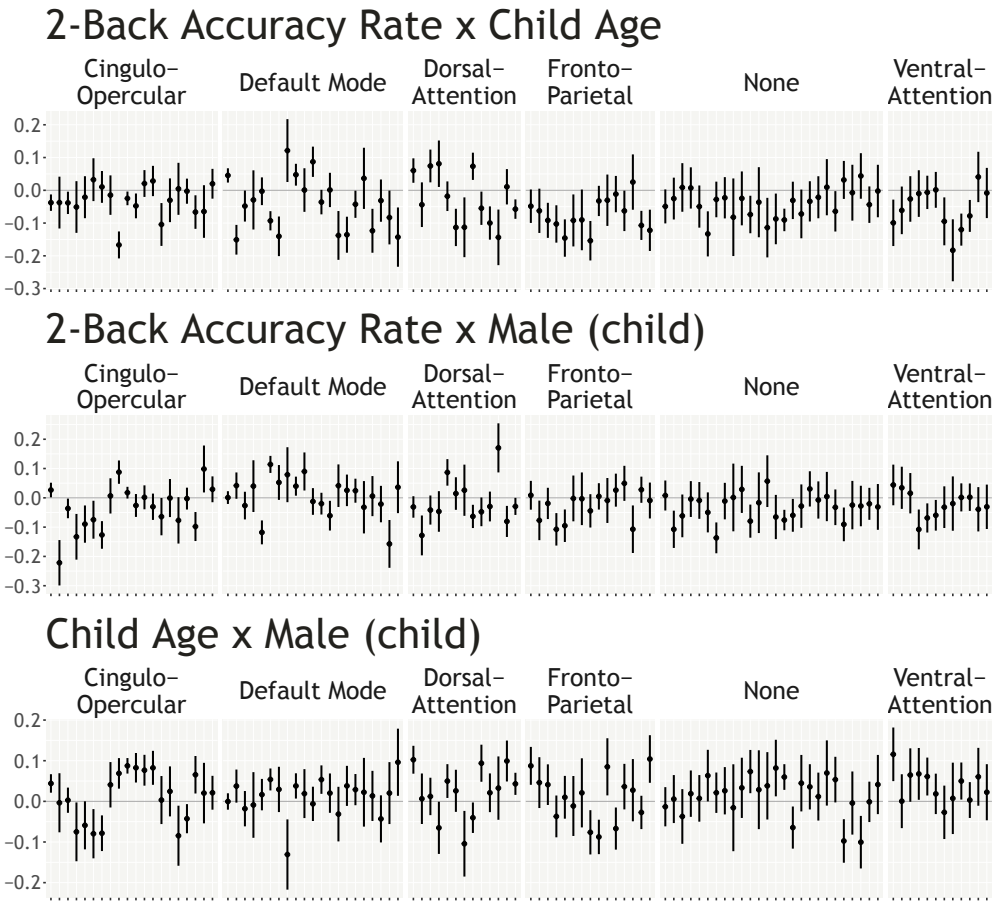


Figure L.4: Regional average coefficients: first-order interaction terms between 2-back accuracy, child age, and child sex. Most effects here are relatively small in magnitude.

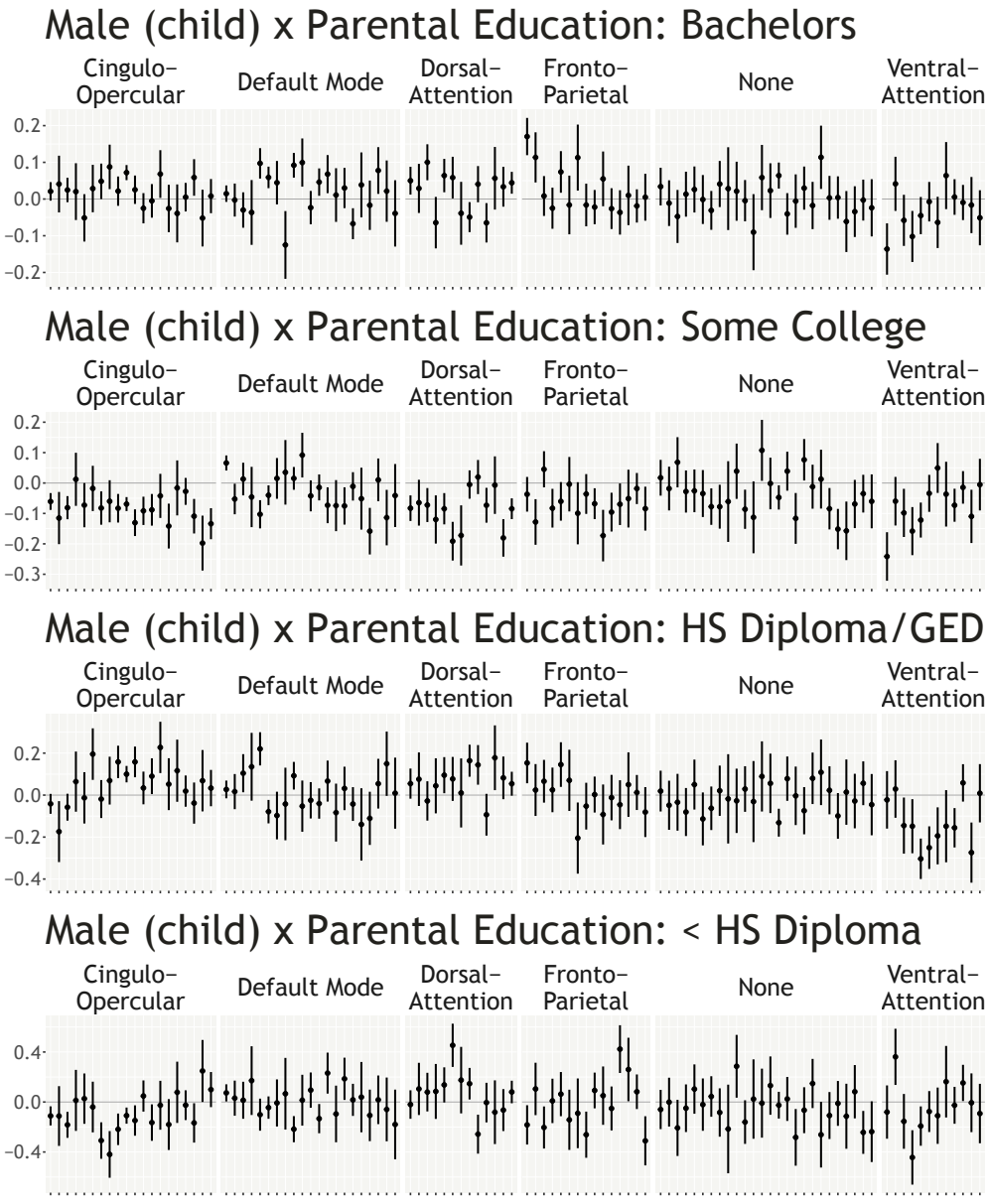


Figure L.5: Regional average coefficients: first-order interaction terms between child sex and parental education. No clear pattern of results is apparent here as with the parental education main effect terms.

