Bayesian Mediation Analysis of Large-scale Complex Imaging Data: Method, Theory and Computation

by

Yuliang Xu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2024

Doctoral Committee:

Professor Jian Kang, Chair
Professor Timothy Johnson
Professor Long Nguyen
Professor Zhenke Wu

Yuliang Xu
yuliangx@umich.edu
ORCID iD: 0000-0002-7255-8930

© Yuliang Xu 2024

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest appreciation to my advisor, Dr.Jian Kang. Without his guidance and support, this dissertation would not have been possible. I will always remember how Dr.Kang encouraged me during several challenging moments throughout my PhD journey. I aspire to emulate the same level of patience and care for students that I have learned from Dr.Kang, integrating it into my future research and collaborations. I would like to thank Dr. Timothy Johnson for his insightful input on the statistical imaging applications, and I consider myself incredibly fortunate to have had the opportunity to collaborate with a senior professor as open-minded and humorous as Dr. Johnson. I also want to express my appreciation for Dr.Zhenke Wu for being such an expert on Bayesian Statistics and serve on my committee. Lastly, I extend my gratitude to Dr. Long Nguyen for his invaluable insights into Bayesian Non-parametric theory. It is truly an honor to have a distinguished professor of Bayesian theory like Dr.Nguyen on my committee.

Other than my committee members, I also want to thank Drs. Florian Gunsilius and Mark Rudelson. Although Dr.Gunsilius is not directly related with my dissertation research, he has opened a new research direction of optimal transport and advised me on a related paper. Dr.Gunsilius has been a great encouragement and aspiration that I always look up to. Dr.Rudelson is a professor in Mathematics and has taught me high-dimensional statistics in a graduate class. After the class, I have approached Dr.Rudelson with some difficulties I encountered during my first dissertation project, and have received valuable suggestions.

Lastly, I want to express my heartfelt gratitude to my family and friends. To my parents, thank you for unwavering belief in me. Your unconditional support has given me the freedom to pursue my own path. I also want to thank my partner for engaging in both enjoyable and occasionally wild math discussions with me. To all my friends in the Biostatistics Department and at the University of Michigan, thank you for being part of my PhD journey. I hope our paths will cross again in the future!

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

In neuroimaging studies, mediation analysis plays a crucial role in understanding the mechanisms through which certain exposures or interventions affect health outcomes. This dissertation develops a novel modeling framework for Bayesian mediation analysis tailored to large-scale and complex imaging data. The framework provides a robust theoretical basis for image mediation analysis and introduces innovative Bayesian inference methods and efficient computational tools. A rigorous theoretical analysis evaluates the method's robustness, considering the impact of unmeasured confounders.

In Chapter 2, we introduce a new spatially varying coefficient structural equation model for Bayesian image mediation analysis (BIMA). Using the potential outcome framework, we define the spatially varying mediation effects of the exposure on outcomes mediated through imaging mediators. We adopt the soft-thresholded Gaussian process (STGP) for prior specifications, which supports sparse and piece-wise smooth functions. We establish posterior consistency for the mediation effects and selection consistency for significant regions impacting the mediation. An efficient posterior computation algorithm for BIMA, scalable to large-scale data, is developed and validated through extensive simulations, showing at least 20% increase in power over existing methods. We apply BIMA to analyze behavioral and fMRI data from the Adolescent Brain Cognitive Development (ABCD) study, focusing on mediation effects of parental education on children's cognitive abilities through working memory brain activities. We identified important mediation regions such as the left Precuneus (involved in the recall of episodic memories), the left Inferior parietal gyrus (involved in sensory processing and sensorimotor integration), and the left Supplementary motor area(involved in motor sequencing).

In Chapter 3, to enhance BIMA's computational efficiency, we develop a general prior with variational inference algorithms for regression models with large-scale imaging data. We introduce a soft-thresholded conditional autoregressive (ST-CAR) prior, which is robust to pre-fixed correlation structures and facilitates active voxel selection. Applying ST-CAR to scalar-on-image and image-on-scalar regression models, we develop coordinate ascent variational inference (CAVI) and stochastic subsampling variational inference (SSVI) algorithms. Simulations demonstrate that the ST-CAR prior excels in selecting active areas with complex

correlations, and CAVI and SSVI offer superior computational performance. We implement these methods in the ABCD study. The SSVI on Image-on-scalar regression brings down the computation time from 86 hours (BIMA) to 7.3 hours.

In Chapter 4, we explore methods to reduce the impact of unobserved confounders on the causal mediation analysis of high-dimensional mediators with spatially smooth structures, such as brain imaging data. The key approach is to incorporate the latent individual effects, which influence the structured mediators, as unobserved confounders in the outcome model, thereby potentially debiasing the mediation effects. We develop BAyesian Structured Mediation analysis with Unobserved confounders (BASMU) framework, and establish its model identifiability conditions. Theoretical analysis is conducted on the asymptotic bias of the Natural Indirect Effect (NIE) and the Natural Direct Effect (NDE) when the unobserved confounders are omitted in mediation analysis. For BASMU, we propose a two-stage estimation algorithm to mitigate the impact of these unobserved confounders on estimating the mediation effect. Extensive simulations demonstrate that BASMU substantially reduces the bias in various scenarios. We apply BASMU to the analysis of fMRI data in the Adolescent Brain Cognitive Development (ABCD) study, focusing on four brain regions previously reported to exhibit meaningful mediation effects. Compared with the existing image mediation analysis method, BASMU identifies two to four times more voxels that have significant mediation effects, with the NIE increased by 41%, and the NDE decreased by 26%.

CHAPTER 1

Introduction

1.1 Background, motivating problem and datasets

Mediation analysis has played an important role in modern medical and biological research, psychological theory, and many areas in social sciences [88, 59, 55, 58]. In the causal inference framework, when the treatment (exposure) variable is fully randomized, we expect to see the causal effect of treatment on the outcome variable. But sometimes researchers are also interested in a third component, the mediator, that acts as a pathway from the treatment to the outcome. In our motivating example with Adolescent Brain Cognitive Development (ABCD) study, task-based functional Magnetic Resonance Imaging (fMRI) data are collected for children of age 9 to 10. We hypothesize that parental education level can have a positive impact on children's IQ score, but we are also interested in knowing whether parental education level can have a positive impact on children's cognitive ability development (reflected from task fMRI images) that further influences their IQ score. The cognitive ability development acts as a pathway that carries part of impact of the exposure to the outcome.

There are some challenges in studying the mediation effect of neuroimaging data. Depending on the resolution of the neuroimages, the number of voxels for one fMRI image can be over ten thousand. In the ABCD study [10], we use the 3D task-fMRI data that contains $N = 1861$ individuals, with $p = 47636$ voxels after preprocessing. The fMRI data contains spatially correlated signals based on complex brain anatomical structure. Moreover, the size of the cognitive signal on each voxel can be very small, and has relative small signal-to-noise ratio even for the active voxels. Hence we need a method that can (1) identify active mediation regions in a 3D high-dimensional mediator, (2) have theoretical guarantee for consistency when the number of sample increases to infinity, (3) provide efficient computation tools that are scalable for large-scale data set. The goal of this dissertation is to provide a solution to meet these challenges.

1.2 Neuroimaging

Neuroimaging techniques, encompassing a wide array of scanning methods, are critical tools in understanding the intricate workings of the human brain. Widely utilized techniques include EEG (Electroencephalography), MEG (Magnetoencephalography), PET (Positron Emission Tomography), and MRI (Magnetic Resonance Imaging). Each of these techniques offers unique insights into the complex neural activities and structural characteristics of the brain. EEG and MEG provide valuable data on brain activity patterns, through the measurement of electrical potentials. PET measures the radioactive tracer distribution. MRI stands out for its comprehensive insight into the brain's structure and function. MRI utilizes a potent magnetic field to align the body's protons, creating detailed 3D images of internal structures. Notably, functional MRI (fMRI) measures variations in blood oxygenation levels, which serve as indicators of brain activity.

Structural and functional brain imaging has been an important tool in clinical diagnostics and neuroscience advances over the last 40 years. Traditional neuroimaging studies usually involve a modest number of subjects (less than 50), recent years many large scale neuroimaging studies have made it possible for tens of thousand subjects [75]. For example, the Human Connectome Project (HCP) [85] has more than 1000 subjects, and the UK Biobank (UKB) [60] has collected more than 10,000 subjects, with the overall aim of over 50,000 subjects. These large-scale neuroimaging data projects have set off the big data era in brain imaging.

There are many statistical challenges in analyzing large scale fMRI data. As discussed in [50], the signal to noise ratio for fMRI data is usually very small compared to regular clinical studies, and small confounding effect can induce false associations [76]. In addition, the brain anatomical structure intrinsically requires complex spatial-temporal correlation structure. The main statistical problems for analyzing fMRI data including voxel-level analysis and network connectivity analysis [111]. The voxel level analysis aims to find active areas that are related with certain tasks or stimulus, and the common practice involves generalized linear models and post-analysis multiple comparison methods. The network analysis aims to find associated groups of ROIs that are related with certain brain functions, and the main methods involve clustering, independent Component Analysis (ICA), and other network and machine learning algorithms.

The ABCD neuroimage data we use in Chapter 2 is the 2-back contrast emotional task fMRI data [15, 4], where the participants are required to perform several rounds of tasks of identifying fearful or happy faces versus neutral faces. The 2-back contrast data refer to the contrast fMRI image of performing the task shown 2 rounds ago compared to a task just shown (0-back). This task contrast fMRI data can reflect the participants' cognitive

ability of working memory, encoding, retrieval, forgetting, recognition [10]. Because of the anatomical structure of human brain, subdividing the human cerebral cortex based on the fMRI images can give us brain parcellation by different cognitive functions [17]. This also allows us to summarize the mediation effect by brain regions.

Our method can also be applied to other types of neuroimage data. As detailed in [10], other than the emotional task contrast data, there are also Monetary Incentive Delay (MID) tasks [44] that are designed to measure the ability of reward processing and motivation control; the Stop Signal Task (SST) [52] to measure the impulse control ability.

Aside from the task-based fMRI data, resting state fMRI (RS-fMRI) data [47, 77] is another research focus in neuroimaging. The RS-fMRI is measured when the participant is at rest (without task or stimulus), and it focuses on the spontaneous low frequency fluctuations in the blood oxygen level dependent signal. RS-fMRI is often used to infer the synchronous activation between spatially-distinct regions, and identify the resting state network. This can help provide diagnostic and disease prognostic information.

1.3 Mediation Analysis

The history of causal mediation analysis can date back to [5], where they first formally proposed the mediation framework under the linear structural equation models (LSEM), and decomposes the total effect of exposure on the outcome into the direct effect and the indirect effect mediated through the pathway mediator. Similar to all other causal inference problems, the identification of mediation effect relies on a set of causal assumptions, and some of them may not be verifiable in practice. Many follow-up works used the LSEM or its extended version to test for the existence of mediation effect [59, 38, 37]. Under the single mediator LSEM framework (denote Y as the outcome, X as the exposure, and M as the mediator), suppose both the exposure and mediator are fully randomized,

$$Y = i_1 + cX + e_1, \quad Y = i_2 + c'X + bM + e_2, \quad M = i_3 + aX + e_3 \tag{1.1}$$

c is the total effect of X on Y, and c' is the direct effect in the presence of the mediator. The indirect effect can be expressed either as a difference of $c - c'$ [57] or as a product of ab [54]. These are referred to as the difference method or the product method. The product method is particularly useful when the mediator is a random process over a spatial support. In this case the coefficient a and b become functions over the support, and the functional product $a(\cdot)b(\cdot)$ can identify areas with active mediation effect, not just the scalar-valued indirect effect. [37] focused on the difference method and proposed the sequential ignorability

assumption to ensure the identification of mediation effect. Its follow-up works [83] and [104] further discussed the sensitivity analysis of a biased mediation effect when the identification assumptions are violated. In this dissertation, we focus on the last 2 equations in the LSEM (1.1) and use the product method to define the indirect effect, under the causal assumptions proposed in [87]. Many recent studies [80, 107, 51, 39] especially for high dimensional complex mediators all adopt this approach. In terms of imaging mediators, the structural equation model based on the last 2 equations (1.1) involves a scalar-on-image regression and an image-on-scalar regression. There are abundant literature on these two classical problems in imaging statistics, and we provide a more detailed review on this in the introduction section of Chapter 2. In Chapter 4, we also provide a limiting bias formula for the natural indirect effect when the no-unmeasured-confounder assumption is violated.

The challenges for imaging mediation analysis originate from the complexity of imaging data: low signal-to-noise ratio, super high dimensionality, complex correlation structure, and difficulty in false discovery control. Recent contributions to high-dimensional mediation analysis try to meet these challenges, including recent works which explored machine learning tools [61]. [109] proposed a multilevel parametric structural equation model to circumvent the no-unmeasured-confounder assumption for a specific data set. [105] extended the mediation problem to where both the outcome and the mediator are high dimensional.

With the development of Bayesian nonparametrics theory and advances in computational power and techniques, more researchers favor applying various flexible Bayesian priors on mediation problem to handle more complex data [99, 79, 80].

1.4 Bayesian Nonparametric Theory

The prior we use in Chapter 2 and 4 is based on a latent Gaussian process prior, which is one type of Bayesian nonparametric priors. Based on the Bayesian nonparametric theory [30], in Chapter 2 we provide posterior consistency proof for the spatially-varying sparse functional parameters in the outcome and mediator models, and further provide proof for the sign consistency in the functional indirect mediator.

Our theory is based on the general consistency theorem proposed in [13]. Since [74] first proposed the general consistency theorem, there have been many different extensions [6, 28]. These theorems provide general sufficient conditions in terms of the existence of proper test statistics, the prior positivity on a neighborhood defined by the Kullback–Leibler divergence. Our consistency theory is also based on verifying these sufficient conditions in [13].

The soft-thresholded Gaussian process prior in Chapter 2 and 4 rely on the nonparametric theory of Gaussian process developed in [29], [84], and [43]. In particular, [29] defined the

sieve space of the Gaussian process with smooth kernels. Based on this definition, they provided a tail bound for the prior probability outside of the sieve space, and an upper bound for the entropy number of the sieve space. These results provide the theoretical foundation on Gaussian process priors in our theory when verifying the existence of test conditions.

When constructing the test statistics, we use the same test as in [84] for the image-on-scalar regression, which is an easier problem compared to the scalar-on-image regression. Unlike separately fitted scalar-on-image (outcome model) and image-on-scalar (mediator model) regression models, in mediation analysis, the mediator model specifies the true generative process for the mediator in the scalar-on-image (outcome model). This prohibits us to make some common assumptions such as mean-zero assumptions for the mediator like in other scalar-on-image literature [43, 70]. Hence we use the chi-square test and similar conditions for the mediator as in [1], and provide a proof that the mediator process satisfy these conditions under some constraints.

1.5 Posterior computation

To make our mediation model applicable for large-scale, high resolution imaging data, we need efficient computational tools. Either the soft-thresholded Gaussian process (STGP) prior used in Chapter 2 and 4 or the soft-thresholded conditional autoregressive (ST-CAR) prior used in Chapter 3 utilizes the soft-thresholding operator on a latent spatially-correlated process. Hence the computational method not only has to provide efficient and accurate estimation for the posterior mean, but also uncertainty quantification for the selection of active mediation areas. To achieve this goal, we explore the MCMC based methods [68] and variational inference methods [9].

To make the latent Gaussian process computationally applicable in high-dimensions, we use the basis decomposition approach [94] to sample the coefficient of basis functions as independent Gaussian priors. The basis decomposition as a dimension reduction method is often used in many imaging models [24, 98]. Although this requires choosing a kernel function suitable for the smoothness of the real data, some sensitivity analysis is needed for the choice of the kernel function. For fMRI imaging applications, as aforementioned [17] we can utilize the brain region atlas and assume inter-region independence structure, which is another way to reduce the correlation matrix into smaller block matrices.

The STGP prior has a complex posterior as a mixture of truncated normal distributions. Traditional Gibbs sampler might be computationally expansive. For efficient posterior sampling, we adapt the Metropolis-adjusted Langevin algorithm (MALA) [69]. MALA is an

MCMC method using the Langevin diffusion as proposal densities. One disadvantage of MALA is that in high-dimensions, if the step size for the gradient of the log posterior remains the same on all dimensions, the posterior space cannot be sufficiently explored. Remedies for this including specifying a pre-conditioning matrix to explore different direction at different rate, and adjust the step-size according to the acceptance rate. [31] proposed the manifold MALA where the pre-conditioning matrix is determined by the Fisher Information matrix. [95] further proposed a position dependent MALA that could yield higher effective sample size than [31]. In our implementation, we use MALA with an approximated gradient since STGP posterior is not directly differentiable, and use the adaptive step size adjusted to the acceptance rate. Both [31] and [95] could potentially improve our implementation. Since the posterior of STGP is a mixture of 0 and some other continuous process, we can directly compute the posterior inclusion probability as a measure for uncertainty quantification of the active mediation areas.

Although MCMC methods allow for the sampling of the entire distribution, variational inference (VI) methods provide a more efficient counterpart at the cost of only approximating the posterior mean. There has been increasing popularity in using VI for high-dimensional posteriors such as imaging data analysis [41, 45], and various scalable extensions of VI [35, 64, 8]. In Chapter 3, we use the mean-field variational inference on ST-CAR prior, and use the posterior mixing probability as the uncertainty quantification for active region selection. We provide a more detailed literature review on different variational inference methods in the introduction of Chapter 3.

1.6 Dissertation contributions

In this dissertation, we contribute to the imaging mediation analysis in three aspects: modeling, theory and computation. For the modeling contribution, we adopt the Soft-thresholded prior in Chapter 2 and propose a Bayesian mediation analysis framework for analyzing high-dimensional imaging mediators. We further extend the framework in Chapter 2 to the case where the unmeasured confounder is allowed to correlate with both the mediator and the outcome, but not the exposure, and we provide a formula for the limiting bias of the indirect effect, and propose a joint model that allows us to estimate the unmeasured confounders. For the theoretical contribution, we provide theoretical guarantees for the sign consistency and L_1 consistency of the functional indirect effect. For the computational contribution, we provide an efficient posterior sampling algorithm based on MALA and apply it to the ABCD data in Chapter 2. We further explore the efficient posterior optimization method, variational inference, and propose a Soft-thresholded conditional autoregressive prior that

can achieve fast convergence and scalable to large-scale data set in Chapter 3.

CHAPTER 2

Bayesian Image Mediation Analysis

2.1 Introduction

Mediation analysis is an important statistical tool that decomposes the total effects of an exposure or treatment variable on an outcome variable into direct effects and indirect effects through mediator variables [56]. Mediation analysis has been widely adopted to gain insights into mechanisms of exposure-outcome effects in many research areas including epidemiology, environmental science, genomics, and neuroimaging. Recent advances in neuroimaging have presented great opportunities and challenges for mediation analysis with large-scale complex neuroimaging data. In many neuroimaging studies, it is of great interest to identify important brain image mediators that mediate the effect of an exposure variable, such as age, social economic status, medical treatment, or substance use, to an outcome variable, such as the cognitive status, disease status.

Our work is motivated by the brain image mediation analysis in the Adolescent Brain Cognitive Development (ABCD) study, the largest long-term study of brain development and child health in the United States. Our objective is to investigate how parental education levels impact the children's general cognitive ability that is mediated through brain function development measured by working memory task fMRI.

We consider voxel-level task fMRI contrast maps as the image mediators which pose several challenges for mediation analysis. First, the number of voxel-level image mediators can be up to 200,000 in a standard brain template, potentially requiring large computational resources for implementing the statistical algorithm. Second, brain image mediators exhibit complex correlation patterns such as the correlations among neighboring voxels and the correlation between brain regions with the same functions. Ignoring or inappropriately accounting for the correlation may introduce bias or lose statistical efficiency in estimating the mediation effects. Third, due to the low signal to noise ratio of brain imaging data, the voxel-level image mediators may have weak or zero effects on the outcome variable. The

standard mediation analysis approach may suffer from low power and high false positive rates when detecting active mediators.

Recent work on high dimensional mediation analysis provides different angles to tackle these challenges, with different statistical models tailored to specific application domains, such as penalized high dimensional survival analysis [53], DNA methylation markers [103, 33]. For imaging applications, [51] first extended the mediation analysis framework into functional data analysis and proposed a model based on least-squares estimation and penalized regression, without considering correlation among individual level noises in the mediators. Built upon this work, [11] proposed a method based on principle component analysis, where high dimensional correlated mediators are mapped to uncorrelated ones through orthogonal transformation. The orthogonal maps are sequentially estimated from maximizing the likelihood of the joint model on each direction of mediators separately. However, interpreting the estimated coefficients relies on untestable assumptions that mediators are randomly assigned to individuals, making the functional causal effect inseparable from the individual level noise.

Aside from the sequential mediator modeling idea, [108] proposed a marginal mediator model with correlated error term, and defined a convex Pathway Lasso penalty to penalize the product term in the indirect effect, instead of penalizing each functional coefficient. The Pathway Lasso method demonstrated strong computation efficiency and accuracy compared to the sequential mediator model, but sparsity in the functional coefficients was not considered. [106] proposed another frequentist approach to high dimensional mediation problems, where sparse principle component analysis is used to map the correlated mediators onto a space of independent mediators, and penalized regression techniques such as the elastic net are employed in the outcome model to enforce sparsity in high dimensional mediators. [61] used machine learning models to map a high dimensional imaging mediator to a single latent variable and treated this single latent variable as a mediator in the classical mediation triangle, sacrificing the interpretability of mediation effects.

Focusing on the temporal mediation effects, [107] proposed Granger mediation analysis, a novel framework for causal mediation analysis of multiple time series, inspired by an fMRI experiment. The framework combines causal mediation analysis and vector autoregressive (VAR) models to address challenges in time-series data, improving estimation bias and statistical power compared to existing approaches.

For Bayesian analysis of mediation effects, [99] presented a pioneering work in both single-level and multi-level models, demonstrating that Bayesian mediation analysis can improve estimation efficiency by incorporating prior knowledge. [78] proposed Bayesian mixture models to account for a large set of correlated mediators with application to biomarker identification. In particular, to deal with the sparsity and correlation in a high dimensional

parameter, a membership parameter was used to indicate whether the signal at a certain location is zero or not, and correlation structure is assumed for this membership parameter. In [79] and [80], different types of Bayesian mixture models were proposed with less focus on the correlation among different locations. In Section 2.5.1 we provide more details to these methods and compare them with our proposed method through simulation studies.

To the best of our knowledge, there is a lack of a Bayesian mediation analysis method for high-dimensional imaging data that can incorporate flexible spatial correlation structure, individual-level spatial noise, and sparsity in the functional coefficients. To fill this gap, we propose a new structural equation model with spatially varying coefficients and adopt the soft-thresholded Gaussian processes [43, STGP] as priors for Bayesian Image Mediation Analysis (BIMA). Under the potential outcome framework, the proposed BIMA framework consists of two spatially varying coefficient models: a scalar-on-image regression model for the joint effect from the exposure and the image mediator on the outcome (the outcome model), and an image-on-scalar regression model for the effect of the exposure on the image mediator (the mediator model). By assigning the STGP priors, we ensure large prior support for the piecewise smooth and sparse spatially varying coefficients in both models, based on which we formally define the spatially varying mediation effects under the potential outcome framework. To accommodate population heterogeneity in imaging data, we introduce spatially varying random effects for each individual in the mediator model, improving the efficiency of estimating the mediation effects. For posterior computation, we develop a modified Metropolis-adjusted Langevin algorithm (MALA) that boosts the computational efficiency via block updating and is scalable to high-dimensional imaging data analysis with many observations.

We perform rigorous theoretical analyses of BIMA. We establish the posterior consistency of all the spatially varying coefficients in the mediator and outcome models under the L_2 empirical norm, leading to the posterior consistency of the spatially varying mediation effects under the L_1 empirical norm. Different from the previous theoretical work on Bayesian scalar-on-image models [43], the image mediation analysis requires us to address the randomness of the functional mediator in the scalar-on-image outcome model while considering the mediator model as the generative model. Hence we proposed a new formulation for functional mediation where the mediator is treated as a random signed measure in the outcome model, and as a random function in the mediator model. This new formulation provides a coherent definition of the natural indirect effect with existing mediation literature while keeping the image mediator bounded in probability in the outcome model.

The rest of the article is structured as follows. In Section 2.2, we introduce the BIMA framework with definitions, models and prior specifications. In Section 2.3, we perform

the theoretical analysis of the proposed methods, where we establish model identifiability and posterior consistency of the spatially varying mediation effects. Then, we develop the posterior computation algorithm in Section 2.4 and perform extensive simulations in Section 2.5. Finally, we apply BIMA to the analysis of the fMRI and cognitive data in the Adolescent Brain Cognitive Development (ABCD) study in Section 2.6 and conclude the paper in Section 2.7.

2.2 Bayesian Image Mediation Analysis

2.2.1 General Notations

Let $\mathbb{R}^d$ denote a d-dimensional Euclidean vector space. Let $\mathcal{S} \subset \mathbb{R}^d$ be a compact support. Let $\mathrm{N}(\mu, \sigma^2)$ represent a normal distribution with mean μ and variance σ^2. Let $L^2(\mathcal{S})$ be the space of square-integrable functions supported on $\mathcal{S}$. Let $\{s_1, \ldots, s_p\}$ be a set of p fixed design points in $\mathcal{S}$. For any function $f(s)$ in $L^2(\mathcal{S})$, let $\|f\|_{q,p} = \left\{p^{-1} \sum_{j=1}^p |f(s_j)|^q\right\}^{1/q}$ be the L_q empirical norm on the fixed grid with p voxels. For any vector $\mathbf{a} = (a_1, \ldots, a_d)^\top \in \mathbb{R}^d$, let $\|\mathbf{a}\|_q = \left\{\sum_{i=1}^d |a_i|^q\right\}^{1/q}$ be the L_q vector norm. For any functions $f, g \in L^2(\mathcal{S})$, define the inner product $\langle f, g\rangle := \int_{\mathcal{S}} f(s)g(s)\lambda(\mathrm{d}s)$ where λ is the Lebesgue measure. The empirical inner product is defined as $\langle f, g\rangle_p := p^{-1} \sum_{j=1}^p f(s_j)g(s_j)$. Let $\mathcal{C}^\rho(\mathcal{S})$ be the order-ρ Hölder space on $\mathcal{S}$ for a positive integer ρ. For a set $\mathcal{B}$, $\bar{\mathcal{B}}$ is used to denote the closure of the set, and $\partial\mathcal{B}$ denotes the boundary. Let $\mathcal{GP}(\nu, \kappa)$ denote a Gaussian Process with mean function $\nu(\cdot)$ and covariance matrix $\kappa(\cdot, \cdot)$.

2.2.2 Spatially-Varying Coefficient Structural Equation Models

Suppose the data consists of n individuals. For individual $i (i = 1, \ldots, n)$, let Y_i denote the outcome variable, X_i denote the exposure variable, $\mathbf{C}_i = (C_{i,1}, \ldots, C_{i,q})^\top \in \mathbb{R}^q$ be a vector of q potential confounding variables. Suppose the imaging data are observed on a compact support $\mathcal{S}$. Let $\{\Delta s_1, \ldots, \Delta s_p\}$ are a partition of $\mathcal{S}$, i.e., $\mathcal{S} = \bigcup_{j=1}^p \Delta s_j$ and $\Delta s_j \cap \Delta s_{j'} = \emptyset$. Let s_j be the center of the voxel Δs_j for $j = 1, \ldots, p$. Let $\mathbf{M}_i = \{M_i(s_j), \ldots, M_i(s_p)\}^\top$ be a vector of observed image intensities, where $M_i(s)$ represent the image intensity function at location $s \in \mathcal{S}$.

To perform image mediation analysis, we consider spatially varying coefficient structural equation models which consist of scalar-on-image regression as the outcome model (2.1) and

image-on-scalar regression as the mediator model (2.2). For $i = 1, \ldots, n$, we assume

$$Y_i = \sum_{j=1}^{p} \beta(s_j)\mathcal{M}_i(\Delta s_j) + \gamma X_i + \boldsymbol{\xi}^\top \mathbf{C}_i + \epsilon_{Y,i}, \quad \epsilon_{Y,i} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_Y^2), \tag{2.1}$$

$$M_i(s_j) = \alpha(s_j)X_i + \boldsymbol{\zeta}^\top(s_j)\mathbf{C}_i + \eta_i(s_j) + \epsilon_{M,i}(s_j), \quad \epsilon_{M,i}(s_j) \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_M^2) \tag{2.2}$$

where $\mathcal{M}_i(\Delta s) = \int_{\Delta s} M_i(s)\lambda(\mathrm{d}s)$ is the total intensity measure over the small partition Δs and $\lambda(\cdot)$ is the Lebesgue measure. Throughout this paper, we assume that the Lebesgue measure on one partition $\lambda(\Delta s_j) = p^{-1}$ for any $j = 1, \ldots, p$.

In the outcome model (2.1), $\beta(s)$ represents the spatially-varying effects of the image mediator on the outcome variable. The scalar coefficient γ is the direct effect of X_i on Y_i. The vector coefficient $\boldsymbol{\xi} \in \mathbb{R}^q$ represents the confounding effects. The random noises $\epsilon_{Y,i}$ are independent and follow a normal distribution with mean zero and variance σ_Y^2.

In the mediator model (2.2), $\alpha(s)$ is the spatially-varying functional parameter of our interest. $\boldsymbol{\zeta}(s) = \{\zeta_1(s), \ldots, \zeta_q(s)\}^\top$ is a vector of the coefficients for the confounders; $\eta_i(s)$ is the spatially-varying individual effect that capture the individual variations unexplained by the exposure variable X_i and the observed confounders $\mathbf{C}_i$; and $\epsilon_{M,i}(s_j)$ is the spatially independent noise term across locations and subjects with constant variance σ_M^2.

2.2.3 Connection to the Wiener process

When $\mathcal{S}$ is one-dimensional, the finite summation $\sum_{j=1}^{p} \beta(s_j)\mathcal{M}_i(\Delta s_j)$ in model (2.1) is an approximation to the continuous integral $\int_{\mathcal{S}} \beta(s)\mathcal{M}_i(\mathrm{d}s)$. In fact, when $\mathcal{S} = [0, 1] \in \mathbb{R}$, the continuous version of model (2.1) and (2.2) can be represented as

$$\begin{aligned} Y_i &= \int_{\mathcal{S}} \beta(s)\mathcal{M}_i(\mathrm{d}s) + \gamma X_i + \boldsymbol{\xi}^\top \mathbf{C}_i + \epsilon_{Y,i}, \\ \mathcal{M}_i(\mathrm{d}s) &= \left\{\alpha(s)X_i + \boldsymbol{\zeta}^\top(s)\mathbf{C}_i + \eta_i(s)\right\}\lambda(\mathrm{d}s) + \sigma_M \mathrm{d}W_{i,M}(s), \end{aligned} \tag{2.3}$$

where $\epsilon_{Y,i} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_Y^2)$ and $W_{i,M}(s)$ is the Wiener process [20].

In neuroimaging applications, we can only observe $M_i(s)$ on fixed grids $\{j = 1, \ldots, p\}$, without loss of generality, we can approximate the values of $M_i(s), \alpha(s), \boldsymbol{\zeta}(s)$ and $\eta_i(s)$ within each Δs_j by the functional values at its center s_j. Therefore the model (2.3) can be approximated by

$$\mathcal{M}_i(\Delta s_j) = \left\{\alpha(s_j)X_i + \boldsymbol{\zeta}^\top(s_j)\mathbf{C}_i + \eta_i(s_j)\right\}\lambda(\Delta s_j) + \varepsilon_{M,i}(\Delta s_j), \tag{2.4}$$

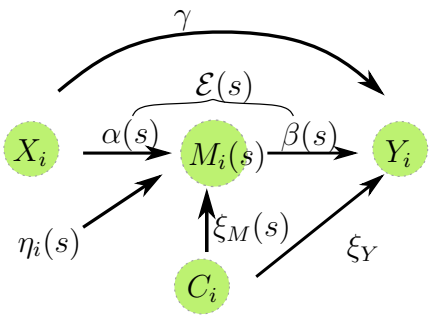


Figure 2.1: Graphical illustration of the structure of the proposed model

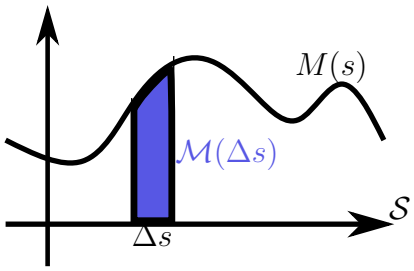


Figure 2.2: Illustration of the definitions of the intensity measure $\mathcal{M}(\Delta s)$ and the intensity function $M(s)$ in one-dimensional support $\mathcal{S}$.

where $\varepsilon_{M,i}(\Delta s_j) \sim \mathrm{N}\{0, \sigma_M^2 \lambda(\Delta s_j)\}$. The advantage of using $\sum_{j=1}^{p} \beta(s_j)\mathcal{M}_i(\Delta s_j)$ in the scalar-on-image model (2.1) compared to other existing formulations [43, 51] can be explained in two ways. First, the finite summation in (2.4) is a natural approximation to the inner-product on $L^2(\mathcal{S})$, hence in mediation analysis, as explained in the next section, we can naturally express the total indirect effect as $\sum_{j=1}^{p} \beta(s_j)\alpha(s_j)\lambda(\Delta s_j)$. Other formulations such as $\beta(s_j)M_i(s_j)/\sqrt{p}$ in [43] do not have this property. Second, the variance of $\varepsilon_{M,i}(\Delta s_j)$ is by design proportional to $\lambda(\Delta s_j)$ instead of $(\lambda(\Delta s_j))^2$. This plays a key role in constructing a test function when showing the posterior consistency in model (2.1), and ensures that we have enough variability in the design matrix in (2.1) to be able to estimate $\beta(s)$. In fact, the $M_i(s_j)/\sqrt{p}$ as used in [43] also has the variance proportional to $1/p$, but they assume the mean part of $M_i(s)$ to be zero for all $s \in \mathcal{S}$, so that $\beta(s_j)\mathbb{E}\{M_i(s_j)\}/\sqrt{p}$ will not explode as $p \to \infty$, but this assumption is not practical in mediation problem. [51] also uses an inner product formulation, but they only assume that all the functional parameters can be represented by finitely many basis functions, and the number of basis does not increase with n or p, whereas in our case, we study all sparse, piece-wise smooth function in $L_2(\mathcal{S})$.

2.2.4 Causal Mediation Analysis

We define the main mediation parameter of interest first.

Definition 1. *Let $\mathcal{E}(s) = \alpha(s)\beta(s)$ be the spatially-varying mediation effect (SVME) function.*

Under the causal inference framework [71], for individual i, we define $Y_{i,(x,\mathbf{m})}$ as the potential outcome variable that would have been observed when the image mediator $\mathbf{M}_i = \mathbf{m}$ and the exposure variable $X_i = x$; and define $\mathbf{M}_{i,(x)}$ as the potential image mediator when the individual i receive exposure x. When the exposure variable X_i changes from x to x',

combining equations (2.1) and (2.4), we represent the natural indirect effect (NIE) and the natural direct effect (NDE) as follows:

$$\text{NIE}(x, x') = \mathbb{E}\left[Y_{i,\{x,\mathbf{M}_{i,(x)}\}} - Y_{i,\{x,\mathbf{M}_{i,(x')}\}} \mid \mathbf{C}_i\right] = \sum_{j=1}^{p} \beta(s_j)\alpha(s_j)\lambda(\Delta s_j)(x - x'), \quad (2.5)$$

$$= \sum_{j=1}^{p} \mathcal{E}(s_j)\lambda(\Delta s_j)(x - x') \quad (2.6)$$

$$\text{NDE}(x, x') = \mathbb{E}\left[Y_{i,\{x,\mathbf{M}_{i,(x')}\}} - Y_{i,\{x',\mathbf{M}_{i,(x')}\}} \mid \mathbf{C}_i\right] = \gamma(x - x'). \quad (2.7)$$

To ensure the above definitions are valid under the causal inference framework, we make the stable unit treatment value assumption (SUTVA) [72] and the following modeling assumptions: for any i, x and $\mathbf{m}$, (1) $Y_{i,(x,\mathbf{m})} \perp X_i \mid \mathbf{C}_i$, (2) $Y_{i,(x,\mathbf{m})} \perp \mathbf{M}_i \mid \{\mathbf{C}_i, X_i\}$, (3) $\mathbf{M}_{i,(x)} \perp X_i \mid \mathbf{C}_i$, (4) $Y_{i,(x,\mathbf{m})} \perp \mathbf{M}_{i,(x')} \mid \mathbf{C}_i$. These assumptions ensure that: (i) NIE and NDE can be identified, and (ii) NIE and NDE can be estimated from observable data. See [87] for the detailed interpretation of the above assumptions.

In image mediation analysis, we are interested in which locations contribute to the NIE or the mediation effects. From (2.6), it is straightforward to see that $\mathcal{E}(s_j)$ represents the contribution of location s_j to the NIE(x, x') for any $x \neq x'$, which is the motivation of Definition 1. For any location $s \in \mathcal{S}$, $\mathcal{E}(s)$ characterizes the impact of the location s on the NIE. Both $\mathcal{E}(s)$ and $p^{-1}\sum_{j=1}^{p}\mathcal{E}(s_j)$ are the parameters of our main interest. It is generally believed that not all brain locations contribute to the mediation effects, and $\mathcal{E}(s)$ is naturally a sparse function when $\alpha(s)$ and $\beta(s)$ are both sparse.

2.2.5 Prior Specifications

To model the sparsity and the spatial smoothness in the spatially varying mediation effects $\mathcal{E}(s)$, we adopt the soft-thresholded Gaussian process (STGP) proposed in [43] for $\alpha(s)$ and $\beta(s)$, separately. For the individual effects $\eta_i(s)$ and confounding effects $\zeta_k(s)$, we assign the regular Gaussian process priors. Let $T_\nu : \mathbb{R} \mapsto \mathbb{R}$ be a soft-thresholded operator defined as $T_\nu(x) := \{x - \text{sgn}(x)\nu\}I(|x| > \nu)$ for any $\nu \geq 0$.

Definition 2 ([43]). *Let $\tilde{f}(s)$ be a Gaussian process (GP) with mean zero and the covariance kernel κ_f, denoted as $\tilde{f} \sim \mathcal{GP}(0, \kappa_f)$. For any $\nu \geq 0$, set $f(s) = T_\nu\{\tilde{f}(s)\}$. Then $f(s)$ is a STGP with covariance kernel κ_f and threshold parameter ν, denoted as $f \sim \mathcal{STGP}(\nu_f, \kappa_f)$.*

In summary, we have the following prior specifications,

$$\beta \sim \mathcal{STGP}(\nu_\beta, \sigma_\beta^2\kappa), \quad \alpha \sim \mathcal{STGP}(\nu_\alpha, \sigma_\alpha^2\kappa), \quad \zeta_k \sim \mathcal{GP}(0, \sigma_\zeta^2\kappa), \quad \eta_i \sim \mathcal{GP}(0, \sigma_\eta^2\kappa), \quad (2.8)$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, q$. As explained in Section 3.2 in [43], given a positive threshold value $\nu > 0$, STGP is flexible to fit a wide range of sparsity levels. The specific values for the thresholding parameters ν_α and ν_β in practice are chosen within a reasonable range according to the effect size of α and β.

The choice of κ, the kernel function for the latent Gaussian process, controls the smoothness of the functional parameters. For the rest of the parameters, the normal priors with mean zero are assigned for $\gamma, \boldsymbol{\xi}$, the inverse-gamma priors are assigned for the variance parameters σ_Y^2, σ_M^2, σ_β^2, σ_α^2 and σ_η^2.

2.3 Theoretical Properties

This section aims to establish the posterior consistency for spatially varying mediation effects $\mathcal{E}(s)$ under the empirical L_1 norm. To achieve this goal, we first show the posterior consistency for $\beta(s)$ in the outcome model (2.1) and $\alpha(s)$ in the mediator model (2.2), respectively. All the derivations and proofs are provided in the Supplementary Material.

2.3.1 Notations and Assumptions

To perform the theoretical analysis, we introduce additional notation. Let $\mathbf{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^n$, $\mathbf{M} = (\mathbf{M}_1, \ldots, \mathbf{M}_n)^\top \in \mathbb{R}^{n \times p}$ and $\mathbf{C} = (\mathbf{C}_1, \ldots, \mathbf{C}_n)^\top \in \mathbb{R}^{n \times q}$. Let $\alpha_0(s)$, $\beta_0(s)$, $\eta_{i,0}(s)$ and $\boldsymbol{\zeta}_0(s)$ represent the corresponding true spatially varying coefficients in the BIMA models (2.1) and (2.2) that generate the observed data $\mathbf{Y}$ and $\mathbf{M}$ given $\mathbf{X}$ and $\mathbf{C}$. Let $\mathcal{E}_0(s) = \alpha_0(s)\beta_0(s)$ represent the true spatially varying mediation effects. We assume that all those true spatially varying coefficients are square-integrable in $L^2(\mathcal{S})$. For matrix A, $\det(A)$ denotes the determinant of A, $\sigma_{\min}(A), \sigma_{\max}(A)$ denote the smallest and the largest singular value of A respectively.

Next, we define a functional space for the sparse and piecewise smooth spatially varying coefficients.

Definition 3 (Sparse functional space)**.** *Define the sparse functional space* $\Theta^{SP} = \{f(s) : s \in \mathcal{S}\}$ *as the collection of spatially-varying coefficient functions that satisfy the three conditions. a) (Continuous)* $f(s)$ *is a continuous function on* $\mathcal{S}$*; b) (Sparse) Assume there exist two disjoint nonempty open sets* $\mathcal{R}_{-1}$ *and* $\mathcal{R}_1$*, and* $\partial\mathcal{R}_{-1} \cap \partial\mathcal{R}_1 = \emptyset$ *such that* $\forall s \in \mathcal{R}_1,\ f(s) > 0$*;* $\forall s \in \mathcal{R}_{-1},\ f(s) < 0$*.* $\mathcal{R}_0 = \mathcal{S} - (\mathcal{R}_1 \cup \mathcal{R}_{-1})$*, and assume* $\mathcal{R}_0$ *has nonempty interior; and c) (Piecewise smooth) For any* $s \in \bar{\mathcal{R}}_1 \cup \bar{\mathcal{R}}_{-1}$*,* $f(s) \in \mathcal{C}^\rho(\bar{\mathcal{R}}_1 \cup \bar{\mathcal{R}}_{-1})$*,* $\rho \geq 1$*.*

This definition has been adopted for specifying the true parameter space of scalar-on-image regression, see Definition 2 in [43]. In BIMA, $\alpha(s)$ and $\beta(s)$ are assumed to be in the

sparse functional space in Definition 3, and later in the proof of Theorem 3, we will show that $\mathcal{E}(s)$ as defined in Definition 1 also belongs to the sparse functional space in Definition 3 when both $\alpha(s)$ and $\beta(s)$ are in this sparse functional space. In brain imaging application, we also consider region parcellation based on the anatomic structure of the human brain. When we allow a partition of R regions in the support $\mathcal{S}$, $\mathcal{S} = \cup_{r=1}^{R}\mathcal{S}_r$, we make the additional assumption that there is no piece-wise smooth area across different regions for α and β, which means that the nonzero areas in α and β only exist within each region $\mathcal{S}_r$, and not across regions or on the region boundaries. Hence, α and β are still continuous functions on $\mathcal{S}$.

Next, we will introduce the parameter space for each of the functional parameters in model (2.1)-(2.2).

Definition 4 (Parameter space)**.** *Let Θ_α, Θ_β, Θ_η, Θ_ζ be the parameter space for α, β, $\{\eta_i\}_{i=1}^n$, $\{\zeta_k\}_{k=1}^q$ respectively, and they are all subsets of the square-integrable space $L^2(\mathcal{S})$. Let $\{\psi_l(s)\}_{l=1}^\infty$ be a set of basis of $L^2(\mathcal{S})$, we specify the following constraints for each parameter space: (a) $\Theta_\alpha \subset \Theta^{SP}$; (b) $\Theta_\beta \subset \Theta^{SP}$, and for any $\beta \in \Theta_\beta$, define $\theta_{\beta,l} = \int_{\mathcal{S}} \beta(s)\psi_l(s)\lambda(ds)$, there exists $L_n = n^{\nu_1}$ where $\nu_1 \in (0,1)$ and $\nu_2 > 0$ such that $\sum_{l=L_n}^{\infty} \theta_{\beta,l}^2 \le L_n^{-\nu_2}$; (c) $\Theta_\eta, \Theta_\zeta \subset C^\rho(\mathcal{S})$; (d) There exists a constant $K > 0$ such that for any f, g in Θ_α, Θ_β, Θ_η, Θ_ζ and $\{\psi_l(s)\}_{l=1}^\infty$, the fixed grid approximation error $|\int_{\mathcal{S}} f(s)g(s)\lambda(ds) - p^{-1}\sum_{j=1}^p f(s_j)g(s_j)| \le Kp^{-2/d}$.*

Remark. In the case of region partition $\mathcal{S} = \cup_{r=1}^{R}\mathcal{S}_r$, we can construct the basis based on each region. Let $\{\psi_{l,r}(s)\}_{l=1}^\infty$ be the basis of $L^2(\mathcal{S}_r)$, and construct $\psi_l(s) = \sum_{r=1}^R \psi_{l,r}(s)I(s \in \mathcal{S}_r)$. The basis decomposition for $f(s) \in L^2(\mathcal{S})$ can be written as $\theta_{f,l} = \int_{\mathcal{S}} \psi_l(s)f(s)\lambda(ds) = \sum_{r=1}^R \int_{\mathcal{S}_r} \psi_l(s)f(s)\lambda(ds) = \sum_{r=1}^R \theta_{f,r,l}$. The decay rate condition in Definition 4 stays the same for $\theta_{f,l}$ because of the finite summation.