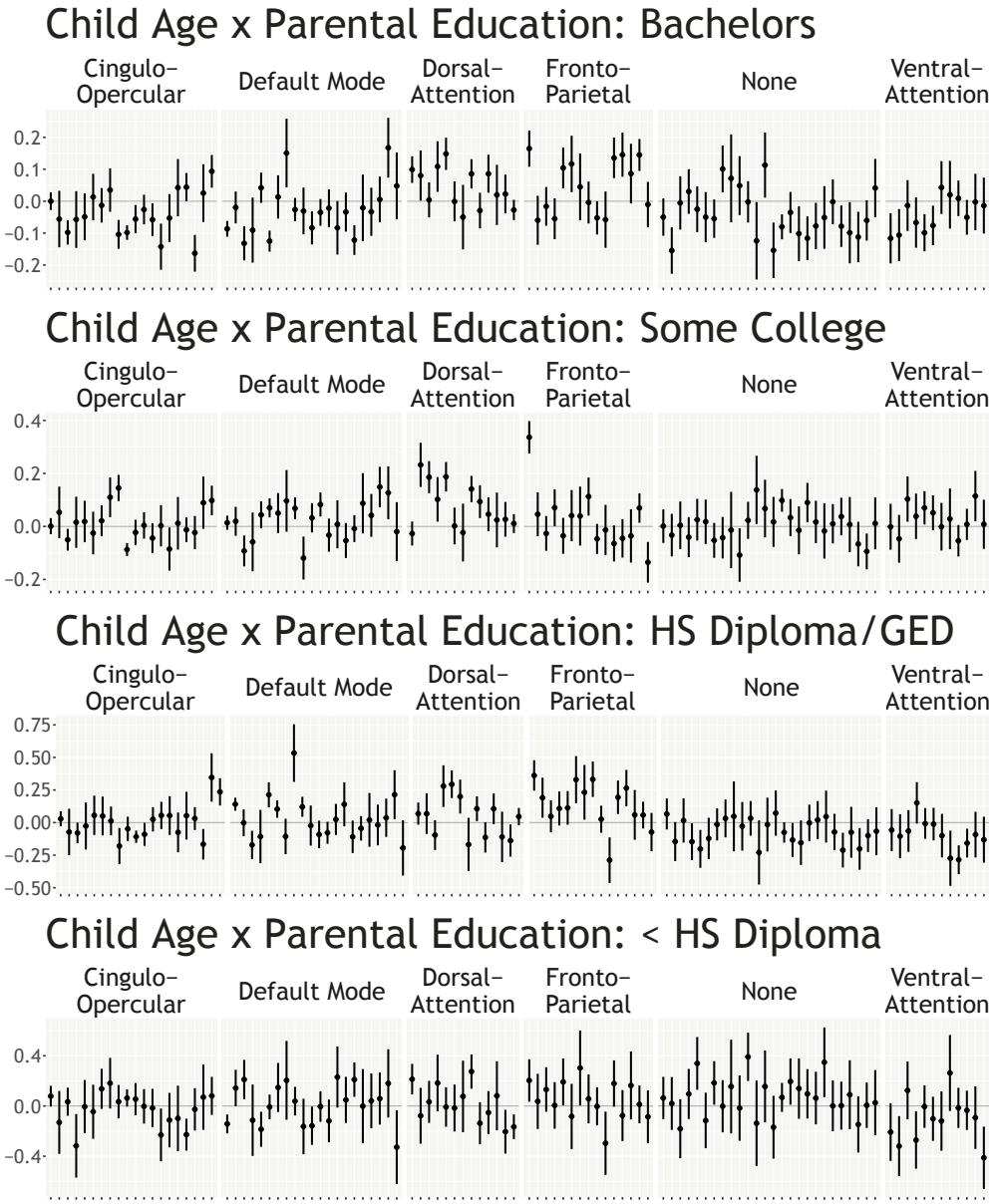


Figure L.6: Regional average coefficients: first-order interaction terms between child age and parental education. The uncertainty in many of these coefficients is relatively large, but there appears to be a consistent pattern of positive interactions in functionally relevant dorsal-attention network regions. Interpretation of this result is somewhat complicated by the general pattern of negative coefficients for the main effects of child age and parental education in these same regions.

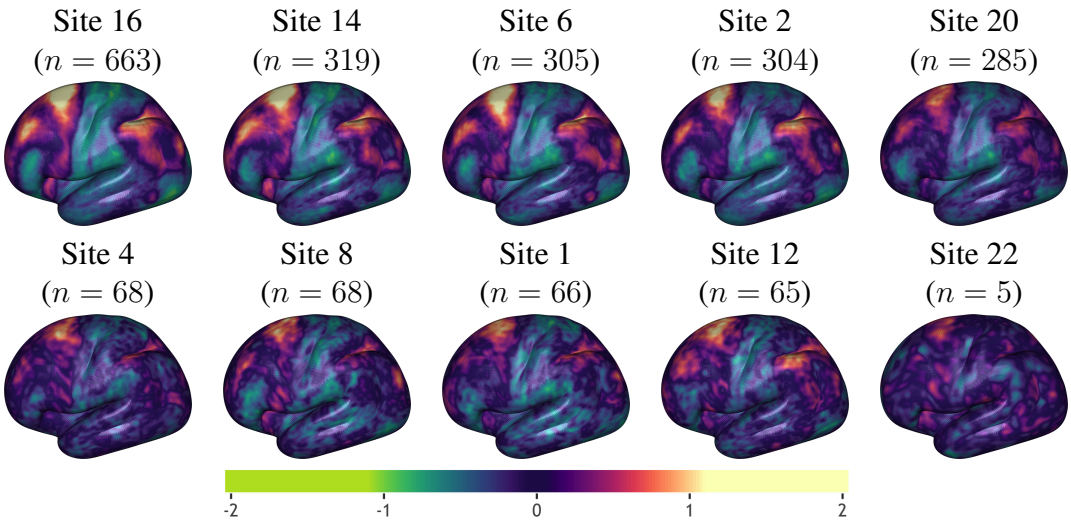


Figure L.7: Site-specific effects for the five largest and five smallest sites in our ABCD study subset. We estimated the site-specific effects as random spatial intercepts using our working model framework. Site effects appear reasonably consistent across the 21 study locations, with of course smoother results evident for the largest sites.

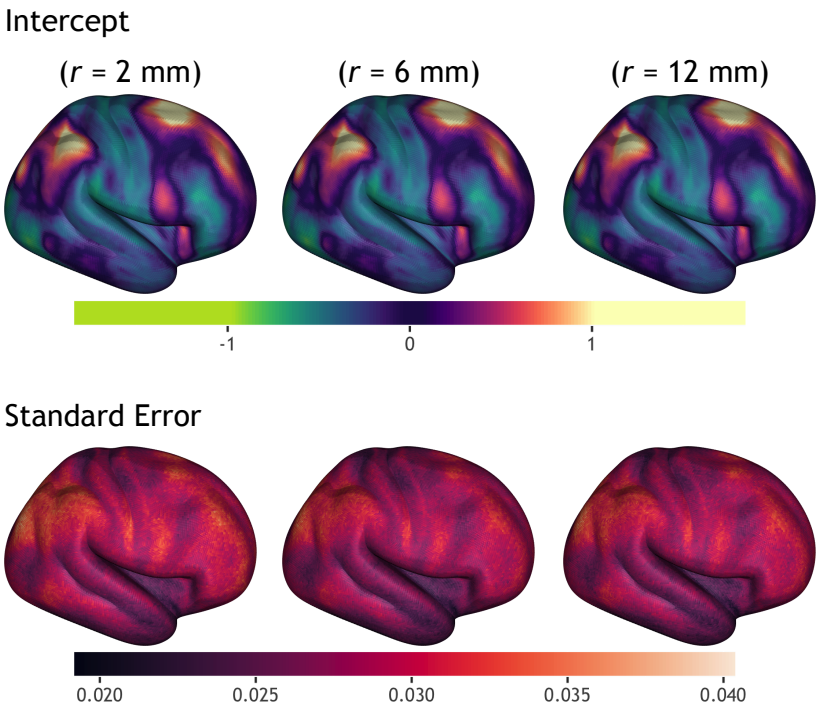


Figure L.8: Sensitivity of model estimation to varying conditional independence neighborhood radii, r. Here, we explore the sensitivity of an intercept-only model for the ABCD study data at varying r.

L.2 ABCD data analysis: MCMC diagnostics and sensitivity analyses

In this section we describe additional sensitivity analyses and MCMC diagnostics we have performed within the scope of the ABCD study data. We noted (in the main text and below in section M.1) that computationally we use a specific sparse precision matrix approximation to induce conditional independence between parameters at locations outside of an r-neighborhood of each other. A natural question in this context is how sensitive the analyses are to the choice of the neighborhood radius r. We briefly explored this question by repeatedly fitting our working model to the ABCD study data, using a spatial intercept as the only predictor, and varying r in the construction of our Vecchia approximation to the prior. Fig. L.8 summarizes the results of this sensitivity analysis. In the figure, the posterior mean estimate (top row) is not visibly sensitive to the choice of r within a 2–12 mm range. The uncertainty in the spatial intercept (bottom row), moreover, is at worst only modestly sensitive to small r.

A related question is how sensitive results are to the correlation function parameters $\boldsymbol{\theta}$. As above, we repeatedly fit our working model using a spatial intercept as the only predictor. For these analyses, we fixed our conditional independence neighborhood radius $r = 8$ mm and used radial

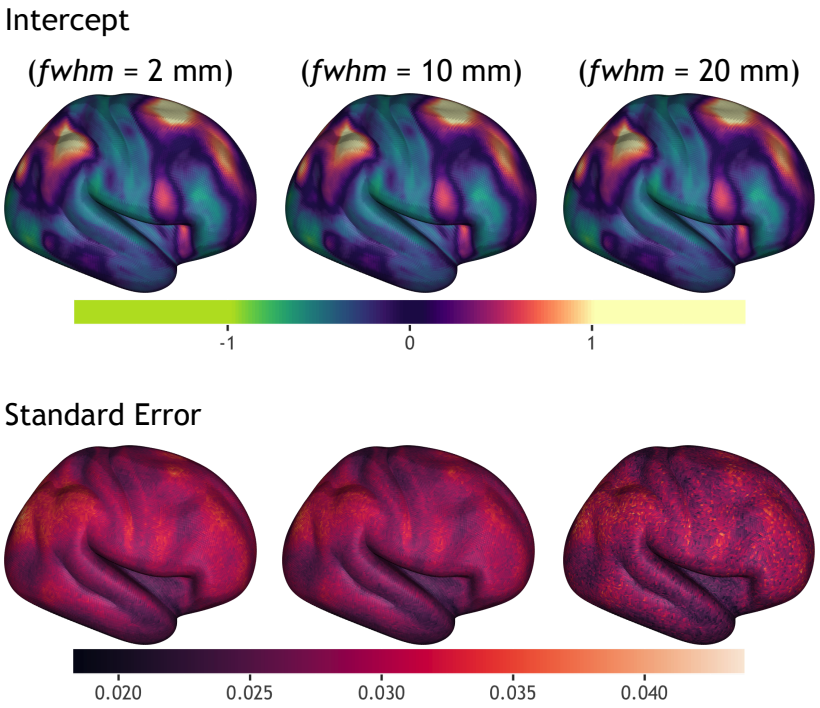


Figure L.9: Sensitivity of model estimation to varying correlation function width. We again explored the sensitivity of an intercept-only model for the ABCD study data, this time for fixed r and correlation function family. Here, we have varied the width of the correlation function to explore the effect on estimation.

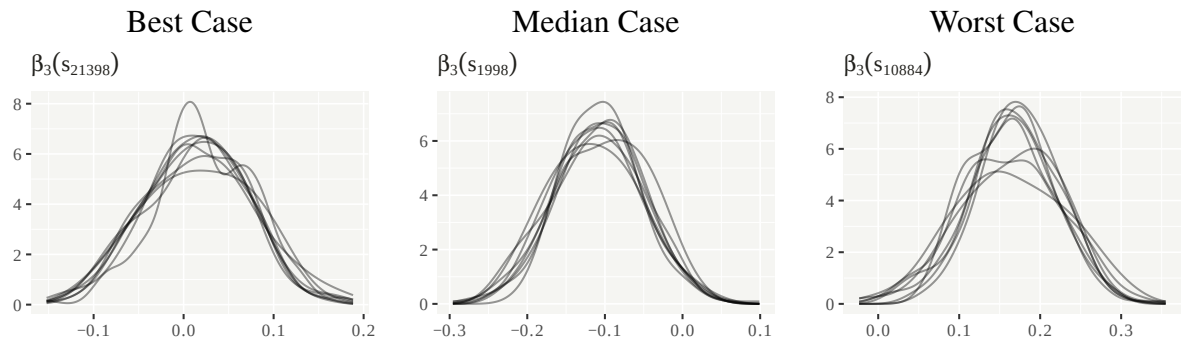


Figure L.10: Density estimates of the posterior distribution of $\beta_3(\cdot)$ for three different vertices and constructed from 8 separate HMC chains. This diagnostic is for the analysis from the main text where $\beta_3(\cdot)$ represents the spatial coefficient function for the linear 2-back accuracy rate term. Selected vertices are rank-ordered from left to right by the corresponding split folded $\hat{R}$ statistic for diagnosing MCMC convergence. The posterior densities appear to have converged reasonably well across the different chains.

basis correlation functions with exponent parameter 1.38 as in the main text. Here we varied only the width of the correlation to probe for sensitivity in the analysis. Fig. L.9 summarizes the results of this analysis across the varying correlation widths. As before, the posterior mean (top row) is not visibly sensitive to the width of the correlation within a 2–20 mm range. The uncertainty in the spatial intercept (bottom row) is again modestly sensitive to the correlation width. The estimate of the spatial standard error for the 20 mm full-width-at-half-maximum correlation appears perhaps deteriorated (bottom right panel).

We also show an example MCMC convergence diagnostic for our analysis of ABCD study data from the main text. Fig. L.10 shows representative posterior density estimates for the linear 2-back accuracy rate coefficient from three vertices, constructed from 8 HMC chains. In the figure, we have rank-ordered the selected vertices by the univariate split folded $\hat{R}$ statistic [170] for MCMC convergence (left to right, $\hat{R} = 1$ to $\hat{R} = 1.01$). The posterior densities show reasonable convergence across the MCMC chains.

Finally, we give an informal comparison of realized estimation differences arising from use of our conditional, marginal, and working model variants in practice. For this comparison, we fit our various models to the real ABCD study data following the protocol described in section 4.4.1. Figs. L.11 and L.12 summarize the results of this comparison due to both modeling and algorithmic differences between the three methods. In particular, Fig. L.11 shows how the posterior means of the $\beta_j(\mathbf{s})$ can be quite similar across our proposed methods despite differences in estimation strategy. Fig. L.12 on the other hand shows that, relative to our working model variant, marginal posterior variances of the $\beta_j(\mathbf{s})$ were systematically larger for the marginal model and smaller for the conditional model in these data. We take these differences at face value here, and note only

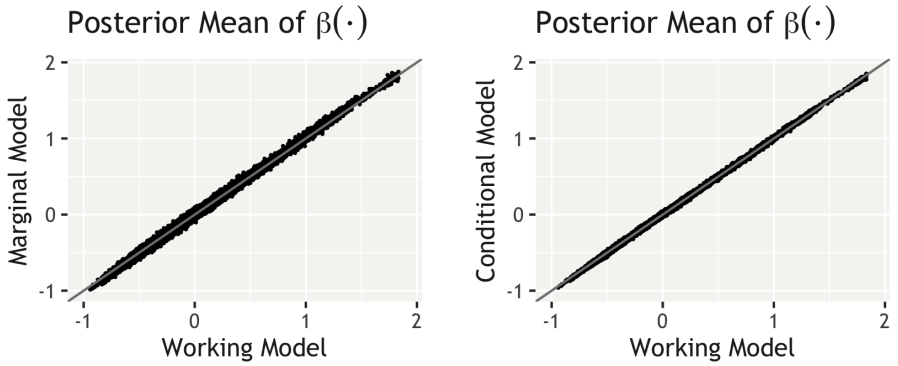


Figure L.11: Comparison of the posterior mean of $\boldsymbol{\beta}(\cdot)$ estimated from posterior samples drawn using each of our proposed conditional, marginal, and working model variants. Gray lines show identity relationships for reference.

that in our simulation studies, both the marginal and working models performed quite well when data were generated directly from the conditional model (see e.g. Table 4.1).

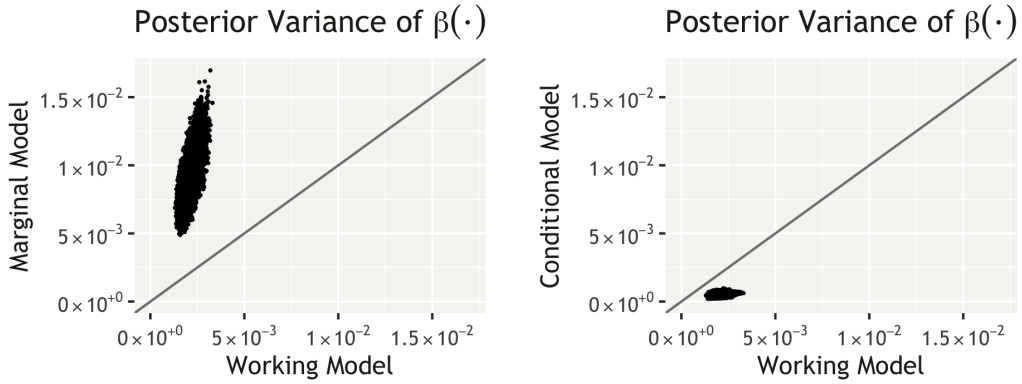


Figure L.12: Comparison of the marginal posterior variances of each $\beta_j(\mathbf{s})$, $j \in 0, \ldots, 23$ and $\mathbf{s} \in S$, estimated from posterior samples drawn using each of our proposed conditional, marginal, and working model variants. Gray lines show identity relationships for reference.