In Definition 4, (a)-(c) define the smoothness and sparse feature of the parameter space, where $\alpha(s)$, $\beta(s)$ are assumed to be piecewise-smooth, sparse and continuous functions, and the individual effect $\eta_i(s)$ and the confounding effects $\zeta_k(s)$ in model (2.2) are only required to be smooth but not necessarily sparse. Definition 4(d) sets an upper bound for the fixed grid approximation error. Assumption 1 below specifies the smoothness of the underlying Gaussian processes and the rate of p as $n \to \infty$.

Assumption 1. *Given the dimension d of $\mathcal{S}$ and a constant τ satisfying $d > 1 + 1/\tau$, $\tau \ge 1$, assume that a) (Smooth Kernel) for each s, the kernel function $\kappa(s,\cdot)$ introduced in the priors (2.8) has continuous partial derivatives up to order $2\rho+2$ for some positive integer ρ, i.e. $\kappa(s,\cdot) \in \mathcal{C}^{2\rho+2}(\mathcal{S})$, and $d + 3/(2\tau) < \rho$; b) (Dimension Limits) $p \ge O(n^{\tau d})$.*

The Assumption 1(a) is the standard condition [29] to ensure the sufficient smoothness of the latent Gaussian processes $\tilde{\beta}(s)$, $\tilde{\alpha}(s), \zeta_k(s)$ and $\eta_i(s)$. The Assumption 1(b) is to specify the order of the number of voxels as the sample size increases, implying that our method can handle high resolution images.

As the mediator model (2.2) involves spatially varying coefficients $\eta_i(s)$ as individual effect parameters, the model identifiability is not trivial and requires some mild conditions on the observations of exposure variables and confounding factors.

Assumption 2. *(a) Each element in* $(\mathbf{X}, \mathbf{C})$ *has a finite fourth moment with sub-Gaussian tails, and* $\sigma_{\min}\{(\mathbf{X}, \mathbf{C})\} > \sqrt{n}$ *almost surely; (b) Conditioning on* $(\mathbf{X}, \mathbf{C})$*, there exists a matrix* $\mathbf{W} = (W_{i,k}) \in \mathbb{R}^{n\times(q+1)}$ *such that* $\det\{\mathbf{W}^\top(\mathbf{X}, \mathbf{C})\} \neq 0$*; and (c) there exists a constant vector* $\mathbf{b} = (b_1, \ldots, b_q)^\top$ *such that for any* $s \in \mathcal{S}$ *and* $k = 1, \ldots, q+1$, $\sum_{i=1}^n W_{i,k}\eta_i(s) = b_k$.

Assumption 2(a) is a reasonable assumption in linear regression with the design matrix $(\mathbf{X}, \mathbf{C})$ [1]. For (b) and (c), one example that can satisfy the above assumption is to set $b = 0 \in \mathbb{R}^{q+1}$, $\mathbf{W} = (\mathbf{X}, \mathbf{C})$, and if we express $\eta_i(s) = \sum_{l=1}^\infty \theta_{\eta,i,l}\psi_l(s)$ as infinite sums of basis in the Hilbert space, then each $(\theta_{\eta,i,l})_{i=1}^n \in \mathbb{R}^n$ is generated from a subspace orthogonal to $\text{span}\{\mathbf{X}, \mathbf{C}_1, \ldots, \mathbf{C}_q\}$. We enforce this assumption in the sampling algorithm by updating $(\theta_{\eta,i,l})_{i=1}^n$ from a constrained multivariate normal distribution.

With Assumption 2, we can establish the model identifiability in (2.2) and show that if the spatially varying coefficients are different from the true value, the mean function of $M_i(s)$, denoted as $\mu_{M,i}(s) := \alpha(s)X_i + \boldsymbol{\zeta}^\top(s)\mathbf{C}_i + \eta_i(s)$, will also be deviated from the true mean function $\mu_{M,i,0}(s) := \alpha_0(s)X_i + \boldsymbol{\zeta}_0^\top(s)\mathbf{C}_i + \eta_{i,0}(s)$.

Let $\Theta_M = \Theta_\alpha \times \Theta_{\boldsymbol{\zeta}} \times (\prod_i \Theta_{\eta,i})$ be the joint parameter space for all parameters in the mean function $\mu_{M,i}(s)$. For any $\epsilon > 0$ and some constant $c_0 > 0$, define the following two subsets of Θ_M as

$$\mathcal{U}_M^c = \left\{\Theta_M : \|\alpha - \alpha_0\|_{2,p}^2 + \sum_{k=1}^q \|\zeta_k - \zeta_{k,0}\|_p^2 + \frac{1}{n}\sum_{i=1}^n \|\eta_i - \eta_i\|_{2,p}^2 > \epsilon^2\right\}$$

$$\mathcal{U}_{M,\mu}^c = \left\{\Theta_M : \frac{1}{n}\sum_{i=1}^n \|\mu_{M,i} - \mu_{M,i,0}\|_{2,p}^2 > c_0\epsilon^2\right\}$$

Proposition 1. *Under Assumptions 2, (a) the mediator model* (2.2) *is identifiable; and (b)* $\mathcal{U}_M^c \subset \mathcal{U}_{M,\mu}^c$ *almost surely with respect to* $(\mathbf{X}, \mathbf{C})$.

2.3.2 Posterior consistency

First, we show joint posterior consistency of all the spatially varying coefficients in the mediator model (2.2) as the number of images $n \to \infty$ and the number of voxels $p \to \infty$.

The following empirical L_2 norm consistency result is proved by verifying conditions in Theorem A.1 in [13]. For the proof of existence of test, we borrow techniques from Proposition 11 in [84].

Theorem 1. *Suppose Assumptions 1-2 hold in the mediator model (2.2). For any $\epsilon > 0$, as $n \to \infty$, we have $\Pi(\mathcal{U}_M^c \mid \mathbf{M}, \mathbf{X}, \mathbf{C}) \to 0$ in P_0^n- probability. This further implies that $\Pi(\|\alpha - \alpha_0\|_{2,p} > \epsilon \mid \mathbf{M}, \mathbf{X}, \mathbf{C}) \to 0$ and $\Pi(n^{-1}\sum_{i=1}^n \|\eta_i - \eta_{i,0}\|_{2,p}^2 > \epsilon^2 \mid \mathbf{M}, \mathbf{X}, \mathbf{C}) \to 0$ in P_0^n- probability.*

Next, we give the L_2 consistency result on $\beta(s)$ with the following notations and assumptions.

For any $f \in L^2(\mathcal{S})$, given the basis $\{\psi_l(s)\}_{l=1}^\infty$ in Definition 4, $f(s) = \sum_{l=1}^\infty \theta_{f,l}\psi_l(s)$, where $\sum_{l=1}^\infty \theta_{f,l}^2 < \infty$. Let $r_L(s) = \sum_{l=L}^\infty \theta_{f,l}\psi_l(s)$ be the remainder term after choosing a cutoff L as the finite sum approximation. Note that the remainder term $\int_{\mathcal{S}} r_L(x)^2 \lambda(\mathrm{d}s) = \sum_{l=L}^\infty \theta_{f,l}^2 \to 0$ as $L \to \infty$ (Appendix E in [30]). We employ the basis expression to show the posterior consistency in model (2.1), especially for studying the role of $\mathcal{M}_i(\Delta s_j)$.

Denote $\tilde{\boldsymbol{\gamma}} = (\gamma, \boldsymbol{\xi}^\top)^\top \in \mathbb{R}^{q+1}$, $\tilde{\mathbf{X}}_i = (X_i, \mathbf{C}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{q+1}$. Let $\beta(s) = \sum_{l=1}^\infty \theta_{\beta,l}\psi_l(s_j)$. Let $\tilde{\mathcal{M}}_{i,l} = \sum_{j=1}^p \psi_l(s_j)\mathcal{M}_i(\Delta s_j)$, and define the $n \times L_n$ matrix $\tilde{\boldsymbol{\mathcal{M}}}_n := (\tilde{\mathcal{M}}_{i,l})_{i=1,\ldots,n, l=1,\ldots,L_n}$. Further, denote $\tilde{\mathbf{W}}_n = (\tilde{\boldsymbol{\mathcal{M}}}_n, \tilde{\mathbf{X}}) \in \mathbb{R}^{n\times(L_n+1+q)}$ as the design matrix.

We state the following assumption for constructing the consistency test in Theorem 2.

Assumption 3. *The least singular value of $\tilde{\mathbf{W}}_n$ satisfies $0 < c_{\min} < \liminf_{n\to\infty} \sigma_{\min}(\tilde{\mathbf{W}}_n)/\sqrt{n}$ with probability $1 - \exp(-\tilde{c}n)$ for some constant $\tilde{c}, c_{\min} > 0$.*

A similar assumption has been made in [1]. One extreme example that satisfies Assumption 3 is when $\tilde{\mathbf{W}}_n$ has mean-zero i.i.d. subgaussian entries. We will also give an example in the Supplementary Material A.2 that satisfies Assumption 3 and follows the generative model (2.2) under some conditions.

Remark. Assumption 3 demonstrates the variability in the design matrix $\tilde{\mathbf{W}}_n$: the posterior consistency of $\beta(s)$ can only be guaranteed when the variability of the design matrix is sufficiently large, implying that the level of complexity of the functional parameter $\beta(s)$ we can possibly estimate is determined by the complexity of the input imaging data.

Theorem 2. *Suppose Assumptions 1 - 3 hold in the outcome model (2.1) and the priors on $\tilde{\boldsymbol{\gamma}}$ satisfy that $\Pi(\|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0\|_2^2 < \epsilon) > 0$ for any $\epsilon > 0$. Then for any $\epsilon > 0$, we have, as*

$n \to \infty$, $\Pi\left(\|\beta - \beta_0\|_{2,p} + \|\tilde{\gamma} - \tilde{\gamma}_0\|_2 > \epsilon \mid \mathbf{Y}, \mathbf{M}, \mathbf{X}, \mathbf{C}\right) \to 0$ *in* P_0^n*- probability. This implies that* $\Pi\left(\|\beta - \beta_0\|_{2,p} > \epsilon \mid \mathbf{Y}, \mathbf{M}, \mathbf{X}, \mathbf{C}\right) \to 0$ *in* P_0^n*- probability.*

In the proof of Theorem 2, especially in constructing the test for $H_0 : \beta(s) = \beta_0(s)$ v.s. $H_1 : \|\beta - \beta_0\|_{2,p} > \epsilon$ through the basis approximation of $\beta(s)$, verifying conditions in the Supplementary Material for $M_i(s)$ in model (2.2) provides insight into the relationship between models (2.1) and (2.2): sufficient variability in $M_i(s)$ ensures posterior consistency of $\beta(s)$.

Theorem 3. *(Posterior consistency of SVME) Under Assumptions 1 - 3, for any* $\epsilon > 0$*, as* $n \to \infty$, $\Pi\left(\|\mathcal{E} - \mathcal{E}_0\|_{1,p} < \epsilon \mid \mathbf{Y}, \mathbf{M}, \mathbf{X}, \mathbf{C}\right) \to 1$ *in* P_0^n*-probability.*

This theorem implies that the posterior distribution of SVME concentrates on an arbitrarily small neighborhood of its true value with probability tending to one when the sample size goes to infinity. Here the sample size refers to the number of images n. By Assumption 1, in this case. the number of voxels p also goes to infinity. This theorem also implies the consistency of estimating NIE using posterior inference by BIMA in the following corollary.

Corollary 1. *(Posterior consistency of NIE) For any* $\epsilon > 0$*, as* $n \to \infty$*,*

$$\Pi\left(p^{-1}\left|\sum_{j=1}^{p}\mathcal{E}(s_j) - \sum_{j=1}^{p}\mathcal{E}_0(s_j)\right| < \epsilon \;\middle|\; \mathbf{Y}, \mathbf{M}, \mathbf{X}, \mathbf{C}\right) \to 1$$

in P_0^n*-probability.*

From Theorem 3, we can further establish the posterior sign consistency of SVME. Consider a minimum effect size $\delta > 0$, define $\mathcal{R}_\delta^+ = \{s : \mathcal{E}_0(s) > \delta\}$ and $\mathcal{R}_\delta^- = \{s : \mathcal{E}_0(s) < -\delta\}$, which represent the true positive SVME region and the true negative SVME region respectively. Let $\mathcal{R}_0 = \{s : \mathcal{E}_0(s) = 0\}$ represent a region of which the true SVME is zero.

Corollary 2. *(Posterior sign consistency of SVME) For any* $\delta > 0$*, let* $\mathcal{R}_\delta = \mathcal{R}_\delta^+ \cup \mathcal{R}_\delta^- \cup \mathcal{R}_0$*, Then as* $n \to \infty$, $\Pi\left[\text{sign}\{\mathcal{E}(s)\} = \text{sign}\{\mathcal{E}_0(s)\}, \forall s \in \mathcal{R}_\delta \mid \mathbf{Y}, \mathbf{M}, \mathbf{X}, \mathbf{C}\right] \to 1$ *in* P_0^n*-probability, where* $\text{sign}(x) = 1$ *if* $x > 0$, $\text{sign}(x) = -1$ *if* $x < 0$ *and* $\text{sign}(0) = 0$.

This corollary ensures that with a large posterior probability BIMA can identify the important regions with significant positive and negative SVMEs that contributes the NIE.

2.4 Posterior Computation

The posterior computation for BIMA is challenging due to the complexity of the nonparametric inference, the high-dimensional parameter space and the non-conjugate prior speci-

fications for the spatially-varying coefficients in the model. To address these challenges, we next construct an equivalent model representation.

2.4.1 Model representation and approximation

We approximate the STGPs and GPs using a basis expansion approach. By Mercer's theorem [94], the correlation kernel function in (2.8) can be decomposed by infinite series of orthonormal basis functions $\kappa(s,s') = \sum_{l=1}^{\infty} \lambda_l \psi_l(s)\psi_l(s')$, and the corresponding GP $g(s) \sim \mathcal{GP}(0, \sigma_g^2\kappa)$ can be expressed as $g(s) = \sum_{l=1}^{\infty} \theta_{g,l}\psi_l(s)$ where $\theta_{g,l} \overset{\text{ind}}{\sim} N(0, \lambda_l\sigma_g^2)$.

In our implementation, we allow region partition to speed up the computation, and assume a region-independence prior kernel structure for the spatially varying parameters $\beta, \alpha, \zeta_k, \eta_i$. In real data analysis, the brain anatomic region parcellation defines the region partition. Assume there are $r = 1, \ldots, R$ regions that form a partition of the support $\mathcal{S}$, denoted as $\mathcal{S}_1, \ldots, \mathcal{S}_R$. The kernel function $\kappa(s_j, s_k) = 0$ for any $s_j \in \mathcal{S}_r, s_k \in \mathcal{S}_{r'}, r \neq r'$, and the prior covariance matrix on the fixed grid has a block diagonal structure. For the whole brain analysis as one region, one can choose $R = 1$.

For the r-th region, let p_r be the number of voxels in $\mathcal{S}_r$, $Q_r = (\psi_l(s_{r,j}))_{l=1,j=1}^{L_r,p_r} \in \mathbb{R}^{L_r \times p_r}$ be the matrix with the (l,j)-th component $\psi_l(s_{r,j})$, $\{s_{r,j}\}_{j=1}^{p_r}$ forms the fixed grid in $\mathcal{S}_r$. Because of the basis approximation with cutoff L_r, Q_r is not necessarily an orthonormal matrix, hence we use QR decomposition to get an approximated orthonormal Q_r, i.e. $Q_r^T Q_r = I_{L_r}$, where I_{L_r} is the identity matrix. With the region partition, the GP priors on the r-th region can be approximated as $g_r = (g(s_{r,1}), \ldots, g(s_{r,p_r}))^T \approx Q_r \boldsymbol{\theta}_{g,r}$, where $\boldsymbol{\theta}_{g,r} \sim \mathcal{N}\left(0, \sigma_g^2 D_r\right)$, D_r is a diagonal matrix with eigenvalues $(\lambda_{r,1}, \ldots, \lambda_{r,L_r})^{\mathrm{T}} \in \mathbb{R}^{L_r}$.

After truncating the expansion at sufficiently large $\{L_r\}_{r=1}^{L}$, STGPs and GPs in the prior specifications (2.8), which all share the same kernel, can be approximated by

$$\begin{aligned}
&\beta_r = T_\nu(\tilde{\beta}_r) \approx T_\nu\left(Q_r \boldsymbol{\theta}_{\tilde{\beta},r}\right), \quad \alpha_r = T_\nu(\tilde{\alpha}_r) \approx T_\nu\left(Q_r \boldsymbol{\theta}_{\tilde{\alpha},r}\right), \\
&\zeta_{k,r} \approx Q_r \boldsymbol{\theta}_{\zeta,k,r}, \quad \eta_{i,r} \approx Q_r \boldsymbol{\theta}_{\eta,i,r},
\end{aligned}$$

where the corresponding basis coefficients follow independent normal priors:

$$\boldsymbol{\theta}_{\tilde{\beta},r} \sim \mathrm{N}_{L_r}(0, \sigma_\beta^2 D_r), \quad \boldsymbol{\theta}_{\tilde{\alpha},r} \sim \mathrm{N}_{L_r}(0, \sigma_\alpha^2 D_r), \quad \boldsymbol{\theta}_{\zeta,k,r} \sim \mathrm{N}_{L_r}(0, \sigma_\zeta^2 D_r), \quad \boldsymbol{\theta}_{\eta,i,r} \sim \mathrm{N}_{L_r}(0, \sigma_\eta^2 D_r).$$

We discuss the details for choosing L_r in Section 4.2. Denote $\mathcal{M}_i(\mathcal{S}_r) = (\mathcal{M}_i(\Delta s_{r,j}))_{j=1}^{p_r} \in$

$\mathbb{R}^{p_r}$, $M_i(\mathcal{S}_r) = (M_i(s_{r,j}))_{j=1}^{p_r} \in \mathbb{R}^{p_r}$, Then the BIMA model can be approximated as follows.

$$Y_i = \sum_{r=1}^{R} T_\nu\left(Q_r \boldsymbol{\theta}_{\tilde{\beta},r}\right) \mathcal{M}_i(\mathcal{S}_r) + \gamma X_i + \boldsymbol{\zeta}_Y^{\mathrm{T}} \mathbf{C}_i + \epsilon_{Y,i},$$

$$M_i(\mathcal{S}_r) = T_\nu\left(Q_r \boldsymbol{\theta}_{\tilde{\alpha},r}\right) X_i + \sum_{k=1}^{q} Q_r \boldsymbol{\theta}_{\zeta,k,r} C_{i,k} + Q_r \boldsymbol{\theta}_{\eta,i,r} + \epsilon_{M_r,i}$$

where $\epsilon_{Y,i} \sim \mathrm{N}(0, \sigma_Y^2)$ and $\epsilon_{M_r,i} \sim \mathrm{N}_{p_r}(0, \sigma_M^2 I_{p_r})$. From the above model representation, both $\boldsymbol{\theta}_{\zeta,k,r}$ and $\boldsymbol{\theta}_{\eta,i,r}$ have conjugate posteriors, but T_ν is not a linear function, and $\boldsymbol{\theta}_{\tilde{\beta},r}$ and $\boldsymbol{\theta}_{\tilde{\alpha},r}$ do not have conjugate posteriors. To overcome this, the Metropolis-adjusted Langevin algorithm (MALA) is used to sample $\boldsymbol{\theta}_{\tilde{\beta},r}$ and $\boldsymbol{\theta}_{\tilde{\alpha},r}$. However, the first-order derivative of the soft-thresholded function $T_\nu(x)$ does not exist at the two change points $x = \pm\nu$. To approximate the first-order derivative, either the derivative of a smooth approximation function or a piece-wise function $d\hat{T}_\nu(x) = I(|x| \geq \nu)$ works in our case. The later one $d\hat{T}_\nu(x) = I(|x| \geq \nu)$ provides better computational efficiency, and is implemented in our algorithm.

2.4.2 Covariance kernel specifications and estimation

We can choose different covariance kernels for the GPs in models (2.1) and (2.2). Given the covariance kernel function $\kappa(\cdot, \cdot)$, to obtain the coefficients λ_l and the basis functions $\psi_l(s)$, Sections 4.3.1 and 4.3.2 in [94] provide the analytic solution for squared exponential kernel, and an approximation method for other kernel functions with no analytic solutions. In practice when $\psi_l(s)$ has no analytical solutions, such as the Matérn kernel, we use eigen decomposition on the covariance matrix, and take the first L eigenvalues as the approximated λ_l, the first L eigenvectors as the approximated $\psi_l(s)$, then apply QR decomposition on the approximated basis functions to obtain orthonormal basis. The limitation of this method is that the covariance matrix is difficult to compute in high dimensions due to precision issues. Hence in high dimensions we split the entire space $\mathcal{S}$ into smaller regions, and compute the basis functions on each region independently. This also aligns with the imaging application with the whole brain atlas. Another benefit is that by splitting the whole parameter space into smaller regions, the sampling space gets smaller and it becomes easier to accept the proposed vector $\beta(s)$ on each region with much less directions to explore. In practice, to choose the number of basis functions L_r for region r with p_r voxels, we first compute the covariance matrix in $\mathbb{R}^{p_r \times p_r}$ with appropriately tuned covariance parameters, get the eigen-value of such covariance matrix, and choose the cutoff such that the summation $\sum_{l=1}^{L_r} \lambda_l$ is over 90% of $\sum_{l=1}^{p_r} \lambda_l$, i.e. the eigenvalues before cutoff account for over 90% of the total eigenvalues.

We provide the detailed sensitivity analysis on choosing the covariance parameters in the Supplementary Material.

2.4.3 The MCMC algorithm

We develop an efficient Markov chain Monte Carlo (MCMC) algorithm for posterior computation. To update parameters $\left\{\boldsymbol{\theta}_{\tilde{\beta},r}, \boldsymbol{\theta}_{\tilde{\alpha},r}\right\}_{r=1}^{R}$, we adopt the Metropolis-adjusted Langevin algorithm (MALA). The step size is tuned during the burn-in period to ensure an acceptance rate between 0.2 and 0.4. The target acceptance rate for each region is set to be proportional to the inverse of the number of basis functions in that region, in order to produce a relatively large effective sample size of the MCMC sample.

To incorporate the block structure with MALA, in each iteration, the proposal $\boldsymbol{\theta}_{\tilde{\beta},r}$ or $\boldsymbol{\theta}_{\tilde{\alpha},r}$ for region $\mathcal{S}_r$ is based on the target posterior density conditional on $\boldsymbol{\theta}_{\tilde{\beta},r'}$ or $\boldsymbol{\theta}_{\tilde{\alpha},r'}$ supported on all other regions where $r' \neq r$. The acceptance ratio is also computed region by region.

MALA has a considerable computational cost especially in high dimensional sampling, where the step size has to be very small to have an acceptance rate reasonably greater than 0. It is important to have a good initial value. To obtain the initial values, we consider a working model with the spatially varying coefficients $\beta(s)$ and $\alpha(s)$ following GP instead of STGP. With the basis expansion approach, we can straightforwardly use Gibbs sampling to obtain the approximated posterior samples of $\beta(s)$ and $\alpha(s)$ of the working model. The posterior mean values of $\beta(s)$ and $\alpha(s)$ estimated from the working model can be used to specify the initial value of the basis coefficients in the MALA algorithm. More detailed discussion on choosing the initial value can be found in Supplementary Material Section S4.

To impose identifiability Assumption 2, the posterior of $\theta_{\eta,i,l}$ is sampled from a constrained multivariate normal distribution, with the constraint $\tilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\theta}_{\boldsymbol{\eta},\boldsymbol{l}} = \mathbf{0}$ where $\boldsymbol{\theta}_{\boldsymbol{\eta},\boldsymbol{l}} = (\theta_{\eta,1,l}, \ldots, \theta_{\eta,n,l})^{\mathrm{T}}$. The algorithm for sampling multivariate normal distribution constrained on a hyperplane follows Algorithm 1 in [16].

For the rest of the parameters, with available conjugate full conditional posteriors, we use Gibbs sampling to update. The algorithm is implemented in Rcpp [21] with RcppArmadillo [22]. The implementation is wrapped as an R package `BIMA`. [1]

2.5 Simulations

To demonstrate the performance of BIMA, two sets of simulation studies are analyzed. In the first simulation study, we compare the performance with other existing Bayesian mediation

[1] Available on Github https://github.com/yuliangxu/BIMA

methods in a low dimensional setting with relative small sample sizes, since some competing methods cannot handle high-dimensional settings efficiently. In the second simulation study, we vary the sample size, noise variance, and image patterns, and conduct a sensitivity analysis on the performance of our method under different settings with different prior specifications.

2.5.1 Comparison with existing methods

To make fully Bayesian inferences in mediation model, there exist a number of methods utilizing threshold priors and mixture models to impose sparsity and model correlation. In this section, we compare BIMA with two recently proposed Bayesian methods: product threshold Gaussian prior (PTG) and Correlated Selection Model (CorS).

Product Threshold Gaussian prior (PTG) [79] constructs prior distribution of the bivariate vector $\{\beta(s_j), \alpha(s_j)\}$ for each location s_j by thresholding a bivariate Gaussian latent vector $\{\tilde{\beta}(s_j), \tilde{\alpha}(s_j)\} \sim \mathrm{N}_2(0, \boldsymbol{\Sigma})$ and their product. i.e.

$$\beta(s_j) = \tilde{\beta}(s_j) \max\left\{I(|\tilde{\beta}(s_j)| > \lambda_1), I(|\tilde{\beta}(s_j)\tilde{\alpha}(s_j)| > \lambda_0)\right\},$$
$$\alpha(s_j) = \tilde{\alpha}(s_j) \max\left\{I(|\tilde{\alpha}(s_j)| > \lambda_2), I(|\tilde{\beta}(s_j)\tilde{\alpha}(s_j)| > \lambda_0)\right\}.$$

PTG model uses the threshold parameters λ_1, λ_2 and λ_0 to control the sparsity in $\beta(s_j)$, $\alpha(s_j)$ and the indirect effect $\beta(s_j)\alpha(s_j)$ respectively, and [79] directly set $\boldsymbol{\Sigma} = \mathrm{diag}\left\{\sigma_\beta^2, \sigma_\alpha^2\right\}$. However, the spatial correlation in spatially-varying coefficients among different locations s_j is not taken into consideration. Hence we anticipate this method to be less suitable for spatially correlated applications such as brain imaging. This method has been implemented in the `R` package `bama` [67]. We set $\lambda_1 = \lambda_2 = \lambda_0 = 0.01$. A total number of 1500 MCMC iterations are performed with 1000 burnins.

Correlated Selection model [78, CorS] adopts a mixture model with four components to specify different sparsity patterns of $\alpha(s_j)$ and $\beta(s_j)$ and incorporate the spatial correlations into prior specifications of mixing weights.

$$[\beta(s_j), \alpha(s_j)]^{\mathrm{T}} \sim \pi_1(s_j)\mathrm{N}_2(0, \mathbf{V}_1) + \pi_2(s_j)\mathrm{N}_2(0, \mathbf{V}_2) + \pi_3(s_j)\mathrm{N}_2(0, \mathbf{V}_3) + \pi_4(s_j)\boldsymbol{\delta}_0,$$

and a membership variable $\gamma(s_j) \in \{1, 2, 3, 4\}$, where $\gamma(s_j) = 1$ indicates $\beta(s_j)\alpha(s_j) \neq 0$, $\gamma(s_j) = 2$ indicates $\beta(s_j) \neq 0, \alpha(s_j) = 0$, $\gamma(s_j) = 3$ indicates $\beta(s_j) = 0, \alpha(s_j) \neq 0$, and $\gamma(s_j) = 4$ indicates $\beta(s_j) = \alpha(s_j) \neq 0$. When $\gamma(s_j) = 1$, $\mathbf{V}_1$ is assigned an inverse Wishart prior. When $\gamma(s_j) = 2$ or 3, $\mathbf{V}_2$ or $\mathbf{V}_3$ only contains σ_β^2 or σ_α^2 on the diagonal and 0 otherwise. Each $\gamma(s_j)$ is assumed to follow a multinomial distribution with probability $\boldsymbol{\pi}(s_j) = \{\pi_1(s_j), \pi_2(s_j), \pi_3(s_j), \pi_4(s_j)\}^\top$ with $\sum_{k=1}^4 \pi_k(s_j) = 1$. For each $m = 1, 2, 3$, let

$\boldsymbol{\pi}_m = \{\pi_m(s_1), \ldots, \pi_m(s_j)\} \in \mathbb{R}^p$. $\text{logit}(\boldsymbol{\pi}_m)$ is assumed to follow a multivariate normal prior with a pre-specified covariance matrix $\sigma_m^2 \mathbf{D} \in \mathbb{R}^{p \times p}$, independently for each $m = 1, 2, 3$. Hence $\mathbf{D}$ is used to reflect the mediator-wise correlation.

We anticipate this method to have good performance in the spatially correlated data application. We use the GitHub implementation of this method (https://github.com/yanys7/Correlated_GMM_Mediation.git). In the simulation study, we set the initial values for all $\alpha(s)$ and $\beta(s)$ to be 0.5, the initial values for $\{\pi_k(s_j), k = 1, 2, 3, 4\}$ to be 0.25, the 2 by 2 scale matrix in Inverse-wishart prior for $\mathbf{V}_1$ to be $[1, 0.5; 0.5, 1]$, and the $p \times p$ matrix $\mathbf{D}$ to be estimated from the input image correlations. A total number of 2000 MCMC iterations are performed with 1000 burn-ins.

Bayesian Image Mediation Analysis (BIMA) adopts a modified square-exponential kernel $\kappa(s, s'; a, b) = \text{cor}\{\beta(s), \beta(s')\} = \exp\{-a(s^2 + s'^2) - b\|s - s'\|^2\}$ with $a = 0.01$ and $b = 10$. We split the input image into four regions. We use Hermite polynomials up to the 10th degree, resulting in 66 basis coefficients to approximate each region. The initial values for all parameters are obtained from Gibbs sampling with Gaussian process priors for α and β. The threshold parameter $\nu = 0.5$ in STGP priors. For the outcome model (2.1), a total of 10^5 iterations are performed, with the acceptance probability tuned to be around 0.2 for each region during the first 80% of burn-in iterations. The mediator model (2.2) follows the same setting, except with a total of 5000 iterations and a burn-in period comprising the first 90%.

Figure 2.3 shows the true image for $\alpha(s)$, $\beta(s)$, and $\mathcal{E}(s)$, i.e. natural indirect effect (NIE). Table 2.1 gives summary statistics of sampled NIE using 3 methods with 100 replicated simulations. The final result of NIE is tuned using the inclusion probability of the sampled NIE for all 3 methods in the following way: for each location s_j, we estimate the empirical probability $\hat{P}(\text{NIE}(s_j) \neq 0)$ from the MCMC sample of NIE, and set a threshold t on $\hat{P}(\text{NIE}(s_j) \neq 0)$: if $\hat{P}(\text{NIE}(s_j) \neq 0) < t$, $\text{NIE}(s_j) = 0$, otherwise $\text{NIE}(s_j)$ equals the posterior sample mean. By tuning t, we can control the FDR to be below 10%. Although we set the target FDR to be 10% for all 3 methods, it is still possible that FDR cannot be tuned to be less than 10% with any $t < 1$ when the sample is very noisy, in which case the largest possible t is used, and the tuned FDR can be larger than 10%. In the extreme case where the largest possible t still maps all location to 0, we get the NAs as shown in Table 2.1. These NA replication results are excluded from the summary statistics in Table 2.1.

From Table 2.1, PTG performs the least ideal in the correlated image setting as shown in Figure 2.3, especially in the estimation for the mediator effect $\beta(s)$. In general, $\beta(s)$ is more challenging to estimate than $\alpha(s)$ for two reasons: i) The mediator model (2.2) has $n \times p$ observations to estimate p dimensional $\alpha(s)$, leading to a higher signal to noise ratio than

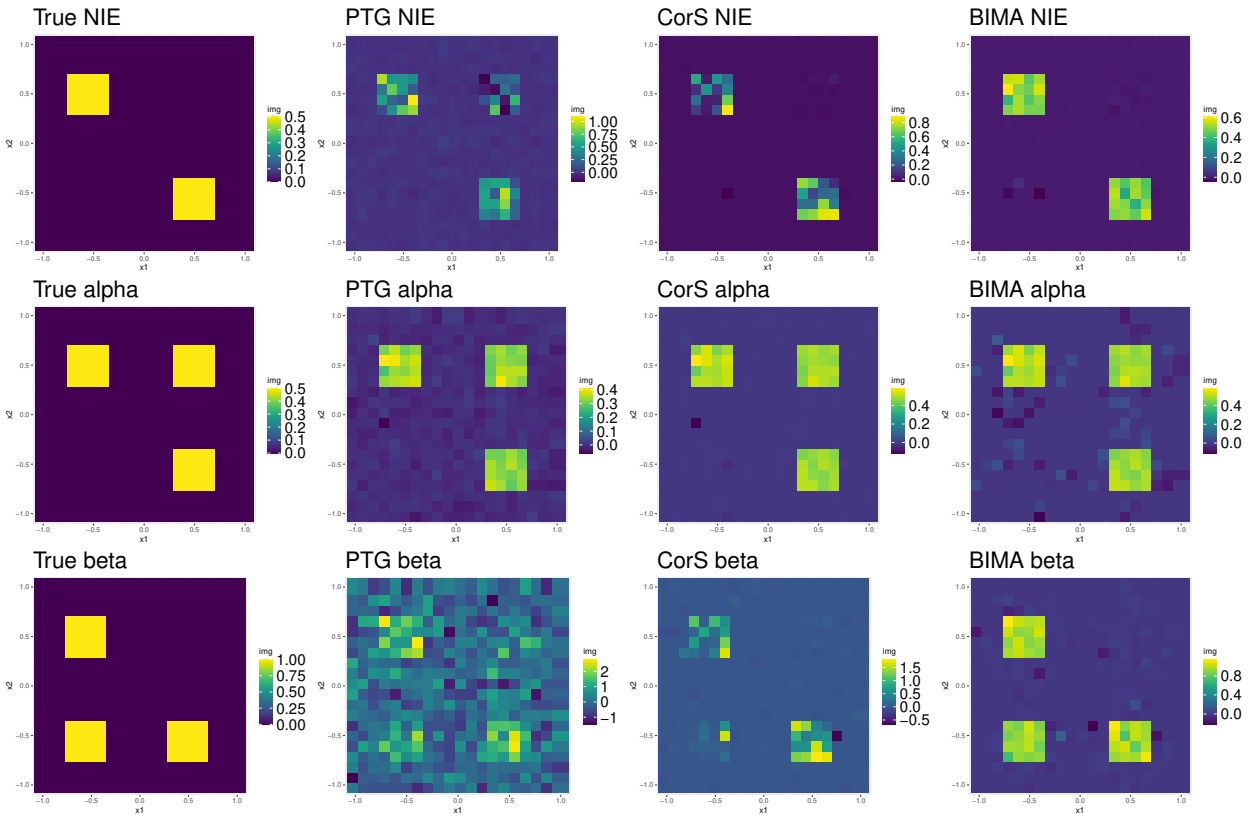


Figure 2.3: Comparison on the posterior mean of the 3 methods with the true images. Rows from top to bottom represent functional NIE $\mathcal{E}(s)$, $\alpha(s)$, $\beta(s)$. Columns from left to right represent true images, posterior mean from PTG model, posterior mean from CorS model, posterior mean from BIMA model.

β in model (2.1); ii) In the outcome model (2.1), M and X are correlated through (2.2), making it more difficult to separate the effect $\beta(s)$ from γ.

CorS model performs very well when n is close to p. However in the higher dimensional setting, when n is much smaller than p, CorS has a lower power than BIMA. BIMA performs well and is stable across all four settings, indicating that it is a suitable method especially for high-dimensional spatially correlated mediators, when n is considerably less than p, such as in brain imaging application. Potential improvement can be made for BIMA when the kernel bases are tuned to accurately represent the smoothness of input mediators.

2.5.2 High-dimensional simulation

To further illustrate the performance of our proposed method, we conduct simulation studies under 4 different settings with 2 sets of patterns as shown in Figure 2.4. Each image is split

Table 2.1: Comparison of posterior inferences on NIE among different methods including PTG, CorS and BIMA based on 100 replications. The standard errors are reported in the brackets

(a) Selection accuracy including the overall accuracy (ACC), false discovery rate (FDR) and true positive rate (TPR). All values are multilied by 100.

Selection Accuracy									
	PTG			CorS			BIMA		
(n, p)	FDR	TPR	ACC	FDR	TPR	ACC	FDR	TPR	ACC
$(200, 400)$	9 (15)	20 (19)	93 (1)	1 (2)	80 (37)	98 (3)	7 (3)	95 (3)	99 (0)
$(300, 400)$	21 (21)	16 (14)	93 (1)	1 (2)	100 (0)	100 (0)	6 (3)	93 (5)	99 (0)
$(200, 676)$	14 (14)	11 (12)	93 (1)	0 (0)	3 (2)	93 (0)	8 (2)	96 (3)	99 (0)
$(300, 676)$	10 (14)	17 (11)	94 (1)	1 (1)	80 (36)	98 (3)	7 (2)	96 (3)	99 (0)

(b) Estimation and computation performance including mean squared errors (MSE) in the true activation region (multiplied by 100) and computation time in seconds.

Estimation and Computation time									
	MSE (Activation)			Time (Seconds)				#of NA	
(n, p)	PTG	CorS	BIMA	PTG	CorS	BIMA (2.1)	BIMA (2.2)	PTG	CorS
$(200, 400)$	24 (1)	5 (10)	2 (1)	251 (7)	26 (3)	27 (2)	28 (1)	31	7
$(300, 400)$	24 (1)	0 (0)	2 (1)	385 (8)	25 (2)	35 (3)	61 (1)	22	0
$(200, 676)$	24 (0)	25 (1)	2 (1)	663 (13)	75 (1)	54 (6)	35 (1)	60	60
$(300, 676)$	24 (0)	5 (9)	1 (1)	1026 (21)	76 (2)	64 (11)	71 (2)	21	11

into 4 regions, each region being a 32×32 grid. The threshold parameter $\nu = 0.5$ in STGP priors. In this simulation, we use Matérn kernel in accordance with the sharp patterns in Figure 2.4.

$$\kappa(s', s; u, \rho) = C_u(\|s' - s\|_2^2/\rho),\ C_u(d) := \frac{2^{1-u}}{\Gamma(u)} \left(\sqrt{2ud}\right)^u K_u(\sqrt{2ud}) \tag{2.9}$$

The number of basis for each region is set to be 20% of the region size. The scale parameter $\rho = 2$, and $u = 1/5$. Due to the high dimension of mediators, we let the MALA algorithm update only $\beta(s)$ for the first 40% of MCMC iterations to get $\beta(s)$ to a stable value, then jointly updating all other parameters in (2.1) using Gibbs Sampling. All other settings are the same as in Section 2.5.1, and the summary statistics for NIE in Table 2.2 are also tuned in the same way using inclusion probability. Table 2.2(b) gives a sensitivity analysis result using different thresholds ν in the STGP priors to show that the estimation is not too sensitive to the choice of ν within a small range.

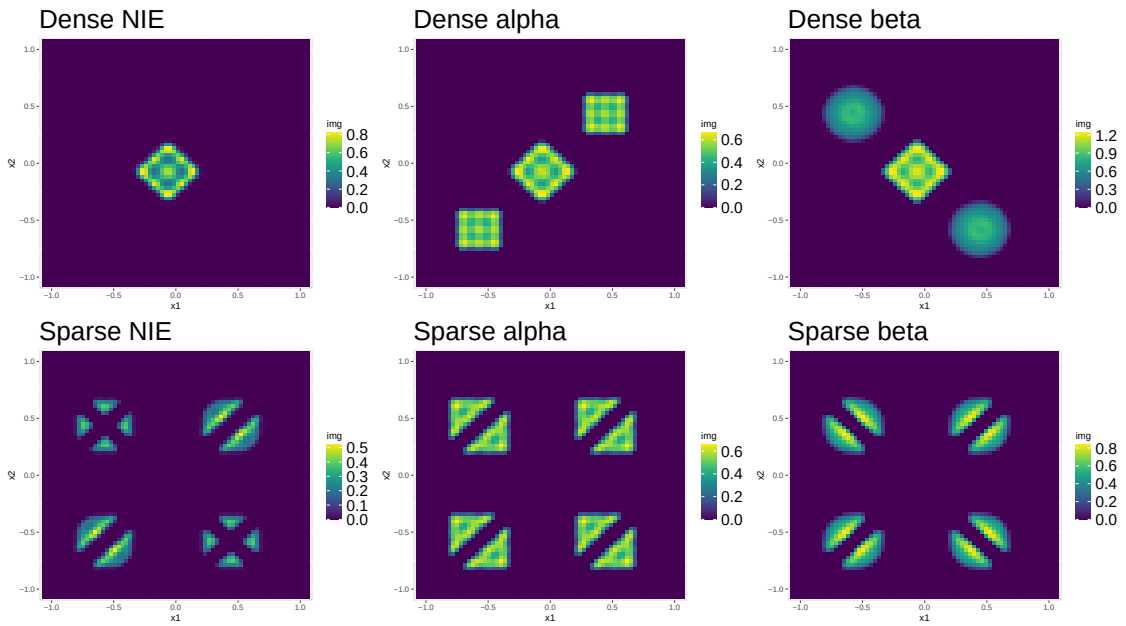


Figure 2.4: Input image pattern for the simulation study. Rows from top to bottom represent dense pattern and sparse pattern. Columns from left to right represent input image NIE, $\alpha(s)$, $\beta(s)$. $p = 4096$.

Table 2.2 demonstrates that our proposed method has stable performance across different settings. In Table 2.2, the mediator model is fully updated and converged including all individual effects $(\eta_i)_{i=1}^n$. Fully updating $(\eta_i)_{i=1}^n$ can take much longer time for the entire model to converge compared to directly setting the individual effects all to 0. In the case all η_i fixed at 0, the estimation for α and $\boldsymbol{\zeta}$ are almost the same compared to updating the full model from the $p = 4096$ simulation studies that we have observed. When $n = 1000$ and $p = 4096$, the computational time of fitting BIMA with running $30,000$ MCMC iterations is less than four hours for both models (2.1) and (2.2). In comparison, the CorS method takes 9.8 hours when $N = 1000, p = 2000$, with 1.5×10^5 iterations. Our approach shows a much better computational efficiency in the high-dimensional setting.

2.6 Analysis of ABCD fMRI Data

In this section, we apply our method to the Adolescent Brain Cognitive Development (ABCD) study release 1 [10]. The 2-back 3mm task fMRI contrast data is used, and the preprocessing method is described in [81]. After preprocessing and removing missing data, the final complete data set consists of $N = 1861$ subjects. The initial size of one image is $61 \times 73 \times 61$, which contains 116 brain regions, but only 90 regions are chosen for this

Table 2.2: High-dimensional simulation results. Selection accuracy (multiplied by 100) includes false discovery rate (FDR), true positive rate (TPR) and overall accuracy (ACC). Computational time (in minutes) are separately reported for fitting model (2.1) (T1) and model (2.2) (T2). The reported values are the average over 100 replications. The standard deviations are reported in the brackets.

(a) Peformance of BIMA in simulations for different sample sizes (n) and the random noise standard deviations in model (2.1) (σ_Y).

Under different generative model, $\lambda = 0.5$							
Pattern	n	σ_Y	FDR	TPR	ACC	T1	T2
Dense	1000	0.1	1 (1)	98 (1)	100 (0)	13 (2)	184 (16)
Sparse	1000	0.1	3 (3)	100 (0)	100 (0)	19 (4)	212 (28)
Dense	**5000**	0.1	1 (2)	97 (4)	100 (1)	54 (15)	1145 (140)
Dense	1000	**0.5**	1 (1)	98 (1)	100 (0)	17 (5)	209 (37)

(b) Sensitivity analysis with different threshold values (ν).

Under different sensitivity parameter ν. Dense pattern, $n = 1000$, $\sigma_Y = 0.1$.					
ν	FDR	TPR	ACC	T1	T2
0.3	5 (1)	99 (0)	99 (0)	20 (5)	198 (18)
0.6	0 (0)	97 (10)	100 (1)	17 (4)	222 (28)

application, and the resulting number of mediators in brain image is $p = 47636$.

We are interested in examining the natural indirect effect (NIE) of parental education level on children's IQ scores, mediated through brain imaging data. Our aim is to explore the varying roles of different brain regions as mediators in the cognitive ability development of a child. Hence the exposure is a binary variable indicating whether the parent has a college or higher degree. The outcome variable is g-score that reflects children's IQ, obtained in the same way as in [81] from the raw data. The confounders in our model include age, gender, race and ethnicity, and household income. For the multi-level variables race and ethnicity (Asian, Black, Hispanic, Other, White), household income (less than 50k, between 50k and 100k, greater than 100k), we use binary coding for each level. Table 2.3 provides the summary statistics of the ABCD data.

In this analysis, we use the Matérn kernel where the hyper-parameters u and ρ are specified for each region according to the estimated covariance matrices. The number of voxels for each region varies from 62 to 1510. To determine the number of basis, we select up to 500

Table 2.3: Summary statistics of the ABCD data stratified by Parent Degree. Mean (standard deviation) are reported for g-Score and Age. Counts are reported for Gender, Income, Race and Ethnicity

Parent degree	Bachelor or higher	No bachelor	Overall
g-Score	0.47 (0.77)	-0.15 (0.80)	0.27 (0.83)
Age	10.09 (0.61)	10.01 (0.63)	10.06 (0.62)
Gender			
Female	611	281	892
Male	635	334	969
Race and Ethnicity			
Asian	30	3	33
Black	47	84	131
Hispanic	151	216	367
White	924	254	1178
Other	94	58	152
Income			
<50K	98	336	434
50∼100K	375	213	558
>=100K	773	66	839
Total	1246	615	1861

locations within a certain range of the centroid for each region. Using these locations, we compute the empirical covariance matrix for each region. The cutoff for the number of basis is then chosen in such a way that it accounts for 90% of the total sum of all the singular values of the estimated covariance matrix. Because the hyper-parameter ν in the STGP prior and the kernel parameters u, ρ in each region are all prefixed, we provide a detailed description of selecting these parameters via testing MSE in the Supplementary Material. The final threshold ν_β for $\beta(s)$ is set to be 0.05, and the final threshold ν_α for $\alpha(s)$ is set to be 0.1. The choice of ν is also based on testing MSE. Detailed sensitivity analysis can be found in the Supplementary Material.

A total of $100,000$ iterations were performed for the outcome model (2.1) with the first 50% as burn-in, and thinning the posterior sample gives us 1000 samples out of the original $50,000$ samples. A total of $40,000$ iterations were performed for the mediator model (2.2) with the first $30,000$ as burn-in, and the posterior sample are thinned every 10 iterations to have 1000 sample. Based on this 1000 posterior sample, Table 2.4 gives a summary of both the overall NIE and NDE and the top 7 regions identified with the largest number of active voxels. The definition of NIE in each region is $\frac{1}{p}\sum_{s\in\mathcal{S}_r}\beta(s)\alpha(s)$, where $\mathcal{S}_r$ is the

collection of all voxels in region r. The rule for selecting the active voxels is based on cutting the posterior inclusion probability (PIP) at 10%. Voxels with PIP values above this threshold are identified as active. The posterior of NDE γ has a mean of 0.27 with the 95% credible interval $(0.20, 0.36)$. The posterior of NIE $\mathcal{E}$ has a mean of 0.0885 with the 95% credible interval $(0.066, 0.111)$. This suggests that parents with college degrees have a positive impact on children's cognitive abilities, and about 25% of the effect is mediated through brain cognitive development. Figure 2.5 shows the estimated active regions and the NIE in coronal view slides.

Table 2.4: Top 7 regions ordered by the number of active voxels with $PIP > 10\%$. Columns 2 to 5 are timed by 100. NIE(+) and NIE(-) are defined as $\frac{1}{p_n}\sum_{s\in\nabla_r}\mathcal{E}(s)I(\mathcal{E}(s)>0)$ and $\frac{1}{p_n}\sum_{s\in\nabla_r}\mathcal{E}(s)I(\mathcal{E}(s)<0)$ for each region r. Average IP is the averaged inclusion probability over all voxels in the entire region.

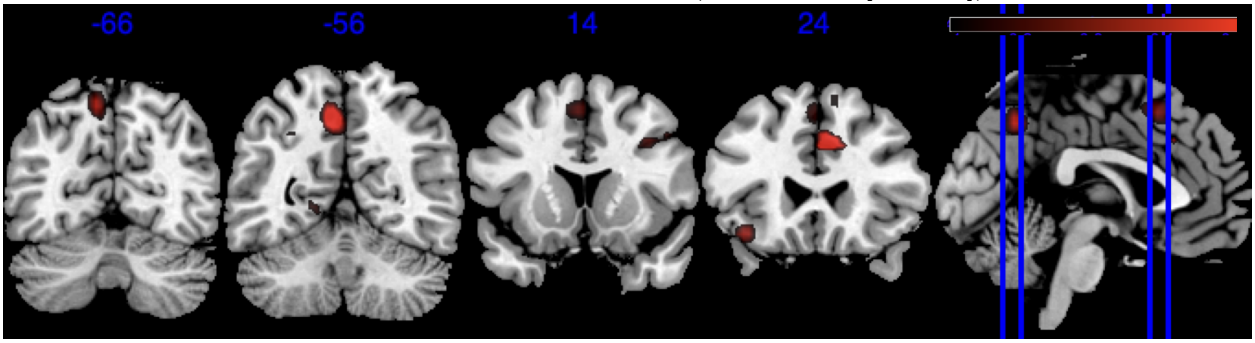
	NIE	**NIE(+)**	**NIE(-)**	**NDE**	**Time (hours) model** (2.1)	**Time (hours) model** (2.2)
Overall	8.85	10.57	-1.72	27.37	1.60	85.93
Region Name (AAL Atlas)	**NIE**	**NIE(+)**	**NIE(-)**	**Average PIP**	**# of active voxels**	**Region Size**
Precuneus_L	3.53	3.53	-0.01	4.98	109	1079
Parietal_Inf_L	2.83	2.83	0.00	5.67	99	696
Postcentral_L	0.21	0.21	0.00	1.98	71	1159
Cingulum_Mid_R	1.82	1.82	0.00	8.98	67	605
Supp_Motor_Area_L	1.14	1.16	-0.02	2.38	52	656
Frontal_Inf_Oper_R	-0.46	0.00	-0.47	1.83	27	421
Frontal_Inf_Orb_L	-0.12	0.02	-0.13	1.98	21	503

2.7 Conclusion and Discussions

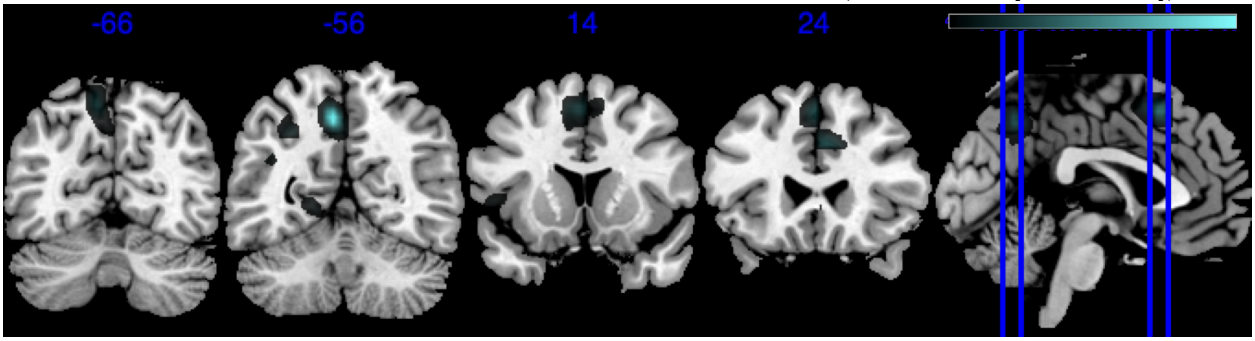
In this paper, we assign soft-thresholded Gaussian process priors on the spatial-varying coefficients in the outcome model and the mediator model. The thresholding parameter controls the sparsity of the functional coefficients, and the soft-thresholded operator provides a continuous mapping from the latent Gaussian process to the sparse coefficients. We extend the mediation analysis framework to incorporate spatial-varying mediators and provide theoretical guarantees on the posterior consistency of the functional natural indirect effect. Our computation approach utilizes the MALA algorithm, which is tailored to the imaging application with block updates.

Through small-scale simulation studies, we compare our method with existing approaches such as the Product Threshold Gaussian prior model (PTG) and the Correlated Selection model (CorS). We demonstrate that our method outperforms existing methods, particularly

Posterior inclusion probability (color range $[0.1, 0.5]$)



Positive posterior mean of the spatial mediation effects (color range $[10^{-5}, 10^{-3}]$)



Negative posterior mean of the spatial mediation effects (color range $[-10^{-4}, -10^{-5}]$)

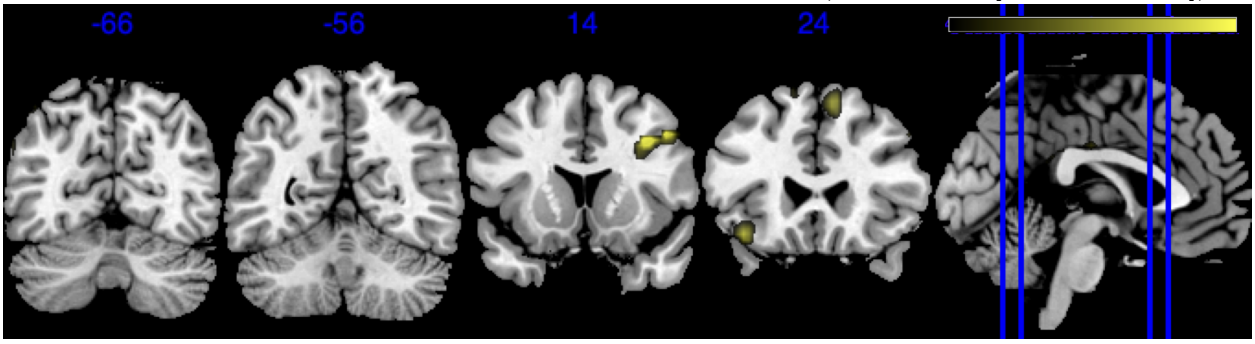


Figure 2.5: Posterior inference on spatially varying indirect effects of parental education on the general cognitive ability that are mediated through the working memory brain activity. The Coronal view slides cutting through 3 of the top 10 regions with largest number of active pixels: the left Precuneus (Precuneus_L), left Inferior parietal gyrus (Parietal_Inf_L) and the left Supplementary motor area (Supp_Motor_Area_L).

in scenarios involving high-dimensional correlated mediators. Furthermore, our implementation is more than four times faster than the CorS method when $N = 1000, p = 4096$. In the larger scale simulation where $p = 4096$, our proposed model performs better when the true signals are smoother and when there is lower variability in the outcome model but higher variability in the mediator model. We also apply our method to mediation analysis on ABCD data, demonstrating its applicability to other mediation problems involving imaging data or other types of high-dimensional mediation data, such as biomarker and genetic data.

However, there are several limitations to our proposed work. While the basis decomposition approach reduces the number of parameters to update, the choice of the appropriate number of basis still depends on the researcher's discretion. The thresholding parameter ν is fixed and determined by the researcher a priori. Further improvements are needed in the sampling algorithm to handle higher-dimensional data, such as 2mm fMRI data with mediators exceeding 2×10^5.

CHAPTER 3

Bayesian Image Regression With Soft-Thresholded Conditional Autoregressive Prior

3.1 Introduction

Regression problems with high-dimensional components have wide-ranging applications. In the context of brain imaging, two prominent regression problems include (i) scalar-on-image (SonI) regression, where the predictor is the image data, and can answer scientific questions such as the impact of brain development on children's IQ score, in the context of ABCD study; and (ii) image-on-scalar (IonS) regression, where the outcome is the image data, and can answer questions such as the impact of the parental education level on different areas of the children's brain development, again in the context of ABCD study. For the imaging component, our primary focus centers on the functional Magnetic Resonance Imaging (fMRI) data, a three-dimensional image commonly used to capture cognitive functions across distinct spatial locations in the human brain. The challenges of involving image data in a regression problem stems from (i) the complex anatomical structure of the human brain, (ii) the low signal-to-noise ratio in fMRI data, and (iii) the computational challenges associated with handling high-dimensional 3D images. In response to these challenges, this paper has two main contributions.

1. We introduce a novel general Soft-thresholded Conditional Autoregressive (ST-CAR) prior, designed to adapt to the correlation structure of the observed data and remain insensitive to user-specified prior correlation structures.

2. We present two variational inference (VI) based algorithms that outperform Gibbs sampler type Markov Chain Monte Carlo (MCMC) algorithms in computational efficiency.

3. The VI implementation of ST-CAR prior retains the inclusion probability that describes how confident we are in the non-null effect of certain voxel, which can be used for downstream uncertainty quantification to control false discoveries.

3.1.1 High-dimensional Regression

Because of the high-dimensional spatially correlated nature of the imaging data, the functional priors for imaging application usually need to be sparse and spatially correlated. For the scalar-on-image (SonI) regression, [92] proposed a frequentist approach where the high-dimensional parameter for the image predictor is penalized by total variation distance, but the smoothness parameters need to be chosen through cross-validation. There are more recent development for high-dimensional regression in the Bayesian regime. [32] and [36] use a combination of Ising prior to control binary selection, and Gaussian Markov Random Field (MRF) to control spatial correlation. In [32], the neighborhood structure in the Gaussian MRF can be learned through the Ising prior, but the correlation among different voxels is only determined by the spatial distance of voxels. [48] proposed a spike-and-slab prior where the binary selection parameter is assigned an Ising prior and the non-zero component is assigned a Dirichlet Process (DP) prior. The Ising prior is used to learn the spatial sparsity, and the DP prior is used to group the effect of active voxels into discrete values. However, the spatial structure of the Ising prior in [48] is assumed to be the same among different voxel pairs. [46] proposed the T-LoHo method for scalar-on-graph regression where the high-dimensional parameter is assigned a tree-based graph partition prior, with the aim of clustering the spatially varying parameter into finite discrete values. [43] proposed a Soft-thresholded prior that is continuous and piecewise smooth with a latent Gaussian process for spatially correlation. But the Soft-thresholded Gaussian Process prior has complex posterior densities for regression problems with normal noise, and the posterior computation can be slow especially for large-scale problems.

For image-on-scalar (IonS) regression where the outcome is a high-dimensional data, the most popular approach is to use low-rank approximation, including using principle component or basis expansion [63, 66], spline function [98, 49] and local polynomial function [110]. A common problem with the frequentist approach of using low-rank approximation is that it can be difficult to make inference on the active area selection when penalizing on the low-rank models, and the choice of the low-rank mapping using basis functions has to reflect the true correlation structure of the high-dimensional data. In the Bayesian regime, [100] proposed a prior composed of latent binary variable for sparsity and latent Gaussian variable for smoothness, however the covariance of the Gaussian variable is pre-fixed. [102] proposed

a machine learning model where the high-dimensional parameter is learned through neural networks, but the focus is on point estimation instead of the joint inference over the entire spatial support.

Moreover, the majority of the above methods, whether based on Gaussian kernel or MRF, the correlation matrix is usually user-defined instead of being learned from the data. Although the smoothness parameters in the Gaussian Kernel or degrees of correlation in MRF can be chosen adaptive to the data through methods like cross-validation, the performance usually depends on the pre-fixed prior correlation structure, and our aim is to propose a prior that is insensitive to the user-specified correlation structure and is able to learn various complex signal patterns. Based on this idea, we develop Soft-threholded conditional autoregressive (ST-CAR) prior, where the Soft-thresholded operator is applied on a latent Gaussian MRF, and the functional parameter can be independent across locations but with spatially-correlated mean. We further develop variational inference algorithms for fast and scalable estimation of the posterior mean.

3.1.2 Approximate Posterior Inference

Posterior sampling for high-dimensional large-scale data set has been challenging for traditional MCMC methods, especially for imaging applications. Variational inference provides an approximation to the posterior mean that avoids sampling of the entire posterior distribution, in exchange for computational efficiency. The main aim of this project is to propose a spatially varying sparse prior that can be applied to various regression problems with imaging component, and develop variational inference algorithms to efficiently obtain the posterior mean estimates for large-scale data set.

As discussed in [9], variational inference techniques approximate the posterior sampling problem by an optimization problem, with the goal to minimize the Kullback-Leibler (KL) divergence between the posterior density and the candidate density function over a family of densities. This allows us to borrow optimization techniques such as stochastic optimization with subsampling, and develop scalable algorithm for massive imaging data. Recent advances in variational inference focus in three directions: (a) scalable algorithms for large scale data [35, 64, 82]; (b) general variational algorithm for more complex models [91]; (c) model-specific applications of variational inference [19, 8]. The stochastic variational inference [35] exploits the global and local exponential family structure for mixture models that are widely applied in topic modeling and genetics applications. The black box variational inference [64] is a general method that estimates the gradient of the evidence lower bound and updates the candidate density by stochastic optimization. Inspired by both stochastic approaches, we

develop a stochastic subsampling variational inference (SSVI) algorithm for the proposed ST-CAR prior.

The ST-CAR prior and its SSVI algorithm can be used as a plug-in method for the functional parameters in general imaging problems such as scalar-on-image regression, image-on-scalar regression, logistic regression with imaging predictor, etc. The ST-CAR prior also has a build-in posterior inclusion probability that can be readily used for signal selection. We present simulation studies on scalar-on-image regression as an example to demonstrate the fast and scalable performance of the SSVI compared to Gibbs sampler, Coordinate Ascent variational inference (CAVI), and the Soft-thresholded prior updated through Metropolis-adjusted Langevin algorithm (MALA).

The rest of the article is structured as follows. In Section 3.2, we introduce ST-CAR prior and its application to the scalar-on-image and image-on-scalar regression models. In Section 3.3, we propose two variational inference algorithms, coordinate ascent variational inference (CAVI) and stochastic subsampling variational inference (SSVI) algorithms, for ST-CAR prior on SonI and IonS regressions. In Section 3.4, we demonstrate the performance of our proposed method using various simulation settings, and compare with existing methods. In Section 3.5, we apply our method to the Adolescent Brain Cognitive Development (ABCD) study and conclude the paper in Section 3.6.

3.2 ST-CAR prior

3.2.1 General notations

Let $\mathrm{N}(\mu, \sigma^2)$ represent a normal distribution with mean μ and variance σ^2. For the index set $\{1, \ldots, p\}$, let $[-j]$ denote the set $\{1, \ldots, p\} \setminus \{j\}$. For a square matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and the largest eigenvalues of A respectively. Let $C^+(A)$ denote the half Cauchy distribution with density function $f(x) = \frac{2}{\pi A} \frac{1}{1+x^2/A^2} I(x \geq 0)$. Let $\mathbb{R}^q$ denote the q-dimensional Euclidean space. I_q is the q by q dimensional identity matrix. $\mathrm{IG}(a, b)$ stands for the inverse-gamma distribution. We use $\mathcal{N}(j)$ to denote a neighborhood index set for j-th component, and $|\mathcal{N}(j)|$ to denote the cardinality of $\mathcal{N}(j)$. For a vector $\boldsymbol{a}$, $\mathrm{diag}\{\boldsymbol{a}\}$ is used to denote the diagonal matrix with diagonal vector $\boldsymbol{a}$. We use $\mathrm{sgn}(x) = I(x > 0) - I(x < 0)$ to denote the sign of x.

3.2.2 ST-CAR prior

We use a similar idea as the soft-thresholded Gaussian Process (STGP) prior [43], and apply a soft-thresholding operator on a latent spatially correlated process. Define the soft-

thresholding operator as $T_\nu(x) := \{x - \text{sgn}(x)\nu\}I(|x| > \nu)$ for any $\nu \geq 0$. We propose a new noisy version of the STGP, referred to as the Soft-thresholded conditional auto-regressive (ST-CAR) prior.

Definition 5. *A sparse, spatially correlated parameter $\beta(s)$ on a fixed grid $s_1, \ldots, s_p$ follows the ST-CAR prior if*

$$\begin{aligned}
&\beta(s_j) \overset{ind}{\sim} \mathrm{N}(T_\nu(\mu_j), \sigma_\beta^2), \quad j = 1, \ldots, p \\
&\mu_j \mid \mu_{\mathcal{N}(j)} \sim \mathrm{N}(\bar{\mu}_{\mathcal{N}(j)}, \tau_{\mu,j}^2), \quad \bar{\mu}_{\mathcal{N}(j)} = \rho_j \sum_{k \in \mathcal{N}(j)} b_{j,k}\mu_k
\end{aligned}$$

Let $\boldsymbol{\beta} = (\beta(s_1), \ldots, \beta(s_p))^{\mathrm{T}}$. We use $\boldsymbol{\beta} \sim$ ST-CAR(ν, B) to denote $\boldsymbol{\beta}$ follows the ST-CAR prior with thresholding parameter ν and the neighborhood matrix B where $(B)_{j,k} = b_{j,k}$.

Here, β is the target spatially varying parameter. The prior mean of β is the soft-thresholded μ, where μ is the latent spatially varying process. The prior mean of μ_j is determined by the correlation coefficient ρ_j, the neighborhood set $\mathcal{N}(j)$, and the neighborhood weights $b_{j,k}$.

The variance parameter σ_β is not identifiable when the ST-CAR prior is applied to SonI or IonS regression. In order to impose sparsity on $\boldsymbol{\beta}$, we use the annealing idea on σ_β, and let σ_β decays to 0 as the iteration increases. This will force the spatially independent $\boldsymbol{\beta}$ to converge to the sparse and spatially correlated $\boldsymbol{\mu}$. The decay rate of σ_β has impact on the variable selection accuracy especially for low signal-to-noise ratio data. A general rule of thumb is to set σ_β relatively large at the beginning to allow for more flexibility and decays to a small value at the end.

The correlation parameter ρ_j can either be pre-fixed at all locations, or updated by

$$\rho_j = \delta_j \tilde{\rho}, \quad \delta_j \sim \mathrm{Ber}(p_j),$$

where $\tilde{\rho}_j$ is pre-fixed. We find these two approaches to have similar result in terms of variable selection accuracy, and the ST-CAR prior is not very sensitive to the choice of ρ or the bandwidth $|\mathcal{N}(j)|$. However, taking the second approach to adaptively update $\boldsymbol{\rho}$ can give us extra information on the correlation structure of the high-dimensional coefficient.

The ST-CAR prior applies the soft-thresholding operator T_ν to a latent process μ_j, and the spatially-correlated structure of μ_j is imposed by setting its mean to a weighted average over a neighborhood $\mathcal{N}(j)$. By applying a binary indicator δ_j to the correlation parameter ρ_j, we are able to adaptively determine whether the value of μ_j is strongly correlated with its neighborhood mean. We adopt the Conditional Auto-Regressive (CAR) covariance structure

[27]. Define a matrix B and a diagonal matrix D_{σ_μ} as

$$(B)_{j,k} = b_{j,k} = \frac{w_{j,k}}{w_{j+}}, \quad (D_{\sigma_\mu})_{j,j} = \tau_{\mu,j} = \frac{\sigma_\mu^2}{w_{j+}} \tag{3.1}$$

where $w_{j,k}$ are the (j,k)-th index of a symmetric matrix W, and $w_{j+} = \sum_{k=1}^{p} w_{j,k}$ is the row(column) summation. The matrix W represents the correlation structure, and in practice we set $w_{j,k} \propto \exp\{-d(s_j, s_k)\}$, exponentially negatively associated with the distance between s_k and s_j. In addition to the CAR structure, we set a bandwidth $|\mathcal{N}(j)|$ on the number of components included in the neighborhood $\mathcal{N}(j)$, such that for each j, $b_{j,k}$ is nonzero only if $w_{j,k}$ is within the first $|\mathcal{N}(j)|$ largest values among $\{w_{j,k}\}_{k=1}^{p}$. Denote $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)^{\mathrm{T}}$, the joint density of $\boldsymbol{\mu}$ takes the form

$$f(\boldsymbol{\mu}|\sigma_\mu) \propto \exp\left[-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}} D_{\sigma_\mu}^{-1}\{I - \mathrm{diag}(\boldsymbol{\rho})\, B\}\,\boldsymbol{\mu}\right]$$

If we denote $\Sigma_\mu^{-1} := D_{\sigma_\mu}^{-1}\{I - \mathrm{diag}(\boldsymbol{\rho})\, B\}$, it can be shown that Σ_μ^{-1} is a symmetric, positive definite matrix when each $\rho_j \in (\lambda_{\min}(B)^{-1}, \lambda_{\max}(B)^{-1})$. Note that B is not a symmetric matrix in general. In the construction in (3.1), $\lambda_{\max}(B) = 1$ and $\lambda_{\min}(B) < 0$. Hence we choose $\rho_j \in [0, 1)$ for any j, and the joint density of $\boldsymbol{\mu}$ is guaranteed to be non-degenerative. One caveat of doing so is that the value of μ_j and the neighborhood mean $\bar{\mu}_{\mathcal{N}(j)}$ is only allowed to be either positively correlated or independent ($\rho_j \geq 0$), but the negative correlation is not taken into consideration. This constraint makes sense in brain imaging applications, because the true signal is assumed to be sparse and piecewise smooth, which excludes the case where the signal across neighboring voxels has a sharp drop from positive to negative values. In general, the ST-CAR prior is suitable for the case where the positive and negative areas do not share boundaries.

The proposed ST-CAR prior enjoys good computational properties as it has a conditional conjugate posterior when applied to a parameter in a regression problem. The main challenge in updating a thresholded parameter is that the thresholding function such as T_ν is a non-linear function. But we can show that μ_j conditional on all other $\mu_{[-j]}$ and β has a mixture of truncated normal distribution as its posterior.

Proposition 2. *Within the ST-CAR prior, the posterior of μ_j can be expressed as a mixture of three truncated normal distributions.*

$$\begin{aligned}&\pi(\mu_j \mid \boldsymbol{\beta}, \mu_{[-j]}, \sigma_\mu, \sigma_\beta) = \\ &P_j^+ \cdot \mathrm{N}_{[\nu,+\infty)}(\mu_j^+, V_j) + P_j^0 \cdot \mathrm{N}_{[-\nu,\nu]}(\bar{\mu}_{\mathcal{N}(j)}, V_0) + P_j^- \cdot \mathrm{N}_{(-\infty,-\nu]}(\mu_j^-, V_j)\end{aligned} \tag{3.2}$$

The expression for $P_j^+, P_j^0, P_j^-, \mu_j^+, \mu_j^-, V_j, V_0$ can be found in the proof of Proposition 2 in the Supplementary.

The proposed ST-CAR prior is a general prior that can be applied to many high-dimensional regression settings, where the coefficient is assumed to be smooth and sparse across their spatial domain. Here, we use scalar-on-image (SonI) and image-on-scalar (IonS) regressions as two examples to illustrate the power of ST-CAR prior. Other potential applications including logistic regression with high-dimensional exposure and other types of generalized linear models.

3.2.3 Application to scalar-on-image (SonI) model

The ST-CAR prior can be applied to various models with sparse and spatially-varying functional parameters. In this section, we demonstrate its advantage using the scalar-on-image (SonI) regression model.

Let $M_i(s_j)$ denote the image intensity at location s_j for individual i, $\mathbf{X}_i \in \mathbb{R}^q$ be a vector-valued confounder variables. Let Y_i denote the scalar-valued outcome for subject i. $i = 1, \ldots, n, j = 1, \ldots, p$.

$$\begin{aligned} Y_i &= \sum_{j=1}^{p} \beta(s_j) M_i(s_j) + \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{X}_i + \epsilon_i \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_Y^2) \\ \boldsymbol{\beta} &\sim \text{ST-CAR}(\nu, B), \quad \boldsymbol{\gamma} \sim \mathrm{N}(0, \sigma_\gamma^2 I_q) \\ \sigma_Y &\sim C^+(1), \quad \sigma_\gamma \sim C^+(1) \end{aligned} \tag{3.3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the high-dimensional spatially-varying coefficient of interest, and $\boldsymbol{\gamma} \in \mathbb{R}^q$ is the vector-valued coefficient for the confounders $\mathbf{X}_i$.

For all of the Half-Cauchy parameters, we use their equivalent conjugate form to update: $\sigma_Y \sim C^+(1)$ is equivalent to $\sigma_Y^2 \sim \mathrm{IG}(1/2, 1/a_Y), a_Y \sim \mathrm{IG}(1/2, 1)$. Because $\boldsymbol{\beta} \sim \text{ST-CAR}(\nu, B)$ essentially assigns spatially independent prior to $\boldsymbol{\beta}$ with varying mean function, we can use singular value decomposition (SVD) on the design matrix $M \in \mathbb{R}^{n \times p}$ to further boost the computation speed[1].

3.2.4 Application to image-on-scalar (IonS) model

The second application we consider is the image-on-scalar (IonS) regression. The spatially varying outcome is denoted as $M_i(s_j)$ for individual $i = 1, \ldots, n$ and location $s_j, j = 1, \ldots, p$.

[1]Details on this derivation can be found in the Appendix

The exposure of interest is denoted as X_i, and the confounder is denoted as $\mathbf{C}_i \in \mathbb{R}^m$. The IonS model is as follows

$$
\begin{aligned}
&M_i(s_j) = \alpha(s_j)X_i + \sum_{k=1}^{m} \xi_k(s_j)C_{i,k} + \eta_i(s_j) + \epsilon_{i,j}, \quad \epsilon_{i,j} \overset{\text{iid}}{\sim} N(0, \sigma_M^2) \\
&\boldsymbol{\alpha} \sim \text{ST-CAR}(\nu, B), \quad \xi_k \overset{\text{iid}}{\sim} \text{GP}(0, \sigma_\xi^2\kappa), k = 1, \ldots, m, \quad \eta_i \overset{\text{iid}}{\sim} \text{GP}(0, \sigma_\eta^2\kappa), i = 1, \ldots, n, \\
&\sigma_M \sim C^+(1), \quad \sigma_\xi \sim C^+(1), \quad \sigma_\eta \sim C^+(1).
\end{aligned} \tag{3.4}
$$

Here, we only assign ST-CAR to α for selecting active region for the exposure. For confounder coefficients ξ_k and the individual effects η_i, we assign Gaussian Process prior with the same kernel function κ for computational convenience. The individual effect parameter η_i separates the spatially correlated noise from the noise term ϵ_i, and avoids setting a dense correlation matrix for the noise term ϵ_i, which speeds up the computation. This is similar to the correlated noise model in [110]. The identifiability of model (3.4) has been shown in [102] under the following sufficient conditions: (1) the design matrix $\tilde{\mathbf{X}} := (\mathbf{X}, \mathbf{C}) \in \mathbb{R}^{n\times(m+1)}$ is a full rank matrix, (2) for any i and any s_j, denote $\boldsymbol{\eta}(s_j) = (\eta_1(s_j), \ldots, \eta_n(s_j)) \in \mathbb{R}^n$, $\tilde{\mathbf{X}}^T\boldsymbol{\eta}(s_j) = 0$. The first condition is easily satisfied when the design matrix $\tilde{\mathbf{X}}$ is not linearly dependent.

For the Gaussian Process prior update of ξ_k and η_i, we use the basis decomposition approach. Leveraging Mercer's theorem, which asserts that for any function $g(s)$ following a Gaussian Process with mean zero and covariance function $\sigma_g^2\kappa(\cdot,\cdot)$, we can utilize the following basis decomposition.

$$
g(s) = \sum_{l=1}^{\infty} \theta_{g,l}\phi_l(s), \quad \theta_{g,l} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_g^2\lambda_l),
$$

where λ_l is the l-th eigen-value, and ϕ_l is the l-th eigen-function (see Section 4.2 in [65]). In practice, we choose a finite L as the cutoff on the number of basis, and approximate $g(s)$ by $\sum_{l=1}^{L} \theta_{g,l}\phi_l(s)$. The number of basis L is chosen such that the summation $\sum_{l=1}^{L} \lambda_l$ is over 90% of $\sum_{l=1}^{p} \lambda_l$. The choice of the kernel function includes exponential square kernel, Matérn kernel and other kernel functions. For the simulation section we use the modified exponential square kernel, $\kappa(s, s'; a, b) = \exp\{-a(s^2 + s'^2) - b(s - s')^2\}$. For the real data analysis with ABCD data, the kernel is a pre-tuned Matérn kernel with region-specific smoothness parameters that can best align with the empirical correlation of the observed image data, same as in [97].

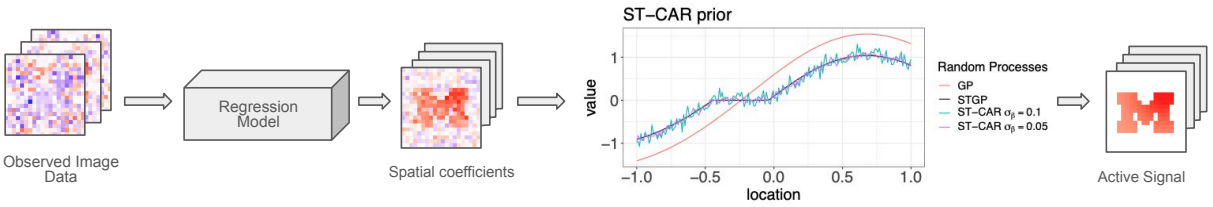


Figure 3.1: Illustration to use ST-CAR prior for regression models with imaging data.

3.3 Posterior Computation

In this section, we use the scalar-on-image regression (3.3) as an example, and introduce three algorithms to estimate the posterior of β: (a) Gibbs sampler (Gibbs), (b) Coordinate Ascent Variational Inference (CAVI), and (c) Stochastic subsampling version of variational inference (SSVI).

Proposition 2 provides the closed-form posterior density for the sparse-mean latent parameter $\{\mu_j\}_{j=1}^p$ as a mixture of 3 truncated normal distributions in the ST-CAR prior of β. All other parameters in this hierarchical model have conjugate posteriors, and Gibbs sampler can be directly applied.

For the neighborhood matrix B in ST-CAR(ν, B), in order to speed up the computation, we use sparse matrix structure in RcppArmadillo [22], and set a fixed bandwidth $|\mathcal{N}(j)|$ for all j. For a given fixed grids $\{s_1, \ldots, s_p\}$ in $\mathbb{R}^d$, we use RANN package [2] to efficiently search for the nearest neighbors in high-dimensional setting.

3.3.1 Coordinate Ascent Variational Inference (CAVI)

The variational inference methods (CAVI, SSVI) are based on the mean-field assumption [9]. If we denote $\theta = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma_Y, \sigma_\gamma)$ as the collection of all parameters. The mean-field variational inference minimizes the evidence lower bound

$$\min_q \mathbb{E}\left[KL(q(\theta) \mid p(\theta|\mathbf{Y}, \mathbf{M}, \mathbf{X}))\right] \quad s.t.\ q(\theta) = q(\boldsymbol{\beta})q(\boldsymbol{\gamma})q(\boldsymbol{\mu})q(\sigma_Y)q(\sigma_\gamma)$$

The conventional Coordinate Ascent Variational Inference (CAVI) algorithm iteratively refines the approximated density q by updating each parameter in successive iterations.

$$\log q^{(t)}(\boldsymbol{\beta}) \propto \mathbb{E}_{q^{(t-1)}(\boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma_Y, \sigma_\gamma)}\left\{\log p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{M}, \mathbf{X}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma_Y, \sigma_\gamma)\right\}$$

Because each parameter in the full hierarchical model has closed-form posterior density, we can directly apply this iterative approach.

One issue with the conventional CAVI is that although it can give a good point estimation as an optimization algorithm, but cannot directly give inference results such as the credible interval, compared with MCMC sampling methods. The novelty in our proposed ST-CAR prior is that we can use the mixing probability in Proposition 2 as the uncertainty quantification measure for selecting significant regions, circumventing the requirement for credible interval based on MCMC samples, while leveraging the computational efficiency provided by CAVI. Proposition 2 gives the posterior probability of μ_j belonging to the positive group $[\nu, \infty)$, zero group $[-\nu, \nu)$, and negative group $(-\infty, -\nu)$. When using CAVI, we can directly compute the Posterior Inclusion Probability (PIP) under q density as $(P_j^+ + P_j^-)$ in (3.2) as a measure of coefficient significance.

3.3.2 Stochastic subsampling variational inference (SSVI)

To make the variational inference method scalable for large data set, we propose a stochastic subsampling version of CAVI, referred as SSVI. The main computational bottleneck of CAVI is to update $\boldsymbol{\beta}$, which is a high-dimensional parameter, and the latent variable $\boldsymbol{\mu}$ further requires complex computation of mixed truncated normal densities. Hence given $\boldsymbol{\mu}$, when updating $\boldsymbol{\beta}$, we randomly select a subsample of data, indexed by $I \subset \{1, \ldots, n\}$, and apply a stochastic gradient update similar to the Stochastic Gradient Langevine Dynamics (SGLD) [93]. Let s_t be the step size at t-th iteration, n be the total number of observations, n_s be the subsample size, and π be the prior density of β_j at the jth voxel,

$$\mathbb{E}_{q^{(t)}}\{\beta_j\} \leftarrow \mathbb{E}_{q^{(t-1)}}\{\beta_j\} + s_t \left(\frac{n}{n_s} \nabla \mathbb{E}_{q^{(t-1)}} \log \sum_{i \in I} p(Y_i, \mathbf{M}_i, \mathbf{X}_i \mid \theta) + \nabla \mathbb{E}_{q^{(t-1)}} \log \pi(\beta_j) \right).$$

This is because under the mean-field assumption, the optimum density $q^*(\beta_j)$ has a closed-form solution: a normal density with mean and variance

$$\begin{aligned}
\mathbb{E}_{q^*}(\beta_j) &= \mathrm{Var}_{q^*}(\beta_j) \times \\
&\quad \left[\mathbb{E}_{q^*}(\sigma_Y^{-2}) \sum_{i=1}^{N} M_{i,j} \left(Y_i - \mathbb{E}_{q^*} \boldsymbol{\gamma}^T \mathbf{X}_i - \sum_{k \in [-j]} \mathbb{E}_{q^*} \beta_k M_{i,k} + \mathbb{E}_{q^*} \left\{ \sigma_\beta^{-2} T_\nu(\mu_j) \right\} \right) \right] \\
\mathrm{Var}_{q^*}(\beta_j) &= \left(\mathbb{E}_{q^*}(\sigma_Y^{-2}) \sum_{i=1}^{N} M_{i,j}^2 + \mathbb{E}_{q^*}(\sigma_\beta^{-2}) \right)^{-1}
\end{aligned}$$

And $\mathbb{E}_{q^*}(\beta_j)$ is also the maximizer to

$$\mathbb{E}_{q^*} \sum_{i=1}^{N} \log p(Y_i, \mathbf{M}_i, \mathbf{X}_i \mid \theta) + \mathbb{E}_{q^*} \log \pi(\beta_j).$$

We require the step size s_t to decrease to 0 as $t \to \infty$. In practice, we use the decay function $s_t = a(b+t)^{-\gamma}$, as suggested in [93].

In practice, we find that in low signal-to-noise ratio (SNR) settings, the CAVI algorithm gives better accuracy. Hence we recommend to use CAVI for SonI model, where the SNR can be very low especially in brain imaging data, and to use SSVI for IonS model, since IonS model has much higher SNR for the coefficient at each voxel.

3.4 Numerical Examples

In this section, we will present the simulation results for SonI (3.3) and IonS (3.4) regressions. Our main goal is to compare the proposed prior ST-CAR with other existing methods, and we will use CAVI as the main algorithm for estimating the spatially varying parameters. This is because Gibbs is usually very slow, and SSVI as a stochastic method tends to be less accurate for low signal-to-noise ratio case, whereas CAVI balances between computational efficiency and selection accuracy, and has overall the best performance. We include a section in the Supplementary that compares the performance of Gibbs, SSVI and CAVI.

3.4.1 Simulation I: Scalar-on-image regression with CAVI

For SonI model (3.3), we compare ST-CAR with 3 other methods: (1) Soft-thresholding Gaussian Process prior (STGP) [43], (2) T-LoHo [46], (3) Elastic Net [112].

For the elastic net result implemented in the glmnet package [25], the mixing parameter α is set to 0.5, and the penalty parameter λ is chosen using cross-validation.

The STGP prior is based on soft-thresholding on the latent Gaussian Process. When $\beta(s) \sim \mathcal{STGP}(\nu, \kappa)$, there exists a corresponding latent Gaussian Process $\tilde{\beta}(s) \sim \mathcal{GP}(0, \kappa)$ such that $\beta(s) = T_\nu(\tilde{\beta})$. This method requires a pre-specified kernel function κ, and the posterior sampling algorithm is Metropolis-adjusted Langevin algorithm (MALA). In this simulation we use the exponential square kernel

$$\kappa(s, s'; a, b) = \text{cor}\{\beta(s), \beta(s')\} = \exp\{-a(s^2 + s'^2) - b(s - s')^2\} \tag{3.5}$$

where $a = 0.01, b = 10$. The implementation is based on BIMA package [2], first developed for [97]. Note that this implementation of STGP allows the users to specify different regions in the image and specify a region-wise independent kernel in order to speed up the computation in high dimensions and boost selection accuracy in each region. Hence for the simulation pattern shown in Figure 3.2, we evenly split the entire 2D region into 4 sub-regions, and use the modified exponential square kernel on each sub-region. The basis function is generated using [42] with 10 degrees of Hermite polynomials for each sub-region. We use the elastic net result as the initial values for $\boldsymbol{\beta}$, and run a total of 10^4 iterations with the last 20% as the converged MCMC sample. The thresholding parameter ν is set to be 0.2. For the variable selection accuracy, we use the Posterior Inclusion Probability (PIP) based on the MCMC sample of $\boldsymbol{\beta}$, defined as $\text{PIP}_j = \sum_{t=1}^{T} I(\beta_j \neq 0)/T$ for the location j with T MCMC sample.

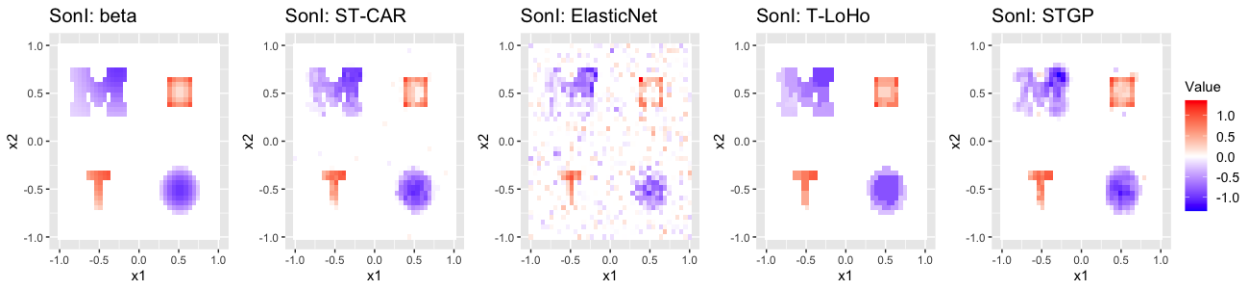
The T-LoHo method is designed for clustering nodes in graph models into finite discrete values, and it shows great performance for this purpose especially under low SNR. However, this method has several limitations when applied to SonI problem with continuous functional value. As a clustering algorithm, T-LoHo can find the active areas accurately, but cannot threshold the small values to 0. For the spatially smooth patterns, T-LoHo can only group them into a few discrete values instead of capturing the smooth transition. In addition, the implementation in the TLOHO package does not provide voxel-level uncertainty quantification measure such as PIP. When comparing the variable selection result, we use the 95% credible interval: β_j is significant only if the 95% CI of β_j does not contain 0. We use the R package on Github for implementation of T-LoHo [3]. This package does not provide the confounder coefficients estimation, hence for the SonI simulation, we set true $\boldsymbol{\gamma} = 0$. We use a total of 50000 MCMC iterations and take the last 10000 as the converged sample.