APPENDIX M

Chapter 4: Details of Posterior Computation

M.1 Posterior computation

M.1.1 Computation for our working model

We begin here with a description of our posterior computation strategy for the working model, and will then proceed by showing how this general plan can be modified to estimate the coefficients of our main conditional and marginal model variants. As in the main text, for fixed spatial domain S, let $\boldsymbol{\beta}_j$ denote the random field $[\beta_j^w(\mathbf{s})]_{\mathbf{s}\in S}$ for $j = 0, \ldots, P-1$, and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^\mathsf{T}, \ldots, \boldsymbol{\beta}_{P-1}^\mathsf{T})^\mathsf{T}$. Let $\boldsymbol{C} = [C\{d(\mathbf{s}, \mathbf{s}')\}]_{\mathbf{s},\mathbf{s}'\in S}$ represent the $(M \times M)$ spatial correlation matrix such that the prior on each $\boldsymbol{\beta}_j$ is equivalently $\mathcal{N}(\boldsymbol{\beta}_j \mid \mathbf{0}, \zeta_j^2\tau^2\boldsymbol{C})$. Similarly, let $\boldsymbol{\Sigma}$ represent the variance of $\epsilon_i^w(\cdot)$, here an $(M \times M)$ diagonal matrix with the $\sigma^2(\mathbf{s})$, $\mathbf{s} \in S$ on the diagonal; let $\boldsymbol{X}$ denote the $(N \times P)$ matrix of participant-level covariates; let $\boldsymbol{y}_i = [y_i(\mathbf{s})]_{\mathbf{s}\in S}$ denote the vectorized outcome image for participant i; and let $\boldsymbol{y} = (\boldsymbol{y}_1^\mathsf{T}, \ldots, \boldsymbol{y}_N^\mathsf{T})^\mathsf{T}$. Finally, let $\boldsymbol{Z} = \mathrm{diag}(\zeta_0^2, \ldots, \zeta_{P-1}^2)$.

To help stabilize our computational steps, we first compute a rank revealing decomposition of the covariate matrix $\boldsymbol{X}$. We will work here with the singular value decomposition (SVD) $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$, though the QR decomposition, etc. would work in the same way. In general, computing the SVD is an $\mathcal{O}(NP^2)$ operation when $P \leq N$; even for relatively large P computing the SVD of $\boldsymbol{X}$ takes a negligible amount of time compared to MCMC. For simplicity, we will assume here that $\boldsymbol{X}$ is full column rank. Let $\boldsymbol{\gamma} = (\boldsymbol{V}^\mathsf{T} \otimes \boldsymbol{I}_M)\boldsymbol{\beta}$ denote our parameter of interest, rotated by $\boldsymbol{V}$. The effective prior on $\boldsymbol{\gamma}$ is simply,

$$\boldsymbol{\gamma} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{V}^\mathsf{T}\boldsymbol{Z}\boldsymbol{V} \otimes \tau^2\boldsymbol{C}\right),$$

which, as noted in the main text, can be efficiently approximated by plugging in a sparse matrix $\tilde{\boldsymbol{C}}^{-1}$ such that $\tilde{\boldsymbol{C}} \approx \boldsymbol{C}$. Given $C(\cdot)$ we can easily construct such a $\tilde{\boldsymbol{C}}^{-1}$ following recipes from

[44]. In turn, the log prior and its gradient can be approximated via,

$$\ln \pi(\boldsymbol{\gamma} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \tau^2) \propto -\frac{1}{2}\boldsymbol{\gamma}^\mathsf{T}\big(\boldsymbol{V}^\mathsf{T}\boldsymbol{Z}^{-1}\boldsymbol{V} \otimes \tau^{-2}\tilde{\boldsymbol{C}}^{-1}\big)\boldsymbol{\gamma}, \tag{M.1}$$

and,

$$\nabla_{\boldsymbol{\gamma}} \ln \pi(\boldsymbol{\gamma} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \tau^2) = -\big(\boldsymbol{V}^\mathsf{T}\boldsymbol{Z}^{-1}\boldsymbol{V} \otimes \tau^{-2}\tilde{\boldsymbol{C}}^{-1}\big)\boldsymbol{\gamma}, \tag{M.2}$$

respectively, where Kronecker identities facilitate evaluation. Similarly, the log likelihood can be rewritten in terms of $\boldsymbol{\gamma}$. Up to the integration constant, the log likelihood of our working model can be written,

$$\ln \pi(\boldsymbol{y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}) = -\frac{1}{2}\gamma^\mathsf{T}\big(\boldsymbol{D}^2 \otimes \boldsymbol{\Sigma}^{-1}\big)\boldsymbol{\gamma} + \boldsymbol{\gamma}^\mathsf{T}\big(\boldsymbol{D}\boldsymbol{U}^\mathsf{T} \otimes \boldsymbol{\Sigma}^{-1}\big)\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^\mathsf{T}\big(\boldsymbol{I}_N \otimes \boldsymbol{\Sigma}^{-1}\big)\boldsymbol{y}. \tag{M.3}$$

From this expression, it can be seen that the part of the log likelihood that includes $\boldsymbol{\gamma}$ depends on the data only through the sufficient statistic $(\boldsymbol{U}^\mathsf{T} \otimes \boldsymbol{I}_M)\boldsymbol{y}$. This implies that, within our working model framework, gradients and Metropolis-Hastings ratios can be computed efficiently with respect to $\boldsymbol{\gamma}$. Similarly, it can be shown that the residual sum of squares depends on the data only through $(\boldsymbol{U}^\mathsf{T} \otimes \boldsymbol{I}_M)\boldsymbol{y}$ and an additional sufficient statistic, $\sum_i \boldsymbol{y}_i^{\circ 2}$, where we use $\boldsymbol{a}^{\circ b} = (a_i^b)$ to denote element-wise or Hadamard exponentiation. This additional fact suggests that $\sigma^2(\cdot)$ can be easily updated without reference to the original data. With these two pieces in hand, we write our posterior computation algorithm to alternate updating $\boldsymbol{\gamma}$ through Hamiltonian Monte Carlo (as discussed in the main text), and drawing Gibbs samples to update all of the variance parameters. Samples of $\boldsymbol{\gamma}$ can easily be rotated back into samples of $\boldsymbol{\beta}$ by applying the reverse transformation, $\boldsymbol{\beta} = (\boldsymbol{V} \otimes I_M)\boldsymbol{\gamma}$. Within each HMC iteration, we update the algorithm's mass matrix via,

$$\boldsymbol{M}(\boldsymbol{Z}, \tau^2) = \boldsymbol{V}^\mathsf{T}\boldsymbol{Z}^{-1}\boldsymbol{V} \otimes \tau^{-2}\tilde{\boldsymbol{C}}_M^{-1}, \tag{M.4}$$

where $\tilde{\boldsymbol{C}}_M^{-1}$ is a sparse matrix again constructed so that $\tilde{\boldsymbol{C}}_M \approx \boldsymbol{C}$. We discussed the logic for doing this in the main text.

M.1.2 Approximation for our "Conditional" model

Our computational strategy for the conditional method relies on the observation that the full conditional distribution of the $\boldsymbol{\omega}_i$ is relatively easy to work with. Although it is too burdensome to fully sample the $\boldsymbol{\omega}_i$ at each iteration of an MCMC routine, it takes only a modest amount of time to find a maximum a posteriori (MAP) estimate of the $\boldsymbol{\omega}_i$ given an estimate of $\boldsymbol{\beta}$. As we have shown above, gradient-based updates are efficient to compute for $\boldsymbol{\beta}$ in our working model. We first obtain an approximate MAP estimate of $\boldsymbol{\beta}$ using our working model with the restriction that

$\sigma^2(\mathbf{s}) \equiv \sigma^2$ for all locations $\mathbf{s} \in S$. An estimate of this parameter can be computed quite quickly using gradient ascent. With estimates of $\boldsymbol{\beta}$, τ^2, and $\boldsymbol{\Sigma}$ in hand, the $\boldsymbol{\omega}_i$ can be set to their conditional posterior mode analytically,

$$\boldsymbol{\omega}_i \leftarrow (\tau^{-2}\boldsymbol{C}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\{\boldsymbol{y}_i - (\boldsymbol{x}_i^\mathsf{T} \otimes \boldsymbol{I}_m)\boldsymbol{\beta}\},$$

using our sparse, conditional independence-type approximation of the matrix $(\tau^{-2}\boldsymbol{C}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}$. Maximizing with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\omega}_i$ can be iterated if necessary for convergence. Once we have a satisfactory estimate of $\boldsymbol{\omega}_i$, we can easily subtract it from $\boldsymbol{y}_i$ and switch to our working model HMC algorithm for inference on $\boldsymbol{\beta}$ if desired.