For ST-CAR prior updated using CAVI algorithm, we use ridge regression result as the initial value for $\boldsymbol{\beta}$, and set the initial value for $\boldsymbol{\mu}$ to be all 0. The thresholding parameter ν is set to be the largest marginal value in $\boldsymbol{\beta}$ estimated from ridge regression. This is because setting ν to be a large value can reduce false discoveries, and $\boldsymbol{\mu}$ is still able to recover the true signal pattern even when starting from all 0 initial values. This algorithm is much less sensitive to the thresholding parameter compared to STGP. The decay rate of σ_β^2 is set to be $0.5(1+t)^{-0.7}$ where t represents the number of iterations. Since we use annealing on σ_β^2 instead of fully conjugate update, we can no longer use ELBO as a stopping rule. Instead, at the $t+1$ iteration, we compute the difference of $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\beta}^{(t+1)}$, defined as $\sum_{j=1}^{p}(\beta_j^{(t)} - \beta_j^{(t+1)})^2/p$, to determine whether the optimization has converged. The tolerance is set to be 10^{-10}. For the neighboring matrix B, we set the number of neighbors as 8, and

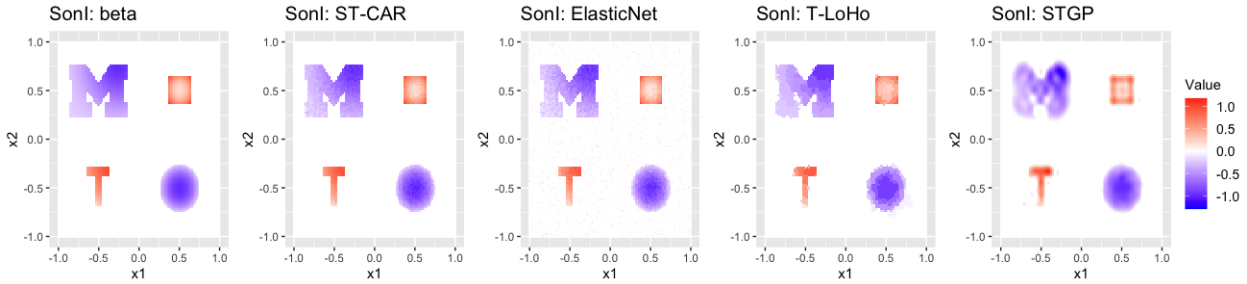
[2] BIMA package https://github.com/yuliangxu/BIMA

[3] TLOHO package https://github.com/changwoo-lee/TLOHO

the correlation parameter $\tilde{\rho}$ is set to be 0.9. The variance parameter σ_μ is fixed at 1 for CAVI update.



(a) Case 1. $n = 600$, $p = 1600$, $\sigma_Y^2 = 1$.



(b) Case 5. $n = 6000$, $p = 10000$, $\sigma_Y^2 = 5$.

Figure 3.2: SonI result illustration for all competing methods. The first figure in each row is the true β signal.

Figure 3.2a and 3.2b provide visual comparison under two simulation settings. The true $\boldsymbol{\beta}$ image is designed to include several challenging patterns where the active area can decay smoothly to almost 0, has complex correlation structure such as the M-shape on the top-left corner, and includes both positive and negative patterns. Case 1 (Figure 3.2a) is the low resolution and low SNR setting, and Case 5 (Figure 3.2b) is the high resolution and high SNR setting.

From this visual comparison, STGP is good at estimating smooth function pattern such as the bottom-left circle, but without further tuning the Gaussian Process kernel, estimating more complex pattern such as the M-shape would be difficult. Here, STGP already takes the region partition into consideration. If we were to generate Gaussian Kernel over the entire support, the result would be more smooth without further tuning the kernel. T-LoHo as a clustering method is good at grouping larger effects together, but as the true signal decays smoothly towards 0, T-LoHo can ignore some small non-zero effects, resulting in a lower statistical power, as shown in the bottom-left circle in Figure 3.2b. Elastic Net can identify the spatial pattern to a certain extent, but its efficacy is limited as it does not leverage

correlation information. Consequently, it may yield a noisy estimation in case 1. Even in case 5, where the point estimation is favorable, Elastic Net can still introduce background noises. ST-CAR can estimate each pattern relatively well without specifying any region partition or tuning the correlation matrix adaptive to different signal patterns. Although some small effects such as the bottom tip of the T-shape can still be missed, ST-CAR provides the best overall performance compared to other priors across different settings without any tuning procedure.

Table 3.1 provides the detailed numerical comparison. The evaluation criteria for estimation accuracy includes (i) Selection accuracy: false discovery rate (FDR), true positive rate (TPR) and overall accuracy (ACC); (ii) Point estimation: root mean squared error (RMSE); (iii) Goodness-of-fit: the predictive mean squared error on the outcome Y_i using training and testing data (train and test pMSE). We also include the computational time comparison averaged over 100 replications. Note that because CAVI is an optimization algorithm and we are able to set a stopping rule, whereas for MCMC sampling algorithms for STGP and T-LoHo, a lot more number of iterations is required. Hence we report both the total time and the number of iterations per second.

For the variable selection result for ST-CAR, Elastic Net and STGP, we use a tuning procedure to find a cut such that the FDR can be controlled blow 10% within a fixed tuning window. For STGP and ST-CAR, the PIP is used to control FDR. For elastic net, $\boldsymbol{\beta}$ is used to control FDR. For T-LoHo, the 95% CI is used without tuning.

Based on Table 3.1, we can see that ST-CAR has the lowest testing pMSE in 3 relatively high SNR cases (Case 2,3,5). For Case 1 and 4 with relatively low SNR, ST-CAR has the second best performance next to T-LoHo. For the computation time in Table 3.1b, ST-CAR has the shortest running time in all settings.

3.4.2 Simulation II: Image-on-scalar regression with SSVI

For IonS model (3.4), we compare ST-CAR with 3 other methods: (1) STGP prior, (2) Scalable Bayesian Image-on-Scalar regression (SBIOS) [96], (3) Mass Univariate Analysis (MUA). For the IonS regression (3.4), estimation of α has a larger SNR compared to estimating β in SonI (3.3), hence we use SSVI for this application for ST-CAR prior. Because we impose GP prior for the confounder parameters ξ_k and individual effect η_i, the GP kernels used in this simulation are all the same for STGP, ST-CAR and SBIOS for fair comparison. We also use region-wise independent kernels for the GP priors in (3.4). The GP kernel is the same as (3.5) with $a = 0.01$ and $b = 10$.

The mass univariate analysis (MUA) is one of the most commonly used method for IonS

Table 3.1: Numeric result for SonI simulation, under 100 replications.

(a) SonI: Comparison of estimation accuracy. The evaluation criteria for estimation includes false discovery rate (FDR), true positive rate (TPR), overall accuracy (ACC), and root mean squared error (RMSE), all multiplied by 100. The evaluation criteria for predictive performance includes training and testing predictive MSE, denoted as Train and Test pMSE respectively

Case 1. $n = 600, p = 1600, \sigma_Y^2 = 1$					Case 2. $n = 600, \boldsymbol{p = 900}, \sigma_Y^2 = 1$				
	ST-CAR	ElasNet	STGP	T-LoHo		ST-CAR	ElasNet	STGP	T-LoHo
FDR	9.52	9.54	10.42	4.41	FDR	9.40	1.82	9.77	3.92
TPR	97.47	44.85	95.37	97.32	TPR	99.81	98.45	94.06	99.34
ACC	98.05	90.87	97.60	98.62	ACC	98.45	99.50	97.63	98.65
RMSE	11.00	30.45	13.44	9.46	RMSE	5.51	7.12	13.00	6.56
Train pMSE	2.29	1.03	5.97	3.61	Train pMSE	1.10	0.33	3.40	1.55
Test pMSE	7.18	60.31	12.54	**6.21**	Test pMSE	**2.02**	3.04	6.36	2.24
Case 3. $\boldsymbol{n = 1000}, p = 1600, \sigma_Y^2 = 1$					Case 4. $n = 600, p = 1600, \boldsymbol{\sigma_Y^2 = 5}$				
	ST-CAR	ElasNet	STGP	T-LoHo		ST-CAR	ElasNet	STGP	T-LoHo
FDR	9.39	0.39	9.89	1.10	FDR	9.64	9.42	9.83	7.09
TPR	100.00	99.08	97.93	99.79	TPR	90.06	37.11	92.00	94.19
ACC	98.42	99.80	98.05	99.76	ACC	97.02	89.82	97.25	97.09
RMSE	4.27	6.53	12.04	5.83	RMSE	14.65	33.31	14.72	12.51
Train pMSE	1.08	0.48	6.47	1.92	Train pMSE	4.85	2.55	9.17	8.56
Test pMSE	**2.03**	3.79	9.92	2.78	Test pMSE	17.76	76.84	19.31	**14.37**
Case 5. $\boldsymbol{n = 6000, p = 10000, \sigma_Y^2 = 5}$					Case 6. $\boldsymbol{n = 6000, p = 10000, \sigma_Y^2 = 10}$				
	ST-CAR	ElasNet	STGP	T-LoHo		ST-CAR	ElasNet	STGP	T-LoHo
FDR	0.46	0.27	17.16	6.84	FDR	1.18	2.47	17.15	4.78
TPR	99.98	99.39	98.52	99.67	TPR	99.91	98.21	98.53	99.52
ACC	99.92	99.86	96.57	95.96	ACC	99.80	99.33	96.58	97.71
RMSE	3.73	6.73	13.78	5.25	RMSE	3.14	3.94	6.44	2.48
Train pMSE	5.48	3.12	80.90	12.78	Train pMSE	10.62	3.81	85.91	18.60
Test pMSE	**9.37**	22.94	86.88	15.54	Test pMSE	**18.43**	41.06	92.25	22.44

(b) Computation time for SonI simulation, averaged over 100 replications.

Computation time	Total time (seconds)			Number of iteratios per second		
Case	ST-CAR	STGP	T-LoHo	ST-CAR	STGP	T-LoHo
Case 1. $n = 600, p = 1600, \sigma_Y^2 = 1$	103.0	503.0	306.5	11.4	208.0	262.6
Case 2. $n = 600, \boldsymbol{p = 900}, \sigma_Y^2 = 1$	24.2	250.8	205.0	42.3	420.5	393.4
Case 3. $\boldsymbol{n = 1000}, p = 1600, \sigma_Y^2 = 1$	111.2	866.2	426.2	10.2	122.9	189.5
Case 4. $n = 600, p = 1600, \boldsymbol{\sigma_Y^2 = 5}$	108.5	486.0	312.5	11.1	212.4	259.8
Case 5. $\boldsymbol{n = 6000, p = 10000, \sigma_Y^2 = 5}$	8034.9	40658.8	11141.1	0.2	2.6	7.3
Case 6. $\boldsymbol{n = 6000, p = 10000, \sigma_Y^2 = 10}$	7811.3	40839.1	11297.8	0.2	2.6	7.2

regression. MUA analyzes IonS as a spatially independent problem, and treats the IonS regression as p independent linear regression problems with exposure X_i and confounders $\mathbf{C}_i$. To select active voxels, we use the Benjamini-Hochberg adjusted p-values [7] to control the false discovery rate. The active voxels selected by MUA have an adjusted p-value below 0.05.

The STGP method is similar to what has been discussed in the SonI regression. For IonS

regression, we use a total of 2×10^4 iterations and take the last 10% as the converged MCMC sample. The thresholding parameter ν is set to be 0.2. We use the point estimates of α and ξ_k from MUA as the initial value for the MALA algorithm.

The Scalable Bayesian Image-on-Scalar regression (SBIOS) [96] is another Bayesian approach where the parameter of interest can be expressed as $\alpha(s) = \tilde{\alpha}(s)\delta(s)$. The latent spatially smooth function $\tilde{\alpha}$ is assigned a GP prior, and the binary selection variable $\delta(s_j)$ is assigned an independent prior $\text{Ber}(p(s_j))$ for each location s_j. SBIOS is designed to analyze a large scale data set by using batch update with stochastic gradient Langevin dynamics algorithm (SGLD). Hence it is more appropriate to be compared with the SSVI implementation of ST-CAR, since both methods are based on stochastic gradient updates of a small random sample drawn from the entire observed data. Different from the idea of SSVI where we simply use stochatic gradient update for an optimization problem, SGLD gives a smooth transition from optimization to MCMC sampling as the step size decays to 0 [93]. Similar to STGP, we can use the MCMC sample of $\delta(s_j)$ to determine the PIP at location j, $\text{PIP}_j = \sum_{t=1}^{T} \delta(s_j)^{(t)} \neq 0/T$ for T MCMC sample. In the simluation, we use 5000 SGLD iterations, with the decay function of the step size set as $s_t = 0.0001 \cdot (10 + t)^{-0.35}$. We use 200 subsample in each iteration. The prior for $\delta(s_j)$ is set to be Ber(0.5) for all locations. The last 20% of iterations is used to compute the point estimation of α and PIP.

The ST-CAR method implemented using SSVI algorithm requires a stochastic gradient update of α. We use a step of 10^{-4} and a subsample of 100 for the SGD optimization. The decay rate function for σ_α^2 is $(1 + t)^{-0.4}$. We use $C^+(1)$ as the prior for σ_μ in (3.1). Because of the randomness in the SGD update, we cannot use the difference between $\alpha^{(t)}$ and $\alpha^{(t+1)}$ or ELBO as a stopping rule. For the simulation, we simply run 10^4 iterations. In practice, the convergence of SSVI can be roughly determined by the convergence of σ_μ^2. For the point estimation and inference of α, we use the averaged values over the last 20% iterations as the posterior mean of α and PIP in order to avoid the randomness from SGD.

Note that updating the individual effects η_i for $i = 1, \ldots, n$ is computational challenging for all Bayesian methods. We choose to update η_i every 100 iterations for ST-CAR, SBIOS, and every 1000 iterations for STGP.

Figure 3.3 provides a visualization of the point estimation for each methods. MUA has the most noisy point estimation since it does not consider the spatial correlation, and there is no sparsity constraint directly imposed other than using the adjusted p-value to determine the level of significance for each voxel location. STGP suffers from the same issue as in Figure 3.2, where the pre-specified kernel is too smooth for the Z-shape and recycle shape (top-left). SBIOS uses the same kernel, but the binary selection parameter $\delta(s_j)$ has a spatially independent prior, and can get a clearer edge compared to STGP and better selection, but

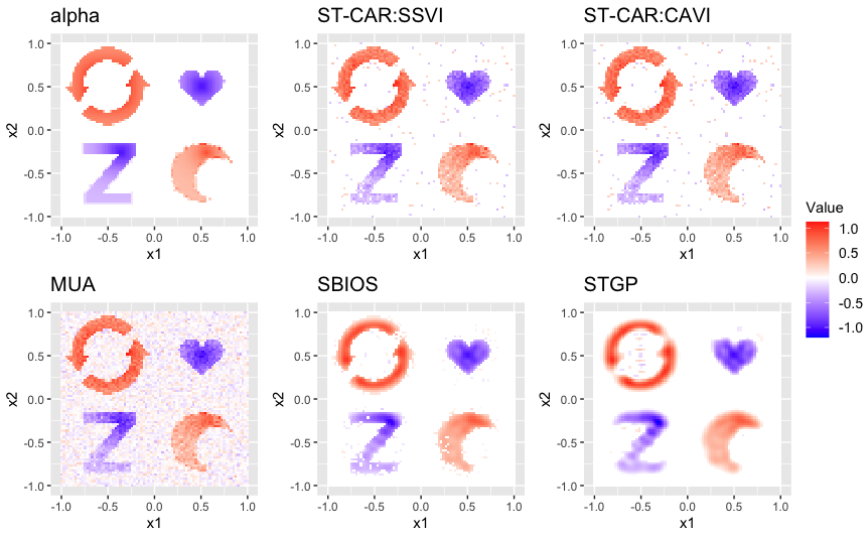


Figure 3.3: Point estimation result of IonS regression for all competing methods, $n = 600$, $p = 6400$, $\sigma_M^2 = 5$. The top-left figure is the true α signal.

the latent GP kernel is still too smooth that the edge of the recycle shape and Z-shape tends to be 0. In the ST-CAR plot, although the functional parameters ξ_k, η_i are all assigned GP prior with the same kernel as STGP and SBIOS, we can still see that ST-CAR is able to give a very clear edge for all 4 shapes. This demonstrates that ST-CAR prior is very flexible to different correlation patterns without much tuning on the neighborhood matrix B or correlation coefficient ρ, especially for the high SNR cases.

Table 3.2a provides the numerical result on IonS model based on 100 replications in six different settings. Because the predictive MSE on the outcome averaged over all voxel locations is very close for all methods, we do not report it here. Instead we focus on the estimation of the coefficient α. The proposed ST-CAR prior with SSVI algorithm gives the lowest RMSE except for case 2 and 6, for which the STGP has the lowest RMSE, although STGP has a much larger FDR in both cases. To control the FDR below 10%, we use the Benjamini-Hochberg (BH) adjusted p-values on MUA and set a threshold such that $\alpha(s_j)$ with the adjusted p-value below 0.1 are selected as active voxels. For ST-CAR, STGP, and SBIOS, we compute the proportion of the active voxels selected by MUA, and apply the same proportion to get the cut off on PIP. In this way, we select roughly the same proportion of voxels as active. Based on the result in Table 3.2a, this selection method can control the FDR for ST-CAR and SBIOS to be below 10%, whereas for STGP, the FDR is still over 10%. The MUA has the worst power (TPR) in all scenarios after controling for FDR. The total running time shown in Table 3.2b also shows a great improvement on the computational speed for the SSVI algorithm when compared with other MCMC sampling type algorithms.

Table 3.2: Numeric result for IonS simulation, under 100 replications.

(a) IonS: Comparison of estimation accuracy. The evaluation criteria includes false discovery rate (FDR), true positive rate (TPR), overall accuracy (ACC), and root mean squared error (RMSE), all multiplied by 100.

Case 1. $n = 600, p = 1600, \sigma_M^2 = 5$					Case 2. $n = 600, p = 900, \sigma_M^2 = 5$				
Criteria	ST-CAR	MUA	STGP	SBIOS	Criteria	ST-CAR	MUA	STGP	SBIOS
FDR	5.8	7.98	16.88	6.3	FDR	4.95	8.23	12.75	4.69
TPR	95.41	93.2	94.18	94.94	TPR	84.34	81.42	85.01	84.61
ACC	97.86	96.95	94.89	97.66	ACC	96.27	95.19	94.9	96.36
RMSE	**7.86**	9.35	10.53	10.79	RMSE	7.88	9.4	**6.85**	7.36
Case 3. $n = 1000, p = 1600, \sigma_M^2 = 5$					Case 4. $n = 600, p = 1600, \sigma_M^2 = 10$				
Criteria	ST-CAR	MUA	STGP	SBIOS	Criteria	ST-CAR	MUA	STGP	SBIOS
FDR	7.1	8.1	19.24	7.97	FDR	5.07	8.06	14.87	4.12
TPR	98.38	97.32	95.57	97.53	TPR	85.31	82.61	90.84	86.19
ACC	98.13	97.69	94.43	97.76	ACC	96.06	94.96	94.88	96.42
RMSE	**6.44**	7.21	10.16	10.35	RMSE	**10.32**	13.14	11.21	11.74
Case 5. $n = 600, p = 6400, \sigma_M^2 = 5$					Case 6. $n = 1000, p = 6400, \sigma_M^2 = 20$				
Criteria	ST-CAR	MUA	STGP	SBIOS	Criteria	ST-CAR	MUA	STGP	SBIOS
FDR	5.97	7.84	20.78	6.04	FDR	5.93	7.93	19.94	2.95
TPR	93.64	91.78	97.19	93.62	TPR	81.83	80.09	96.64	84.46
ACC	97.35	96.55	93.9	97.33	ACC	95.06	94.32	94.18	96.16
RMSE	**8.52**	9.19	9.79	10.25	RMSE	13.84	14.19	**9.76**	10.72

(b) Computation time for IonS simulation, averaged over 100 replications.

Computation time	Total time (seconds)			Number of iteratios per second		
Case	ST-CAR	STGP	SBIOS	ST-CAR	STGP	SBIOS
Case 1. $n = 600, p = 1600, \sigma_M^2 = 5$	73.7	588.4	717.4	137.5	3.4	7.3
Case 2. $n = 600, \boldsymbol{p = 900}, \sigma_M^2 = 5$	55	381.6	874.5	186	5.3	6.3
Case 3. $\boldsymbol{n = 1000}, p = 1600, \sigma_M^2 = 5$	117.3	1062.1	1968.4	88.9	1.9	3.1
Case 4. $n = 600, p = 1600, \boldsymbol{\sigma_M^2 = 10}$	82.9	621.8	1214.1	122.7	3.2	4.9
Case 5. $n = 600, p = 6400, \boldsymbol{\sigma_M^2 = 5}$	409.2	2190.3	1049.8	24.7	0.9	5.3
Case 6. $n = 600, p = 6400, \boldsymbol{\sigma_M^2 = 20}$	596.7	5090.2	1733.2	17	0.4	3.2

On average, STGP takes 7.6 times long compared to SSVI, and SBIOS takes 10.4 times long compared to SSVI. In the Supplemental Material, we also provide additional result of using CAVI under ST-CAR prior and compare the performance with SSVI. SSVI still slightly outperforms CAVI in the IonS regression in terms of both estimation and computation speed.

For a more comprehensive comparison between different estimation algorithms (Gibbs, CAVI, SSVI) under ST-CAR prior, we include a small low-dimensional comparison for the SonI regression in the Supplemental Material. The result suggests that CAVI tends to have better estimation accuracy in SonI regression where the signal-to-noise ratio is low.

3.5 Application to ABCD Study

In this section, we use our method to analyze the Adolescent Brain Cognitive Development (ABCD) study release 1 data [10]. The ABCD study is a long-term study on the brain development of children in the United States. In this real data analysis, we use the 2-back 3mm task fMRI contrast data [81]. The scientific questions of interest are: (i) whether the brain signals in different regions have different impact on the children's IQ score (SonI); (ii) whether parents with higher education degree has an impact on the children's cognitive ability development (IonS). For the task fMRI data, after preprocessing, we have $p = 47636$ voxels and $n = 1861$ subjects in total.

To answer (i) with SonI model (3.3), we use the children's IQ score as the scalar outcome Y_i, and use the task fMRI data as the high-dimensional predictor $M_i(s_j)$, where s_j stands for voxel locations in the brain. The confounders include parental education level (binary, 1 if the parent has a bachelor degree or higher), age, gender, race and ethnicity (Asian, Black, Hispanic, Other, White), and household income (less than 50k, between 50k and 100k, greater than 100k). The coefficient of interest is β in (3.3). We expect β to be very sparse and has small effect, since the interpretation for $\beta(s_j) = b$ is that one unit increase in the brain signal in location s_j is associated with b amount of change in the children's IQ score, and the range of the standardized IQ score is $(-2.84, 3.26)$, a small range compared to the large number of predictors $p = 47636$.

To answer (ii) with IonS model (3.4), we use the task fMRI data as the outcome, and use the parental education level as the exposure. The confounders include age, gender, race and ethnicity, and household income. For the IonS model, for ξ_k, η_i that are assigned GP priors, we use region-independent kernel structure. The interpretation for $\alpha(s_j) = a$ in (3.4) is that, parent with bachelor degrees or higher is associated with a amount of change in the brain signal at location s_j. Hence we expect the effect size of α to be relatively larger than that of β.

In ST-CAR prior, the two most important tuning parameters are the thresholding parameter ν, and the initial value for σ_β^2 which controls how close β is to the latent sparse μ. In theory [43], the choice of ν does not have a huge impact as long as the initial values are close enough to the truth, or the MCMC sampling algorithm can run long enough to fully explore the parameter space. Because we are using VI algorithms, it is important to start with a good initial value. Hence we perform a sensitivity analysis to select the best ν and initial σ_β^2 in terms of the smallest testing pMSE. The entire data set is split into 70% training data and 30% testing data. Based on the sensitivity analysis results in Table 3.3 and Appendix Table B.4 and B.3, we choose $\nu = 0.007$, the initial value for σ_β^2 to be 10^{-5},

bandwidth 9 and decay rate $\gamma = 0.35$ in the decay rate function of σ_β^2 for SonI, and choose $\nu = 0.005$, the initial value for σ_α^2 to be 0.1, bandwidth 26 and decay rate $\gamma = 0.45$ for IonS. Although varying ν, bandwidth and decay rate have little influence on the results. Table 3.3 also reflects that our method has better testing pMSE compared to the competing methods Elastic Net, STGP and MUA. Due to computational limitation of other methods, we chose not to run all competing methods on the real data.

Table 3.3: Sensitivity Analysis on SonI and IonS regressions.

(a) SonI for varying ν and initial value for σ_β^2. The Elastic Net and STGP results are shown as a comparison. Bandwidth=9 in the ST-CAR model. Additional sensitivity analysis where bandwidth=26 and varying decay rate γ for σ_β^2 is available in the Appendix.

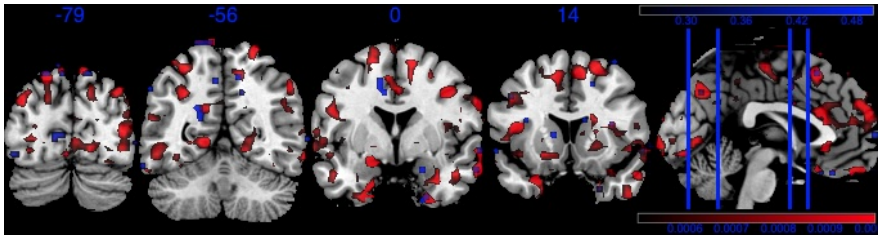
σ_β^2	10^{-5}	10^{-5}	10^{-5}	10^{-5}	5×10^{-5}	10^{-5}	5×10^{-6}	ElasNet	STGP
ν	0.003	0.005	0.007	0.01	0.005	0.005	0.005		
test pMSE	0.58	0.5	**0.48**	0.48	0.61	0.5	0.55	0.53	0.5
Test R^2	0.16	0.28	**0.30**	0.30	0.12	0.28	0.20	0.23	0.28
train pMSE	0.17	0.24	0.27	0.29	0.11	0.24	0.21	0.45	0.49
Train R^2	0.75	0.65	0.61	0.58	0.84	0.65	0.70	0.35	0.29

(b) IonS for varying ν, initial value for σ_α^2, and decay rate γ for σ_α^2. The total test pMSE is the summation of all voxel-level pMSE. Bandwidth is 26. Additional sensitivity result where bandwidth=9 is available in the appendix.

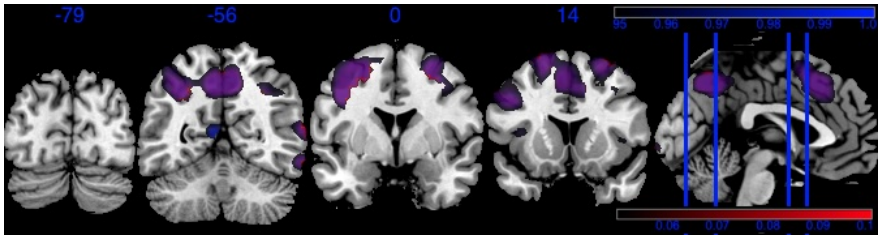
initial σ_α^2	1	0.1	0.01	0.1	0.1	0.1	0.1	0.1	0.1	
ν	0.005	0.005	0.005	0.001	0.01	0.05	0.005	0.005	0.005	MUA
Decay rate γ	0.35	0.35	0.35	0.35	0.35	0.35	0.25	0.45	0.55	
total test pMSE	47357.74	47351.78	47354.06	47354.13	47351	47354.1	47354.12	**47350.7**	47352.68	47487.05

We use CAVI on SonI, which takes 1.7 hours to run, and use SSVI on IonS, which takes 7.3 hours to run. Due to the vast sparsity and low SNR in β, the computational time of SonI is similar to STGP (1.6 hours). But the IonS model with SSVI algorithm shows a huge computational improvement compared to STGP (85.9 hours).

We present the final data analysis result in both visual illustration in Figure 3.4, and numeric values in Table 3.4. Figure 3.4 is a visualization on the positive significant voxels in SonI and IonS. The color range for the plots are between $[a, b]$, where only voxels with values greater than a are shown, and voxels with values greater than b are shown in the brightest color. From Figure 3.4a, due to the low SNR in SonI, both the effect size and PIP are small, and only a small amount of voxels with large effect size aligns with the mapping of PIP greater than 0.25. In comparison, α in IonS has larger effect size, and as shown in Figure 3.4b, the large effect areas aligns well with the mapping of PIP greater than 0.98.



(a) SonI: values of β (Red) w. color range [0.0005,0.001], and values of PIP (overlaying blue) w. color range [0.25, 0.5].



(b) IonS: values of α (Red) w. color range [0.05,0.1], and values of PIP (overlaying blue) w. color range [0.95,1].

Figure 3.4: Visual illustration of β in SonI and α in IonS.

In Table 3.4, we show the region level numeric result. Note that, although both SonI and IonS have a small amount of negative effects, they are very close to 0 compared to the positive effect scale, hence we only report the positive effect here. From Table 3.4, for SonI, *Precuneus_L* is the region with the largest positive effect, which means brain development in this region can have the most positive effect on the children's IQ score. This aligns with the previous study in [97] and scientific findings [90] that Precuneus is related with memory tasks. For IonS, *Frontal_Mid* region in both the left and right hemispheres have the largest positive effect, and have been shown to play a key role in the development of literacy (left *Frontal_Mid*) and numeracy (right *Frontal_Mid*) in previous findings [23].

3.6 Discussion and Conclusion

In this work, we have proposed the ST-CAR prior, which is a general and flexible prior that could be applied to any regression problems with imaging component. Variational inference algorithms are proposed for the ST-CAR prior. Especially, we implemented the coordinate ascent variational inference (CAVI) as a baseline VI algorithm that can provide good estimation accuracy in low SNR settings, and we proposed a novel stochastic subsampling variational inference (SSVI) algorithm that is more efficient computationally to be applied to high SNR settings. We demonstrated the use the ST-CAR prior in both scalar-on-image

Table 3.4: Numeric result for the top 10 regions sorted by number of significant positive voxels in SonI and IonS. For SonI, *sig count* is the number of significant voxels ($\text{PIP}_j \geq 0.25$) in each region, *pos_sig count* is the number of significant voxels with $\beta(s_j) \geq 0.0005$, and *pos sum* is $\sum_{j \in \mathcal{S}_r} \beta(s_j) I(\beta(s_j) > 0)$, the sum of positive effect for all voxels in region r. The IonS result has the same interpretation, except the cutoff for significant voxels is $\text{PIP}_j \geq 0.95$, and the cutoff for positive effect in *pos_sig count* is 0.05.

SonI					IonS				
region name	region code	sig count	pos_sig count	pos sum	region name	region code	sig count	pos_sig count	pos sum
Precuneus_L	67	12	12	0.25	Parietal_Inf_L	61	382	357	38.94
Temporal_Sup_R	82	12	9	0.16	Precuneus_L	67	377	312	37.26
Temporal_Inf_R	90	18	9	0.18	Precentral_L	1	305	293	33.55
Precuneus_R	68	12	8	0.14	Precuneus_R	68	316	285	36.24
Temporal_Inf_L	89	14	8	0.15	Frontal_Mid_R	8	322	270	43.12
Occipital_Mid_L	51	6	6	0.12	Frontal_Mid_L	7	272	244	43.06
Parietal_Inf_L	61	8	6	0.15	Supp_Motor_Area_L	19	215	205	23.82
Frontal_Sup_Orb_R	6	6	5	0.08	Parietal_Sup_L	59	224	167	18.63
Frontal_Mid_L	7	11	5	0.15	Temporal_Mid_R	86	175	154	27.18
Frontal_Mid_Orb_R	10	5	5	0.08	Frontal_Sup_L	3	155	147	21.81

and image-on-scalar regression models. Through comparisons in numeric studies, we find our proposed method has better performance in terms of estimation and computation, compared with existing methods such as T-LoHo, STGP, and SBIOS. The proposed method is applied to the ABCD study with task fMRI image data, and identifies the left Precuneus as a significant region to contribute the children's IQ development, and the development of the middle frontal gyrus as the significant region that can be most positively impacted by parental education level.

CHAPTER 4

Bayesian Structured Mediation Analysis With Unobserved Confounders

4.1 Introduction

In the emerging field of causal inference with complex data, high-dimensional mediation analysis is increasingly important, particularly with the surge in brain imaging and connectome datasets [51, 12]. We propose a causal mediation framework to account for the unobserved confounding effects for such high-dimensional complex mediators with certain correlation structures, referred to as structured mediators. The structured mediators include a broad family of applications such as spatial climate data and health data with repeated measurements. Our method development is motivated by the brain imaging application. Using the functional Magnetic Resonance Imaging (fMRI) data from the Adolescent Brain Cognitive Development (ABCD) study, we examine the relationship between parental education and children's general cognitive ability, seeking to identify the neural mediation pathways that underlie this causal link. Despite the advances in imaging mediation [97], the influence of unobserved confounders, such as stress levels or nutrient intake, have largely been ignored. Traditional high-dimensional mediation analyses, including Bayesian Imaging Mediation Analysis [BIMA, 97], [51, 80], and [61], rely on the no-unobserved-confounder assumption, which is unverifiable in real data [62]. Existing mediation studies have proposed sensitivity analyses to account for the violation of this assumption [38, 83, 18]. The sensitivity analysis approach can provide a range for the Natural Indirect Effect (NIE) and the Natural Direct Effect (NDE) with a known scale of the unobserved confounders [14, 104], or assumes binary outcome and unobserved confounders [18]. However, it is almost impossible to attain the accurate scale or data format of the unobserved confounders in practice. In this work, we directly estimate the structured unmeasured confounder effects. As long as the unobserved confounders have spatially smooth effects on the mediator, our method can estimate them and debias the mediation effects.

We propose a new framework for BAyesian Structured Mediation analysis with Unobserved confounders (BASMU). BASMU aims to relax the no-unobserved-confounding assumption by assuming that the unobserved confounders exist and correlate with both the mediator and the outcome. This assumption can be viewed as one case that violates the sequential ignorability assumption proposed in [38], where the predictor (treatment) is fully randomized but the mediator is not. [38] provided parametric sensitivity analysis under linear structural equation models with scalar mediators. Building upon the same sequential ignorability assumptions, [83] extended this idea to the nonparametric mediator. A detailed comparision among [38], [83], [34] and [37] is provided in [83]. [18] studied the sensitivity bound on the direct and indirect effect of mediation with unobserved confounders for binary treatment and outcomes. A recent and more interpretable bias analysis method proposed in [14] uses partial R^2 to analyze the bias when the unobserved confounder is omitted in linear regression. However, they restricted the problem to a scalar predictor in linear regression. [104] extended this idea to mediation analysis where the mediator can be a multi-dimensional vector, and proposed a matrix version of partial R^2. In addition to the potential bias due to omitting unobserved confounders, there are unique challenges in brain imaging mediation analysis. The fMRI data of the human brain as a mediator can be in high dimensions with a complex spatial structure, and only small areas have active mediation effects where the rest of brain voxels contribute near-zero effects. Hence, instead of the sensitivity analysis on the average treatment effect, we are more interested in detecting which brain areas would become active or nonactive mediators after accounting for the unobserved confounders.

Different from all the aforementioned sensitivity analysis-oriented methods, our proposed mediation framework allows the estimation of unobserved confounders by estimating the individual effect parameters in the mediator model. To our knowledge, BASMU is the first attempt to adjust the unobserved confounders for structural mediation analysis directly.

Following the same framework as in [5], where they first formally proposed the mediation framework under the linear structural equation models (LSEM), we propose a new LSEM where the unobserved confounders are specified in both the outcome model and mediator model. When the structural mediators satisfy certain assumptions, we can estimate the unobserved confounders from the mediator model and adjust the estimated unobserved confounders in the outcome model.

Our proposed BASMU framework advances existing methodologies by directly estimating unobserved confounder effects in high-dimensional mediation analysis, overcoming limitations of traditional approaches that rely on unverifiable no-unobserved confounder assumptions. This approach enables direct estimation of individual effects on the outcome and provides an asymptotic bias analysis when unobserved confounders are omitted. We propose

a two-stage algorithm to estimate high-dimensional unobserved individual effects using only the mediator model. This method allows a more flexible prior on the unobserved confounder coefficient and performs better across all simulation settings. Our framework increases the detection of significant mediation effects, as demonstrated in the analysis of brain imaging data from the ABCD study.

In Section 4.2, we introduce the structured mediation framework with unobserved confounders, along with identifiability assumptions and asymptotic bias analysis. Section 4.3 details the two-stage estimation algorithm. Section 4.4 presents simulation studies comparing BIMA and BASMU, while Section 4.5 applies our algorithm to ABCD data. We conclude with a discussion in Section 4.6.

4.2 Bayesian Image Mediation with Unobserved Confounders Framework

Structured mediators refer to a broad range of multivariate mediators with latent correlation structures. Examples include imaging data and climate data with spatial correlation, gene expression data that share the same biological pathways, Electronic Health Records data that correlate with multiple measurements, or patients with similar conditions. Our proposed method targets such high-dimensional correlated data as mediators, and we give the following generic definition of the structured mediator that has a smooth mean function and a completely independent noise term.

Definition 6. *For a given support $\mathcal{S}$, for any $s \in \mathcal{S}$, the structured mediator for subject i is defined as $M_i(s) = f_i(s) + \epsilon_i(s), \epsilon_i(s) \sim N(0, \sigma_M^2)$, $\epsilon_i(s)$ is independent of $\epsilon_i(s')$ for any $s \neq s'$, and $f_i : \mathcal{S} \mapsto \mathbb{R}$ is a real-value function that satisfies certain smoothness conditions.*

Here, the domain $\mathcal{S}$ can be a subset of the vector space or a spatial domain such as a three-dimensional brain in fMRI applications. The number of elements in $\mathcal{S}$, or its size, is strictly larger than one. The function $f_i(s)$ has a certain smoothness structure, for example, $f_i(s)$ can be represented by a lower order basis. One counter-example to the structured mediator is a single mediator, in which case we cannot separate the unobserved confounding effect from the independent random noise.

Motivated by the ABCD study, we are interested in the mediation effect of children's brain development in the causal relation of parental education level on children's general cognitive ability. Let s_j be the j-th voxel location in the brain. Let $M_i(s_j)$ denote the image intensity at location s_j for subject i, and X_i be a scalar-valued exposure variable of interest. Let $\mathbf{C}_i \in \mathbb{R}^q$ denote the observed confounders for individual i. Let Y_i denote the

scalar-valued outcome for subject i. $i = 1, \ldots, n, j = 1, \ldots, p$. Let $\mathcal{S} \in \mathbb{R}^d$ be a compact support, $L^2(\mathcal{S})$ be the square-integrable functional space on $\mathcal{S}$, and $\{\Delta s_j\}_{j=1}^p$ be an even partition on $\mathcal{S}$ such that $\mathcal{S} = \bigcup_{j=1}^p \Delta s_j$ and $\Delta s_j \cap \Delta s_{j'} = \emptyset$. We also assume the Lebesgue measure on a pixel partition to be $\lambda(\Delta s_j) = p^{-1}$. Let s_j be the center of the partition Δs_j. We use the abbreviation $\eta_i = \{\eta_i(s_j)\}_{j=1}^p$ to denote the functional value on the fixed grid, and similarly for α, β, ξ_k. Let $\mathcal{GP}(0, \kappa)$ denote a Gaussian Process with mean function 0 and covariance function $\kappa(\cdot, \cdot)$. For a set A, let $|A|$ be the cardinality of A.

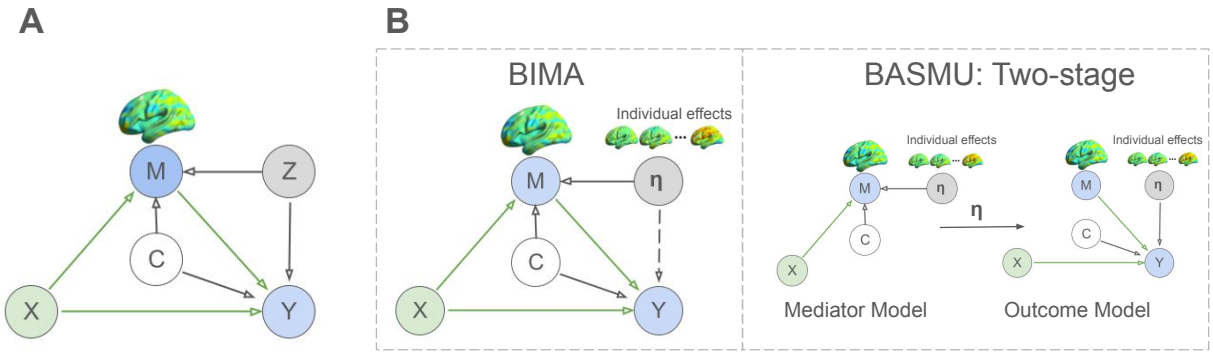


Figure 4.1: **Model Overview.** The green arrows represent the causal mediation triangle, where $X \to Y$ is the Natural Direct Effect (NDE), and the mediation pathway $X \to M \to Y$ is the Natural Indirect Effect (NIE). **A.** Directed Acyclic Graph (DAG) for structured mediation with unobserved confounders. Here, Z stands for the unobserved confounders. **B.** Causal graph representation of BIMA and BASMU with Two-Stage estimation.

Define the mediator model

$$M_i(s_j) = \alpha(s_j)X_i + \sum_{k=1}^{q} \xi_k(s_j)C_{i,k} + \eta_i(s_j) + \epsilon_{M,i}(s_j), \quad \epsilon_{M,i}(s_j) \overset{\text{iid}}{\sim} N(0, \sigma_M^2). \tag{4.1}$$

We use α to denote the impact of the treatment X_i on the image mediator M_i, ξ_k the functional-coefficient for the vector-valued confounders $C_i \in \mathbb{R}^q$, η_i the spatially-varying individual-specific parameter, and $\epsilon_{M,i}$ the spatially-independent noise term.

Observe that if there exists unobserved confounder $\mathbf{Z}_i$ of unknown dimension m, there would be a term $\sum_{r=1}^m \xi_{m,z}(s_j)Z_{r,i}$ which plays the same role as $\sum_{k=1}^q \xi_k(s_j)C_{i,k}$. Since $\mathbf{Z}_i$ is unobservable, if $\xi_{m,z}(s_j)$ is a spatially-varying effect, we can replace the term $\sum_{r=1}^m \xi_{m,z}(s_j)Z_{r,i}$ by the individual effects $\eta_i(s_j)$, as long as they have the same spatially-varying structure so that $\eta_i(s_j)$ is separable from the independent noise $\epsilon_{M,i}(s_j)$. The crucial choice of whether to consider the unobserved confounding effect of $\eta_i(s_j)$ on the outcome leads to the introduction of two distinct modeling frameworks in the following subsections.

4.2.1 Structured Mediation When Omitting Unobserved Confounders

If we ignore the impact of individual effects on the outcome, the potentially biased outcome model can be defined as

$$Y_i = \sum_{j=1}^{p} \beta(s_j) M_i(\Delta s_j) + \gamma X_i + \sum_{k=1}^{q} \zeta_k C_{i,k} + \epsilon_{Y,i}, \quad \epsilon_{Y,i} \overset{i.i.d}{\sim} N(0, \sigma_Y^2). \tag{4.2}$$

We refer to model (4.1) and (4.2) as the Bayesian Image Mediation Analysis (BIMA) framework [97], where the unobserved confounders are omitted. Here, β denotes the effect of the image mediator M_i on the outcome Y_i, γ the scalar-valued direct effect, and ζ_k the coefficient for the kth observed confounder C_k.

In BIMA, with the stable unit treatment value assumption (SUTVA) [72], we follow the mediation assumption proposed in [86]: for any i, endogenous x and $\mathbf{m}$,

$$\begin{aligned} &\text{(i) } Y_{i,(x,\mathbf{m})} \perp X_i \mid \{\mathbf{C}_i\}, \text{ (ii) } Y_{i,(x,\mathbf{m})} \perp \mathbf{M}_i \mid \{\mathbf{C}_i, X_i\}, \\ &\text{(iii) } \mathbf{M}_{i,(x)} \perp X_i \mid \{\mathbf{C}_i\}, \text{ (iv) } Y_{i,(x,\mathbf{m})} \perp \mathbf{M}_{i,(x')} \mid \{\mathbf{C}_i\}. \end{aligned} \tag{4.3}$$

The first three assumptions ensure that the observed confounder C_i controls the outcome-treatment confounding, the outcome-mediator confounding, and the mediator-treatment confounding respectively. The fourth assumption ensures the outcome-mediator confounders are not affected by the underlying endogenous treatment x.

Under the causal inference framework [71] with no unobserved confounders, the average treatment effect between x and x' is $\mathbb{E}\left[Y_{i,\{x,\mathbf{M}_{i,(x)}\}} - Y_{i,\{x',\mathbf{M}_{i,(x')}\}}\right]$, and can be decomposed into the NIE

$$\mathbb{E}\left[Y_{i,\{x,\mathbf{M}_{i,(x)}\}} - Y_{i,\{x,\mathbf{M}_{i,(x')}\}}\right] = \sum_{j=1}^{p} \beta(s_j)\alpha(s_j)\lambda(\Delta s_j)(x - x'),$$

and the NDE $\mathbb{E}\left[Y_{i,\{x,\mathbf{M}_{i,(x')}\}} - Y_{i,\{x',\mathbf{M}_{i,(x')}\}}\right] = \gamma(x - x')$. Our primary interest is in estimating the spatially-varying NIE, defined as $\mathcal{E}(s) = \alpha(s)\beta(s)$, and the scalar-valued NIE is $\mathcal{E} = \sum_{j=1}^{p} \alpha(s_j)\beta(s_j)\lambda(\Delta s_j)$. For spatial mediation analysis, we are not only interested in the scalar-valued NIE $\mathcal{E}$, but also the spatial areas in $\mathcal{S}$ where $\mathcal{E}(s)$ is nonzero.

4.2.2 BAyesian Structured Mediation analysis with Unobserved confounders (BASMU)

Define a full outcome model

$$Y_i = \sum_{j=1}^{p} \beta(s_j) M_i(\Delta s_j) + \gamma X_i + \sum_{k=1}^{q} \zeta_k C_{i,k} + \sum_{j=1}^{p} \nu(s_j)\eta_i(s_j)\lambda(\Delta s_j) + \epsilon_{Y,i}, \tag{4.4}$$

where $\epsilon_{Y,i} \overset{\text{iid}}{\sim} N(0, \sigma_Y^2)$ and $M_i(\Delta s_j) = \mathbb{E}\{M_i(s_j)\}\lambda(\Delta s_j) + \epsilon_{M,i}(\Delta s_j)$ with $\epsilon_{M,i}(\Delta s_j) \overset{\text{ind}}{\sim} N(0, \sigma_M^2 \lambda(\Delta s_j))$.

The structural equation models (4.1) and (4.4) together form the BASMU framework. Under the BASMU framework, we include the effect of η_i on Y_i to represent the impact of unobserved confounders, denoted as ν.

4.2.3 Assumptions and Identification of BASMU

With the presence of unobserved confounders η_i as shown in the BASMI model (4.1) and (4.4), the assumptions (4.3) are violated. Instead, we impose the following mediation assumptions,

$$\begin{aligned} &\text{(i) } Y_{i,(x,\mathbf{m})} \perp X_i \mid \{\mathbf{C}_i, \eta_i\}, \text{ (ii) } Y_{i,(x,\mathbf{m})} \perp \mathbf{M}_i \mid \{\mathbf{C}_i, X_i, \eta_i\}, \\ &\text{(iii) } \mathbf{M}_{i,(x)} \perp X_i \mid \{\mathbf{C}_i, \eta_i\}, \text{ (iv) } Y_{i,(x,\mathbf{m})} \perp \mathbf{M}_{i,(x')} \mid \{\mathbf{C}_i, \eta_i\} \end{aligned} \tag{4.5}$$

The set of assumptions in (4.5) are the same as in [18]. But the scope of [18] is restricted to the sensitivity analysis of a binary outcome with binary exposure and scalar mediator, whereas we take advantage of the individual effect η_i as the unobserved confounders, and propose the full outcome model in (4.4) to reduce the bias in β and the total indirect effect $\sum_{j=1}^{p} \alpha(s_j)\beta(s_j)$.

The joint identifiability of models (4.1) and (4.4) is non-trivial, especially with the introduction of η_i and ν. We impose a set of model identifiability assumptions.

Define column vectors $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}} \in \mathbb{R}^n$, $\mathbf{C} = (\mathbf{C}_1, \ldots, \mathbf{C}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times q}$. Let $\{\psi_l(s)\}_{l=1}^{\infty}$ be a set of basis of $L^2(\mathcal{S})$. Denote the basis coefficients $\theta_{\eta_i,l} = \int_{\mathcal{S}} \eta_i(s)\psi_l(s)\lambda(\mathrm{d}s)$, and $\boldsymbol{\theta}_{\eta,l} = (\theta_{\eta_i,l}, \ldots, \theta_{\eta_n,l})$.

Assumption 4. *(i) Conditional on* $(\mathbf{X}, \mathbf{C})$*, there exists a constant matrix* $\mathbf{W} = (W_{i,k}) \in \mathbb{R}^{n\times(q+1)}$ *such that* $\det\{\mathbf{W}^\top(\mathbf{X}, \mathbf{C})\} \neq 0$*; (ii) There exists a constant vector* $\mathbf{b} = (b_1, \ldots, b_q)^\top$ *such that for any* $s \in \mathcal{S}$ *and* $k = 1, \ldots, q+1$, $\sum_{i=1}^{n} W_{i,k}\eta_i(s) = b_k$*; (iii) The design matrix after basis decomposition* $\{\mathbf{X}, \mathbf{C}, \boldsymbol{\theta}_{\eta,1}, \ldots, \boldsymbol{\theta}_{\eta,L}\} \in \mathbb{R}^{n\times(L+1+q)}$ *is assumed to be full rank,*

and for any subset $\mathcal{S}_m \subset \mathcal{S}$ where $|\mathcal{S}_m| = m$, the design matrix before basis decomposition $\left\{\mathbf{X}, \mathbf{C}, \{\boldsymbol{\eta}(s_k)\}_{s_k \in \mathcal{S}_m}\right\} \in \mathbb{R}^{n \times (m+1+q)}$ is also assumed to be full rank. (iv) The unobserved confounding effect ν is either low-rank (i.e. $\nu(s) = \sum_{l=1}^{L} \theta_{\nu,l}\psi_l(s)$, $L = o(n)$), or is sparse (i.e. $\boldsymbol{\nu} = (\nu(s_1), \ldots, \nu(s_p)) \in \{v \in \mathbb{R}^p : \|v\|_0 = m\}$, $m = o(n)$).

Remark. Assumption 4 is to guarantee the identifiability of the BASMU model. Assumption 4 (i) and (ii) are used in the proof of identifiability of $\{\eta_i\}_{i=1}^n$ in the mediator (4.1), which guarantee the identifiability of all parameters in the mediator model (4.1). Assumption 4 (iii) and (iv) are to ensure that the design matrix in the outcome model (4.4) is full-rank and ν is either sparse or low-rank, which guarantee the identifiability of $\beta, \gamma, \{\zeta_k\}_{k=1}^q, \nu, \sigma_Y$ in the outcome model (4.4) given $\{\eta_i\}_{i=1}^n$. One example of Assumption 4 is $\mathbf{b} = 0$ and $\mathbf{W} = (\mathbf{X}, \mathbf{C})$, and $\boldsymbol{\theta}_{\eta,l} \in \mathbb{R}^n$ are sampled from a subspace of $\mathbb{R}^n$ orthogonal to span $\{\mathbf{X}, \mathbf{C}\}$.

Let $\boldsymbol{\theta}_{\text{all}} = \{\alpha, \{\xi_k\}_{k=1}^q, \{\eta_i\}_{i=1}^n, \sigma_M, \beta, \gamma, \{\zeta_k\}_{k=1}^q, \nu, \sigma_Y\}$ be the collection of all parameters in model (4.1) and (4.4).

Proposition 3. *Under Assumption 4, the BASMU model in* (4.1) *and* (4.4) *is jointly identifiable, i.e. given density function $\prod_i f(Y_i, \mathbf{M}_i; \boldsymbol{\theta}|X_i, \mathbf{C}_i)$, $\prod_i f(Y_i, \mathbf{M}_i; \boldsymbol{\theta}_{all}|X_i, \mathbf{C}_i) = \prod_i f(Y_i, \mathbf{M}_i; \boldsymbol{\theta}^*_{all}|X_i, \mathbf{C}_i)$ implies $\boldsymbol{\theta}_{all} = \boldsymbol{\theta}^*_{all}$.*

Proposition 3 shows that as long as ν has a latent low dimensional representation, or ν is sparse, the proposed BASMU model (4.1) and (4.4) are jointly identifiable. In the next section, we analyze the bias induced by ignoring the unobserved confounder in (4.4).

4.2.4 Bayesian Bias Analysis With Omitted Unobserved Confounders

Based on the consistency result for (4.1) in [97], the posterior mean of α is a consistent estimator. Hence the main focus is on the asymptotic bias of the posterior mean of β under (4.2) as a point estimator. We assign Gaussian Process (GP) prior on $\beta \sim \mathcal{GP}(0, \sigma_\beta^2\kappa)$. By Mercer's theorem, for a given kernel $\kappa(s, s') = \sum_{l=1}^{\infty} \lambda_l\psi_l(s)\psi_l(s')$, we can represent $\beta(s) = \sum_{l=1}^{L} \theta_{\theta_\beta,l}\psi_l(s)$, $L = o(n)$ where $\theta_{\beta,l} \overset{\text{ind}}{\sim} N(0, \lambda_l\sigma_\beta^2)$.

We use a superscript '0' to denote a true parameter value, e.g., θ^0. Denote the true unobserved confounder term $U^0 := (\boldsymbol{\eta}^0)^{\mathrm{T}} \nu^0 \in \mathbb{R}^n$, and given estimators $\hat{\boldsymbol{\eta}}$ and $\hat{\nu}$, denote $\hat{U} := (\hat{\boldsymbol{\eta}})^{\mathrm{T}} \hat{\nu} \in \mathbb{R}^n$. The result below discusses the asymptotic bias of β under model (4.2) where the unobserved confounder is omitted, and model (4.4) where the unobserved confounder is considered.

For the structured mediator $M_i(\Delta s_j) = \mathbb{E}\{M_i(s_j)\}\lambda(\Delta s_j) + \epsilon_{M,i}(\Delta s_j)$, $\epsilon_{M,i}(\Delta s_j) \overset{\text{ind}}{\sim} N(0, \sigma_M^2\lambda(\Delta s_j))$, denote $E_i(s) := \mathbb{E}\{M_i(s)\}$. In addition, denote the basis coefficients under

GP basis decomposition as $\theta_{E,i,l} = \int_{\mathcal{S}} E_i(s)\psi_l(s)\lambda(\mathrm{d}s)$, where ψ_l is the same basis function in the GP prior of β.

Assumption 5. *(i) For any l, l', $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\theta_{E,i,l}\theta_{E,i,l'} = H_{l,l'}$, where $H_{l,l'}$ is some finite constant; (ii) For any l, $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\theta_{E,i,l}U_i^0 = h_l^0$, where h_l^0 is a finite constant, U_i^0 is the i-th element in U^0.*

Assumption 6. *Conditional on $\hat{\boldsymbol{\eta}}$ and $\hat{\nu}$, $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\theta_{E,i,l}\hat{U}_i \xrightarrow{p} \hat{h}_l$ where $\hat{h}_l$ is a random variable that only depends on $\hat{\boldsymbol{\eta}}$ and $\hat{\nu}$.*

One example of Assumption 5 is to view $E_i(s)$ as i.i.d samples from some unknown process $E(s)$ (for example, Gaussian Process) with finite first and second moments, and $H_{l,l'}$ is the finite covariance for the basis coefficients at l and l'. If we view elements in U^0 as n i.i.d samples drawn from a distribution U with finite second moments, h_l^0 is the covariance between the l-th basis coefficient of $E(s)$ and U, and is also finite (Holder's inequality). The same example applies to Assumption 6.

In the Proposition below, we use $\hat{\theta}_\beta^B$ to denote the point estimator under the BIMA outcome model and use $\hat{\theta}_\beta^F$ to denote the point estimator under the full BASMU outcome model that takes account into the unobserved confounders. Denote $\tilde{M}_{i,l} = \int_S M_i(s)\psi_l(s)\lambda(\mathrm{d}s)$ and $\tilde{M} \in \mathbb{R}^{n\times L}$, $(\tilde{M})_{i,l} = \tilde{M}_{i,l}$, and denote $A := \tilde{M}^{\mathrm{T}}\tilde{M} \in \mathbb{R}^{L\times L}$ where $L = o(n)$. Denote $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ as the smallest and largest singular values of A respectively.

Proposition 4. *Assume that A satisfy $0 < c_{\min} < \liminf_{n\to\infty}\sigma_{\min}(A)/n \le \limsup_{n\to\infty}\sigma_{\max}(A)/n \le c_{\max} < \infty$ with probability $1 - \exp\{-c_0 n\}$ for some positive constant $c_0, c_{\min}, c_{\max}$. In addition, in the GP prior basis coefficients $\theta_{\beta,l} \overset{ind}{\sim} N(0, \sigma_\beta^2\lambda_l)$, assume that $\lambda_L > c_\lambda n^{-1+a_\lambda}$ for some positive constant c_λ, a_λ. We can draw the following conclusions given Assumptions 5 and 6:*

(i) When $U^0 = \mathbf{0}$, i.e. no unobserved confounder, the asymptotic bias of $\hat{\theta}_\beta^B$ is 0.

(ii) Given Assumption 5 and 6, and assume that the true unobserved confounder term $\sum_{i=1}^{n}(U_i^0)^2/n$ is finite, then the bias of the posterior mean of θ_β under BIMA model (4.2) bias$(\hat{\theta}_\beta^B) \xrightarrow{p} (H + \sigma_M^2 I_L)^{-1}h^0$, and the bias under the full model (4.4) bias$(\hat{\theta}_\beta^F) \xrightarrow{p} (H + \sigma_M^2 I_L)^{-1}\left(h^0 - \hat{h}\right)$.

Remark. The result of Proposition 4 is conditional on the true values of γ and ζ_k in (4.4) for simplicity of the analysis. A similar bias analysis result can be drawn on the NDE γ if we treat $\mathbf{X}$ as one additional column in $\tilde{M}$, and the bias of γ is the corresponding element in bias$(\hat{\theta}_\beta^B)$ and bias$(\hat{\theta}_\beta^F)$.

Implications. Proposition 4 (i) can be seen as a corollary of (ii). Based on Proposition 4 (ii), we can expect that: (a) the bias of BIMA depends on the scale of the unobserved

term h^0, and is nonzero unless $h^0 = 0$; (b) the bias of BASMU depends on the estimation of $\hat{\eta}$ and $\hat{\nu}$; (c) larger random noise σ_M in the mediator model may reduce the bias, because larger random noise makes the observed mediator $M_i(s)$ less correlated with the individual effect $\eta_i(s)$.

4.3 Two-stage Estimation

The most natural way to estimate model (4.1) and (4.4) is to use a fully Bayesian approach to update all parameters iteratively. This involves updating the individual effects η_i jointly from both models (4.1) and (4.4) in every iteration. Due to the large parameter space of $\{\eta_i\}_{i=1}^n$ to search from, the joint estimation approach usually takes very long to converge. For every iteration when $\{\eta_i\}_{i=1}^n$ is updated, the new $\{\eta_i\}_{i=1}^n$ can have a huge impact on the likelihood of (4.4), and all other parameters in (4.4) need longer iterations to converge to stable values, hence the joint estimation can make estimating the outcome model (4.4) very unstable. Because the posterior of $\{\eta_i\}_{i=1}^n$ is mainly dominated by the mediator model (4.1), sampling $\{\eta_i\}_{i=1}^n$ based solely on (4.1) can already give a consistent estimation (see remark of Assumption 4), hence instead of the fully Bayesian joint estimation approach, we propose the two-stage estimation.

In the two-stage estimation, we compute the posterior of model (4.1) and (4.4) separately. First, we draw posterior samples based on model (4.1) based on the priors

$$\alpha \sim \mathcal{GP}(0, \sigma_\alpha^2 \kappa), \xi_k \sim \mathcal{GP}(0, \sigma_\xi^2 \kappa), \eta_i \sim \mathcal{GP}(0, \sigma_\eta^2 \kappa), \tag{4.6}$$

and compute the posterior mean of η_i conditional only on model (4.1), denoted as $\hat{\eta}_i$. Using $\hat{\eta}_i$ as part of the fixed design matrix in (4.4), draw MCMC samples from (4.4) conditioning on $\eta_i = \hat{\eta}_i$, based on the following prior for ν,

$$\nu(s) = g(s)\delta(s), \quad g(s) \overset{\text{ind}}{\sim} N(0, \sigma_\nu^2), \quad \delta(s) \overset{\text{ind}}{\sim} \text{Ber}(1/2). \tag{4.7}$$

The two-stage estimation uses a flexible prior on ν with a spatial independent structure, and the selection variable δ allows sparsity in ν. Through simulation studies, we find that $\hat{\eta}_i$ can estimate η_i well when p is reasonably larger than L, i.e. when the kernel for η_i is smooth. Estimating ν is still challenging for two-stage estimation, given that η_i can be mis-specified. However due to the flexible prior on ν, even when ν cannot be fully recovered from the two-stage estimation, the estimation for β can still be greatly improved compared to BIMA. For fast posterior computation, we use Singular Value Decomposition (SVD) on $\{\eta_i\}_{i=1}^n$. The detailed two-stage algorithm is provided in the Appendix C.3.

So far, we have compared two model frameworks, BIMA and BASMU [1], the first of which omits the unobserved confounders completely. Figure 4.1B provides a visual illustration of the structure of the two methods. The two-stage estimation is expected to give good point estimation results, although not fully Bayesian inference. The flexible prior on ν still allows us to debias β, as long as $\hat{\eta}_i$ is not too far from the truth.

4.4 Simulation Study

We compare the performance of BIMA and BASMU through extensive simulation studies. For α, β, and ν, we simulate 2D $p = 40 \times 40$ images (true signals are shown as in Figure C.1 and Figure C.4a). For the GP priors in (4.6) and β, we use Matérn kernel with $\rho = 2, \tau = 0.2, d = 2$,

$$\kappa(s', s; \tau, \rho) = C_\nu(\|s' - s\|_2^2/\rho),\ C_\tau(d) := \frac{2^{1-\tau}}{\Gamma(\tau)} \left(\sqrt{2\tau} d\right)^\tau K_\tau(\sqrt{2\tau} d). \tag{4.8}$$

The GP prior parameters $\alpha, \beta, \eta_i, \xi_k$ use the same basis decomposition of the kernel function in (4.8), denoted as $\kappa(s', s) = \sum_{l=1}^{\infty} \lambda_l \psi_l(s') \psi_l(s)$. For example, $\beta(s)$ is approximated by $\sum_{l=1}^{L} \theta_{\beta,l} \phi_l(s)$ with the prior $\theta_{\beta,l} \overset{\text{ind}}{\sim} N(0, \lambda_l \sigma_\beta^2)$. Set $L = 120$ basis coefficients as the cutoff. We use the Metropolis-Adjusted Langevin Algorithm (MALA) for updating α and β, and the Gibbs sampler for the rest of the parameters. For the outcome models in both BIMA and BASMU, we use a total of 2×10^4 iterations with the last 10% used as MCMC samples. The mediator model (4.1) uses 10^3 iterations with the last 10% used as MCMC samples.

Table 4.1 summarizes the settings for six cases, varying σ_η, σ_M, and n to show the theoretical implications of Proposition 4. We simulate three signal patterns for ν (dense, sparse, and zero) as shown in Appendix Figure C.4a. Each case has 100 replications. The dense ν signal is simulated using low dimensional basis coefficients mapped to p-dimension through the same Matérn kernel in (4.8).

We present the scalar-valued NIE $\mathcal{E}$ in terms of the bias, variance and MSE over 100 replications as in Table 4.1, and visualize the spatial MSE (Appendix Figure C.2) and Bias (Appendix Figure C.3) on the posterior mean of $\beta(s)$ as a point estimator over 100 replications. Each voxel s_j in Figure C.2 represents the MSE of the posterior mean of $\beta(s_j)$.

Results in Table 4.1 show that BASMU generally archives the lowest MSE for $\mathcal{E}$, except for Case 3 where no unobserved confounding effect is present. To verify the theoretical implications by Proposition 4, comparing Case 2 and 4, as n increases, the MSE of BASMU decreases, whereas the BIMA model has increased MSE and bias. This is because $\|h^0 - \hat{h}\|_2$

[1] The BASMU R package can be found on the GitHub page https://github.com/yuliangxu/BASMU

Table 4.1: Simulation result of the scalar NIE $\mathcal{E}$ averaged over 100 replications. The smaller MSE of $\mathcal{E}$ is bolded in each case. The default generative parameter settings are $\sigma_\eta = 0.5$, $n = 300$, $\sigma_M = 2$.

	BIMA	**BASMU**		**BIMA**	**BASMU**		**BIMA**	**BASMU**
Case 1	dense ν		Case 3	all 0 ν		Case 5	dense ν, $\sigma_\eta = 1$	
Bias	-2.72	1.06	Bias	-0.5	-0.17	Bias	13.31	2.21
Var	3.59	4.11	Var	2.31	4.49	Var	3.35	3.2
MSE	10.97	**5.20**	MSE	**2.53**	4.48	MSE	180.36	**8.06**
Case 2	sparse ν		Case 4	sparse ν, $n = 600$		Case 6	dense ν, $\sigma_M = 4$	
Bias	7.56	1.87	Bias	10.77	2.29	Bias	-6.33	-0.57
Var	3.42	3.76	Var	1.54	1.54	Var	13.43	12.63
MSE	60.51	**7.22**	MSE	117.5	**6.79**	MSE	53.36	**12.82**

decreases as $\hat{U} \to U^0$ when n increases. Comparing Case 1 and 5, as σ_η increases, U^0 increases, the MSE for BASMU has little changes compared to the huge increase in MSE and bias for BIMA due to the increased scale of U^0. Comparing Case 1 and 6, as σ_M increases, the bias for BASMU decreases. In fact, from the spatial MSE and bias in Figure C.3 Case 6 compared to Case 1, both BIMA and BASMU have an overall decreased MSE and bias in β, though the decreased bias area does not overlap with the true nonzero signal regions in α and β, hence not fully reflected on the result of scalar NIE. Figure C.2 and C.3 show straight-forward evidence that the two-stage estimation of BASMU can indeed reduce the bias of $\beta(s)$ and have a lower MSE over varying spatial locations in all scenarios. In Appendix Table C.1 we also provide the NDE result with similar implications.

4.5 Analysis of ABCD Data

For the real data analysis, we use the ABCD study release 1 [10] as an example. The scientific question of interest is the mediation effect of children's brain development on the impact of parental education level on children's general cognitive ability. The structured mediator is the task fMRI data of the children. The outcome is children's general cognitive ability, the exposure is the parental education level, a binary indicator of whether or not the parent has a bachelor's or higher degree. The confounders include age, gender, race and ethnicity (Asian, Black, Hispanic, Other, White), and household income (less than 50k, between 50k and 100k, greater than 100k). For the multi-level variables race and ethnicity, and household income, we use binary coding for each level. The potential unobserved confounders can be the stress level of the participants, nutrient supply, and other genetic factors that might impact brain development.

The task fMRI data is 2-back 3mm task contrast data, and the preprocessing method is described in [81]. Each voxel in the 3D image mediator represents the brain signal intensity when the subject tries to remember the tasks they performed 2 rounds ago.

A previous mediation analysis using the same data set has been conducted in [97] that ignored unobserved confounders. In this analysis, we use the top four regions (shown in Table 4.2) previously identified in [97] with the most significant mediation effect, and perform a BASMU analysis with the brain image data on these four regions. After preprocessing, we have $n = 1861$ subjects and $p = 3539$ voxels as mediators. The GP kernel in use is the same Matérn kernel as in [97]. BIMA and BASMU share the same mediator model result with a total of 10^4 iterations. For the outcome model in BIMA, we use 3×10^4 total of iterations. For BASMU, due to the SVD at each iteration, the computational speed of the two-stage algorithm is relatively slow for the real data. Hence we set $\delta(s)$ in (4.7) to 1 at the beginning and run for 10^4 iterations first to get the initial values and use these initial values to run the two-stage algorithm for 2×10^3 iterations with the last 20% as the MCMC sample. The total running times are 54 minutes for the mediator model, 61 minutes for the BIMA outcome model, and about 4 hours for the BASMU outcome model. All real data analyses are performed on a laptop with an Apple M1 chip and 8GB memory.

Table 4.2: Comparison of ABCD data analysis under BIMA and BASMU. The top table reports the active voxel selection, from column 3 to 8: number of active voxels selected by BIMA/BASMU (brackets: percentage of selected voxels over the total number of voxels), number of commonly selected voxels, number of voxels only selected by BIMA/BASMU, and the total number of voxels in each region. The bottom table reports the numeric values of the NIE, from column 3 to 8: summation of NIE over the region under BIMA/BASMU, summation of NIE over voxels with positive effect under BIMA/BASMU, summation of NIE over voxels with negative effect under BIMA/BASMU.

Selection of active mediation voxels $\mathcal{E}(s_j)$						
Region code and name	BIMA	BASMU	common	BIMA_only	BASMU_only	size
34 Cingulum_Mid_R	80 (13%)	**342 (57%)**	68	12	274	605
57 Postcentral_L	108 (9%)	**246 (21%)**	92	16	154	1159
61 Parietal_Inf_L	138 (20%)	**387 (56%)**	131	7	256	696
67 Precuneus_L	150 (14%)	**404 (37%)**	137	13	267	1079
Effect size of $\mathcal{E}$						
Region code and name	BIMA NIE	BASMU NIE	BIMA NIE (+)	BASMU NIE (+)	BIMA NIE (-)	BASMU NIE (-)
34 Cingulum_Mid_R	0.007	0.007	0.022	0.032	-0.015	-0.025
57 Postcentral_L	0.007	**0.021**	0.068	0.087	-0.062	-0.065
61 Parietal_Inf_L	0.009	0.000	0.163	0.163	-0.154	-0.162
67 Precuneus_L	0.051	0.075	0.107	0.140	-0.057	-0.065

Table 4.2 shows the comparison of NIE between BIMA and BASMU. We use the criteria of whether the 95% credible interval includes 0 for active mediation voxel selection. In the bottom table of NIE size, the total, positive, and negative effects are separately reported for

each method. The NDE under BIMA is 0.247, with a 95% credible interval $(0.166, 0.329)$. The NDE under BASMU is 0.183, with a 95% credible interval $(0.145, 0.218)$. The NIE over all locations is 0.073 for BIMA with (0.012, 0.127) as the 95% credible interval, and 0.103 for BASMU with (0.043, 0.155) as the 95% credible interval. To check the model fitting, the R^2 for the BIMA outcome model (4.2) is 0.41, and the R^2 for the BASMU outcome model (4.4) is 0.42. Figure 4.2 provides a visual illustration of the selected active mediation voxels. Appendix Figure C.4b gives a scatter plot of each estimated $\mathcal{E}(s_j)$ between BIMA and BASMU.

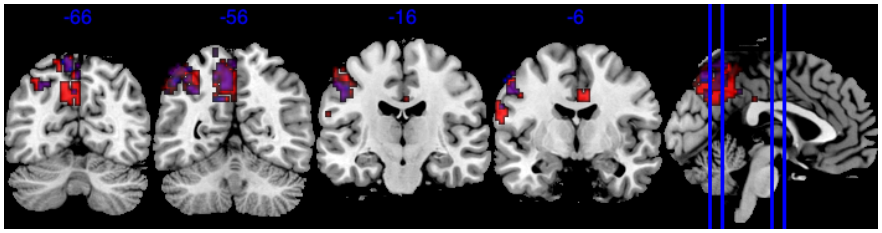


Figure 4.2: Coronal view of active NIE $\mathcal{E}$ areas. The blue areas are active mediation voxels selected by BIMA, the red areas are selected by BASMU, and the overlaying purple areas are commonly selected by both methods.

Based on Table 4.2, BASMU tends to select more active mediation voxels in all top four regions compared to BIMA. Especially in region 34 (right middle cingulum), an area for integrating memory information, BASMU selects three times more mediation voxels, although the scalar NIE for this region remains unchanged. In terms of effect size, BASMU in region 57 (left postcentral), an area for episodic memory retrieval, has a larger scale in NIE. In the scatter plot in Appendix Figure C.4b, we can see that the large positive voxels are usually selected by both methods, whereas BASMU gives more selection on smaller positive effects. In the coronal view of the brain image NIE Figure 4.2, the center of the large active areas are usually selected by both methods, whereas the edge of those large areas tends to differ between BIMA and BASMU. In summary, after accounting for the unobserved confounders, we tend to select more active mediation voxels, and the effect of parental education on children's general cognitive ability mediated through children's brain activity takes a larger proportion in the total effect after adjusting for the unobserved confounders. The NDE also tends to decrease after adjusting for unobserved confounders.

4.6 Conclusion and Discussion

In this work, based on the BIMA [97], we propose the BASMU framework for structured mediators to account for the unobserved confounder effects. We utilize the individual effects

as the unobserved confounders and incorporate them into the outcome model. We provide rigorous proof for the theoretical analysis on the asymptotic bias of the outcome model, and the identifiability of the BASMU model. For the estimation step, due to the complexity of the BASMU model, we propose the two-stage estimation algorithm. While full Bayesian inference from joint estimation is challenging, our two-stage estimation method yields reasonably accurate point estimates for β and NIE, as evidenced by extensive simulation results. Alternative ways of the two-stage algorithm include bootstrap, or updating η_i and updating all other parameters until convergence and then iteratively updating η_i in loops until full convergence. We apply BASMU in the ABCD study and find the mediation effect takes a larger proportion after adjusting for the unobserved confounders. The limitation of this work is that the mediator has to be spatially smooth or satisfy certain pre-fixed correlation structures for the individual effects to be estimated. In practice, unobserved confounders with more complex correlation structures may not be fully accounted for under the current BASMU framework.

CHAPTER 5

Future Work

The preceding chapters have discussed the Bayesian Image Mediation problem and proposed novel methods with theoretical guarantees or computational advantages. However, there are still several aspects of this problem that have not been the main focus of this dissertation, but are important to address in the future. In this chapter, we provide some extended discussion in using FDR control method in mediation signal selection, and other distribution-free approaches to handle unobserved confounder in the mediation problem.

5.1 Variable Selection Procedure for FDR Control

In the scope of Chapters 2 and 3, the proposed thresholding priors can yield sparsity in the posterior samples, and we use Posterior Inclusion Probability (PIP) to determine how likely a voxel can be an active signal. However, there is a lack of discussion on how to use PIP to control False Discoveries. For the Image-on-scalar regression in Chapter 3, we use multiple comparison p-value correction method on the result of the Mass Univariate Analysis, and use the proportion of active voxels as the cutoff to choose a cutoff on PIP, so that in the simulation study, the ST-CAR prior result all has FDR below 10%. This can serve as one naive way of FDR control, but for Scalar-on-Image regressions, there are no straight forward ways to choose such a threshold on PIP.

One possible future direction is to use the knock-off idea [3]. The knock-off idea permutes the order of the observed data to find false discoveries. For example, in Scalar-on-Image regression, for one target location s_j that may be an active signal, if we only permute the individual indices i and replace $M_i(s_j)$ by the permuted $M_{(i)}(s_j)$ but keep other locations of observed data the same, and rerun the analysis. If the new result still shows location s_j as an active voxel, it is likely a false discovery. Using this approach, we can determine an interval threshold (l, r) on the PIP, where if $\text{PIP}_j < l$, it is deemed as a non-active voxel, and if $\text{PIP}_j > r$, it is deemed as an active voxel. During some preliminary simulation test, if

we start from a conservative threshold ($\text{PIP}_j > 0.9$) to determine the active voxels, permute on voxels with $\text{PIP}_j < 0.9$, we can let l be the smallest PIP on voxels that become active from non-active after the permutation, and let r be the smallest PIP voxels that become non-active from active after the permutation. We find this approach to work well in selecting true negatives and true positives in some simple simulation case. The type I and type II error using this approach requires further investigation into the knock-off and variable selection literature.

5.2 Distribution-free Approach for Unobserved Confounders

One limitation of Chapter 4 is that the unobserved confounders must follow certain smoothness assumptions to be estimated. This assumes the unobserved confounders must follow certain distribution, although it can be very flexible if we choose a flexible Gaussian kernel. Nonetheless, recent literature [40] has utlized the distributionally robust optimization. Different from traditional sensitivity analysis based on point-wise estimation, [40] proposes a new sensitivity analysis approach based on a non-parametric model, and gives an estimation on the bound of the odds ratio of selection bias caused by the unmeasured confounder. This new direction has the potential for distribution-free sensitivity analysis on the image mediation analysis with unmeasured confounders.

APPENDIX A

Chapter 2: Appendix

A.1 Proof

A.1.1 Proof of Proposition 1

Proof of Proposition 1. In this proof we omit the notations $\mu_{M,i}$ to μ_i for simplicity. First we show the identifiability of model (2.2), namely part (a) in Proposition 1.

Consider two parameter sets $\Theta_M = \{\alpha, \{\zeta_k\}_{k=1}^q, \{\eta_i\}_{i=1}^n\}$ and $\Theta'_M = \{\alpha', \{\zeta'_k\}_{k=1}^q, \{\eta'_i\}_{i=1}^n\}$ Suppose the probability distributions of $\mathbf{M}$ given $\mathbf{X}$ and $\mathbf{C}$ under Θ_M and Θ'_M are equal, i.e.,

$$\pi(\mathbf{M} \mid \mathbf{X}, \mathbf{C}, \Theta_M) = \pi(\mathbf{M} \mid \mathbf{X}, \mathbf{C}, \Theta'_M),$$

where $\mathbf{X}$ and $\mathbf{C}$ satisfy the Assumption 2. Note that $\mathbf{M} = \{M_i(s)\}$. The joint distributions of two multi-dimensional random variables are the same implies that the corresponding marginal distributions of any element of the two random variables are also the same. Hence we have for any $i \in \{1, \ldots, n\}$ and any $s \in \mathcal{B}$,

$$\pi(M_i(s) \mid \mathbf{X}, \mathbf{C}, \Theta_M) = \pi(M_i(s) \mid \mathbf{X}, \mathbf{C}, \Theta'_M).$$

Since $M_i(s)$ follows a normal distribution, for $i \in \{1, \ldots, n\}$ and any $s \in \mathcal{B}$,

$$\mu'_i(s) = \mu_i(s) \text{ and } \sigma'^2_M = \sigma^2_M,$$

where $\mu_i(s) = \alpha(s)X_i + \eta_i(s) + \sum_{k=1}^q \boldsymbol{\zeta}_k(s)C_{i,k}$ and $\mu'_i(s) = \alpha'(s)X_i + \eta'_i(s) + \sum_{k=1}^q \boldsymbol{\zeta}_k(s)C_{i,k}$. Consider the decomposition of $\mu_i(s)$, $\mu'_i(s)$, $\alpha(s), \alpha'(s)$, $\eta_i(s)$ and $\eta'_i(s)$.

$$\mu_i(s) = \sum_{l=1}^{\infty} \theta_{\mu,i,l}\psi_l(s), \quad \alpha(s) = \sum_{l=1}^{\infty} \theta_{\alpha,l}\psi_l(s), \quad \eta_i(s) = \sum_{l=1}^{\infty} \theta_{\eta,i,l}\psi_l(s), \quad \zeta_k(s) = \sum_{l=1}^{\infty} \theta_{\zeta,k,l}\psi_l(s)$$

$$\mu_i'(s) = \sum_{l=1}^{\infty} \theta_{\mu',i,l}\psi_l(s), \quad \alpha'(s) = \sum_{l=1}^{\infty} \theta_{\alpha',l}\psi_l(s), \quad \eta_i'(s) = \sum_{l=1}^{\infty} \theta_{\eta',i,l}\psi_l(s), \quad \zeta_k'(s) = \sum_{l=1}^{\infty} \theta_{\zeta',k,l}\psi_l(s),$$

where the basis coefficients are satisfied with the following identities.

$$\theta_{\mu,i,l} = \theta_{\alpha,l}X_i + \theta_{\eta,i,l} + \sum_{k=1}^{q} \theta_{\zeta,k,l}C_{i,k}, \quad \text{and} \quad \theta_{\mu',i,l} = \theta_{\alpha',l}X_i + \theta_{\eta',i,l} + \sum_{k=1}^{q} \theta_{\zeta',k,l}C_{i,k}.$$

Since $\mu_i(s) = \mu_i'(s)$ for any $i \in \{1, \ldots, n\}$ and any $s \in \mathcal{B}$, then for any $l \geq 1$, $\theta_{\mu,i,l} = \theta_{\mu',i,l}$. Then we have $(\theta_{\alpha,l} - \theta_{\alpha',l})X_i + \theta_{\eta,1,l} - \theta_{\eta',1,l} + \sum_{k=1}^{q} (\theta_{\zeta,k,l} - \theta_{\zeta',k,l})\, C_{1,k} = 0$. According to the Assumption 2, for $t = 1, \ldots, q+1$, $\sum_{i=1}^{n} W_{i,t}(\theta_{\eta,i,l} - \theta_{\eta',i,l}) = 0$. Let $\mathbf{b}_l = (\theta_{\alpha_1,l} - \theta'_{\alpha_1,l}, \theta_{\zeta,1,l} - \theta_{\zeta',1,l}, \ldots, \theta_{\zeta,q,l} - \theta_{\zeta',q,l}, \theta_{\eta,1,l} - \theta'_{\eta,1,,l}, \ldots, \theta_{\eta,n,,l} - \theta'_{\eta,n,,l})^\top$ for any $l \geq 1$ and

$$\mathbf{A} = \begin{pmatrix} \mathbf{0}_{(q+1)\times 1} & \mathbf{0}_{(q+1)\times q} & \mathbf{W}^\top \\ \mathbf{X} & \mathbf{C} & \mathbf{I}_n \end{pmatrix},$$

where $\mathbf{b}_l$ is of dimension $(q + 1 + n) \times 1$ and $\mathbf{A}$ is of dimension $(n + q + 1) \times (n + q + 1)$. Then we have the linear system: $\mathbf{A}\mathbf{b}_l = \mathbf{0}_{(n+q+1)\times 1}$.

Denote $\tilde{\boldsymbol{X}} = (\mathbf{X}_{n\times 1}, \mathbf{C}_{n\times q}) \in \mathbb{R}^{n\times(q+1)}$. Note that $\det(\mathbf{A}) = \det(\mathbf{0} - \mathbf{W}^\top \mathbf{I}_n^{-1}\tilde{\boldsymbol{X}}) \det(\mathbf{I}_n) = \det(\mathbf{W}^{\mathrm{T}}\tilde{\boldsymbol{X}}) \neq 0$ by Assumption 2. This implies that $\mathbf{0}_{n+1+q}$ is the unique solution of $\mathbf{A}\mathbf{b}_l = \mathbf{0}_{n+1+q}$. Thus

$$\theta_{\alpha,l} = \theta_{\alpha',l}, \qquad \theta_{\eta,i,l} = \theta_{\eta',i,l}, \qquad \theta_{\zeta,k,l} = \theta_{\zeta',k,l}$$

This further implies that for any s and any i,

$$\alpha(s) = \alpha'(s), \qquad \eta_i(s) = \eta_i'(s), \qquad \zeta_k(s) = \zeta_k'(s)$$

This proves the identifiability of model (2.2). Next, we show the statement in (b) in Proposition 1. Part (b) will be used in the proof of Theorem 1.

By directly setting $\mathbf{W} = \tilde{\boldsymbol{X}}$, and $\sum_{i=1}^{n} W_{i,t}\eta_i(s) = 0$ for $t = 1, \ldots, q+1$, we know that $\sum_{i=1}^{n} \tilde{X}_{i,t}\eta_i(s) = 0$ for $t = 1, \ldots, q+1$. For each s, let $\tilde{\boldsymbol{\alpha}}(s) = \{\alpha(s), \zeta_1(s), \ldots, \zeta_q(s)\}^{\mathrm{T}} \in \mathbb{R}^{q+1}$ and $\tilde{\boldsymbol{\alpha}}'(s) = \{\alpha'(s), \zeta_1'(s), \ldots, \zeta_q'(s)\}^{\mathrm{T}} \in \mathbb{R}^{q+1}$. Let $\tilde{\mathbf{b}}_l = (\theta_{\alpha_1,l} - \theta'_{\alpha_1,l}, \theta_{\zeta,1,l} - \theta_{\zeta',1,l}, \ldots, \theta_{\zeta,q,l} - \theta_{\zeta',q,l})^{\mathrm{T}}$ and $\mathbf{g}_l = (\theta_{\eta,1,l} - \theta'_{\eta,1,,l}, \ldots, \theta_{\eta,n,,l} - \theta'_{\eta,n,,l})^{\mathrm{T}}$. Then $\tilde{\mathbf{X}}_i^{\mathrm{T}}\{\tilde{\boldsymbol{\alpha}}(s) - \tilde{\boldsymbol{\alpha}}'(s)\} = \sum_{l=1}^{\infty} \tilde{\mathbf{X}}_i^{\mathrm{T}}\tilde{\mathbf{b}}_l\psi_l(s)$

and $\tilde{\eta}_i(s)-\tilde{\eta}_i'(s) = \sum_{l=1}^{\infty} g_{l,i}\psi_l(s)$. Since $\int_{\mathcal{S}}\{\mu_i(s)-\mu_i'(s)\}^2\lambda(\mathrm{d}s)$ is finite, by Fubini's theorem,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{S}}\{\mu_i(s)-\mu_i'(s)\}^2\lambda(\mathrm{d}s)\\
&=\frac{1}{n}\int_{\mathcal{S}}\sum_{i=1}^{n}\left\{\tilde{\boldsymbol{X}}_i^{\mathrm{T}}(\tilde{\boldsymbol{\alpha}}(s)-\tilde{\boldsymbol{\alpha}}'(s))\right\}^2\lambda(\mathrm{d}s)+\frac{1}{n}\int_{\mathcal{S}}\sum_{i=1}^{n}\{\eta_i(s)-\eta_i'(s)\}^2\lambda(\mathrm{d}s)\\
&=\frac{1}{n}\int_{\mathcal{S}}\sum_{i=1}^{n}\left\{\left(\sum_{l=1}^{\infty}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{b}}_l\psi_l(s)\right)^2+\left(\sum_{l=1}^{\infty}g_{l,i}^2\psi_l(s)\right)^2\right\}\lambda(\mathrm{d}s)\\
&=\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{l=1}^{\infty}(\tilde{\mathbf{X}}_i^{\mathrm{T}}\tilde{\mathbf{b}}_l)^2+\sum_{l=1}^{\infty}g_{l,i}^2\right\}\\
&=\frac{1}{n}\sum_{l=1}^{\infty}\|\tilde{\mathbf{X}}\tilde{\mathbf{b}}_l\|_2^2+\frac{1}{n}\sum_{l=1}^{\infty}\|\mathbf{g}_l\|_2^2.
\end{aligned}
$$

By Assumption 2(a) that $\sigma_{\min}(\tilde{\mathbf{X}})>\sqrt{n}$, $\|\tilde{\mathbf{X}}\tilde{\mathbf{b}}_l\|_2^2\geq\sigma_{\min}^2(\tilde{\mathbf{X}})\|\tilde{\mathbf{b}}_l\|_2^2\geq n\|\tilde{\mathbf{b}}_l\|_2^2$. Hence

$$
\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{S}}\{\mu_i(s)-\mu_i'(s)\}^2\lambda(ds)\geq\sum_{l=1}^{\infty}\|\tilde{\mathbf{b}}_l\|_2^2+\frac{1}{n}\sum_{l=1}^{\infty}\|\mathbf{g}_l\|_2^2.
$$

Note that the empirical norm $\|f\|_{2,p}$ is a finite grid approximation of the Hilbert space inner product $\sqrt{\int_{\mathcal{S}}f^2(s)\lambda(\mathrm{d}s)}$. By Definition 4(d), the approximation error is given by $err(f)=\left|\|f\|_{2,p}^2-\int_{\mathcal{S}}f^2(s)\lambda(\mathrm{d}s)\right|\leq Kp^{-2/d}$.

$$
\begin{aligned}
\|\alpha-\alpha'\|_{2,p}^2&=\sum_{l=1}^{\infty}(\theta_{\alpha,l}-\theta_{\alpha',l})^2+err(\alpha-\alpha')\\
\|\zeta_k-\zeta_k'\|_{2,p}^2&=\sum_{l=1}^{\infty}(\theta_{\zeta_k,l}-\theta_{\zeta_k',l})^2+err(\zeta_k-\zeta_k'),\ k=1,...,q\\
\|\eta_i-\eta_i'\|_{2,p}^2&=\sum_{l=1}^{\infty}(\theta_{\eta_i,l}-\theta_{\eta_i',l})^2+err(\eta_i-\eta_i'),\ i=1,...,n
\end{aligned}
$$

For n large enough such that $Kp^{-2/d}<\frac{1}{q+3}\epsilon^2$, the following inequality

$$
\|\alpha(s)-\alpha'(s)\|_{2,p}^2+\sum_{k=1}^{q}\|\zeta_k(s)-\zeta_k'(s)\|_{2,p}^2+\frac{1}{n}\sum_{i=1}^{n}\|\eta_i(s)-\eta_i'(s)\|_{2,p}^2>\epsilon^2
$$

implies that there exists constant $c_1'\sum_{l=1}^{\infty}\|\tilde{\mathbf{b}}_l\|_2^2+n^{-1}\sum_{l=1}^{\infty}\|\mathbf{g}_l\|_2^2>c_1'\epsilon^2$ which further implies

that there exists constant c_0,

$$\frac{1}{n}\sum_{i=1}^{n}\|\mu_i(s)-\mu_i'(s)\|_{2,p}^2 > c_0\epsilon^2$$

Hence Proposition 1(b) follows.

□

A.1.2 Proof of Theorem 1

Theorem 1 is proved by checking the conditions in Theorem A.1 in [13].