M.1.3 Approximation for our "Marginal" model

Rather than fix a point estimate of the $\boldsymbol{\omega}_i$ as above, our strategy for the marginal model will be to instead obtain a fixed estimate of the correlated error variance—$\boldsymbol{H} = \tau^2\boldsymbol{C} + \boldsymbol{\Sigma}$ in the main text—and use this estimate in our general HMC algorithm (described above). To compute with the marginal method, we first obtain an initial estimate of $\boldsymbol{\beta}$ using gradient ascent in our working model approximation as above. With this estimate in hand, we can estimate the marginal or sill variance $(\tau^2 + \sigma^2(\mathbf{s}))$ for each location $\mathbf{s}$ using the standard formula $\sum_i \{y_i(\mathbf{s}) - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}(\mathbf{s})\}^2/(N - 1)$. Then, again following [44], it is straightforward to construct a conditional independence-type approximation $\tilde{\boldsymbol{H}}^{-1}$ such that $\tilde{\boldsymbol{H}} \approx \boldsymbol{H}$, and so that $\tilde{\boldsymbol{H}}$ contains our estimates of the spatial sills on the diagonal. To work with MCMC, $\tilde{\boldsymbol{H}}$ can simply be substituted in place of $\boldsymbol{\Sigma}$ in our general HMC outline above. For computational savings, we do not update $\tilde{\boldsymbol{H}}$ over MCMC iterations when we work with the model in this way.

M.2 Estimation of $\boldsymbol{\theta}$ through maximum marginal likelihood

In general spatial kriging applications, it is common to estimate $\boldsymbol{\theta}$ by maximum marginal likelihood [e.g. 110, 135]. This can be done, for example by integrating out the mean model parameters and optimizing the resulting marginal likelihood with respect to the covariance and correlation parameters. Retaining the vector-based notation from our posterior computation sections and integrating the $\boldsymbol{\beta}_j$ and $\boldsymbol{\omega}_i$ out of equation (1) in the main text, the marginal log likelihood (less the integration constant) for our spatial regression model is,

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{Z}, \tau^2) = -\frac{1}{2}\sum_i \ln\det\boldsymbol{\Omega}_i + \boldsymbol{y}_i^\mathsf{T}\boldsymbol{\Omega}_i^{-1}\boldsymbol{y}_i, \tag{M.5}$$

where $\boldsymbol{\Omega}_i = \tau^2\big(1 + \sum_j \zeta_j^2 x_{ij}^2\big)\boldsymbol{C} + \boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}$ is the $(M \times M)$ sparse matrix with $[\sigma^2(\mathbf{s})]_{\mathbf{s}\in S}$ on the diagonal. Equation (M.5) can of course be maximized directly, but at the cost of also solving for $M + P + 1$ additional parameters in $\boldsymbol{\Sigma}$, τ^2, and the ζ_j^2. Also, from a practical point of view, it is somewhat undesirable that the marginal variance of $\boldsymbol{y}_i$ depends on $\boldsymbol{x}_i$, implying the need to re-optimize (M.5) every time a covariate is added to or removed from the model. Conceptually, it does not make much sense to imagine that the spatial correlation structure of the model mean parameters may change depending on the inclusion or exclusion of given covariates.

Instead of working with (M.5) directly, we choose to estimate $\boldsymbol{\theta}$ by optimizing the marginal log likelihood for a surrogate simpler model. To estimate $\boldsymbol{\theta}$, we replace (M.5) above with,

$$\tilde{f}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}, \tau^2) = -\frac{N}{2}\ln\det(\tau^2\boldsymbol{C} + \boldsymbol{\Sigma}) + -\frac{1}{2}\sum_i \boldsymbol{y}_i^{\mathsf{T}}(\tau^2\boldsymbol{C} + \boldsymbol{\Sigma})^{-1}\boldsymbol{y}_i, \tag{M.6}$$

which, incidentally, is the unnormalized marginal likelihood for our working model with an intercept as the only predictor. Equation (M.6) can be evaluated approximately either through use of a conditional independence-type approximation of the matrix $(\tau^2\boldsymbol{C} + \boldsymbol{\Sigma})^{-1}$, or by down-sampling the $\boldsymbol{y}_i$ to a more manageable number of spatial locations. We chose the former option in the present paper, and in practice mean-center each image $\boldsymbol{y}_i$ prior to optimization. While this approach can work well, we have noticed anecdotally that it can also tend to underestimate the width of the correlation function. Obtaining a good estimate of $\boldsymbol{\theta}$ in more complex settings—as in (M.5)—remains an open research question. We do not, however, expect inference on $\boldsymbol{\beta}(\cdot)$, for example, to be overly sensitive to the choice of $\boldsymbol{\theta}$, given a reasonable number of observations.

Finally, we have used the gradient-free optimization routine BOBYQA [129] to maximize (M.6), which, surprisingly, improved performance over gradient-based optimizers (both run time and stability). The BOBYQA algorithm works by iteratively constructing a quadratic approximation to the objective function at a set of interpolation points, which are themselves updated as a trust region is progressively estimated [129]. The algorithm may fail if, for example, (M.6) exhibits local behavior that cannot be well approximated by a quadratic function.

BIBLIOGRAPHY

[1] Santiago Aja-Fernández, Tomasz Pie, Gonzalo Vegas-Sánchez-Ferrero, et al. Spatially variant noise estimation in mri: A homomorphic approach. *Medical image analysis*, 20(1):184–197, 2015.

[2] Huda Akil, Maryann E Martone, and David C Van Essen. Challenges and opportunities in mining neuroscience data. *science*, 331(6018):708–712, 2011.

[3] Edson Amaro Jr and Gareth J Barker. Study design in fmri: basic principles. *Brain and cognition*, 60(3):220–232, 2006.

[4] Neculai Archip, Olivier Clatz, Stephen Whalen, Dan Kacher, Andriy Fedorov, Andriy Kot, Nikos Chrisochoides, Ferenc Jolesz, Alexandra Golby, Peter M Black, et al. Non-rigid alignment of pre-operative mri, fmri, and dt-mri with intra-operative mri for enhanced visualization and navigation in image-guided neurosurgery. *Neuroimage*, 35(2):609–624, 2007.

[5] Margaret Armstrong. Improving the estimation and modelling of the variogram. In *Geostatistics for natural resources characterization*, pages 1–19. Springer, 1984.

[6] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

[7] Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

[8] Deanna M Barch, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, et al. Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.

[9] JW Belliveau, DN Kennedy, RC McKinstry, BR Buchbinder, RMt Weisskoff, MS Cohen, JM Vevea, TJ Brady, and BR Rosen. Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, 254(5032):716–719, 1991.

[10] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[11] Veronica J Berrocal, Alan E Gelfand, and David M Holland. Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics*, 68(3):837–848, 2012.

[12] Helmut Bertalanffy, Ludwig Benes, Takahito Miyazawa, Olaf Alberti, Adrian M Siegel, and Ulrich Sure. Cerebral cavernomas in the adult. review of the literature and analysis of 72 surgically treated patients. *Neurosurgical review*, 25(1):1–53, 2002.

[13] Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541, 1995.

[14] David Blackwell, James B MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[15] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[16] Jerzy Bodurka, F Ye, N Petridou, Kevin Murphy, and Peter A Bandettini. Mapping the mri voxel volume in which thermal noise matches physiological noise—implications for fmri. *Neuroimage*, 34(2):542–549, 2007.

[17] F DuBois Bowman, Brian Caffo, Susan Spear Bassett, and Clinton Kilts. A bayesian hierarchical framework for spatial modeling of fmri data. *NeuroImage*, 39(1):146–156, 2008.

[18] Alexander Bowring, Fabian JE Telschow, Armin Schwartzman, and Thomas E Nichols. Confidence sets for cohens d effect size images. *NeuroImage*, 226:117477, 2021.

[19] Stefan Brodoehl, Christian Gaser, Robert Dahnke, Otto W Witte, and Carsten M Klingner. Surface-based analysis increases the specificity of cortical activation patterns and connectivity results. *Scientific reports*, 10(1):1–13, 2020.

[20] Giorgio Carpaneto and Paolo Toth. Primal-dual algrorithms for the assignment problem. *Discrete Applied Mathematics*, 18(2):137–153, 1987.

[21] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[22] BJ Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018.

[23] David F. Cechetto and Jane C. Topolovec. Cerebral cortex. In Vilayanur S Ramachandran, editor, *Encyclopedia of the Human Brain Set/VS Ramachandran/2002 Academic Press-Elsevier Science USA.*, pages 663–679. 2002 Academic Press-Elsevier Science USA, New York, 2002.

[24] Gilles Celeux, Florence Forbes, Christian P Robert, and D Mike Titterington. Deviance information criteria for missing data models. *Bayesian analysis*, 1(4):651–673, 2006.

[25] Andrew A Chen, Joanne C Beer, Nicholas J Tustison, Philip A Cook, Russell T Shinohara, Haochang Shou, and Alzheimer's Disease Neuroimaging Initiative. Mitigating site effects in covariance for machine learning in neuroimaging data. *Human brain mapping*, 43(4):1179–1195, 2022.