For simplicity, throughout the proof of Theorem 1, we use the following notations: $\theta = \{\alpha, \{\zeta_k\}_{k=1}^q, \{\eta_i\}_{i=1}^n\}$, and the true parameters denoted as $\theta_0 = \{\alpha_0, \{\zeta_k^0\}_{k=1}^q, \{\eta_i^0\}_{i=1}^n\}$. In addition, let $\mu_i(s) = \alpha(s)X_i + \sum_{k=1}^q \zeta_k(s)C_{i,k} + \eta_i(s)$ be the mean function given $\{X_i, \{C_{i,k}\}_{k=1}^q\}_{i=1}^n$, and $\mu_i^0(s)$ be the mean function under the true parameters.

Conditional on $\{X_i, \{C_{i,k}\}_{k=1}^q\}_{i=1}^n$, for individual i and location s_j, $M_i(s_j)$ follows independent distributions across $i = 1, \ldots, n, j = 1, \ldots, p$, with density function $\pi(M_i(s_j); \theta) = \phi(\mu_i(s_j), \sigma^2)$, where $\phi(\mu_i(s_j), \sigma^2)$ is used to denote the normal density with mean $\mu_i(s_j)$ and variance σ^2. Let $\Lambda_{i,j}(\theta_0, \theta) := \log\{\pi(M_i(s_j); \theta_0)/\pi(M_i(s_j); \theta)\}$.

First, we verify the prior positivity condition as follows.

Lemma 1. *(Prior positivity condition) There exists a set B, $\Pi(B) > 0$ such that*

1. $\liminf_{\{n,p\}\to\infty} \Pi\left\{\theta \in B : (np)^{-1}\sum_{i=1}^n\sum_{j=1}^p \mathbb{E}_{\theta_0}\{\Lambda_{i,j}(\theta_0,\theta)\} < \epsilon\right\} > 0$ *for all* $\epsilon > 0$*; and*

2. $(np)^{-2}\sum_{i=1}^n\sum_{j=1}^p \mathrm{Var}_{\theta_0}\{\Lambda_{i,j}(\theta_0,\theta)\} \to 0$, *as* $n \to \infty$ *and* $p \to \infty$, *for all* $\theta \in B$.

Proof. Define

$$\|\theta-\theta_0\|_\infty = \max\left\{\sup_{s\in\mathcal{S}}|\alpha(s)-\alpha_0(s)|, \max_k \sup_{s\in\mathcal{S}}|\zeta_k(s)-\zeta_k^0(s)|, \max_i \sup_{s\in\mathcal{S}}|\eta_i(s)-\eta_i^0(s)|\right\}. \tag{A.1}$$

For constant $\delta > 0$, consider

$$B_\delta = \{\theta \in \Theta : \|\theta-\theta_0\|_\infty < \delta\}.$$

Since the prior distributions for the above parameters are independent, to show $\Pi(B_\delta) > 0$, we only need to show that the prior of each term in (A.1) being upper bounded by a constant has a positive probability.

By Theorem 4 in [29], for any $i = 1, \ldots, n$, $k = 1, \ldots, q$,

$$\Pi\left(\sup_{s\in\mathcal{S}} |\eta_i(s) - \eta_i^0(s)| < \delta\right) > 0, \quad \Pi\left(\sup_{s\in\mathcal{S}} |\zeta_k(s) - \zeta_k^0(s)| < \delta\right) > 0.$$

By Lemma 2 in [43], for any threshold $\nu > 0$ and any true $\alpha_0(s) \in \Theta_\alpha$, there exists $\tilde{\alpha}(s)$ in the RKHS of $\kappa(\cdot,\cdot)$ such that $\alpha_0 = T_\nu(\tilde{\alpha}_0)$. Note that the soft-thresholding function $T_\nu(x)$ is a 1-Lipschitz continuous function of x, and by Theorem 4 in [29], we have $\Pi\left(\sup_{s\in\mathcal{S}} |\tilde{\alpha}(s) - \tilde{\alpha}_0(s)| < \delta\right) > 0$, which implies $\Pi\left(\sup_{s\in\mathcal{S}} |T_\nu(\tilde{\alpha}(s)) - T_\nu(\tilde{\alpha}_0(s))| < \delta\right) > 0$. Hence for any $\theta \in B_\delta$, where $\Pi(B_\delta) > 0$, we have

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\Lambda_{i,j}(\theta_0,\theta)\right] =& \mathbb{E}\left[\mathbb{E}_{\theta_0}\left\{\Lambda_{i,j}(\theta_0,\theta) \mid \mathbf{X}, \mathbf{C}\right\}\right] \\
=& -\frac{1}{2\sigma_M^2}\mathbb{E}\left[\mathbb{E}_{\theta_0}\left\{(M_i(s_j) - \mu_i^0(s_j))^2 \mid \mathbf{X}, \mathbf{C}\right\}\right] \\
&+ \frac{1}{2\sigma_M^2}\mathbb{E}\left[\mathbb{E}_{\theta_0}\left\{(M_i(s_j) - \mu_i^0(s_j) + \mu_i^0(s_j) - \mu_i(s_j))^2 \mid \mathbf{X}, \mathbf{C}\right\}\right] \\
=& \mathbb{E}\left[\frac{1}{2\sigma_M^2}(\mu_i^0(s_j) - \mu_i(s_j))^2\right]
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{1}{2\sigma_M^2}\{\mu_i^0(s_j) - \mu_i(s_j)\}^2 \\
&\leq \frac{1}{2\sigma_M^2}\left[\{\alpha(s_j) - \alpha_0(s_j)\}X_i + \sum_{k=1}^{q}\{\zeta_k(s_j) - \zeta_k^0(s_j)\}C_{i,k} + \{\eta_i(s_j) - \eta_i^0(s_j)\}\right]^2 \\
&\leq \frac{2}{\sigma_M^2}\left[X_i^2\{\alpha(s_j) - \alpha_0(s_j)\}^2 + \sum_{k=1}^{q}\{\zeta_k(s_j) - \zeta_k^0(s_j)\}^2 C_{i,k}^2 + \{\eta_i(s_j) - \eta_i^0(s_j)\}^2\right]
\end{aligned}$$

By choosing a constant $K_{\max}$ such that $\max_i\{\mathbb{E}\left\{|X_i|^2\right\}, \max_k \mathbb{E}\left\{|C_{i,k}|^2\right\}\} \leq K_{\max}$, then for any $\theta \in B_\delta$, $(np)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}_{\theta_0}\left\{\Lambda_{i,j}(\theta_0,\theta)\right\} < 2\sigma_M^{-2}K_{\max}(2+q)\delta^2$, hence for a small enough ϵ such that $0 < \epsilon < \sigma_M^{-2}K_{\max}(2+q)\delta^2$,

$$\begin{aligned}
&\liminf_{\{n,p\}\to\infty} \Pi\left\{\theta \in B_\delta : \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}_{\theta_0}\left(\Lambda_{i,j}(\theta_0,\theta)\right) < \epsilon\right\} \\
&\geq \Pi\left\{\|\theta - \theta_0\|_\infty \leq \sqrt{\left(\frac{2}{\sigma_M^2}K_{\max}(2+q)\right)^{-1}\epsilon}\right\} > 0.
\end{aligned}$$

To show the second condition, we only need to show that for any i, j and any $\theta \in B_\delta$, the

variance $\mathrm{Var}_{\theta_0}\{\Lambda_{i,j}(\theta_0,\theta)\}$ is bounded by some constant.

$$\begin{aligned}\mathrm{Var}_{\theta_0}\{\Lambda_{i,j}(\theta_0,\theta)\} &= \mathbb{E}\{\mathrm{Var}_{\theta_0}\{\Lambda_{i,j}(\theta_0,\theta)\mid \mathbf{X},\mathbf{C}\}\} + \mathrm{Var}\{\mathbb{E}_{\theta_0}\{\Lambda_{i,j}(\theta_0,\theta)\mid \mathbf{X},\mathbf{C}\}\}\\ &= \mathbb{E}\left\{\frac{1}{\sigma_M^2}\left(\mu_i^0(s_j)-\mu_i(s_j)\right)^2\right\} + \mathrm{Var}\left\{\frac{1}{2\sigma_M^2}(\mu_i^0(s_j)-\mu_i(s_j))^2\right\}\\ &\le \max\left\{\frac{4}{\sigma_M^2}K_{\max}(2+q)\delta^2, \frac{4}{\sigma_M^4}K_{\max,V}(2+q)\delta^4\right\} < \infty,\end{aligned}$$

where $K_{\max,V} \ge \max_i\left\{\mathrm{Var}(X_i^2), \max_k \mathrm{Var}(C_{i,k}^2)\right\}$.

□

Before the test construction, we add a useful lemma on the tail probability of the maximum of sub-Gaussian random variables.

Lemma 2. *Let $X_i, i=1,\dots,N$ be sub-Gaussian random variables. Let σ_i^2 be the constant such that $\mathbb{P}(|X_i|>t)\le 2\exp(-t^2/\sigma_i^2)$ for any $t>0$ and $i=1,\dots,N$. Let $\tilde{\sigma}_N^2 = \bigvee_{i=1}^N \sigma_i^2$. Then for any $t>0$, $\mathbb{P}\left(\max_i |X_i| > \sqrt{\tilde{\sigma}_N^2 \log 2N + t}\right) \le \exp(-t)$.*

Proof. Let $u = \sqrt{\tilde{\sigma}_N^2 \log 2N + t}$,

$$\mathbb{P}\left(\max_i |X_i| > u\right) \le \sum_i \mathbb{P}(|X_i|>u) \le 2N\exp\left\{-u^2/\tilde{\sigma}_N^2\right\} = \exp(-t).$$

□

Next, we construct a test that satisfies the Type I and Type II error bound on a specified sieve space.

Lemma 3. *(Existence of tests) There exist test functions $\{\Phi_{np}\}$, subset $\mathcal{U}_n, \Theta_n \subset \Theta$, and constant $K_1, K_2, c_1, c_2 > 0$ such that*

(a) $\mathbb{E}_{\theta_0}\Phi_{np} \to 0$, *as* $n\to\infty$ *and* $p\to\infty$;

(b) $\sup_{\theta\in\mathcal{U}_n^c\cap\Theta_n}\mathbb{E}_\theta(1-\Phi_{np}) \le K_1 e^{-c_1 np}$;

(c) $\Pi(\Theta_n^c) \le K_2 e^{-c_2 np}$.

Proof. Define the sieve space of θ as Θ_n, which be decomposed into product of the following

parameter space:

$$\Theta_n = \Theta_{\alpha,n} \times \prod_{k=1}^{q} \Theta_{\zeta,k,n} \times \prod_{i=1}^{n} \Theta_{\eta,i,n}$$

$$\Theta_{\alpha,n} = \left\{ \alpha \in \Theta_\alpha : \sup_{s \in \mathcal{R}_1 \cup \mathcal{R}_{-1}} \|D^\omega \alpha(s)\|_\infty < \sqrt{np}, \|\omega\|_1 \leq \rho \right\}$$

$$\Theta_{\zeta,k,n} = \left\{ \zeta_k \in \Theta_\zeta : \sup_{s \in \mathcal{S}} \|D^\omega \zeta_k(s)\|_\infty < np, \|\omega\|_1 \leq \rho \right\}, k = 1, \ldots, q$$

$$\Theta_{\eta,i,n} = \left\{ \eta_i \in \Theta_\eta : \sup_{s \in \mathcal{S}} \|D^\omega \eta_i(s)\|_\infty < np, \|\omega\|_1 \leq \rho \right\}, i = 1, \ldots, n$$

where $D^\omega f(s)$ stands for $(\partial^{\|\omega\|_1}/\partial^{\omega_1}, \ldots, \partial^{\|\omega\|_1}/\partial^{\omega_d}) f(s)$ for any $\omega = (\omega_1, \ldots, \omega_d)^{\mathrm{T}}$ with $\omega_j (j = 1, \ldots, d)$ being positive intergers and $s \in \mathbb{R}^d$.

To show the conditions (a) and (b), we use Lemma 8.27(i) in [30], by viewing $\mathbf{M} \sim N_{np}(\boldsymbol{\mu}, \sigma^2 I)$, $\boldsymbol{\mu} = \{\mu_i(s_j)\}_{i=1,j=1}^{n,p} \in \mathbb{R}^{np}$. By Lemma 8.27(i), for any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_0 \in \mathbb{R}^{np}$, there exists $\Phi(\boldsymbol{\mu}_1)$ such that for any $\boldsymbol{\mu}$ where $\|\boldsymbol{\mu} - \boldsymbol{\mu}_1\|_2 \leq \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2/2$,

$$\mathbb{E}_{\boldsymbol{\mu}_0} \Phi(\boldsymbol{\mu}_1) \vee \mathbb{E}_{\boldsymbol{\mu}} \{1 - \Phi(\boldsymbol{\mu}_1)\} \leq \exp\left\{-c_1 \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2^2 / \sigma_M^2\right\}$$

Because the type II error in condition (b) does not depend on a single $\boldsymbol{\mu}_1$, to remove the dependence on $\boldsymbol{\mu}_1$, and to use a neighborhood $\mathcal{U}_n$ defined by the empirical norm as the distance metric instead of the Euclidean norm, we use the same technique as the one in Proposition 11 in [84]. For any $r \geq 1$, any integer $j \geq 1$, define shells for $\boldsymbol{\mu}$

$$\mathcal{C}_{j,r} := \{\Theta_n : jr \leq \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|_2 \leq (j+1)r\}$$

Denote $\mathcal{P}(\mathcal{C}_{j,r}, jr/2, \|\cdot\|_2)$ as the largest packing number of $\mathcal{C}_{j,r}$ with Euclidean distance $jr/2$, and denote the corresponding $jr/2$-separated set of $\mathcal{C}_{j,r}$ as $\mathcal{P}_j$. Note that $\mathcal{P}_j$ is also a $jr/2$-covering set of $\mathcal{C}_{j,r}$. Hence for any $\boldsymbol{\mu} \in \mathcal{C}_{j,r}$, there exists $\boldsymbol{\mu}_1 \in \mathcal{P}_j$ such that

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}_1\|_2 \leq \frac{jr}{2} \leq \frac{1}{2} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2.$$

Choose $\Phi_j = \max_{\boldsymbol{\mu}_1 \in \mathcal{P}_j} \{\Phi(\boldsymbol{\mu}_1)\}$, then for any $\boldsymbol{\mu} \in \mathcal{C}_{j,r}$, conditioning on $\mathbf{X}, \mathbf{C}$,

$$\mathbb{E}_{\boldsymbol{\mu}_0, \sigma_0} \Phi_j \leq 2\mathcal{P}(\mathcal{C}_{j,r}, \frac{jr}{2}, \|\cdot\|_2) \exp\left\{-c_1 [(jr)^2] / \sigma_M^2\right\}$$

$$\mathbb{E}_{\boldsymbol{\mu}, \sigma} (1 - \Phi_j) \leq \exp\left\{-c_1 [(jr)^2] / \sigma_M^2\right\}$$

Denote $\mathcal{N}(\Theta_n, r, \|\cdot\|_\infty)$ as the smallest covering number for the set Θ_n with radius r and

distance function $\|\cdot\|_\infty$. Now we need an upper bound on $\log \mathcal{N}(\Theta_n, r, \|\cdot\|_\infty)$. Note that by Lemma 2 in [29] and a similar approach in Lemma A1 in [43], there exist constants $K_\alpha, K_\zeta, K_\eta$, such that $\log \mathcal{N}(\Theta_{\alpha,n}, r, \|\cdot\|_\infty) \leq K_\alpha (np)^{d/(2\rho)} r^{-d/\rho}$, and $\log \mathcal{N}(\Theta_{\eta,i,n}, r, \|\cdot\|_\infty) \leq K_\eta (np)^{d/\rho} r^{-d/\rho}$, $\log \mathcal{N}(\Theta_{\zeta,k,n}, r, \|\cdot\|_\infty) \leq K_\zeta (np)^{d/\rho} r^{-d/\rho}$. Hence there exists constant K_0,

$$\begin{aligned}
&\log \mathcal{N}(\Theta_n, r, \|\cdot\|_\infty) \\
&\leq \log \mathcal{N}(\Theta_{\alpha,n}, r, \|\cdot\|_\infty) + \sum_{k=1}^{q} \log \mathcal{N}(\Theta_{\zeta,k,n}, r, \|\cdot\|_\infty) + \sum_{i=1}^{n} \log \mathcal{N}(\Theta_{\eta,i,n}, r, \|\cdot\|_\infty) \\
&\leq K_0 n (np)^{d/\rho} r^{-d/\rho}
\end{aligned}$$

Conditioning on $(\mathbf{X}, \mathbf{C})$, denote

$$\Theta_n^* := \left\{ \boldsymbol{\mu} \in \mathbb{R}^{np} : \mu_{ij} = \alpha(s_j) X_i + \sum_{k=1}^{q} \zeta_k(s_j) C_{i,k} + \eta_i(s_j), \theta \in \Theta_n \right\}.$$

Now we first show that conditioning on $(\mathbf{X}, \mathbf{C})$, given $c_n^* = \max_i \{|X_i|, \|\mathbf{C}_i\|_\infty\}_i$,

$$\log \mathcal{N}(\Theta_n^*, r/\left(4\sqrt{np}\right), \|\cdot\|_\infty) \leq \log \mathcal{N}(\Theta_n, r/\left(4c_n^*\sqrt{np}\right), \|\cdot\|_\infty).$$

Denote $\mathcal{S}_{\mu,n}^*$ as a $(c_n^* r)$-covering set of Θ_n^* under $\|\cdot\|_\infty$. $\mathcal{S}_{\mu,n}^*$ is constructed in the following way: for any $\boldsymbol{\mu} \in \Theta_n^*$, there exists a corresponding $\theta_\mu = (\alpha, \{\zeta_k\}_{k=1}^q, \{\eta_i\}_{i=1}^n) \in \Theta_n$ such that $\mu_{ij} = \alpha(s_j) X_i + \sum_{k=1}^q \zeta_k(s_j) C_{i,k} + \eta_i(s_j)$, hence there exists $\theta_{\mu,1} \in \mathcal{N}_{\mu,n}$ where $\mathcal{N}_{\mu,n}$ is the smallest covering set with cardinality $\mathcal{N}(\Theta_n, r, \|\cdot\|_\infty)$, and there exists corresponding $\boldsymbol{\mu}_1 \in \Theta_n^*$ given θ_1.

$$|\mu_{1,ij} - \mu_{ij}| \leq |\left(\alpha(s_j) - \alpha_1(s_j)\right) X_i| + \sum_{k=1}^{q} |\left(\zeta_k(s_j) - \zeta_{1,k}(s_j)\right) C_{i,k}| + |\eta_i(s_j) - \eta_{1,i}(s_j)| \leq c_n^* r$$

. Hence $\mathcal{S}_{\mu,n}^*$ can be constructed as a collection of all such $\boldsymbol{\mu}_1$. Let $|\mathcal{S}_{\mu,n}^*|$ be the cardinality of such $\mathcal{S}_{\mu,n}^*$. By the construction of $\mathcal{S}_{\mu,n}^*$, $|\mathcal{S}_{\mu,n}^*| \leq \mathcal{N}(\Theta_n, r, \|\cdot\|_\infty)$.

Since $\|\cdot\|_{2,np} \leq \|\cdot\|_\infty$, we have

$$\begin{aligned}
\log \mathcal{P}(\Theta_n^*, r/2, \|\cdot\|_2) &\leq \log \mathcal{N}(\Theta_n^*, r/4, \|\cdot\|_2) = \log \mathcal{N}(\Theta_n^*, r/\left(4\sqrt{np}\right), \|\cdot\|_{2,np}) \\
&\leq \log \mathcal{N}(\Theta_n^*, r/\left(4\sqrt{np}\right), \|\cdot\|_\infty) \leq \log |\mathcal{S}_{\mu,n}^*| \\
&\leq \log \mathcal{N}(\Theta_n, r/\left(4c_n^*\sqrt{np}\right), \|\cdot\|_\infty) \\
&\leq K_0 (4c_n^*)^{d/\rho} n (np)^{3d/(2\rho)} r^{-d/\rho} \qquad \text{(A.2)}
\end{aligned}$$

Denote event $A = \left[c_n^* < a\sqrt{\log\{n\}}\right]$ and I_A be its indicator, where a is an absolute constant, Lemma 2 implies that $\mathbb{P}(I_{A^c}) \to 0$ as $n \to \infty$, where A^c denotes the complement of A. Hence given A, $\log \mathcal{P}(\Theta_n^*, r/2, \|\cdot\|_2) \leq K_a (\log n)^{d/(2\rho)} n(np)^{3d/(2\rho)} r^{-d/\rho}$.

Then for any $\boldsymbol{\mu} \in \cup_{j\geq 1}\mathcal{C}_{j,r}, \sigma \in \cup_{j\geq 1}\mathcal{C}_{j,\epsilon}$, define $\Phi = \sum_{j\geq 1} \Phi_j I(\boldsymbol{\mu} \in \mathcal{C}_{j,r})$, for some constants K_2, K_3, conditioning on $\mathbf{X}, \mathbf{C}$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}_0}\Phi &\leq \sum_{j\geq 1} 2\mathcal{P}(\mathcal{C}_{j,r}, jr/2, \|\cdot\|_2) \exp\{-c_1[(jr)^2]/\sigma_M^2\} \\
&\leq 2\mathcal{P}(\Theta_n^*, r/2, \|\cdot\|_2) \sum_{j\geq 1} \exp\{-c_1[j(r)^2]/\sigma_M^2\} \\
&\leq 2\mathcal{P}(\Theta_n^*, r/2, \|\cdot\|_2) K_2 \exp\left(-\frac{c_1 r^2}{4\sigma_M^2}\right) \\
&\leq K_3 \mathcal{P}(\Theta_n^*, r/2, \|\cdot\|_2) \exp\left(-\frac{c_1 r^2}{\sigma_M^2}\right) \\
\mathbb{E}_{\boldsymbol{\mu}}(1-\Phi) &\leq \sum_{j\geq 1} \exp\{-c_1[(jr)^2]\left(2\sigma_M^2\right)^{-1}\} \\
&\leq K_3 \exp\left\{-\frac{c_1 r^2}{\sigma_M^2}\right\}
\end{aligned}$$

Choose $r = \sqrt{np}\epsilon$, for any $\epsilon > 0$, we can choose n, p large enough such that $r > 1$. By Proposition 1(b), $\mathcal{U}_M^c \subset \mathcal{U}_{M,1}^c$ almost surely, where

$$\begin{aligned}
\mathcal{U}_M^c &= \left\{\Theta : \|\alpha(s) - \alpha_0(s)\|_{2,p}^2 + \sum_{k=1}^q \|\zeta_k(s) - \zeta_{k,0}(s)\|_p^2 + \frac{1}{n}\sum_{i=1}^n \|\eta_i(s) - \eta_i(s)\|_{2,p}^2 > \epsilon^2\right\} \\
\mathcal{U}_{M,1}^c &= \{\Theta : \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|_{2,np} > \sqrt{c_0}\epsilon\}
\end{aligned}$$

Then for any $\theta \in \Theta_n \cap \mathcal{U}_{M,1}^c$, note that $(\log n)^{d/(2\rho)}(np)^{d/\rho} < n^{d/(2\rho)}(np)^{d/\rho} < p$ given Assumption 1, $\rho > d + 3/(2\tau)$.

$$\begin{aligned}
&\mathbb{E}\left\{\mathbb{E}_{\boldsymbol{\mu}_0,\sigma_0}\left\{\Phi \mid \mathbf{X}, \mathbf{C}\right\}\right\} \leq \mathbb{E}_A\left\{\mathcal{P}(\Theta_n^*, \sqrt{np}c_0\epsilon/2, \|\cdot\|_2)\right\} K_4 \exp\{-c_1'' np\epsilon^2\} + \mathbb{E}\left\{I_{A^c}\right\} \\
&\quad \leq K' \exp\left\{c_1'''(\log n)^{d/(2\rho)} n(np)^{d/\rho}\epsilon^{-d/\rho} - c_1'' np\epsilon^2\right\} \overset{p\to\infty}{\to} 0 \\
&\mathbb{E}_{\Theta_n\cap\mathcal{U}_M^c}(1-\Phi) \leq E_{\Theta_n\cap\mathcal{U}_{M,1}^c}(1-\Phi) \leq K'' \exp\{-c_2' np\epsilon^2\}
\end{aligned}$$

To verify (c), $\Pi(\Theta_n^c) \leq \Pi(\Theta_{\alpha,n}^c) + \sum_{i=1}^n \Pi(\Theta_{\eta,i,n}^c) + \sum_{k=1}^q \Pi(\Theta_{\zeta,k,n}^c)$. Theorem 5 in [29] ensures that $\Pi(\Theta_{\eta,i,n}^c) \leq K_3 e^{-c_3(np)^2}, \Pi(\Theta_{\zeta,k,n}^c) \leq K_3 e^{-c_3(np)^2}$, Lemma 4 in [43] ensures that

$\Pi(\Theta_{\alpha,n}^c) \leq K_\alpha e^{-c_\alpha np}$. Hence

$$\begin{aligned}\Pi(\Theta_n^c) &\leq K_\alpha e^{-c_\alpha np} + K_3 e^{-\left(c_3(np)^2 - \log(n+q)\right)} \\ &\leq K_2 e^{-c_2 np}\end{aligned}$$

□

The proof for Theorem 1 is complete. Note that this can be easily extended to the marginal consistency of α alone by conditioning on other parameters at the true value.

A.1.3 Proof of Theorem 2

Similar to Theorem 1, we verify the conditions in Theorem A.1 in [13].

Let θ_0 denote the set of true parameters $\{\beta_0, \gamma_0, \boldsymbol{\xi}_0\}$ that generate the outcome variable Y_i given $\mathcal{M}_i$, X_i and $\mathbf{C}_i$. Let $\theta = (\beta, \gamma, \boldsymbol{\xi}) \in \Theta_\beta \times \mathbb{R}^{q+1}$ denote any parameter in the parameter space, where Θ_β is defined in Definition 4.

Lemma 4. *(Prior positivity condition) Under model (2.1), define* $\Lambda_i(\theta_0, \theta) = \log\{\pi(Y_i;\theta_0)/\pi(Y_i;\theta)\}$, *there exists a set* $B \subset \Theta$ *such that* $\Pi(B) > 0$ *and for any* $\theta \in B$*:*

(a) $\liminf_{n\to\infty} \Pi[\theta \in B : n^{-1}\sum_{i=1}^n \mathbb{E}_{\theta_0}\{\Lambda_i(\theta_0,\theta)\} < \epsilon] > 0$ *for any* $\epsilon > 0$

(b) $n^{-2}\sum_{i=1}^n \mathrm{Var}_{\theta_0}\{\Lambda_i(\theta_0,\theta)\} \to 0$

Proof. For one individual i, the density

$$\pi_i(Y_i, \mathcal{M}_i, X_i, \mathbf{C}_i; \theta) = \pi_i(Y_i|\mathcal{M}_i, X_i, \mathbf{C}_i; \theta)\pi_i(\mathcal{M}_i, X_i, \mathbf{C}_i).$$

Here, with the abbreviated notation $\tilde{\boldsymbol{\gamma}} = (\gamma, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{q+1}$, and $\tilde{\mathbf{X}}_i = (X_i, \mathbf{C}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{q+1}$. Hence given $\left\{\tilde{\mathbf{X}}_i\right\}_{i=1}^n$ and $\{\mathcal{M}_i(\Delta s_j)\}_{i=1,j=1}^{n,p}$, and denote $\boldsymbol{\mathcal{M}}_i = \{\mathcal{M}_i(\Delta s_j)\}_{j=1}^p$,

$$Y_i \overset{\text{ind}}{\sim} N\left(\sum_{j=1}^p \beta(s_j)\mathcal{M}_i(\Delta s_j) + \tilde{\boldsymbol{\gamma}}^{\mathrm{T}}\tilde{\mathbf{X}}_i, \sigma_Y^2\right).$$

Given the true parameters $\beta_0(s)$, $\tilde{\boldsymbol{\gamma}}_0$, define the subset

$$B_\delta = \left\{\Theta : \sup_j |\beta(s_j) - \beta_0(s_j)|^2 \leq \delta, \|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0\|_2^2 \leq \delta\right\}$$

If we denote the mean of Y_i under true parameters as $\mu_{i,0}$, otherwise as μ_i, the log-likelihood ratio for $\theta_0 = (\beta_0(s), \tilde{\boldsymbol{\gamma}}_0)$ versus $\theta = (\beta(s), \tilde{\boldsymbol{\gamma}})$ can be written as

$$\begin{aligned}\Lambda_i(D_{n,i}; \theta_0, \theta) &= \log\{\pi_i(Y_i; \beta_0(s), \tilde{\boldsymbol{\gamma}}_0)\} - \log\{\pi_i(Y_i; \beta(s), \tilde{\boldsymbol{\gamma}})\} \\ &= -\frac{1}{2\sigma_Y^2}(Y_i - \mu_{i,0})^2 + \frac{1}{2\sigma_Y^2}(Y_i - \mu_i)^2\end{aligned}$$

Hence,

$$\begin{aligned}K_{i,n}(\theta_0, \theta) &:= \mathbb{E}_{\theta_0}(\Lambda_i(D_{n,i}; \theta_0, \theta)) = \mathbb{E}\left\{\mathbb{E}_{\theta_0}\left(\Lambda_i | \boldsymbol{\mathcal{M}}_i, \tilde{\mathbf{X}}_i\right)\right\} \\ &= \mathbb{E}\left\{\frac{1}{2\sigma_Y^2}(\mu_i - \mu_{i,0})^2\right\} \\ &\leq \mathbb{E}\left[\frac{1}{2\sigma_Y^2}\left\{(\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0)^{\mathrm{T}}\tilde{\mathbf{X}}_i + \sum_{j=1}^{p}(\beta(s_j) - \beta_0(s_j))\mathcal{M}_i(\Delta s_j)\right\}^2\right]\end{aligned}$$

Note that by equation (2.4), given $\tilde{\mathbf{X}}_i$, $\mathcal{M}_i(\Delta s_j) \sim N(\mu_i(s_j)\lambda(\Delta s_j), \sigma_M^2\lambda(\Delta s_j))$ with its second moment as $\sigma_M^2\lambda(\Delta s_j) - (\mu_i(s_j)\lambda(\Delta s_j))^2$. When $\lambda(\Delta s_j) = 1/p$, the second moment is $\sigma_M^2/p - (\mu_i(s_j))^2/p^2$, and its 4th moment is of the order $O(p^{-4})$. Hence $\mathbb{E}\left\{\|\boldsymbol{\mathcal{M}}_i\|_2^2 \big| \tilde{\mathbf{X}}_i\right\}$ can be upper bounded by a constant, and so does $\mathrm{Var}\left\{\|\boldsymbol{\mathcal{M}}_i\|_2^2 \big| \tilde{\mathbf{X}}_i\right\}$. For the finite dimensional vector $\tilde{\mathbf{X}}_i$ with finite 4-th moment (Assumption 2(a)), there is a finite bound $\mathbb{E}\|\tilde{\mathbf{X}}_i\|_2^2 < K_0$.

For any $(\tilde{\boldsymbol{\gamma}}, \beta(s)) \in B_\delta$,

$$K_{i,n}(\theta_0, \theta) \leq \frac{1}{2\sigma_Y^2}\mathbb{E}\left\{\delta\|\tilde{\mathbf{X}}_i\|_2^2\|\boldsymbol{\mathcal{M}}_i\|_2^2\right\}$$

Hence we have $K_{i,n}(\theta_0, \theta) \leq \delta K'$ for some constant $K' > 0$. Similarly, denote $Z_i =$

$(Y_i - \mu_{i,0})/\sigma_Y$ as the standard normal variable under H_0,

$$
\begin{aligned}
V_{i,n}(\theta_0, \theta) = & \operatorname{Var}\left\{\mathbb{E}_{\theta_0}(\Lambda_i \mid \tilde{\mathbf{X}}_i, \boldsymbol{\mathcal{M}}_i)\right\} + \mathbb{E}\left\{\operatorname{Var}_{\theta_0}(\Lambda_i \mid \tilde{\mathbf{X}}_i, \boldsymbol{\mathcal{M}}_i)\right\} \\
\operatorname{Var}\left\{\mathbb{E}_{\theta_0}(\Lambda_i|\tilde{\mathbf{X}}_i, \boldsymbol{\mathcal{M}}_i)\right\} = & \operatorname{Var}\left\{\frac{1}{2\sigma_Y^2}(\mu_i - \mu_{i,0})^2\right\} \\
\leq & \frac{1}{4\sigma_Y^4}\operatorname{Var}\left[\left\{(\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0)^{\mathrm{T}}\tilde{\mathbf{X}}_i + \sum_{j=1}^{p}(\beta(s_j) - \beta_0(s_j))\mathcal{M}_i(\Delta s_j)\right\}^2\right] \\
\leq & \frac{1}{\sigma_Y^4}\operatorname{Var}\left[\delta\|\tilde{\mathbf{X}}_i\|_2^2 + \delta\|\boldsymbol{\mathcal{M}}_i\|_2^2\right] \\
\leq & \frac{1}{\sigma_Y^4}\operatorname{Var}\left\{\delta\|\tilde{\mathbf{X}}_i\|_2^2 + \mathbb{E}\left(\delta\|\boldsymbol{\mathcal{M}}_i\|_2^2\middle|\tilde{\mathbf{X}}_i\right)\right\} + \\
& \frac{1}{\sigma_Y^4}\mathbb{E}\left\{\operatorname{Var}\left(\delta\|\tilde{\mathbf{X}}_i\|_2^2 + \delta\|\boldsymbol{\mathcal{M}}_i\|_{2,p}^2\middle|\tilde{\mathbf{X}}_i\right)\right\} \\
< & \infty
\end{aligned}
$$

For the second term,

$$
\begin{aligned}
\mathbb{E}_{\theta_0}\left\{\operatorname{Var}_{\theta_0}(\Lambda_i|\tilde{\mathbf{X}}_i, \boldsymbol{\mathcal{M}}_i)\right\} = & \mathbb{E}\left[\operatorname{Var}_{\theta_0}\left\{-\frac{1}{2}Z_i^2 + \frac{1}{2}\left(Z_i + \frac{\mu_{i,0} - \mu_i}{\sigma_Y}\right)^2\middle|\tilde{\mathbf{X}}_i, \boldsymbol{\mathcal{M}}_i\right\}\right] \\
= & \mathbb{E}\left[\operatorname{Var}_{\theta_0}\left\{\frac{\mu_{i,0} - \mu_i}{\sigma_Y}Z_i\middle|\tilde{\mathbf{X}}_i, \boldsymbol{\mathcal{M}}_i\right\}\right] \\
= & \mathbb{E}\left\{\frac{1}{\sigma_Y^2}(\mu_i - \mu_{i,0})^2\right\} \\
\leq & \frac{1}{\sigma_Y^2}\mathbb{E}\left(\delta\|\tilde{\mathbf{X}}_i\|_2^2 + \delta\|\boldsymbol{\mathcal{M}}_i\|_2^2\right) < \infty
\end{aligned}
$$

Hence for any $\beta \in B_\delta$,

$$
\frac{1}{n^2}\sum_{i=1}^{n} V_{n,i}(\beta_0, \beta) \to 0
$$

For any $0 < \epsilon < \delta K'$,

$$
\begin{aligned}
& \Pi\left((\beta, \tilde{\boldsymbol{\gamma}}, \sigma_Y) \in B_\delta : \frac{1}{n}\sum_{i=1}^{n} K_{n,i} < \epsilon\right) \\
& \geq \Pi\left(\sup_j |\beta_0(s_j) - \beta(s_j)| < \sqrt{\epsilon/K'}, \|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0\|_2^2 < \epsilon/K'\right) > 0.
\end{aligned}
$$

The last inequality follows from Theorem 1 in [43] and the assumption that for any $\epsilon > 0$, $\Pi(\|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0\|_2^2 < \epsilon) > 0$.

□

Verifying the Existence of test condition

To verify the existence of test condition, we need the basis expansion expression of model (2.1). Recall model (2.1), we abbreviate the scalar and vector covariates and denote $\tilde{\boldsymbol{\gamma}} = (\gamma, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{q+1}$, $\tilde{\mathbf{X}}_i = (X_i, \mathbf{C}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{q+1}$. Let $\tilde{\mathcal{M}}_{i,l} = \sum_{j=1}^{p} \psi_l(s_j)\mathcal{M}_i(\Delta s_j)$, and define the $n \times L_n$ matrix $\tilde{\boldsymbol{\mathcal{M}}}_n := (\tilde{\mathcal{M}}_{i,l})_{i=1,..,N,l=1,...,L_n}$.

$$
\begin{aligned}
Y_i &= \sum_{j=1}^{p} \beta(s_j)\mathcal{M}_i(\Delta s_j) + \tilde{\boldsymbol{\gamma}}^{\mathrm{T}}\tilde{\mathbf{X}}_i + \epsilon_i \\
&= \sum_{j=1}^{p} \left\{ \sum_{l=1}^{\infty} \theta_{\beta,l}\psi_l(s_j) \right\} M_i(\Delta s_j) + \tilde{\boldsymbol{\gamma}}^{\mathrm{T}}\tilde{\mathbf{X}}_i + \epsilon_i \\
&= \sum_{l=1}^{\infty} \theta_{\beta,l} \sum_{j=1}^{p} \psi_l(s_j)\mathcal{M}_i(\Delta s_j) + \tilde{\boldsymbol{\gamma}}^{\mathrm{T}}\tilde{\mathbf{X}}_i + +\epsilon_i \\
&= \sum_{l=1}^{\infty} \theta_{\beta,l}\tilde{\mathcal{M}}_{i,l} + \tilde{\boldsymbol{\gamma}}^{\mathrm{T}}\tilde{\mathbf{X}}_i + \epsilon_i \\
&= (\tilde{\boldsymbol{\mathcal{M}}}_n, \tilde{\mathbf{X}}_n) \begin{pmatrix} \boldsymbol{\theta}_\beta \\ \tilde{\boldsymbol{\gamma}} \end{pmatrix} + r_{L_n,i} + \epsilon_i
\end{aligned} \tag{A.3}
$$

The remainder term $r_{L_n,i} = \sum_{l=L_n}^{\infty} \theta_{\beta,l} \sum_{j=1}^{p} \psi_l(s_j)\mathcal{M}_i(\Delta s_j)$.

Before verifying the existence of test condition, we introduce the following lemma

Lemma 5. *Let independent residual terms*

$$
r_{L_n,i} = \sum_{l=L_n}^{\infty} \theta_{\beta,l} \sum_{j=1}^{p} \psi_l(s_j)\mathcal{M}_i(\Delta s_j)
$$

as defined in (A.3) *across* $i = 1, \ldots, n$. *Denote the event* $A_{L_n} = [|r_{L_n,i}| < t]$. *Then for any given* i, *and for some sufficiently large positive constant* t , $\mathbb{P}[A_{L_n} i.o.] = 1$.

Proof. Denote the mean function in (2.4) of $\mathcal{M}_i(\Delta s_j)$ as $\mu_i(s_j)$. Then $\mathcal{M}_i(\Delta s_j) = p^{-1}\mu_i(s_j) + p^{-1/2}Z_{i,j}$ where $Z_{i,j}$ is independent standard normal variable across $i = 1, \ldots, n, j = 1, \ldots, p$. Let $\tilde{\mathcal{M}}_{i,l} = \sum_{j=1}^{p} \mathcal{M}_i(\Delta s_j)\psi_l(s_j)$. Then

$$
r_{L_n,i} = \sum_{l=L_n}^{\infty} \theta_{\beta,l}\tilde{\mathcal{M}}_{i,l} = \sum_{l=L_n}^{\infty} \theta_{\beta,l}\frac{1}{p}\sum_{j=1}^{p} \mu_i(s_j)\psi_l(s_j) + \sum_{l=L_n}^{\infty} \theta_{\beta,l}\frac{1}{\sqrt{p}}\sum_{j=1}^{p} \psi_l(s_j)Z_{i,j}
$$

which implies that $r_{L_n,i}$ follows a normal distribution with mean

$$\mu_{L_n,i,r} = \sum_{l=L_n}^{\infty} \theta_{\beta,l} \frac{1}{p} \sum_{j=1}^{p} \mu_i(s_j)\psi_l(s_j)$$

and variance

$$\sigma^2_{L_n,r} = \frac{1}{p} \sum_{j=1}^{p} \left(\sum_{l=L_n}^{\infty} \theta_{\beta,l}\psi_l(s_j) \right)^2.$$

Let $\theta_{M,i,l} = p^{-1} \sum_{j=1}^{p} \mu_i(s_j)\psi_l(s_j)$. Since $\sum_{l=L_n}^{\infty} \theta^2_{M,i,l} \to 0$ for any i, and $\sum_{l=L_n}^{\infty} \theta^2_{\beta,l} \to 0$ as $L_n \to \infty$, the mean $\mu_{L_n,i,r} \to 0$ as $n \to \infty$.

Given the orthonormality of the basis, and denote $\beta_{L_n}(s) = \sum_{l=1}^{L_n} \theta_{\beta,l}\psi_l(s)$ as the finite basis smooth approximation of $\beta(s)$, write

$$\sigma^2_{L_n,r} = \int_{\mathcal{S}} |\beta(s) - \beta_{L_n}(s)|^2 \mathrm{d}\lambda(s) + r_p = \sum_{l=L_n}^{\infty} \theta^2_{\beta,l} + r_p,$$

where the approximation error $r_p = \left| \int_{\mathcal{S}} |\beta(s) - \beta_{L_n}(s)|^2 \mathrm{d}\lambda(s) - p^{-1} \sum_{j=1}^{p} |\beta(s_j) - \beta_{L_n}(s_j)|^2 \right|$. From Definition 4(d) $r_p < K_\beta p^{-2/d}$, where $K_\beta > 0$ is a constant. Hence $\sigma^2_{L_n,r} \to 0$ as $n \to \infty$.

For large enough n, $\mu_{L_n,i,r}$ is bounded for all i. By the normal tail bound (Proposition 2.1.2 in [89]), for $Z \sim N(0,1)$, $\mathbb{P}(Z > t) \leq \frac{1}{t\sqrt{2\pi}} \exp\{-t^2/2\}$. Then we have

$$\mathbb{P}(r_{L_n,i} > t) \leq \frac{\sigma_{L_n,r}}{t - \mu_{L_n,i,r}} \exp\left\{ -\frac{(t - \mu_{L_n,i,r})^2}{2\sigma^2_{L_n,r}} \right\} \leq a_n = C\sigma_{L_n,r} \exp(-c'/\sigma^2_{L_n,r}). \tag{A.4}$$

By Definition 4, $a_n \leq \exp(-c' n^{\nu_1\nu_2}) < n^{-1}$, hence $\sum_{i=1}^{n} \mathbb{P}(r_{L_n,i} > t) < \infty$.

$$\mathbb{P}(A^c_{L_n}) = \mathbb{P}(|r_{L_n,i}| > t) \leq \mathbb{P}(r_{L_n,i} > t) + \mathbb{P}(r_{L_n,i} < -t)$$

For the $\mathbb{P}(r_{L_n,i} < -t)$ part, we only need to replace $t - \mu_{L_n,i,r}$ by $t + \mu_{L_n,i,r}$ in (A.4), and the same conclusion follows, $\sum_{i=1}^{n} \mathbb{P}(r_{L_n,i} < -t) < \infty$. By Borel-Cantelli Lemma, we can draw the conclusion.

□

Lemma 6. *(Existence of tests) There exist test functions Φ_n, subsets $\mathcal{U}_n, \Theta_n \subset \Theta$, and constant $K_1, K_2, c_1, c_2 > 0$ such that*

(a) $\mathbb{E}_{\theta_0}\Phi_n \to 0$;

(b) $\sup_{\theta \in \mathcal{U}_n^c \cap \Theta_n} \mathbb{E}_\theta(1 - \Phi_n) \leq K_1 e^{-c_1 n}$;

(c) $\Pi(\Theta_n^c) \leq K_2 e^{-c_2 n}$.

Proof. To verify the existence of tests, we define the sieve space of β as

$$\Theta_{p,n} := \left\{ \beta \in \Theta_\beta : \sup_{s \in \mathcal{R}_1 \cup \mathcal{R}_{-1}} \|D^\omega \beta(s)\|_\infty < p^{1/(2d)}, \|\omega\|_1 \leq \rho \right\}$$

The construction of the test follows a similar idea as in Lemma 1 in [1]. For any $\epsilon > 0$, denote

$$\mathcal{U}^c = \{\beta \in \Theta_\beta, \tilde{\boldsymbol{\gamma}} \in \Theta_{\tilde{\boldsymbol{\gamma}}} : \|\beta - \beta_0\|_{2,p} + \|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0\|_2 > \epsilon\}.$$

Following the notations and new formulation of model (2.1) in (A.3) under the basis decomposition, we create the test as follows. Denote $\boldsymbol{\theta}_\beta = (\theta_{\beta,1}, \ldots, \theta_{\beta,L_n})^{\mathrm{T}}$, $\boldsymbol{\theta}_w = (\boldsymbol{\theta}_\beta^\top, \tilde{\boldsymbol{\gamma}}^\top)^\top$ as the vector of parameters.