[26] Heng Chen, Lucina Q Uddin, Xiaonan Guo, Jia Wang, Runshi Wang, Xiaomin Wang, Xujun Duan, and Huafu Chen. Parsing brain structural heterogeneity in males with autism spectrum disorder reveals distinct clinical subtypes. *Human brain mapping*, 40(2):628–637, 2019.

[27] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

[28] Roberto Cordella, Francesco Acerbi, Morgan Broggi, Davide Vailati, Vittoria Nazzi, Marco Schiariti, Giovanni Tringali, Paolo Ferroli, Angelo Franzini, and Giovanni Broggi. Intraoperative neurophysiological monitoring of the cortico-spinal tract in image-guided mini-invasive neurosurgery. *Clinical Neurophysiology*, 124(6):1244–1254, 2013.

[29] Robert Cox, John Ashburner, Hester Breman, Kate Fissell, Christian Haselgrove, Colin Holmes, Jack Lancaster, David Rex, Stephen Smith, Jeffrey Woodward, et al. A (sort of) new image data format standard: Nifti-1: We 150. *NeuroImage*, 22, 2004.

[30] Robert W Cox. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

[31] Robert W Cox and James S Hyde. Software tools for analysis and visualization of fmri data. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 10(4-5):171–178, 1997.

[32] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7, 2013.

[33] Noel Cressie. Spatial prediction and ordinary kriging. *Mathematical geology*, 20(4):405–421, 1988.

[34] Noel Cressie and Gary Glonek. Median based covariogram estimators reduce bias. *Statistics & probability letters*, 2(5):299–304, 1984.

[35] Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016.

[36] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

[37] Peter J Diggle. Binary mosaics and the spatial pattern of heather. *Biometrics*, pages 531–539, 1981.

[38] Frank H Duffy and Heidelise Als. Autism, spectrum or clusters? an eeg coherence study. *BMC neurology*, 19(1):1–13, 2019.

[39] Joke Durnez, Beatrijs Moerkerke, Andreas Bartsch, and Thomas E Nichols. Alternative-based thresholding with application to presurgical fmri. *Cognitive, Affective, & Behavioral Neuroscience*, 13(4):703–713, 2013.

[40] Amanda K Easson, Zainab Fatima, and Anthony R McIntosh. Functional connectivity-based subtypes of individuals with and without autism spectrum disorder. *Network Neuroscience*, 3(2):344–362, 2019.

[41] SW Feldstein-Ewing and M Luciana. The adolescent brain cognitive development (abcd) consortium: Rationale, aims, and assessment strategy. *Developmental Cognitive Neuroscience*, 32:1–164, 2018.

[42] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[43] Jeffrey A Fessler and Scott D Booth. Conjugate-gradient preconditioning methods for shift-variant pet image reconstruction. *IEEE Transactions on Image Processing*, 8(5):688–699, 1999.

[44] Andrew O Finley, Abhirup Datta, Bruce D Cook, Douglas C Morton, Hans E Andersen, and Sudipto Banerjee. Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414, 2019.

[45] Andrew O Finley, Huiyan Sang, Sudipto Banerjee, and Alan E Gelfand. Improving the performance of predictive process modeling for large datasets. *Computational statistics & data analysis*, 53(8):2873–2884, 2009.

[46] Bruce Fischl, Martin I Sereno, and Anders M Dale. Cortical surface-based analysis: Ii: inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999.

[47] Bruce Fischl, Martin I Sereno, Roger BH Tootell, and Anders M Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8(4):272–284, 1999.

[48] Matteo Frigo and Steven G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. Special issue on "Program Generation, Optimization, and Platform Adaptation".

[49] Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

[50] Karl J Friston, Peter Jezzard, and Robert Turner. Analysis of functional mri time-series. *Human brain mapping*, 1(2):153–171, 1994.

[51] KJ Friston, O Josephs, E Zarahn, AP Holmes, S Rouquette, and J-B Poline. To smooth or not to smooth?: Bias and efficiency in fmri time-series analysis. *Neuroimage*, 12(2):196–208, 2000.

[52] Tianfan Fu, Luo Luo, and Zhihua Zhang. Quasi-newton hamiltonian monte carlo. In *UAI*, 2016.

[53] Montserrat Fuentes and Adrian E Raftery. Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1):36–45, 2005.

[54] H Garavan, H Bartsch, K Conway, A Decastro, RZ Goldstein, S Heeringa, T Jernigan, A Potter, W Thompson, and D Zahs. Recruiting the abcd sample: Design considerations and procedures. *Developmental cognitive neuroscience*, 32:16–22, 2018.

[55] Alan E Gelfand, Hyon-Jung Kim, CF Sirmans, and Sudipto Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003.

[56] Alan E Gelfand, Athanasios Kottas, and Steven N MacEachern. Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

[57] Alan E Gelfand, Li Zhu, and Bradley P Carlin. On the change of support problem for spatio-temporal data. *Biostatistics*, 2(1):31–45, 2001.

[58] Andrew Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, 27(15):2865–2873, 2008.

[59] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

[60] Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

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

[62] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[63] Evan M Gordon, Timothy O Laumann, Babatunde Adeyemo, Jeremy F Huckins, William M Kelley, and Steven E Petersen. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral cortex*, 26(1):288–303, 2016.

[64] Katherine Gotham, Andrew Pickles, and Catherine Lord. Trajectories of autism severity in children using standardized ados scores. *Pediatrics*, 130(5):e1278–e1284, 2012.

[65] Adrian R Groves, Michael A Chappell, and Mark W Woolrich. Combined spatial and non-spatial prior for inference on mri time-series. *Neuroimage*, 45(3):795–809, 2009.

[66] Gaël Guennebaud, Benoît Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[67] Donald J Hagler Jr, SeanN Hatton, M Daniela Cornejo, Carolina Makowski, Damien A Fair, Anthony Steven Dick, Matthew T Sutherland, BJ Casey, Deanna M Barch, Michael P Harms, et al. Image processing and analysis methods for the adolescent brain cognitive development study. *Neuroimage*, 202:116091, 2019.

[68] Sven Haller and Andreas J Bartsch. Pitfalls in fmri. *European radiology*, 19(11):2689–2706, 2009.

[69] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[70] David I Hastie, Silvia Liverani, and Sylvia Richardson. Sampling from dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and computing*, 25(5):1023–1037, 2015.

[71] Matthew J Heaton, Abhirup Datta, Andrew O Finley, Reinhard Furrer, Joseph Guinness, Rajarshi Guhaniyogi, Florian Gerber, Robert B Gramacy, Dorit Hammerling, Matthias Katzfuss, et al. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, 2019.

[72] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

[73] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[74] Chris C Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168, 2006.

[75] Seok-Jun Hong, Sofie L Valk, Adriana Di Martino, Michael P Milham, and Boris C Bernhardt. Multidimensional neuroanatomical subtyping of autism spectrum disorder. *Cerebral Cortex*, 28(10):3578–3588, 2018.

[76] Seok-Jun Hong, Joshua T Vogelstein, Alessandro Gozzi, Boris C Bernhardt, BT Thomas Yeo, Michael P Milham, and Adriana Di Martino. Toward neurosubtypes in autism. *Biological psychiatry*, 88(1):111–128, 2020.

[77] Leroy Hood and Stephen H Friend. Predictive, personalized, preventive, participatory (p4) cancer medicine. *Nature reviews Clinical oncology*, 8(3):184–187, 2011.

[78] Michal Hrdlicka, Iva Dudova, Irena Beranova, Jiri Lisy, Tomas Belsan, Jiri Neuwirth, Vladimir Komarek, Ludvika Faladova, Marketa Havlovicova, Zdenek Sedlacek, et al. Subtypes of autism by cluster analysis based on structural mri data. *European child & adolescent psychiatry*, 14(3):138–144, 2005.

[79] Vanessa Hus, Andrew Pickles, Edwin H Cook Jr, Susan Risi, and Catherine Lord. Using the autism diagnostic interviewrevised to increase phenotypic homogeneity in genetic studies of autism. *Biological psychiatry*, 61(4):438–448, 2007.

[80] Susanne M Jaeggi, Martin Buschkuehl, Walter J Perrig, and Beat Meier. The concurrent validity of the n-back task as a working memory measure. *Memory*, 18(4):394–412, 2010.

[81] Johan Martijn Jansma, Nick F Ramsey, Richard Coppola, and René S Kahn. Specific versus nonspecific brain activity in a parametric n-back task. *Neuroimage*, 12(6):688–697, 2000.

[82] R Joanne Jao Keehn, Sangeeta Nair, Ellyn B Pueschel, Annika C Linke, Inna Fishman, and Ralph-Axel Müller. Atypical local and distal patterns of occipito-frontal functional connectivity are related to symptom severity in autism. *Cerebral Cortex*, 29(8):3319–3330, 2019.