For any $\epsilon > 0$, to test the hypothesis

$$H_0 : \{\beta(s), \tilde{\boldsymbol{\gamma}}\} = \{\beta_0(s), \tilde{\boldsymbol{\gamma}}_0\}, \quad \text{v.s.} \quad H_1 : \{\beta(s), \tilde{\boldsymbol{\gamma}}\} \in \mathcal{U}^c.$$

Define test function

$$\Phi_n = I\left\{ \left\| \left(\tilde{\mathbf{W}}_n^{\mathrm{T}} \tilde{\mathbf{W}}_n \right)^{-1} \tilde{\mathbf{W}}_n^{\mathrm{T}} \mathbf{Y} - \boldsymbol{\theta}_{\boldsymbol{w}}^{\mathbf{0}} \right\|_2 > \frac{\epsilon}{2} \right\},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}} \in \mathbb{R}^n$. Let $\boldsymbol{Z} \sim N(0, I_n)$ be a standard normal vector. As defined in the main text above Assumption 3, $\tilde{\mathbf{W}}_n = \left(\tilde{\boldsymbol{\mathcal{M}}}_n, \tilde{\mathbf{X}} \right) \in \mathbb{R}^{n \times (L_n + 1 + q)}$, and $\boldsymbol{\theta}_{\boldsymbol{w}}^{\mathbf{0}}$ denotes the true value of $\boldsymbol{\theta}_{\boldsymbol{w}}$.

Let $\boldsymbol{R}_n = (r_{L_n,1}, \ldots, r_{L_n,n})^{\mathrm{T}} \in \mathbb{R}^n$ be the remainder term. Then under H_0,

$$\left(\mathbf{Y} - \tilde{\mathbf{W}}_n \boldsymbol{\theta}_{\boldsymbol{w}}^{\mathbf{0}} \right) = \boldsymbol{R}_n^0 + \boldsymbol{Z}\sigma_Y.$$

Let $H := \left(\tilde{\mathbf{W}}_n^{\mathrm{T}} \tilde{\mathbf{W}}_n \right)^{-1} \tilde{\mathbf{W}}_n^{\mathrm{T}}$.

$$H\mathbf{Y} - \boldsymbol{\theta}_{\boldsymbol{w}}^{\mathbf{0}} = H\left(\mathbf{Y} - \tilde{\mathbf{W}}_n \boldsymbol{\theta}_{\boldsymbol{w}}^{\mathbf{0}} \right) + H\tilde{\mathbf{W}}_n \boldsymbol{\theta}_{\boldsymbol{w}}^{\mathbf{0}} - \boldsymbol{\theta}_{\boldsymbol{w}}^{\mathbf{0}} = H\left(\mathbf{Y} - \tilde{\mathbf{W}}_n \boldsymbol{\theta}_{\boldsymbol{w}}^{\mathbf{0}} \right)$$

Denote the singular value decomposition of $\tilde{\mathbf{W}}_n$ as $\tilde{\mathbf{W}}_n = U\Lambda V^{\mathrm{T}}$ where $UU^{\mathrm{T}} = I_n$, $VV^{\mathrm{T}} = I_{L_n}$, Λ is at most rank L_n, and the smallest singular value is $\sigma_{\min,n}$. Let $\sigma_{\min,n} := \sigma_{\min}(\tilde{\mathbf{W}})$. For some positive constant $c_{\min}$, denote event

$$\Sigma = [\sigma_{\min,n} > c_{\min}\sqrt{n}].$$

Let $\chi^2(a,b)$ denote the non-central χ^2 distribution with non-central parameter a and degree b. Then under event Σ,

$$\begin{aligned}
&\left\|H\left(\mathbf{Y}-\tilde{\mathbf{W}}_n\boldsymbol{\theta}_{\boldsymbol{w}}^{\mathbf{0}}\right)\right\|_2^2=\left\|H\left(\boldsymbol{R}_n^0+\boldsymbol{Z}\sigma_Y\right)\right\|_2^2\\
&=\left(\boldsymbol{R}_n^0+\boldsymbol{Z}\sigma_Y\right)^{\mathrm{T}}U\Lambda^{-2}U^{\mathrm{T}}\left(\boldsymbol{R}_n^0+\boldsymbol{Z}\sigma_Y\right)\\
&\leq\left(\boldsymbol{R}_n^0+\boldsymbol{Z}\sigma_Y\right)^{\mathrm{T}}\sigma_{\min,n}^{-2}\begin{pmatrix}I_{L_n}&0\\0&0_{n-L_n}\end{pmatrix}\left(\boldsymbol{R}_n^0+\boldsymbol{Z}\sigma_Y\right)\\
&=\sigma_Y^2\sigma_{\min,n}^{-2}\left(\boldsymbol{R}_n^0/\sigma_Y+\boldsymbol{Z}\right)^{\mathrm{T}}\begin{pmatrix}I_{L_n}&0\\0&0_{n-L_n}\end{pmatrix}\left(\boldsymbol{R}_n^0/\sigma_Y+\boldsymbol{Z}\right)\sim\sigma_Y^2\sigma_{\min,n}^{-2}\chi^2(L_n,u_n)
\end{aligned}$$

$u_n=\frac{1}{\sigma_Y^2}\left\|\begin{pmatrix}I_{L_n}&0\\0&0_{n-L_n}\end{pmatrix}\boldsymbol{R}_n^0\right\|_2^2$ is the non-central parameter in the non-central χ^2 distribution of order L_n. Each element in R_n^0 is the residual term $r_{L_n,i}$.

Several results are available for the upper bound of noncentral χ^2 tail probability, here we use Theorem 7 in [101], when $x>L_n+2u_{n,\sigma_Y}$, for some constant c,

$$\mathbb{P}\left\{\chi^2(L_n,u_n)-(L_n+u_n)>x\right\}<\exp\left(-cx\right)$$

Hence if $\epsilon^2/(4\sigma_Y^2)\sigma_{\min,n}>L_n+2u_{n,\sigma_Y}$,

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left\{\Phi_nI(\Sigma)\right\}&\leq\mathbb{P}\left\{\sigma_Y^2\sigma_{\min,n}^{-2}\chi^2(L_n,u_{n,\sigma_Y})>\frac{\epsilon^2}{4}\right\}=\mathbb{P}\left\{\chi^2(L_n,u_{n,\sigma_Y})>\frac{\epsilon^2}{4\sigma_Y^2}\sigma_{\min,n}^2\right\}\\
&\leq\exp\left\{-c\left(\frac{\epsilon^2}{4\sigma_Y^2}\sigma_{\min,n}^2-L_n-u_{n,\sigma_Y}\right)\right\}.
\end{aligned}$$

By Lemma 5, for sufficiently large n, $|r_{L_n,i}|<c_0$ with probability 1. Note that $L_n+u_{n,\sigma_Y}<(1+c_0^2/\sigma_Y^2)\,L_n$, given $\sigma_{\min,n}>\sqrt{n}c_{\min}>0$, for sufficiently large n, there exists a constant $c'>0$ such that $\epsilon^2/(4\sigma_Y^2)\sigma_{\min,n}^2-L_n-u_{n,\sigma_Y}>c'n$. Hence by Assumption 3, $\mathbb{E}_{\beta_0}\left\{\Phi_nI(\Sigma)\right\}\leq\exp\left\{-c'n\right\}$ and $\mathbb{E}_{\beta_0}(\Phi)=\mathbb{E}_{\beta_0}\left\{\Phi I(\Sigma)\right\}+\mathbb{E}_{\beta_0}\left\{\Phi I(\Sigma^c)\right\}\leq\exp\left\{-c'n\right\}+\exp\left\{-\tilde{c}n\right\}\leq\exp\left\{-\tilde{c}'n\right\}$, for $n>2\log(2)/\tilde{c}'$, where $\tilde{c}'=\min\{\tilde{c},c'\}/2$.

To find the upper bound of the Type II error, let $\tilde{r}_p=\int_{\mathcal{S}}\{\beta(s)-\beta_0(s)\}^2\lambda(ds)-\|\beta(s)-\beta_0(s)\|_{2,p}^2$ and $r_{L_n}=\sum_{l=L_n}^{\infty}\theta_{\beta,l}^2$. Then $\tilde{r}_p\to0$ as $p\to\infty$ and $r_{L_n}\to0$ as $n\to\infty$. Note that

$$\int_{\mathcal{S}}\{\beta(s)-\beta_0(s)\}^2\lambda(ds)=\int_{\mathcal{S}}\left\{\sum_{l=1}^{\infty}(\theta_{\beta,l}-\theta_{\beta^0,l})\psi_l(s)\right\}^2\lambda(\mathrm{d}s)=\|\boldsymbol{\theta}_{\boldsymbol{\beta}}-\boldsymbol{\theta}_{\boldsymbol{\beta}^{\mathbf{0}}}\|_2^2+r_{L_n},$$

where $\boldsymbol{\theta}_{\boldsymbol{\beta}}, \boldsymbol{\theta}_{\boldsymbol{\beta}^0} \in \mathbb{R}^{L_n}$. By $\|\boldsymbol{\theta}_{\boldsymbol{w}} - \boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{w}}\|_2^2 = \|\boldsymbol{\theta}_{\boldsymbol{\beta}} - \boldsymbol{\theta}_{\boldsymbol{\beta}^0}\|_2^2 + \|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0\|_2^2$,

$$\|\beta(s) - \beta_0(s)\|_{2,p}^2 + \|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0\|_2^2 = \|\boldsymbol{\theta}_{\boldsymbol{w}} - \boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{w}}\|_2^2 - \tilde{r}_p + r_{L_n}.$$

For a sufficiently large n and p, we have $\tilde{r}_p < \epsilon^2/16$ and $r_{L_n} < \epsilon^2/16$. Then $r_{L_n} - r_p < \epsilon^2/8$. Thus, when $\|\beta(s) - \beta_0(s)\|_{2,p}^2 + \|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0\|_2^2 > \epsilon^2/2$, $\|\boldsymbol{\theta}_{\boldsymbol{w}} - \boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{w}}\|_2^2 > 3\epsilon^2/8$.

Recall

$$\mathcal{U}^c = \{\beta \in \Theta_\beta, \tilde{\boldsymbol{\gamma}} \in \Theta_{\tilde{\boldsymbol{\gamma}}} : \|\beta - \beta_0\|_p + \|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_0\|_2 > \epsilon\}.$$

Define the sieve space $\Theta_n := \Theta_{p,n} \times \Theta_{\boldsymbol{\gamma}}$.

$$\begin{aligned}
\sup_{\mathcal{U}^c \cap \Theta_n} \mathbb{E}_\beta (1 - \Phi_n) I(\Sigma) &= \sup_{\mathcal{U}^c \cap \Theta_n} \mathbb{P}\left\{ \left\| H\mathbf{Y} - \boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{w}} \right\|_2 \leq \frac{\epsilon}{2} \right\} \\
&\leq \sup_{\mathcal{U}^c \cap \Theta_n} \mathbb{P}\left\{ \left| \|H\mathbf{Y} - \boldsymbol{\theta}_{\boldsymbol{w}}\|_2 - \left\| \boldsymbol{\theta}_{\boldsymbol{w}} - \boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{w}} \right\|_2 \right| \leq \frac{\epsilon}{2} \right\} \\
&\leq \sup_{\mathcal{U}^c \cap \Theta_n} \mathbb{P}\left\{ \|H\mathbf{Y} - \boldsymbol{\theta}_{\boldsymbol{w}}\|_2 > -\frac{\epsilon}{2} + \left\| \boldsymbol{\theta}_{\boldsymbol{w}} - \boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{w}} \right\|_2 \right\} \\
&\leq \sup_{\mathcal{U}^c \cap \Theta_n} \mathbb{P}\left\{ \|H\mathbf{Y} - \boldsymbol{\theta}_{\boldsymbol{w}}\|_2 > c_1 \epsilon \right\},
\end{aligned}$$

where $c_1 = \left(\sqrt{3/8} - 1/2 \right)$.

Lastly, by Lemma 4 in [43], for some constant c_2, $\Pi(\Theta_n^c) \leq K_2' e^{-c_2 p^{1/d}} \leq K_2 e^{-c_2 n}$ with Assumption 1(b) that $p \geq O(n^{\tau d})$. □

A.1.4 Proof of Theorem 3

Proof. First we show that, conditioning on all other parameters, the joint posterior of $\alpha(s)$ and $\beta(s)$ can be factored into the marginal posteriors of $\alpha(s)$ and $\beta(s)$. Let $\mathbf{D} = \{\mathbf{Y}, \mathbf{M}, \mathbf{X}, \mathbf{C}\}$. For simplicity, we omit "(s)" in $\alpha(s)$ and $\beta(s)$ in the following derivation.

$$\begin{aligned}
\Pi(\alpha, \beta \mid \mathbf{D}) &= \frac{\Pi(\mathbf{D} \mid \alpha, \beta)\pi(\alpha, \beta)}{\Pi(\mathbf{D})} \\
&= \frac{\Pi(\mathbf{M}, \mathbf{Y} \mid \alpha, \beta, \mathbf{X}, \mathbf{C})\pi(\alpha)\pi(\beta)\pi(\mathbf{X}, \mathbf{C})}{\Pi(\mathbf{Y} \mid \mathbf{M}, \mathbf{X}, \mathbf{C})\Pi(\mathbf{M} \mid \mathbf{X}, \mathbf{C})\pi(\mathbf{X}, \mathbf{C})} \\
&= \frac{\Pi(\mathbf{Y} \mid \mathbf{M}, \alpha, \beta, \mathbf{X}, \mathbf{C})\Pi(\mathbf{M} \mid \alpha, \beta, \mathbf{X}, \mathbf{C})\pi(\alpha)\pi(\beta)}{\Pi(\mathbf{Y} \mid \mathbf{M}, \mathbf{X}, \mathbf{C})\Pi(\mathbf{M} \mid \mathbf{X}, \mathbf{C})} \\
&= \frac{\Pi(\mathbf{Y} \mid \mathbf{M}, \beta, \mathbf{X}, \mathbf{C})\pi(\beta)}{\Pi(\mathbf{Y} \mid \mathbf{M}, \mathbf{X}, \mathbf{C})} \frac{\Pi(\mathbf{M} \mid \alpha, \mathbf{X}, \mathbf{C})\pi(\alpha)}{\Pi(\mathbf{M} \mid \mathbf{X}, \mathbf{C})} \\
&= \Pi(\beta \mid \mathbf{D})\Pi(\alpha \mid \mathbf{D})
\end{aligned}$$

Now,

$$
\begin{aligned}
&\Pi\left(\|\alpha\beta - \alpha_0\beta_0\|_{1,p} > \epsilon \mid \mathbf{D}\right) \\
&\quad = \Pi\left(\|(\beta-\beta_0)(\alpha-\alpha_0) + \alpha_0(\beta-\beta_0) + \beta_0(\alpha-\alpha_0)\|_{1,p} > \epsilon \mid \mathbf{D}\right) \\
&\quad \leq \Pi\left(\|(\beta-\beta_0)(\alpha-\alpha_0)\|_{1,p} + \|\alpha_0(\beta-\beta_0)\|_{1,p} + \|\beta_0(\alpha-\alpha_0)\|_{1,p} > \epsilon \mid \mathbf{D}\right) \\
&\quad \leq \Pi\left(\|(\beta-\beta_0)(\alpha-\alpha_0)\|_{1,p} > \epsilon \mid \mathbf{D}\right) + \Pi\left(\|\beta_0(\alpha-\alpha_0)\|_{1,p} > \epsilon \mid \mathbf{D}\right) + \\
&\qquad \Pi\left(\|\alpha_0(\beta-\beta_0)\|_{1,p} > \epsilon \mid \mathbf{D}\right) \qquad \text{(A.5)}
\end{aligned}
$$

Given that both α_0 and β_0 are defined on a compact set $\mathcal{S} \in \mathbb{R}^d$ (Definition 4), there exists $K > 0$ such that $\|\alpha_0\|_\infty \leq K$ and $\|\beta_0\|_\infty \leq K$, by Theorem 1, 2, and the norm inequality $\|\cdot\|_{1,p} \leq \|\cdot\|_{2,p}$, the last two terms in (A.5) goes to 0 in $P^{(n)}_{\alpha_0,\beta_0}$-probability as $n \to \infty$.

For any $\delta > 0$,

$$
\begin{aligned}
&\Pi\left(\|(\beta-\beta_0)(\alpha-\alpha_0)\|_{1,p} > \epsilon \mid \mathbf{D}\right) \\
&\quad \leq \Pi\left(\|\beta-\beta_0\|_{2,p}\|\alpha-\alpha_0\|_{2,p} > \epsilon \mid \mathbf{D}\right) \\
&\quad \leq \Pi\left(\|\beta-\beta_0\|_{2,p}\|\alpha-\alpha_0\|_{2,p} > \epsilon \mid \mathbf{D}, \|\alpha-\alpha_0\|_{2,p} > \delta\right)\Pi\left(\|\alpha-\alpha_0\|_{2,p} > \delta \mid \mathbf{D}\right) + \\
&\qquad \Pi\left(\|\beta-\beta_0\|_{2,p}\|\alpha-\alpha_0\|_{2,p} > \epsilon \mid \mathbf{D}, \|\alpha-\alpha_0\|_{2,p} < \delta\right)\Pi\left(\|\alpha-\alpha_0\|_{2,p} < \delta \mid \mathbf{D}\right) \\
&\quad \leq \Pi\left(\|\alpha-\alpha_0\|_{2,p} > \delta \mid \mathbf{D}\right) + \Pi\left(\|\beta-\beta_0\|_{2,p}\delta > \epsilon \mid \mathbf{D}, \|\alpha-\alpha_0\|_{2,p} < \delta\right) \\
&\quad = \Pi\left(\|\alpha-\alpha_0\|_{2,p} > \delta \mid \mathbf{D}\right) + \Pi\left(\|\beta-\beta_0\|_{2,p}\delta > \epsilon \mid \mathbf{D}\right).
\end{aligned}
$$

As $n \to \infty$, $\Pi\left(\|\alpha-\alpha_0\|_{2,p} > \delta \mid \mathbf{D}\right) \to 0$ and $\Pi\left(\|\beta-\beta_0\|_{2,p}\delta > \epsilon \mid \mathbf{D}\right) \to 0$ in $P^{(n)}_{\alpha_0,\beta_0}$-probability, which implies that $\Pi\left(\|(\beta-\beta_0)(\alpha-\alpha_0)\|_{1,p} > \epsilon \mid \mathbf{D}\right) \to 0$

□

A.1.5 Proof of Corollary 2

Proof. The proof of the sign consistency is similar to Theorem 3 in [43].

To show Corollary 2, for simplicity, denote $\mathcal{E}(s) := \alpha(s)\beta(s)$ and $\mathcal{E}_0(s) := \alpha_0(s)\beta_0(s)$, $\forall s \in \mathcal{S}$ as the true function of the total effect. Since both $\alpha(s)$ and $\beta(s)$ satisfy Definition 3, we use the notations

$$
\mathcal{R}^f_i := \Big\{ s \in \mathcal{S} : \operatorname{sgn}\{f(s)\} = i \Big\},\ f \in \{\alpha, \beta\},\ i \in \{-1, 0, 1\},
$$

and by Definition 3, $\mathcal{R}^\alpha_{\pm 1}, \mathcal{R}^\beta_{\pm 1}$ are open sets. Define $\mathcal{R}^\mathcal{E}_1 = \left(\mathcal{R}^\alpha_1 \cap \mathcal{R}^\beta_1\right) \cup \left(\mathcal{R}^\alpha_{-1} \cap \mathcal{R}^\beta_{-1}\right)$, $\mathcal{R}^\mathcal{E}_{-1} = \left(\mathcal{R}^\alpha_{-1} \cap \mathcal{R}^\beta_1\right) \cup \left(\mathcal{R}^\alpha_1 \cap \mathcal{R}^\beta_{-1}\right)$, $\mathcal{R}^\mathcal{E}_0 = \mathcal{S} - (\mathcal{R}^\mathcal{E}_1 \cup \mathcal{R}^\mathcal{E}_{-1})$, $\mathcal{R}^\mathcal{E}_{\pm 1}$ are open sets. To show $\mathcal{R}^\mathcal{E}_0$

has nonempty interior, if we denote $\bar{A} := S - A$ as the complementary set of A in S, we only need to show

$$\left(\overline{\mathcal{R}_1^\alpha \cup \mathcal{R}_{-1}^\alpha}\right) \cup \left(\overline{\mathcal{R}_1^\beta \cup \mathcal{R}_{-1}^\beta}\right) \subseteq R_0^{\mathcal{E}}$$

where the LHS has nonempty interior by the Definition 3. $\mathcal{R}_0^{\mathcal{E}} = \overline{\mathcal{R}_1^{\mathcal{E}}} \cap \overline{\mathcal{R}_{-1}^{\mathcal{E}}}$,

$$\begin{aligned}\overline{\mathcal{R}_1^{\mathcal{E}}} &= \overline{\left(\mathcal{R}_1^\alpha \cap \mathcal{R}_1^\beta\right)} \cap \overline{\left(\mathcal{R}_{-1}^\alpha \cap \mathcal{R}_{-1}^\beta\right)} \\ &= \left(\overline{\mathcal{R}_1^\alpha \cup \mathcal{R}_{-1}^\alpha}\right) \cup \left(\overline{\mathcal{R}_1^\beta \cup \mathcal{R}_{-1}^\beta}\right) \cup \left(\overline{\mathcal{R}_1^\beta \cup \mathcal{R}_{-1}^\alpha}\right) \cup \left(\overline{\mathcal{R}_1^\alpha \cup \mathcal{R}_{-1}^\beta}\right)\end{aligned}$$

Similarly we can show $\left(\overline{\mathcal{R}_1^\alpha \cup \mathcal{R}_{-1}^\alpha}\right) \cup \left(\overline{\mathcal{R}_1^\beta \cup \mathcal{R}_{-1}^\beta}\right) \subseteq \overline{\mathcal{R}_{-1}^{\mathcal{E}}}$, hence $\mathcal{R}_0^{\mathcal{E}}$ has nonempty interior. The parameter space of $\mathcal{E}$, $\Theta_{\mathcal{E}}$ satisfies Definition 3.

Now denote $\mathcal{S}_0 = \{s \in \mathcal{S} : \mathcal{E}_0(s) = 0\}$, $\mathcal{S}_+ = \{s \in \mathcal{S} : \mathcal{E}_0(s) > 0\}$, $\mathcal{S}_- = \{s \in \mathcal{S} : \mathcal{E}_0(s) < 0\}$. Notice that $\mathcal{R}_{\pm 1}^{\mathcal{E}_0} \subseteq \mathcal{S}_\pm$, and $\mathcal{S}_0 \subseteq \mathcal{R}_0^{\mathcal{E}_0}$. The key difference is that $\mathcal{S}_{0,\pm}$ are not necessarily open sets.

For any $\mathcal{A} \subseteq \mathcal{S}$ and any integer $m \geq 1$, let Q_p be the discrete measure that assigns $1/p$ mass to each fixed design points in $\{s_j\}_{j=1}^p$, define

$$\mathcal{F}_m(\mathcal{A}) := \left\{\mathcal{E} \in \Theta_{\mathcal{E}} : \int_{\mathcal{A}} |\mathcal{E}(s) - \mathcal{E}_0(s)| dQ_p(s) < \frac{1}{m}\right\}.$$

Note that for any $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{S}$, we have $\mathcal{F}_m(\mathcal{S}) \subseteq \mathcal{F}_m(\mathcal{B}) \subseteq \mathcal{F}_m(\mathcal{A})$.

$$\Pi(\mathcal{F}_m(\mathcal{S}_0) \mid \mathbf{D}) \geq \Pi(\mathcal{F}_m(\mathcal{S}) \mid \mathbf{D}) \to 1, \text{ as } n \to \infty.$$

By the monotone continuity of probability measure,

$$\Pi\{\mathcal{E}(s) = \mathcal{E}_0(s) = 0 \mid \mathbf{D}\} = \Pi\{\mathcal{E}(s) = 0, s \in \mathcal{S}_0 \mid \mathbf{D}\} = \lim_{m\to\infty} \Pi\{\mathcal{F}_m(\mathcal{S}_0) \mid \mathbf{s}\} = 1, \text{ as } n \to \infty.$$

Now to show the consistency of the positive sign, for any small $\delta > 0$, denote $\mathcal{S}_+^\delta := \{s \in \mathcal{S} : \mathcal{E}_0(s) > \delta\}$. Because $\mathcal{E}_0$ is a continuous function, its preimage $\mathcal{E}^{-1}((\delta, \infty))$ supported on $\mathbb{R}^d$ is also an open set, $\mathcal{S}_+^\delta = \mathcal{R}_1^{\mathcal{E}} \cap \mathcal{E}^{-1}((\delta, \infty))$ hence is also an open set.

For any $s_0 \in \mathcal{S}_+^\delta$, we can find a small open ball $B(s_0, r_1) = \{s : \|s - s_0\|_2 < r_1\} \subseteq \mathcal{S}_+^\delta$. By the continuity of $\mathcal{E}$ and $\mathcal{E}_0$, for any large m, there exists $r_2 > 0$ such that $\|s - s_0\|_2 < r_2$ implies $|\mathcal{E}(s) - \mathcal{E}(s_0)| < 1/m$. Let $r = \min\{r_1, r_2\}$.

For any open subset B in $\mathcal{S}$, Definition 4(d) implies that for any large m, there exists N_m

such that for any $n > N_m$,

$$\left| \int_B |\mathcal{E} - \mathcal{E}_0| Q_p(\mathrm{d}s) - \int_B |\mathcal{E} - \mathcal{E}_0| \lambda(\mathrm{d}s) \right| < \frac{V_m}{2m},$$

where we denote $V_m = \lambda\{B(s_0, r)\} \to 0$ as $m \to \infty$.

Hence for any small $\delta > 0$, notice that

$$\begin{aligned}
& \frac{1}{V_m} \int_{B(s_0,r)} |\mathcal{E}(s) - \mathcal{E}_0(s)| \lambda(\mathrm{d}s) < \frac{1}{m} \\
\Rightarrow \quad & \frac{1}{V_m} \int_{B(s_0,r)} \mathcal{E}(s) \lambda(\mathrm{d}s) > \frac{1}{V_m} \int_{B(s_0,r)} \mathcal{E}_0(s) \lambda(\mathrm{d}s) - \frac{1}{m} \\
\Rightarrow \quad & \frac{1}{V_m} \int_{B(s_0,r)} \mathcal{E}(s) \lambda(\mathrm{d}s) > \delta - \frac{1}{m} \\
\Rightarrow \quad & \exists s_1 \in B(s_0, r), \ s.t. \ \mathcal{E}(s_1) > \delta - \frac{1}{m} \\
\Rightarrow \quad & \mathcal{E}(s_0) + \frac{1}{m} > \delta - \frac{1}{m}, \forall s_0 \in S_+^\delta.
\end{aligned}$$

Hence we have

$$\begin{aligned}
& \Pi\left\{\forall s_0 \in \mathcal{S}_+^\delta, \mathcal{E}(s_0) > 0 \mid \mathbf{D}\right\} \geq \Pi\left\{\forall s_0 \in \mathcal{S}_+^\delta, \mathcal{E}(s_0) \geq \delta \mid \mathbf{D}\right\} \\
= & \lim_{m\to\infty} \Pi\left\{\forall s_0 \in \mathcal{S}_+^\delta, \ \mathcal{E}(s_0) > \delta - \frac{2}{m} \Big| \mathbf{D}\right\} \\
\geq & \lim_{m\to\infty} \Pi\left\{\int_{B(s_0,r)} |\mathcal{E}(s) - \mathcal{E}(s_0)| \lambda(\mathrm{d}s) < \frac{V_m}{m} \Big| \mathbf{D}\right\} \\
\geq & \lim_{m\to\infty} \Pi\left\{\int_{B(s_0,r)} |\mathcal{E}(s) - \mathcal{E}(s_0)| \mathrm{d}Q_p(s) < \frac{V_m}{2m} \Big| \mathbf{D}\right\} = 1,
\end{aligned}$$

The proof for the consistency of the negative sign is similar to the positive sign.

□

A.2 Example for Assumption 3

In this section, we give an example that demonstrates the generative model (2.2) satisfies Assumption 3 under some stronger assumptions.

Assumption 7. *When viewing the mediator model* (2.2) *as the true generative model of* $\tilde{\mathbf{W}}_n$*, assume*

1. *for any* $s \in \mathcal{S}$, $\sum_{i=1}^n X_i \epsilon_{M,i}(s) = 0$ *and* $\sum_{i=1}^n C_{k,i} \epsilon_{M,i}(s) = 0$, $k = 1, \ldots, q$, *with*

probability one;

2. *for the chosen basis* $\{\psi_l(s)\}_{l=1}^{\infty}$*, the individual effects* $\eta_i(s)$ *can be viewed as one realization of the random Gaussian process* $\eta_i \sim \mathcal{GP}(0, \sigma_\eta \kappa)$*, and can be decomposed as* $\eta_i(s) = \sum_{l=1}^{\infty} \theta_{\eta,i,l}\psi_l(s)$ *where* $\theta_{\eta,i,l} \overset{ind}{\sim} \mathrm{N}(0, \sigma_\eta^2 \lambda_l)$*;*

Proposition 5. *Under Assumption 7, the least singular value of* $\tilde{\mathbf{W}}_n$ *satisfies*

$$0 < c_{\min} < \liminf_{n\to\infty} \sigma_{\min}(\tilde{\mathbf{W}}_n)/\sqrt{n}$$

with probability $1 - \exp(-\tilde{c}n)$ *for some constant* $\tilde{c}, c_{\min} > 0$*.*

Recall the notations in (A.3), $\tilde{\mathbf{W}}_n = (\tilde{\boldsymbol{\mathcal{M}}}_n, \tilde{\mathbf{X}}) \in \mathbb{R}^{n\times(L_n+1+q)}$, and $\tilde{\mathbf{X}}_i = (X_i, \mathbf{C}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{q+2}$.

The proof of Proposition 5 needs to show that the least singular value of $\tilde{\mathbf{W}}_n$, denoted as $\sigma_{\min}(\tilde{\mathbf{W}}_n)$ satisfies that

$$\mathbb{P}\left(\sigma_{\min}(\tilde{\mathbf{W}}_n) < c\sqrt{n} \mid \mathbf{X}, \mathbf{C}\right) \leq e^{-c'n}$$

Proof. Given (2.4) for $\mathcal{M}(\Delta s)$ and $\lambda(\Delta s_j) = \frac{1}{p}$, we can write

$$\tilde{\mathcal{M}}_{i,l} = \tilde{\theta}_{\alpha,l} X_i + \sum_{k=1}^{q} \tilde{\theta}_{\zeta,k,l} C_{i,k} + \theta_{\eta,i,l} + \tilde{\varepsilon}_{i,l}$$

where $\tilde{\varepsilon}_{i,l} \sim N\{0, (\sigma_M^2/p)\sum_{j=1}^{p} \psi_l(s_j)^2\}$, and each $\tilde{\theta}_{\alpha,l} = \langle \alpha, \psi_l \rangle_p$, $\tilde{\theta}_{\zeta,k,l} = \langle \zeta_k, \psi_l \rangle_p$. Hence we can write

$$\tilde{\boldsymbol{\mathcal{M}}}_n = \tilde{\mathbf{X}}\boldsymbol{\theta}_M + \boldsymbol{\Theta}_E$$

Here, $\boldsymbol{\theta}_M = \begin{pmatrix} \tilde{\theta}_{\alpha,1}, \ldots, \tilde{\theta}_{\alpha,L_n} \\ \tilde{\theta}_{\zeta_1,1}, \ldots, \tilde{\theta}_{\zeta_1,L_n} \\ \cdots \\ \tilde{\theta}_{\zeta_q,1}, \ldots, \tilde{\theta}_{\zeta_q,L_n} \end{pmatrix} \in \mathbb{R}^{(q+1)\times L_n}$, $(\boldsymbol{\Theta}_E)_{i,l} = \langle \eta_i, \psi_l \rangle_p + \tilde{\varepsilon}_{i,l}$, .

By Assumption 2(c) and Assumption 7.1, we have that $\boldsymbol{\Theta}_E^{\mathrm{T}}\tilde{\mathbf{X}} = \mathbf{0}$. Denote $\mathbf{A}_n = \tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}$,

then

$$\begin{aligned}\tilde{\mathbf{W}}_n^{\mathrm{T}}\tilde{\mathbf{W}}_n &= \begin{pmatrix} \tilde{\boldsymbol{\mathcal{M}}}_n^{\mathrm{T}}\tilde{\boldsymbol{\mathcal{M}}}_n & \tilde{\boldsymbol{\mathcal{M}}}_n^{\mathrm{T}}\tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\boldsymbol{\mathcal{M}}}_n & \tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}} \end{pmatrix} \\ &= \begin{pmatrix} \left(\tilde{\mathbf{X}}\boldsymbol{\theta}_M + \boldsymbol{\Theta}_E\right)^{\mathrm{T}}\left(\tilde{\mathbf{X}}\boldsymbol{\theta}_M + \boldsymbol{\Theta}_E\right) & \left(\tilde{\mathbf{X}}\boldsymbol{\theta}_M + \boldsymbol{\Theta}_E\right)^{\mathrm{T}}\tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^{\mathrm{T}}\left(\tilde{\mathbf{X}}\boldsymbol{\theta}_M + \boldsymbol{\Theta}_E\right) & \tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\theta}_M^{\mathrm{T}}\mathbf{A}_n\boldsymbol{\theta}_M + \boldsymbol{\Theta}_E^{\mathrm{T}}\boldsymbol{\Theta}_E & \boldsymbol{\theta}_M^{\mathrm{T}}\mathbf{A}_n \\ \mathbf{A}_n\boldsymbol{\theta}_M & \mathbf{A}_n \end{pmatrix}.\end{aligned}$$

Furthermore,

$$\left(\tilde{\mathbf{W}}_n^{\mathrm{T}}\tilde{\mathbf{W}}_n\right)^{-1} = \begin{pmatrix} \left(\boldsymbol{\Theta}_E^{\mathrm{T}}\boldsymbol{\Theta}_E\right)^{-1} & -\left(\boldsymbol{\Theta}_E^{\mathrm{T}}\boldsymbol{\Theta}_E\right)^{-1}\boldsymbol{\theta}_M^{\mathrm{T}} \\ -\boldsymbol{\theta}_M\left(\boldsymbol{\Theta}_E^{\mathrm{T}}\boldsymbol{\Theta}_E\right)^{-1} & \mathbf{A}_n^{-1} + \boldsymbol{\theta}_M\left(\boldsymbol{\Theta}_E^{\mathrm{T}}\boldsymbol{\Theta}_E\right)^{-1}\boldsymbol{\theta}_M^{\mathrm{T}} \end{pmatrix}.$$

This implies that the Schur complement of $\mathbf{A}_n$ in $\left(\tilde{\mathbf{W}}_n^{\mathrm{T}}\tilde{\mathbf{W}}_n\right)^{-1}$ is $\left(\boldsymbol{\Theta}_E^{\mathrm{T}}\boldsymbol{\Theta}_E\right)^{-1}$. Denote $\|\cdot\|$ as the operator norm. Notice that $\frac{1}{\sigma_{\min}^2(\tilde{\mathbf{W}}_n)} = \|\tilde{\mathbf{W}}_n^{-1}\|^2 = \left\|\left(\tilde{\mathbf{W}}_n^{\mathrm{T}}\tilde{\mathbf{W}}_n\right)^{-1}\right\|$. By Lemma 7, $\sigma_{\min}(\Theta_E)$ has a lower bound $c\sqrt{n}$ with probability $1 - e^{-c'n}$.

$$\begin{aligned}\left\|\left(\tilde{\mathbf{W}}_n^{\mathrm{T}}\tilde{\mathbf{W}}_n\right)^{-1}\right\|^2 &\leq \left\|\left(\boldsymbol{\Theta}_E^{\mathrm{T}}\boldsymbol{\Theta}_E\right)^{-1}\right\|^2 + 2\left\|\left(\boldsymbol{\Theta}_E^{\mathrm{T}}\boldsymbol{\Theta}_E\right)^{-1}\boldsymbol{\theta}_M^{\mathrm{T}}\right\|^2 + \left\|\mathbf{A}_n^{-1} + \boldsymbol{\theta}_M\left(\boldsymbol{\Theta}_E^{\mathrm{T}}\boldsymbol{\Theta}_E\right)^{-1}\boldsymbol{\theta}_M^{\mathrm{T}}\right\|^2 \\ &\leq \frac{1}{\sigma_{\min}^4(\boldsymbol{\Theta}_E)}\left(1 + \|\boldsymbol{\theta}_M\|^2\right)^2 + \left\|\mathbf{A}_n^{-1}\right\|^2\end{aligned}$$

Note that $\sum_{l=1}^{\infty}\theta_{\alpha,l}^2 < \infty$ and $\sum_{l=1}^{\infty}\theta_{\zeta_k,l}^2 < \infty, k = 1,..,q$, hence $\|\boldsymbol{\theta}_M\|$ is bounded by a constant. With Assumption 2.(a), $\sigma_{\min}(\mathbf{A}_n) > n$. Hence with probability $1 - e^{-c'n}$,

$$\frac{1}{\sigma_{\min}^4\left(\tilde{\mathbf{W}}_n\right)} \leq C\frac{1}{\sigma_{\min}^4(\boldsymbol{\Theta}_E)} + \frac{1}{n^2} \leq \frac{C'}{n^2}$$

Hence $\sigma_{\min}\left(\tilde{\mathbf{W}}_n\right) > c\sqrt{n}$ with probability $1 - e^{-c'n}$. □

Lemma 7. *Under model* (2.2),

$$\tilde{\boldsymbol{\mathcal{M}}}_n = \tilde{\mathbf{X}}\boldsymbol{\theta}_M + \boldsymbol{\Theta}_E$$

Then under assumptions 1-7, the smallest singular value $\sigma_{\min}(\boldsymbol{\Theta}_E)$ *satisfies that, for some* $c_1, c_2 > 0$,

$$\mathbb{P}\left\{\sigma_{\min}(\boldsymbol{\Theta}_E) < c_1\sqrt{n}\,|\boldsymbol{X}, \boldsymbol{C}\right\} \leq e^{-c_2 n} \tag{A.6}$$

Proof. To show (A.6), we can write

$$\boldsymbol{\Theta}_E = \tilde{\boldsymbol{\Theta}}_\eta + \tilde{\boldsymbol{\Theta}}_E + \boldsymbol{R}_p.$$

To unpack each matrix, we give the (i,l)th element in each matrix: $\left(\tilde{\boldsymbol{\Theta}}_\eta\right)_{i,l} = \int_{\mathcal{S}} \eta_i(s)\psi_l(s)\lambda(\mathrm{d}s)$, $\left(\tilde{\boldsymbol{\Theta}}_E\right)_{i,l} \sim N(0, \sigma_M^2 \int_{\mathcal{S}} \psi_l(s)^2 \lambda(\mathrm{d}s))$. Note that we view $\eta_i(s)$ as independent copies of Gaussian processes, and by Assumption 7(b), $\left(\tilde{\boldsymbol{\Theta}}_\eta\right)_{i,l} \sim N(0, \lambda_l \sigma_\eta^2)$.

The remainder term $\boldsymbol{R}_p$ is the approximation error between the continuous integrals and their fixed grid approximation. Denote the fixed grid approximations as $\left(\boldsymbol{\Theta}_\eta^*\right)_{i,l} = \frac{1}{p}\sum_{j=1}^p \eta_i(s_j)\psi_l(s_j)$, $(\boldsymbol{\Theta}_E^*)_{i,l} \sim N(0, \frac{1}{p}\sum_{j=1}^p \psi_l^2(s_j))$, and $\boldsymbol{R}_p = \left\{\boldsymbol{\Theta}_\eta^* - \tilde{\boldsymbol{\Theta}}_\eta\right\} + \left\{\boldsymbol{\Theta}_E^* - \tilde{\boldsymbol{\Theta}}_E\right\}$, and $|\left(\boldsymbol{R}_p\right)_{i,l}| \leq K p^{-2/d}$ almost surely for all i, l,

We need to show

(i) $\sigma_{\min}\left(\tilde{\boldsymbol{\Theta}}_E\right)$ has a lower bound $c\sqrt{n}$ with probability $1 - e^{-\tilde{c}n}$.

(ii) $\sigma_{\min}\left(\tilde{\boldsymbol{\Theta}}_E + \tilde{\boldsymbol{\Theta}}_\eta\right)$ has a lower bound $c\sqrt{n}$ with probability $1 - e^{-\tilde{c}n}$.

(iii) Adding the error term $\boldsymbol{R}_p$ does not change this lower bound.

To show (i), let $\boldsymbol{Z}$ be an $L_n \times n$ dimensional random matrix where the entries are i.i.d standard normal variables. Then by Theorem 1 in [73],

$$\mathbb{P}\left\{\sigma_{\min}(\boldsymbol{Z}) < \epsilon\left(\sqrt{n} - \sqrt{L_n - 1}\right)\right\} \leq (C\epsilon)^{n - L_n + 1} + e^{-cn}$$

Because we have $L_n = o(n)$ (Assumption 7.3), hence we use a relaxed lower bound, for some $c_0, c_0' > 0$,

$$\mathbb{P}\left(\sigma_{\min}(\boldsymbol{Z}) < c_0\sqrt{n}\right) \leq e^{-c_0' n}.$$

Because ψ_l forms an orthonormal basis, $\int_{\mathcal{S}} \psi_l^2(s)\lambda(\mathrm{d}s) = 1$, $\tilde{\boldsymbol{\Theta}}_E = \sigma_M Z$.

To show (ii), note $\tilde{\boldsymbol{\Theta}}_\eta = \sigma_\eta \Lambda Z$. Λ is the diagonal matrix with element λ_l. $\tilde{\boldsymbol{\Theta}}_\eta + \tilde{\boldsymbol{\Theta}}_E = D_E Z$ where D_E is a diagonal matrix with lth element $\sqrt{\sigma_\eta^2 \lambda_l + \sigma_M^2}$. For any $x \in \mathbb{R}^{L_n}$,

$$\begin{aligned}\sigma_{\min}(D_E Z) &= \min_{\|x\|_2 = 1} \|Z^{\mathrm{T}} D_E^{\mathrm{T}} x\|_2 = \min_{\|x\|_2 = 1} \frac{\|Z^{\mathrm{T}} D_E^{\mathrm{T}} x\|_2}{\|D_E^{\mathrm{T}} x\|_2} \|D_E^{\mathrm{T}} x\|_2 \geq \min_{\|y\|_2 = 1} \|Z^{\mathrm{T}} y\|_2 \min_{\|x\|_2 = 1} \|D_E^{\mathrm{T}} x\|_2 \\ &= \sigma_{\min}(Z)\, \sigma_{\min}(D_E)\end{aligned}$$

Hence $\sigma_{\min}\left(\tilde{\boldsymbol{\Theta}}_\eta + \tilde{\boldsymbol{\Theta}}_E\right) = \sigma_{\min}(D_E Z) \geq \sigma_{\min}(Z)\, \sigma_{\min}(D_E) \geq \sqrt{\sigma_\eta^2 \lambda_{L_n} + \sigma_M^2}\, \sigma_{\min,n}(Z)$. Since $\lambda_{L_n} \to 0$ as $n \to \infty$, σ_M^2 is the leading term.

To show (iii), by Weyl's inequality, $\sigma_{\min}\left(\tilde{\boldsymbol{\Theta}}_\eta + \tilde{\boldsymbol{\Theta}}_E + \boldsymbol{R}_p\right) \geq \sigma_{\min}\left(\tilde{\boldsymbol{\Theta}}_\eta + \tilde{\boldsymbol{\Theta}}_E\right) - \sigma_{\max}(\boldsymbol{R}_p)$. Since we have $\max_{i,l}|(\boldsymbol{R}_p)| \leq Kp^{-2/d}$, by Assumption 1 and 7, $\sigma_{\max}(\boldsymbol{R}_p) \leq K\sqrt{n \times L_n}p^{-2/d} \leq n^{\frac{\nu_1+1}{2}-2\tau} \to 0$ as $n \to \infty$ (Assumption 1). □

A.3 Sensitivity Analysis in Data Application

In this section, we provide details on data preprocessing and selecting the kernel parameters and the prior parameter λ in both models (2.1) and (2.2).

To get an appropriate kernel for the real data, we choose the Matérn kernel parameters based on the smoothness of the image mediators. The input images are standardized across subject. To get parameters in the Matérn kernel function as defined in (2.9), we tune (ρ, u) on a grid in the following way: First, the empirical sample correlations of the image predictors are computed, then the parameters (ρ, u) are obtained using grid search so that the estimated correlation from the kernel function can best align with the empirical correlation computed from the image mediators. The kernel parameters are chosen region-by-region. We refer to this set of kernel parameters as the optimal kernel.