[83] Ajay Jasra, Chris C Holmes, and David A Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005.

[84] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[85] Terry L Jernigan, Betty Casey, Duncan Clark, Ian Colrain, Anders Dale, Thomas Ernst, Raul Gonzalez, Mary Heitzeg, Krista Lisdahl, Monica Luciana, Bonnie Nagel, Elizabeth Sowell, Lindsay Squeglia, Susan Tapert, and Deborah Yurgeluntodd. Adolescent brain cognitive development study (abcd) 2.0.1 release #721, 2019.

[86] Steven G Johnson. *The NLopt nonlinear optimization package*. http://github.com/stevengj/nlopt.

[87] Edward G Jones and Alan Peters. *Cerebral Cortex: Comparative Structure and Evolution of Cerebral Cortex, Part II*, volume 8. Springer Science & Business Media, 2012.

[88] Ivana Jovčevska, Nina Kočevar, and Radovan Komel. Glioma and glioblastoma-how much do we (not) know? *Molecular and clinical oncology*, 1(6):935–941, 2013.

[89] William Kahan. Pracniques: further remarks on reducing truncation errors. *Communications of the ACM*, 8(1):40, 1965.

[90] Matthias Katzfuss and Joseph Guinness. A general framework for vecchia approximations of gaussian processes. *Statistical Science*, 36(1):124–141, 2021.

[91] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[92] Mark E Ladd, Peter Bachert, Martin Meyerspeer, Ewald Moser, Armin M Nagel, David G Norris, Sebastian Schmitter, Oliver Speck, Sina Straub, and Moritz Zaiss. Pros and cons of ultra-high-field mri/mrs for human application. *Progress in nuclear magnetic resonance spectroscopy*, 109:1–50, 2018.

[93] Nan Laird. Random effects and the linear mixed model. In *Analysis of Longitudinal and Cluster-Correlated Data*, pages 79–95. Institute of Mathematical Statistics, 2004.

[94] I Large, H Bridge, B Ahmed, S Clare, James Kolasinski, WW Lam, KL Miller, Tim Bjørn Dyrby, AJ Parker, JET Smith, et al. Individual differences in the alignment of structural and functional markers of the v5/mt complex in primates. *Cerebral Cortex*, 26(10):3928–3944, 2016.

[95] Ann Le Couteur, Gyles Haden, Donna Hammal, and Helen McConachie. Diagnosing autism spectrum disorders in pre-school children using two standardised assessment instruments: the adi-r and the ados. *Journal of autism and developmental disorders*, 38(2):362–372, 2008.

[96] Megan H Lee, Christopher D Smyser, and Joshua S Shimony. Resting-state fmri: a review of methods and clinical applications. *American Journal of neuroradiology*, 34(10):1866–1872, 2013.

[97] Guangfei Li, Yu Chen, Thang M Le, Wuyi Wang, Xiaoying Tang, and Chiang-Shan R Li. Neural correlates of individual variation in two-back working memory and the relationship with fluid intelligence. *Scientific reports*, 11(1):1–13, 2021.

[98] Martin A Lindquist, Stephan Geuter, Tor D Wager, and Brian S Caffo. Modular preprocessing pipelines can reintroduce artifacts into fmri data. *Human brain mapping*, 40(8):2358–2376, 2019.

[99] Martin A Lindquist, Ji Meng Loh, and Yu Ryan Yue. Adaptive spatial smoothing of fmri images. *Statistics and its Interface*, 3(1):3–13, 2010.

[100] Zhuqing Liu, Andreas J Bartsch, Veronica J Berrocal, and Timothy D Johnson. A mixed-effects, spatially varying coefficients model with application to multi-resolution functional magnetic resonance imaging data. *Statistical methods in medical research*, 28(4):1203–1215, 2019.

[101] Zhuqing Liu, Veronica J Berrocal, Andreas J Bartsch, and Timothy D Johnson. Pre-surgical fmri data analysis using a spatially adaptive conditionally autoregressive model. *Bayesian analysis (Online)*, 11(2):599, 2016.

[102] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *nature*, 412(6843):150–157, 2001.

[103] C Lord, M Rutter, PC DiLavore, S Risi, K Gotham, and SL Bishop. Autism diagnostic observation schedule,(ados-2) modules 1-4. *Los Angeles, California: Western Psychological Services*, 2012.

[104] Catherine Lord, Sharon Storoschuk, Michael Rutter, and Andrew Pickles. Using the adi-r to diagnose autism in preschool children. *Infant Mental Health Journal*, 14(3):234–252, 1993.

[105] M Luciana, JM Bjork, BJ Nagel, DM Barch, R Gonzalez, SJ Nixon, and MT Banich. Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (abcd) baseline neurocognition battery. *Developmental cognitive neuroscience*, 32:67–79, 2018.

[106] Falk Lüsebrink, Alessandro Sciarra, Hendrik Mattern, Renat Yakupov, and Oliver Speck. T1-weighted in vivo human whole brain mri dataset with an ultrahigh isotropic resolution of 250 μm. *Scientific data*, 4(1):1–12, 2017.

[107] Pulong Ma and Anindya Bhadra. Beyond matérn: On a class of interpretable confluent hypergeometric covariance functions. *Journal of the American Statistical Association*, (just-accepted):1–27, 2022.

[108] Steven N MacEachern. Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, volume 1, pages 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999, 1999.

[109] Enes Makalic and Daniel F Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2015.

[110] Kanti V Mardia and Roger J Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, 1984.

[111] Amanda F Mejia, Yu Yue, David Bolin, Finn Lindgren, and Martin A Lindquist. A bayesian general linear modeling approach to cortical surface fmri data analysis. *Journal of the American Statistical Association*, 115(530):501–520, 2020.

[112] Michal Mikl, Radek Mareček, Petr Hluštík, Martina Pavlicová, Aleš Drastich, Pavel Chlebus, Milan Brázdil, and Petr Krupa. Effects of spatial smoothing on fmri group inferences. *Magnetic resonance imaging*, 26(4):490–503, 2008.

[113] RL Moseley, RJF Ypma, RJ Holt, D Floris, LR Chura, MD Spencer, Simon Baron-Cohen, John Suckling, Edward Bullmore, and Mikail Rubinov. Whole-brain functional hypoconnectivity as an endophenotype of autism in adolescents. *Neuroimage: clinical*, 9:140–152, 2015.

[114] Peter Muller, Giovanni Parmigiani, and Kenneth Rice. Fdr and bayesian multiple comparisons rules. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, 2006. Working Paper 115.

[115] Jeanette A Mumford and Thomas Nichols. Simple group fmri modeling and inference. *Neuroimage*, 47(4):1469–1475, 2009.

[116] Radford M Neal. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[117] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002.

[118] Christopher Nimsky, Oliver Ganslandt, Michael Buchfelder, and Rudolf Fahlbusch. Intraoperative visualization for resection of gliomas: the role of functional neuronavigation and intraoperative 1.5 t mri. *Neurological research*, 28(5):482–487, 2006.

[119] NobelPrize.org. Nobel Prize Outreach AB 2002. The Nobel Prize in Physiology or Medicine 2003. https://www.nobelprize.org/prizes/medicine/2003/summary/, April 2002.

[120] Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

[121] Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, 25(1):46–59, 2005.

[122] Panagiotis Papastamoulis and George Iliopoulos. An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, 19(2):313–331, 2010.

[123] Santiago Pelegrina, M Teresa Lechuga, Juan A García-Madruga, M Rosa Elosúa, Pedro Macizo, Manuel Carreiras, Luis J Fuentes, and M Teresa Bajo. Normative data on the n-back task for children and young adolescents. *Frontiers in psychology*, 6:1544, 2015.

[124] William D Penny, Nelson J Trujillo-Barreto, and Karl J Friston. Bayesian fmri time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, 2005.

[125] Donald B Plewes and Walter Kucharczyk. Physics of mri: a primer. *Journal of magnetic resonance imaging*, 35(5):1038–1054, 2012.

[126] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

[127] Michael JD Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis*, pages 51–67. Springer, 1994.

[128] Michael JD Powell. The newuoa software for unconstrained optimization without derivatives. In *Large-scale nonlinear optimization*, pages 255–297. Springer, 2006.

[129] Michael JD Powell. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26, 2009.

[130] Franziska Preusse, Elke Van Der Meer, Gopikrishna Deshpande, Frank Krueger, and Isabell Wartenburger. Fluid intelligence allows flexible recruitment of the parieto-frontal network in analogical reasoning. *Frontiers in human neuroscience*, 5:22, 2011.

[131] Fernando A Quintana, Peter Müller, Alejandro Jara, and Steven N MacEachern. The dependent dirichlet process and related models. *Statistical Science*, 37(1):24–41, 2022.

[132] Carl E Rasmussen and Zoubin Ghahramani. Infinite mixtures of gaussian process experts. In *Advances in neural information processing systems*, pages 881–888, 2002.

[133] Carl Edward Rasmussen and Joaquin Quinonero-Candela. Healing the relevance vector machine through augmentation. In *Proceedings of the 22nd international conference on Machine learning*, pages 689–696, 2005.

[134] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[135] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, chapter Model Selection and Adaptation of Hyperparameters, pages 105–128. Number 5. MIT press Cambridge, MA, 2006.

[136] Lu Ren, Lan Du, Lawrence Carin, and David Dunson. Logistic stick-breaking process. *Journal of Machine Learning Research*, 12(Jan):203–239, 2011.

[137] Martin Reuter, H Diana Rosas, and Bruce Fischl. Highly accurate inverse consistent registration: a robust approach. *NeuroImage*, 53(4):1181–1196, 2010.

[138] Bruce R Rosen and Robert L Savoy. fmri at 20: has it changed the world? *Neuroimage*, 62(2):1316–1324, 2012.

[139] Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

[140] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[141] Håvard Rue, Andrea Riebler, Sigrunn H Sørbye, Janine B Illian, Daniel P Simpson, and Finn K Lindgren. Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017.

[142] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*, chapter Inference, pages 133–160. Number 12. Cambridge university press, 2003.

[143] Michael Rutter, A Le Couteur, Catherine Lord, et al. Autism diagnostic interview-revised. *Los Angeles, CA: Western Psychological Services*, 29(2003):30, 2003.

[144] Ashish Kaul Sahib, Klaus Mathiak, Michael Erb, Adham Elshahabi, Silke Klamer, Klaus Scheffler, Niels K Focke, and Thomas Ethofer. Effect of temporal resolution and serial autocorrelations in event-related functional mri. *Magnetic resonance in medicine*, 76(6):1805–1813, 2016.

[145] Suchi Saria and Anna Goldenberg. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems*, 30(4):70–75, 2015.

[146] Matthias Seeger, Christopher Williams, and Neil Lawrence. Fast forward selection to speed up sparse gaussian process regression. *Artificial Intelligence and Statistics 9*, 2003.

[147] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[148] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[149] Per Sidén, Anders Eklund, David Bolin, and Mattias Villani. Fast bayesian whole-brain fmri analysis with spatial 3d priors. *NeuroImage*, 146:211–225, 2017.

[150] Michael A Silva, Alfred P See, Walid I Essayed, Alexandra J Golby, and Yanmei Tie. Challenges and techniques for presurgical brain mapping with functional mri. *NeuroImage: Clinical*, 17:794–803, 2018.

[151] Michael A Silver and Sabine Kastner. Topographic maps in human frontal and parietal cortex. *Trends in cognitive sciences*, 13(11):488–495, 2009.

[152] Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, et al. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23:S208–S219, 2004.

[153] Stephen M Smith and Thomas E Nichols. Statistical challenges in big data human neuroimaging. *Neuron*, 97(2):263–268, 2018.

[154] Stephen M Smith, Diego Vidaurre, Christian F Beckmann, Matthew F Glasser, Mark Jenkinson, Karla L Miller, Thomas E Nichols, Emma C Robinson, Gholamreza Salimi-Khorshidi, Mark W Woolrich, et al. Functional connectomics from resting-state fmri. *Trends in cognitive sciences*, 17(12):666–682, 2013.

[155] Anne V Snow, Luc Lecavalier, and Carrie Houts. The structure of the autism diagnostic interview-revised: diagnostic and phenotypic implications. *Journal of Child Psychology and Psychiatry*, 50(6):734–742, 2009.

[156] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van der Linde. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493, 2014.

[157] Michael L Stein et al. Spatial variation of total column ozone on a global scale. *The Annals of Applied Statistics*, 1(1):191–210, 2007.

[158] ML Stein. *Interpolation of spatial data. Springer series in statistics*. Springer-Verlag New York, 1999.

[159] Christoph Stippich. *Clinical functional MRI: presurgical functional neuroimaging*. Springer, 2015.

[160] Shu-Chih Su, Brian Caffo, Elizabeth Garrett-Mayer, and Susan Spear Bassett. Modified test statistics by inter-voxel variance shrinkage with an application to fmri. *Biostatistics*, 10(2):219–227, 2009.

[161] Kaustubh Supekar, Lucina Q Uddin, Amirah Khouzam, Jennifer Phillips, William D Gaillard, Lauren E Kenworthy, Benjamin E Yerys, Chandan J Vaidya, and Vinod Menon. Brain hyperconnectivity in children with autism and its links to social deficits. *Cell reports*, 5(3):738–747, 2013.

[162] Siyi Tang, Nanbo Sun, Dorothea L Floris, Xiuming Zhang, Adriana Di Martino, and BT Thomas Yeo. Reconciling dimensional and categorical models of autism heterogeneity: a brain connectomics and behavioral study. *Biological psychiatry*, 87(12):1071–1082, 2020.

[163] Benjamin M Taylor and Peter J Diggle. Inla or mcmc? a tutorial and comparative evaluation for spatial prediction in log-gaussian cox processes. *Journal of Statistical Computation and Simulation*, 84(10):2266–2284, 2014.

[164] Ming Teng, Farouk S Nathoo, and Timothy D Johnson. Bayesian analysis of functional magnetic resonance imaging data with spatially varying auto-regressive orders. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):521–541, 2019.

[165] Bertrand Thirion. *Handbook of Neuroimaging Data Analysis*, chapter Functional neuroimaging group studies, pages 335–354. Number 12. Chapman & Hall/CRC, 2016.

[166] John Darrell Van Horn and Arthur W Toga. Multi-site neuroimaging trials. *Current opinion in neurology*, 22(4):370, 2009.

[167] Joost A Agelink van Rentergem, Marie K Deserno, and Hilde M Geurts. Validation strategies for subtypes in psychiatry: a systematic review of research on autism spectrum disorder. *Clinical psychology review*, 87:102033, 2021.

[168] Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.

[169] Aldo V Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312, 1988.

[170] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc (with discussion). *Bayesian analysis*, 16(2):667–718, 2021.

[171] Nora D Volkow, George F Koob, Robert T Croyle, Diana W Bianchi, Joshua A Gordon, Walter J Koroshetz, Eliseo J Pérez-Stable, William T Riley, Michele H Bloch, Kevin Conway, et al. The conception of the abcd study: From substance use to a broad nih collaboration. *Developmental cognitive neuroscience*, 32:4–7, 2018.

[172] Bo Wang and D Michael Titterington. Inadequacy of interval estimates corresponding to variational bayesian approximations. In *International Workshop on Artificial Intelligence and Statistics*, pages 373–380. PMLR, 2005.

[173] Guoqing Wang, John Muschelli, and Martin A Lindquist. Moderated t-tests for group-level fmri analysis. *NeuroImage*, 237:118141, 2021.

[174] Lianming Wang and David B Dunson. Fast bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011.

[175] Andrew S Whiteman, Andreas J Bartsch, Jian Kang, and Timothy D Johnson. Bayesian inference for brain activity from functional magnetic resonance imaging collected at two spatial resolutions. *Annals of Applied Statistics*, 2022. Accepted. arXiv preprint arXiv:2103.13131.

[176] Andrew S Whiteman, Jian Kang, and Timothy D Johnson. Bayesian inference for group-level cortical surface image-on-scalar-regression with gaussian process priors, 2022+. Submitted.

[177] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

[178] Andrew TA Wood and Grace Chan. Simulation of stationary gaussian processes in [0, 1] d. *Journal of computational and graphical statistics*, 3(4):409–432, 1994.

[179] Mark W Woolrich, Brian D Ripley, Michael Brady, and Stephen M Smith. Temporal autocorrelation in univariate linear modeling of fmri data. *Neuroimage*, 14(6):1370–1386, 2001.

[180] Mark William Woolrich, Mark Jenkinson, J Michael Brady, and Stephen M Smith. Fully bayesian spatio-temporal modeling of fmri data. *IEEE transactions on medical imaging*, 23(2):213–231, 2004.

[181] K Worsley. Random field theory. *Statistical parametric mapping: the analysis of functional brain images*, pages 232–245, 2011.

[182] Keith J Worsley, Chien Heng Liao, John Aston, V Petre, GH Duncan, F Morales, and Alan C Evans. A general statistical analysis for fMRI data. *NeuroImage*, 15(1):1–15, 2002.

[183] Bin Zhang. Regression clustering. In *Third IEEE International Conference on Data Mining*, pages 451–458. IEEE, 2003.

[184] Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.

[185] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.