Table A.1: Predictive MSE for different kernels

	Optimal Kernel	90% of ρ	$u = 1, \rho = 15$	$u = 0.2, \rho = 80$	110% of ρ
Test MSE	0.515	0.516	0.547	0.539	0.507

To test and compare the performances of different kernels, we split the data into 50% as training data and 50% as testing data. Because the performance of different kernels can be directly compared through testing MSE using the outcome model (2.1) , we conduct a sensitivity analysis using model (2.1) to select an appropriate set of kernel parameters. The optimal kernel is obtained in the aforementioned way. To test the sensitivity of the kernel, we fix u to be the same as the optimal u, but change ρ to be 90% and 110% of the optimal ρ. Another 2 kernels where u, ρ are constant across different regions are also included in the comparison. The comparison result is in Table A.1. Based on Table A.1, the case 110% of the optimal ρ seems to give a slightly better prediction performance, hence we choose this kernel for model (2.1). The kernel in model (2.2) remains to be the optimal kernel we choose.

We use the same 2-fold cross validation method to select an appropriate value of ν in the prior of $\beta(s)$. Based on Table A.2, if we select a very small $\nu = 0.01$, there is severe overfitting issue; if ν gets too large, the testing accuracy also decreases. Hence based on this 2-fold testing result, $\nu = 0.05$ appears to be the most appropriate thresholding parameter.

ν	0.01	0.05	0.07	0.1
Training MSE	0.0003	0.3621	0.4043	0.4693
Test MSE	1.8444	0.5079	0.5120	0.5254

Table A.2: Training and test MSE for model (2.1) under different prior thresholding parameter ν for the coefficient $\beta(s)$.

The running time for fitting model (2.1) based on 50% of the data is only within 1 hour, so this testing procedure under the current data scale is not very computationally expansive.

Value of ν	0.05	0.08	0.1	0.5
Averaged test MSE	1.008132	1.008075	1.007796	1.007751
Value of ν	1	1.5	1.7	2.0
Averaged test MSE	1.007740	1.007611	1.007532	1.007711

Table A.3: Averaged testing MSE over all voxels under different value of ν for model (2.2).

A similar sensitivity analysis is conducted for model (2.2) to select ν in the prior of $\alpha(s)$. Estimating the individual effect $\{\eta_i(s)\}_{i=1}^N$ can be very time-consuming, hence the individual effects are set to 0 only for the sensitivity analysis. From table A.3, the difference in the testing MSE among different values of ν is very small. Hence we choose $\nu = 0.1$ conservatively to be able to include more activation voxels without compromising the predictive ability.

A.4 Discussion on MALA initial values

As discussed in section 4.3 in the main text, we can use Gibbs sampler to fit the outcome and mediator model first, and then use the posterior mean of β and α as the initial value for MALA algorithm. In the real data analysis, for the mediator model (2.2), we directly use the posterior mean of $\boldsymbol{\theta}_\alpha$ as the initial value for $\boldsymbol{\theta}_\alpha$ in the MALA algorithm. For the outcome model (2.1), we use the Lasso regression to estimate β first, then add ν to locations where $\beta(s) > 0$, and subtract ν from $\beta(s)$ when $\beta(s) < 0$, to get a hard-threholded version of the latent GP $\tilde{\beta}$. The last step is to use basis on $\tilde{\beta}$ to get the initial values for $\boldsymbol{\theta}_\beta$ in MALA.

APPENDIX B

Chapter 3: Appendix

B.1 Posterior Derivation

For the fully conjugate posterior derivations, the hierarchical model of applying the sparse-mean prior on the scalar-on-image regression can be written as

$$
\begin{aligned}
Y_i = \sum_{j=1}^{p} \beta(s_j) M_i(s_j) + \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{X}_i + \epsilon_i \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_Y^2), \quad \sigma_Y \sim C^+(1) \\
\beta(s_j) \overset{\text{ind}}{\sim} \mathrm{N}(T_\nu(\mu_j), \sigma_\beta^2), \quad \sigma_\beta \sim C^+(1) \\
\mu_j \mid \mu_{[-j]} \sim \mathrm{N}(\bar{\mu}_{\mathcal{N}(j)}, \tau_j^2), \quad \tau_{\mu,j}^2 = \sigma_\mu^2 / w_{j+}, \quad \sigma_\mu^2 \sim C^+(1) \\
\boldsymbol{\gamma} \sim \mathrm{N}(0, \sigma_\gamma^2 I_q), \quad \sigma_\gamma \sim C^+(1)
\end{aligned}
$$

Based on this hierarchical model, we can derive the posterior distributions of each parameter. Note that the posterior for most the parameters are straight forward, except for μ_j, which involves the soft-thresholding operator T_ν. The posterior of μ_j can be expressed in terms of mixture of truncated normal distributions with 3 component.

B.1.1 Proof of Proposition 2

Proof. The posterior of μ_j can be expressed as

$$
\begin{aligned}
\log \pi(\mu_j \mid \boldsymbol{\beta}, \mu_{[-j]}, \sigma_\mu, \sigma_\beta) \propto -\frac{1}{2\sigma_\beta^2} \left(\beta(s_j) - T_\nu(\mu_j)\right)^2 - \frac{w_{j+}}{2\sigma_\mu^2} \left(\mu_j - \bar{\mu}_{\mathcal{N}(j)}\right)^2 \qquad \text{(B.1)} \\
\pi(\mu_j \mid \dots) = P_j^+ \cdot \mathrm{N}_{[\nu,+\infty)}(\mu_j^+, V_j) + P_j^0 \cdot \mathrm{N}_{[-\nu,\nu]}(\bar{\mu}_{\mathcal{N}(j)}, V_0) + P_j^- \cdot \mathrm{N}_{(-\infty,-\nu]}(\mu_j^-, V_j)
\end{aligned}
$$

where $\mathrm{N}_{[a,b]}(\mu, \sigma^2)$ is notation for the truncated normal distribution supported on $[a, b]$ with mean μ and variance σ^2. The middle component is just the truncated normal on $[-\nu, \nu]$ with

the prior mean $\bar{\mu}_{\mathcal{N}(j)}$ and variance $V_0 = \frac{w_{j+}}{\sigma_\mu^2}$. For the other two component,

$$V_j = \left(\frac{1}{\sigma_\beta^2} + \frac{w_{j+}}{\sigma_\mu^2} \right)^{-1}, \quad \mu_j^+ = V_j \left\{ \frac{1}{\sigma_\beta^2} \left(\beta(s_j) + \nu \right) + \frac{w_{j+}}{\sigma_\mu^2} \bar{\mu}_{\mathcal{N}(j)} \right\},$$
$$\mu_j^- = V_j \left\{ \frac{1}{\sigma_\beta^2} \left(\beta(s_j) - \nu \right) + \frac{w_{j+}}{\sigma_\mu^2} \bar{\mu}_{\mathcal{N}(j)} \right\}$$

The density of this 3 component mixture can be expressed as

$$\pi(\mu_j \mid \dots) = \frac{1}{Z_j} (Z_j^+ f_j^+ + Z_j^0 f_j^0 + Z_j^- f_j^-)$$

where Z_j^+, Z_j^0, Z_j^-, Z_j represent different normalizing constant, and f_j^+, f_j^0, f_j^- represent the density functions of 3 truncated normal distributions $\mathrm{N}_{[\nu,+\infty)}(\mu_j^+, V_j), \mathrm{N}_{[-\nu,\nu]}(\bar{\mu}_{\mathcal{N}(j)}, V_0), \mathrm{N}_{(-\infty,-\nu]}(\mu_j^-, V_j)$ respectively. Hence the mixing probabilities can be represented as

$$P_j^+ = \frac{Z_j^+}{Z_j}, \quad , P_j^0 = \frac{Z_j^0}{Z_j}, \quad P_j^- = \frac{Z_j^-}{Z_j}.$$

Now denote $\tilde{f}_*, * \in \{-, 0, +\}$ as the RHS in (B.1) supported on $x \in (-\infty, -\nu), [-\nu, \nu], (\nu, +\infty)$ respectively.

$$\begin{aligned}
\log(Z_j^+) &= \log \tilde{f}_j^+ - \log f_j^+, \quad x \in (\nu, +\infty) \\
&= -\frac{1}{2\sigma_\beta^2} \left(\beta(s_j) + \nu - \mu_j \right)^2 - \frac{w_{j+}}{2\sigma_\mu^2} \left(\mu_j - \bar{\mu}_{\mathcal{N}(j)} \right)^2 \\
&\quad - \left\{ \log \frac{1}{\sqrt{V_j}} - \frac{1}{2V_j} \left(\mu_j - \mu_j^+ \right)^2 - \log \left(1 - \Phi \left(\frac{\nu - \mu_j^+}{\sqrt{V_j}} \right) \right) \right\}
\end{aligned}$$

$$
\begin{aligned}
\log(Z_j^0) &= \log \tilde{f}_j^0 - \log f_j^0, \quad x \in [-\nu, +\nu] \\
&= -\frac{1}{2\sigma_\beta^2} \left(\beta(s_j)\right)^2 - \frac{w_{j+}}{2\sigma_\mu^2} \left(\mu_j - \bar{\mu}_{\mathcal{N}(j)}\right)^2 \\
&- \left\{ \log \frac{1}{\sqrt{V_0}} - \frac{1}{2V_0} \left(\mu_j - \bar{\mu}_{\mathcal{N}(j)}\right)^2 - \log \left(1 - \Phi\left(\frac{\nu - \bar{\mu}_{\mathcal{N}(j)}}{\sqrt{V_0}}\right)\right)\right\} \\
\log(Z_j^-) &= \log \tilde{f}_j^- - \log f_j^-, \quad x \in (-\infty, -\nu) \\
&= -\frac{1}{2\sigma_\beta^2} \left(\beta(s_j) - \mu_j - \nu\right)^2 - \frac{w_{j+}}{2\sigma_\mu^2} \left(\mu_j - \bar{\mu}_{\mathcal{N}(j)}\right)^2 \\
&- \left\{ \log \frac{1}{\sqrt{V_j}} - \frac{1}{2V_j} \left(\mu_j - \mu_j^-\right)^2 - \log \left(1 - \Phi\left(\frac{\nu - \mu_j^-}{\sqrt{V_j}}\right)\right)\right\}
\end{aligned}
$$

Hence the entire density function is complete. □

B.1.2 Variational inference: Q-densities for scalar-on-image regression

In the following derivation, we denote the vector $Y \in \mathbb{R}^n$, matrix $M \in R^{n\times p}$, $X \in R^{n\times q}$ to denote the outcome and design matrices.

B.1.2.1 Q-density for β using SVD

First, we use Singular Value Decomposition (SVD) on M and re-express the scalar-on-image regression model as follows.

Let the compact SVD of $M \in \mathbb{R}^{n\times p}$ be $M = UDV^T$ where $U \in \mathbb{R}^{n\times n}, V^T \in \mathbb{R}^{n\times p}$, and $U^TU = UU^T = I_n$, $V^TV = I_n$. Let $\tilde{\beta} = \beta - T_\nu(\mu)$, $\tilde{Y} = Y - MT_\nu(\mu) - X\gamma = M\tilde{\beta} + \epsilon$.

Now apply the rotation matrix U on both sides, $\tilde{Y}^* = U^T\tilde{Y} = DV^T\beta + \epsilon$. The q-density for $\tilde{\beta}$ is now a normal density with mean and variance

$$
\begin{aligned}
\mathrm{Var}_q(\tilde{\beta}|\sim) &= \left(\mathbb{E}_q\left(\frac{1}{\sigma_\beta^2}\right) I_p + \mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right) VD^TDV^T\right)^{-1}, \\
\mathbb{E}_q(\tilde{\beta}|\sim) &= \mathrm{Var}_q(\tilde{\beta}|\sim)\left(\mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right) VD^T\tilde{Y}^*\right).
\end{aligned}
$$

Note that $\mathbb{E}_q(\tilde{\beta}|\sim)$ can be further simplified,

$$\mathbb{E}_q(\tilde{\beta}|\sim) = VD\left(\frac{1}{\tau^2}I_n + D^2\right)^{-1}\tilde{Y}^*$$

where $\tau^2 = \frac{\mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right)}{\mathbb{E}_q\left(\frac{1}{\sigma_\beta^2}\right)}$. Then $\mathbb{E}_q(\beta|\sim) = \mathbb{E}_q(\tilde{\beta}|\sim) + T_\nu(\mu)$.

The q-density for σ_β is as follows. Note that we use the hierarchical expression to sample half-Cauchy prior $\sigma_\beta \sim C^+(1)$: $\sigma_\beta^2 \sim IG(\frac{1}{2}, \frac{1}{a_\beta})$, $a_\beta \sim IG(\frac{1}{2}, 1)$.

$$\log \pi(\sigma_\beta^2|\sim) \propto -\frac{1}{2\sigma_\beta^2}\sum_{j=1}^{p}\left(\beta(s_j) - T_\nu(\mu_j)\right)^2 - \frac{p}{2}\log(\sigma_\beta^2) - \left(\frac{1}{2}+1\right)\log\sigma_\beta^2 - \frac{1}{a_\beta}\frac{1}{\sigma_\beta^2}$$

Hence the q-density for σ_β^2 follows $IG\left(\frac{p+1}{2}, \frac{1}{2}\sum_{j=1}^{p}\mathbb{E}_q\left(\beta(s_j) - T_\nu(\mu_j)\right)^2 + \frac{1}{a_\beta}\right)$. Note that $\mathbb{E}_q\|\beta - T_\nu(\mu)\|_2^2 = \mathrm{Tr}(\mathrm{Var}(\beta)) + \|\mathbb{E}_q(\beta) - T_\nu(\mu)\|_2^2$, and the marginal variance $\mathrm{Var}_q(\beta(s_j)) = \left[\mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right)\sum_{i=1}^{n}M_i(s_j)^2 + \mathbb{E}_q\left(\frac{1}{\sigma_\beta^2}\right)\right]^{-1}$.

The q-density for a_β is $\mathrm{IG}\left(1, 1 + \mathbb{E}\left\{\frac{1}{\sigma_\beta^2}\right\}\right)$.

B.1.2.2 Q-density for γ

The q-density of γ follows the multivariate normal distribution with mean and variance

$$\mathrm{Var}_q(\gamma|\sim) = \left\{\mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right)X^TX + \mathbb{E}_q\left(\frac{1}{\sigma_\gamma^2}I_q\right)\right\}^{-1}$$

$$\mathbb{E}(\gamma|\sim) = \mathrm{Var}_q(\gamma|\sim)\left\{\mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right)\sum_i\left(Y_i - M_i^T\beta\right)X_i\right\}$$

To speed up the computation, we use eigen-decomposition $X^TX = Q\Lambda_X Q^T$, and the variance update can be written as

$$\mathrm{Var}_q(\gamma|\sim) = Q\mathrm{diag}\left\{\mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right)\Lambda_X + \mathbb{E}_q\left(\frac{1}{\sigma_\gamma^2}\right)\right\}^{-1}Q^T.$$

Similarly, the q-density for σ_γ^2 follows $\mathrm{IG}\left(\frac{q+1}{2}, \frac{1}{2}\mathbb{E}_q\|\gamma\|_2^2 + \mathbb{E}_q\left(\frac{1}{a_\gamma}\right)\right)$.

The q-density for a_γ is $\mathrm{IG}\left(1, 1 + \mathbb{E}\left\{\frac{1}{\sigma_\gamma^2}\right\}\right)$.

B.1.2.3 Q-density for σ_Y

$$\sigma_Y^2 \overset{q}{\sim} \text{IG}\left(\frac{n+1}{2}, \frac{1}{2}\mathbb{E}_q\|Y - M\beta - X\gamma\|_2 + \mathbb{E}_q\left(\frac{1}{a_Y}\right)\right)$$
$$a_Y \overset{q}{\sim} \text{IG}\left(1, 1 + \mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right)\right)$$

B.1.2.4 ELBO derivation

$$\begin{aligned}
\text{ELBO} =& \mathbb{E}_q\left\{\log \pi(Y \mid M, X, \beta, \gamma, \sigma_\beta^2, \sigma_\gamma^2, \sigma_Y^2)\right\} \\
& - \mathbb{E}_q\left\{\log q(\beta) + \log q(\gamma) + \log q(\sigma_\beta^2) + \log q(\sigma_\gamma^2) + \log q(\sigma_Y^2)\right\} \\
=& \mathbb{E}_q\left\{\log \pi(Y| \sim)\right\} \\
& + \mathbb{E}_q\left\{\log \pi(\beta| \sim) - \log q(\beta)\right\} + \mathbb{E}_q\left\{\log \pi(\mu)| \sim) - \log q(\pi(\mu))\right\} \\
& + \mathbb{E}_q\left\{\log \pi(\sigma_\beta^2| \sim) - \log q(\sigma_\beta^2)\right\} + \mathbb{E}_q\left\{\log \pi(a_\beta| \sim) - \log q(a_\beta)\right\} \\
& + \mathbb{E}_q\left\{\log \pi(\gamma| \sim) - \log q(\gamma)\right\} + \mathbb{E}_q\left\{\log \pi(\sigma_\gamma^2| \sim) - \log q(\sigma_\gamma^2)\right\} \\
& + \mathbb{E}_q\left\{\log \pi(a_\gamma| \sim) - \log q(a_\gamma)\right\} \\
& + \mathbb{E}_q\left\{\log \pi(\sigma_Y^2| \sim) - \log q(\sigma_Y^2)\right\} + \mathbb{E}_q\left\{\log \pi(a_Y| \sim) - \log q(a_Y)\right\}
\end{aligned}$$

In the implementation, we separately compute each part of the ELBO and add them together.

$$\begin{aligned}
\text{ELBO}_{\text{logL}} =& \mathbb{E}_q\left\{\log \pi(Y \mid M, X, \beta, \gamma, \sigma_\beta^2, \sigma_\gamma^2, \sigma_Y^2)\right\} \\
=& \frac{n}{2}\mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right) - \frac{1}{2}\mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right)\mathbb{E}_q\|Y - M\beta - X\gamma\|_2^2
\end{aligned}$$

Here, denote $\mathbb{E}_q\text{SSE} = \mathbb{E}_q\|Y - M\beta - X\gamma\|_2^2$,

$$\mathbb{E}_q\text{SSE} = \|Y - M\mathbb{E}_q\beta - X\mathbb{E}_q\gamma\|_2^2 + \text{tr}\left\{M^T M\text{Var}_q(\beta)\right\} + \text{tr}\left\{X^T X\text{Var}_q(\gamma)\right\}.$$

With the eigen decomposition on $X^T X$,

$$\text{tr}\left\{X^T X\text{Var}_q(\gamma)\right\} = \text{tr}\left\{\Lambda_X \text{diag}\left\{\mathbb{E}_q\left(\frac{1}{\sigma_Y^2}\right)\Lambda_X + \mathbb{E}_q\left(\frac{1}{\sigma_\gamma^2}\right)\right\}^{-1}\right\}$$

B.2 Additional Simulation Results

In the first simulation A1, we provide a low dimensional comparison on the SonI regression between three different implementation of the ST-CAR model, the Gibbs sampler, the CAVI algorithm, and the SSVI algorithm.

In the second simulation A2, we provide the additional to results to the simulation II IonS with a further comparison between CAVI and SSVI in high-dimensional settings.

B.2.1 Simulation A1: Low dimensional comparison (SonI)

We compare the proposed `Gibbs`, `CAVI`, `SSVI` with ST-CAR prior to the classical penalized regression method `glmnet`, and a Bayesian method where β is assigned a Soft-thresholded Gaussian Process prior [43] implemented in the `BIMA` package.

The Frequentist penalized regression is implemented using R package `glmnet`[26] with lasso penalty ($\alpha = 1$), using 10-fold cross-validation.

The `BIMA` method requires a pre-specified kernel function, and the posterior sampling algorithm is Metropolis-adjusted Langevin algorithm (MALA). In this simulation we sue the exponential square kernel

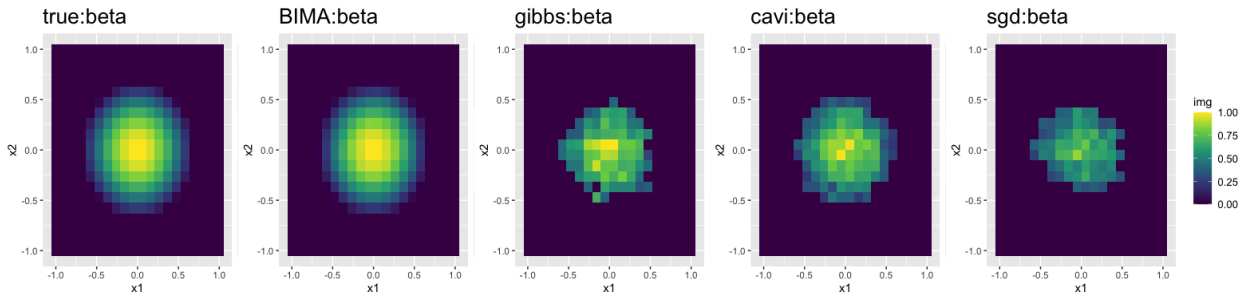
$$\kappa(s, s'; a, b) = \text{cor}\{\beta(s), \beta(s')\} = \exp\{-a(s^2 + s'^2) - b(s - s')^2\}$$

where $a = 0.01, b = 10$, and used $L = 66$ basis functions.

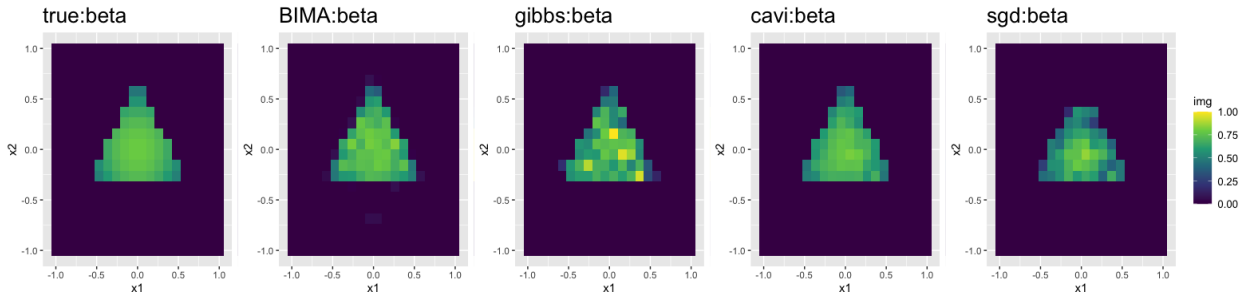
For the four Bayesian methods (`Gibbs`, `CAVI`, `SSVI`, `BIMA`), we set the thresholding parameter $\nu = 0.1$. To evaluate the variable selection accuracy, for the variational inference ST-CAR methods (`CAVI`, `SSVI`), we use the mixing probabilities shown in 3.2 to define the posterior inclusion probability(PIP)

$$PIP(\beta(s_j)) = 1 - P_j^0$$

where both `CAVI` and `SSVI` can trace the mixing probability P_j^0. We use the converged value at the last iteration of P_j^0 in `CAVI` to compute PIP. Since `SSVI` is a stochastic method, we use the averaged P_j^0 over the last 2000 iterations to compute its PIP. For the MCMC methods (`Gibbs`,`BIMA`), we directly use the posterior sample of $T_\nu(\mu_j)$ (for `Gibbs`) or $\beta(s_j)$ (for `BIMA`) being nonzero over the last 20% of iterations as the posterior inclusion probability. For the final selection reported in Table B.1, we use the true generating image β, and set a threshold t on PIP: if $PIP(\beta(s_j)) < t$, $\beta(s_j) = 0$, otherwise $\beta(s_j)$ equals the posterior sample mean or the variational mean. By tuning t, we can control the FDR to be below 10%.



(a) Case 1



(b) Case 2

Figure B.1: Illustration of estimated β under 2 different cases.

Simulation I provides a relatively low-dimensional small-scale example, where $n = 200, p = 400$. We simulate two testing image cases for β in (3.3), as shown in Fig B.1a-B.1b. In case 1, the true image intensity has a smooth transition from 1 to 0, and the voxels around the edge of the signal tend to have low signal-to-noise ratio, but the signal region is a smooth round shape that can be easily estimated by smooth Gaussian process. In case 2, the true image of β is a sharp triangular shape, but the edge voxels of the signal has a sharp contrast to 0, with higher signal to noise ratio compared to case 1.

Table B.1 provides the simulation results of the posterior mean estimates of β, with mean and standard deviation computed across 100 replications. SSVI has the best time efficiency across 4 Bayesian methods.

B.2.2 Simulation A2: High-dimensional Comparison between CAVI and SSVI (IonS)

In the second simulation A2, we provide the additional to results to the simulation II IonS with a further comparison between CAVI and SSVI in high-dimensional settings. Table B.2 provides the additional result on the performance of CAVI compared to SSVI.

Table B.1: Simulation results based on 100 replications, with standard deviation in the bracket. All values are timed by 100 except for time (in seconds). FDR (false discovery rate) is the proportion of times that zero coefficients are identified as nonzero among all identified nonzero coefficients. Power is the proportion of times that nonzero coefficients are identified as nonzero among all nonzero coefficients. Accuracy is the proportion of times the prediction is correct. RMSE is the root mean square error over all voxels.

Case1	`Gibbs`	`CAVI`	`SSVI`	`BIMA`	`glmnet`
FDR	5.4 (3)	7.8 (2)	7.8 (2)	13.5 (1)	3.0 (3)
Power	80.0 (7)	94.4 (3)	84.9 (4)	100.0 (0)	24.1 (7)
Accuracy	92.6 (2)	95.9 (1)	93.3 (1)	95.3 (0)	77.0 (2)
RMSE	9.1 (2)	5.4 (1)	11.0 (1)	0.5 (0)	19.7 (3)
time	97.7 (5)	43.2 (10)	12.7 (0)	29.4 (1)	1.2 (0)

(a) Case 1

Case2	`Gibbs`	`CAVI`	`SSVI`	`BIMA`	`glmnet`
FDR	8.0 (3)	3.7 (0)	2.0 (2)	16.6 (3)	0.0 (0)
Power	100.0 (0)	100.0 (0)	97.0 (2)	100.0 (0)	94.7 (4)
Accuracy	98.8 (0)	99.5 (0)	99.3 (0)	97.4 (1)	99.3 (1)
RMSE	4.2 (1)	1.9 (0)	7.3 (1)	2.2 (0)	1.8 (1)
time	101.4 (11)	15.9 (5)	12.9 (1)	21.7 (1)	1.2 (0)

(b) Case 2

B.3 Additional Real Data result

Table B.3 provides additional sensitivity analysis results on SonI when the bandwidth is 26. Table B.4 provides additional sensitivity analysis results on IonS when the bandwidth is 9 and the decay rate parameter γ is 0.35.

Table B.2: Additional Simulation results to Simulation II. Comparison between CAVI and SSVI for ST-CAR prior, based on 100 replications.

	FDR		TPR		ACC	
Case	SSVI	CAVI	SSVI	CAVI	SSVI	CAVI
Case 1. $n = 600, p = 1600, \sigma_M^2 = 5$	5.8	6.89	95.41	95.15	97.86	96.94
Case 2. $n = 600, \boldsymbol{p = 900}, \sigma_M^2 = 5$	4.95	6.23	84.34	84.05	96.27	95.32
Case 3. $\boldsymbol{n = 1000}, p = 1600, \sigma_M^2 = 5$	7.1	7.91	98.38	98.31	98.13	97.32
Case 4. $n = 600, p = 1600, \boldsymbol{\sigma_M^2 = 10}$	5.07	5.64	85.31	84.79	96.06	95.85
Case 5. $n = 600, p = 6400, \boldsymbol{\sigma_M^2 = 5}$	5.97	11.72	93.64	94.18	97.35	91.3
Case 6. $n = 600, p = 6400, \boldsymbol{\sigma_M^2 = 20}$	5.93	8.57	81.83	83.58	95.06	91.73
	RMSE		Total time (seconds)		Number of iteratios per second	
Case	SSVI	CAVI	SSVI	CAVI	SSVI	CAVI
Case 1. $n = 600, p = 1600, \sigma_M^2 = 5$	7.86	9.13	73.7	239.3	137.5	41.3
Case 2. $n = 600, \boldsymbol{p = 900}, \sigma_M^2 = 5$	7.88	9.17	55	155.8	186	63.8
Case 3. $\boldsymbol{n = 1000}, p = 1600, \sigma_M^2 = 5$	6.44	7.11	117.3	429.8	88.9	23.1
Case 4. $n = 600, p = 1600, \boldsymbol{\sigma_M^2 = 10}$	10.32	12.7	82.9	255.4	122.7	39.2
Case 5. $n = 600, p = 6400, \boldsymbol{\sigma_M^2 = 5}$	8.52	8.99	409.2	1282.8	24.7	6.7
Case 6. $n = 600, p = 6400, \boldsymbol{\sigma_M^2 = 20}$	13.84	13.72	596.7	2641.5	17	3.4

Table B.3: Additional sensitivity analysis for SonI when the bandwidth is 26, on three parameters: (i) the initial value of σ_β^2, (ii) the thresholding parameter ν in ST-CAR prior, (iii) the decay rate γ in the decay rate function for σ_β^2 where $(\sigma_\beta^2)^{(t)} = a(b+t)^{-\gamma}$.

σ_β^2	5×10^{-6}	1×10^{-5}	5×10^{-5}	1×10^{-4}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
ν	0.007	0.007	0.007	0.007	0.005	0.01	0.012	0.007	0.007
γ	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.25	0.45
test pMSE	0.54	0.51	0.56	0.64	0.52	0.52	0.51	0.52	0.51
train pMSE	0.29	0.37	0.31	0.22	0.33	0.4	0.41	0.36	0.37

Table B.4: Additional sensitivity analysis for IonS when the bandwidth is 9, $\gamma = 0.35$ in the decay rate function, on two parameters: (i) the initial value of σ_β^2, (ii) the thresholding parameter ν in ST-CAR prior

σ_α^2	1	0.1	0.01	0.1	0.1
λ	0.01	0.01	0.01	0.005	0.05
total test pMSE	47362.8	47351.85	47356.9	47354.97	47353.58

APPENDIX C

Chapter 4: Appendix

C.1 Proof of Proposition 3

Proof. Denote $\theta = (\alpha, \beta, \xi, \zeta, \eta, \nu, \sigma_M, \sigma_Y)$ as the collection of all parameters. First of all, based on model (4.1) and the fact that for the intensity measure of $M_i(\Delta s_j)$ over a small voxel partition Δs_j follows Gaussian distribution $M_i(\Delta s_j) = \mathbb{E}\{M_i(s_j)\}\lambda(\Delta s_j) + \epsilon_{M,i}(\Delta s_j)$, the mean and variance function of $M_i(\Delta s_j)$ can be uniquely identified, hence σ_M and the mean function $\mathbb{E}\{M_i(s_j) \mid \theta, X_i, C_i\}$ are both uniquely identifiable. We denote the mean function $\mathbb{E}\{M_i(s_j) \mid \theta, X_i, C_i\}$ as

$$A_{i,j}(\alpha, \xi, \eta_i) = X_i\alpha(s_j) + \sum_{k=1}^{q} C_{i,k}\xi_k(s_j) + \eta_i(s_j)$$

Similarly, σ_Y is also uniquely identifiable. In the derivation below, for simplicity we denote

$$B_i(\gamma, \zeta, \nu, \eta_i) = X_i\gamma + \sum_{k=1}^{q} C_{i,k}\zeta_k + \sum_{j=1}^{p} \nu(s_j)\eta_i(s_j)\lambda(\Delta s_j)$$

Let $\mathbf{M}_i = \{M_i(\Delta s_j)\}_{j=1}^{p}$. Conditional on the covariates $X_i, C_{i,k}$, the joint distribution of $Y_i, \mathbf{M}_i$ can be expressed as

$$\pi\left(Y_i, \mathbf{M}_i \mid X_i, \{C_i\}_{k=1}^{m}, \theta\right) = \pi\left(Y_i \mid \mathbf{M}_i, X_i, \{C_i\}_{k=1}^{m}, \theta\right) \prod_{j=1}^{p} \pi\left(M_i(\Delta s_j) \mid X_i, \{C_i\}_{k=1}^{m}, \theta\right)$$

$$= \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left\{-\frac{1}{2\sigma_Y^2}\left(Y_i - \sum_{j=1}^{p} M_i(\Delta s_j)\beta(s_j) - B_i(\gamma, \zeta, \nu, \eta_i)\right)^2\right\}$$

$$\times \prod_{j=1}^{p}\left(\frac{1}{\sqrt{2\pi\sigma_M^2\lambda(\Delta s_j)}}\right) \times \exp\left\{-\frac{1}{2\sigma_M^2}\sum_{j=1}^{p}\frac{1}{\lambda(\Delta s_j)}\left(M_i(\Delta s_j) - A_{i,j}(\alpha, \xi, \eta_i)\right)^2\right\}$$

Suppose that Y_i takes value y_i and $M_i(\Delta s_j)$ takes value $m_{i,j}$ in the joint density function, note that y_i and $m_{i,j}$ can be any real values. We can write the log joint density function over the i-th individual $i = 1, \ldots, n$ as

$$
\begin{aligned}
&\sum_{i=1}^{n} \left\{ \log \pi(y_i, \{m_{i,j}\}_{j=1}^{p} \mid X_i, \{C_{i,k}\}_{k=1}^{q}, \theta) \right\} \\
\propto &\sum_{i=1}^{n} \Bigg\{ -\frac{1}{2\sigma_Y^2} y_i^2 - \frac{1}{2\sigma_Y^2} \left(\sum_{j=1}^{p} m_{i,j} \beta(s_j) \right)^2 - \frac{1}{2\sigma_Y^2} B_i^2(\gamma, \zeta, \nu, \eta_i) \\
&+ \frac{1}{\sigma_Y^2} y_i \left(\sum_{j=1}^{p} m_{i,j} \beta(s_j) \right) + \frac{1}{\sigma_Y^2} B_i(\gamma, \zeta, \nu, \eta_i) y_i + \frac{1}{\sigma_Y^2} \left(\sum_{j=1}^{p} m_{i,j} \beta(s_j) \right) B_i(\gamma, \zeta, \nu, \eta_i) \\
&- \frac{p}{2\sigma_M^2} \sum_{j=1}^{p} (m_{i,j})^2 - \frac{p}{2\sigma_M^2} A_{i,j}^2(\alpha, \xi, \eta) + \frac{p}{2\sigma_M^2} \sum_{j=1}^{p} m_{i,j} A_{i,j}(\alpha, \xi, \eta_i) \Bigg\}
\end{aligned}
$$

This is a polynomial of $y_i, m_{i,1}, \ldots, m_{i,p}$, and we only need to match the coefficients of the first-order, second-order, and interaction terms to identify the unique coefficients. Hence σ_Y^2 can be uniquely determined by the quadratic term $\sum_{i=1}^{n} y_i^2$, similarly $\{\beta(s_j)\}_{j=1}^{p}$ uniquely determined by the interaction terms $\{\sum_{i=1}^{n} y_i m_{i,j}\}_{j=1}^{p}$, and $B_i(\gamma, \zeta, \nu, \eta_i)$ uniquely determined by the first-order term y_i. Given that $\{\beta(s_j)\}_{j=1}^{p}$ and σ_Y^2 are uniquely identifiable, we can also uniquely determine σ_M^2 from the coefficient of $\sum_{i=1}^{n} \sum_{j=1}^{p} m_{i,j}^2$. Given the identified $\{\beta(s_j)\}_{j=1}^{p}$, σ_Y^2 and σ_M^2, $A_{i,j}(\alpha, \xi, \eta_i)$ is also identified from the coefficient of the first-order $m_{i,j}$.

Now we have shown the identifiability of $\sigma_Y^2, \sigma_M^2, \{\beta(s_j)\}_{j=1}^{p}$, $A_{i,j}(\alpha, \xi, \eta_i)$ and $B_i(\gamma, \zeta, \nu, \eta_i)$.

Next, we need to show that the rest of the parameters $(\alpha, \xi, \zeta, \eta, \nu)$ can also be uniquely identified from $A_{i,j}(\alpha, \xi, \eta_i)$ and $B_i(\gamma, \zeta, \nu, \eta_i)$. Given Assumption 4.1-2, the identifiability of α, ξ, η in $A_{i,j}(\alpha, \xi, \eta_i)$ directly follows from Proposition 1 in [97].

To show the identifiability of γ, ζ, ν in $B_i(\gamma, \zeta, \nu, \eta_i)$, note that given $B_i(\gamma, \zeta, \nu, \eta_i)$ and η_i are identifiable for $i = 1, \ldots, n$, comparing $B_i(\gamma', \zeta', \nu', \eta_i) = B_i(\gamma, \zeta, \nu, \eta_i), i = 1, \ldots, n$ to reach the identifiability of γ, ζ, ν is equivalent to solving a linear system (given that the design matrix is full rank, Assumption 4.3) with n equations and $p + 1 + q$ variables. Hence under assumption (ii) where nu is sparse, $\nu \in \Theta^{\text{SP}}$, when n is large enough, the number of nonzero elements in ν will be smaller than $n - q - 1$, hence γ, ζ, ν are all identifiable.

Under assumption (i), ν is spatially-correlated and can be decomposed using L number

of basis. Let $\nu(s) = \sum_{l=1}^{L} \theta_{\nu,l}\psi_l(s)$, and

$$\int_{\mathcal{S}} \nu(s)\eta_i(s)\lambda(\mathrm{d}s) = \int_{\mathcal{S}} \sum_{l=1}^{L} \theta_{\nu,l}\psi_l(s)\eta_i(s)\lambda(\mathrm{d}s) = \sum_{l=1}^{L} \theta_{\nu,l}\theta_{\eta_i,l}$$

Hence

$$B_i(\gamma, \zeta, \nu, \eta_i) = X_i\gamma + \sum_{k=1}^{q} C_{i,k}\zeta_k + \sum_{l=1}^{L} \theta_{\nu,l}\theta_{\eta_i,l}$$

Based on Assumption 4.3, with the design matrix $B = (\mathbf{X}, \mathbf{C}_1, \ldots, \mathbf{C}_q, \boldsymbol{\theta}_{\eta,1}, \ldots, \boldsymbol{\theta}_{\eta,L}) \in \mathbb{R}^{n\times(1+L+q)}$. And $B_i(\gamma, \zeta, \nu, \eta_i) - B_i(\gamma', \zeta', \nu', \eta_i) = 0$ for $i = 1, \ldots, n$ can be written as

$$B \cdot \left\{ (\gamma, \zeta, \theta_{\nu,1}, \ldots, \theta_{\nu,L})^{\mathrm{T}} - \left(\gamma', \zeta', \theta'_{\nu,1}, \ldots, \theta'_{\nu,L}\right)^{\mathrm{T}} \right\} = \mathbf{0}$$

By Assumption 4.3, $\det(B) > 0$, hence $(\gamma, \zeta, \theta_{\nu,1}, \ldots, \theta_{\nu,L}) = \left(\gamma', \zeta', \theta'_{\nu,1}, \ldots, \theta'_{\nu,L}\right)$, therefore γ, ζ, ν are also identifiable. Similarly, if $\nu \in \Theta^{\mathrm{SP}}$, the design matrix becomes $\left\{\mathbf{X}, \mathbf{C}_1, \ldots, \mathbf{C}_q, \{\boldsymbol{\eta}(s_k)\}_{s_k\in\mathcal{S}_m}\right\} \in \mathbb{R}^{n\times(m+1+q)}$ where $\mathcal{S}_m = \{s : \nu(s) \neq 0\}$, and is also full rank by Assumption 4. □

C.2 Proof of Proposition 4

Proof. Throughout this proof, we use the notation $o_p(1)$ as follows: if $X_n = o_p(1)$, $X_n \xrightarrow{p} 0$ as $n \to \infty$.

Using the decomposition on $\beta(s) = \sum_{l=1}^{L} \theta_{\theta_\beta,l}\psi_l(s)$, $\tilde{M}_{i,l} = \int_S M_i(s)\psi_l(s)\lambda(\mathrm{d}s)$, the full outcome model can be decomposed as

$$Y_i = \sum_{l=1}^{L} \theta_{\beta,l}\tilde{M}_{i,l} + \gamma X_i + \sum_{k=1}^{q} \zeta_k C_{i,k} + \sum_{j=1}^{p} \eta_i(s_j)\nu(s_j) + \epsilon_{Y,i}$$

where $\epsilon_{Y,i} \overset{\text{iid}}{\sim} N(0, \sigma_Y^2)$.

With the prior specification $\theta_{\beta,l} \overset{\text{ind}}{\sim} N(0, \sigma_\beta^2\lambda_l)$, denote diagonal matrix $D \in \mathbb{R}^{L\times L}$, where $(D)_{l,l'} = \lambda_l I(l = l')$. Denote $\tilde{M}_i = \left(\tilde{M}_{i,l}, \ldots, \tilde{M}_{i,L}\right)^{\mathrm{T}} \in \mathbb{R}^L$. The posterior mean of θ_β is

$$\begin{aligned}
\mathrm{Var}\left\{\theta_\beta | \sim\right\} &= \left(\frac{1}{\sigma_\beta^2}D^{-1} + \frac{1}{\sigma_Y^2}\sum_{i=1}^{n} \tilde{M}_i\tilde{M}_i^{\mathrm{T}}\right)^{-1} \\
\mathbb{E}\left\{\theta_\beta | \sim\right\} &= \mathrm{Var}\left\{\theta_\beta | \sim\right\} \left\{ \frac{1}{\sigma_Y^2}\sum_{i=1}^{n} \left(Y_i - \gamma X_i - \sum_{k=1}^{q} \zeta_k C_{i,k} - \sum_{j=1}^{p} \eta_i(s_j)\nu(s_j) \right) \tilde{M}_i \right\}
\end{aligned}$$

To simplify these two bias expressions, we denote $\tilde{M} \in \mathbb{R}^{n\times L}$ with each row being $\tilde{M}_i^{\mathrm{T}}$. Let $A := \tilde{M}^T\tilde{M} \in \mathbb{R}^{L\times L}$.

Denote the point estimator $\hat{\theta}_\beta^F = \mathbb{E}\{\theta_\beta|\sim\}$ for the posterior mean under full model (4.4). Denote θ^0 as the true parameters. Conditional on the estimator $\hat{\eta}$ and $\hat{\nu}$, the bias of $\hat{\theta}_\beta^F$ can be written as

$$\begin{aligned}
\text{bias}(\hat{\theta}_\beta^F) &= \mathbb{E}\left(\hat{\theta}_\beta^F - \theta_\beta^0\right) \\
&= \mathbb{E}_{Y|\{X,C,M\}}\left(\hat{\theta}_\beta^F\right) - \theta_\beta^0 \\
&= \text{Var}\{\theta_\beta|\sim\}\frac{1}{\sigma_Y^2}\sum_{i=1}^{n}\left\{\left(\theta_\beta^0\right)^{\mathrm{T}}\tilde{M}_i + \left(\nu^0\right)^{\mathrm{T}}\eta_i^0 - (\hat{\nu})^{\mathrm{T}}\hat{\eta}_i\right\}\tilde{M}_i - \theta_\beta^0
\end{aligned}$$

Similarly, if we denote the point estimator using BIMA model as $\hat{\theta}_\beta^B$,

$$\text{bias}(\hat{\theta}_\beta^B) = \text{Var}\{\theta_\beta|\sim\}\frac{1}{\sigma_Y^2}\sum_{i=1}^{n}\left\{\left(\theta_\beta^0\right)^{\mathrm{T}}\tilde{M}_i + \left(\nu^0\right)^{\mathrm{T}}\eta_i^0\right\}\tilde{M}_i - \theta_\beta^0$$

Recall the notation $\tilde{M} \in \mathbb{R}^{n\times L}$ and $A := \tilde{M}^T\tilde{M} \in \mathbb{R}^{L\times L}$, we can simplify $\text{bias}(\hat{\theta}_\beta^F)$. By the singular value assumption of A in Proposition 4, with probability $1-\exp\{-c_0 n\}$, A is full rank. Conditioning on the event that A is full rank hence invertible, denote $\boldsymbol{\eta}^0 \in \mathbb{R}^{n\times p}$ with the i-th row being η_i^{T}.

$$\begin{aligned}
\text{bias}(\hat{\theta}_\beta^B) &= \left[\frac{\sigma_Y^2}{\sigma_\beta^2}D^{-1} + A\right]^{-1}\left[A\theta_\beta^0 + \left(\tilde{M}\right)^{\mathrm{T}}\left(\boldsymbol{\eta}^0\right)^{\mathrm{T}}\nu^0\right] - \theta_\beta^0 \\
&= \left[\frac{\sigma_Y^2}{\sigma_\beta^2}D^{-1}A^{-1} + I_L\right]^{-1}\left[\theta_\beta^0 + A^{-1}\left(\tilde{M}\right)^{\mathrm{T}}\left(\boldsymbol{\eta}^0\right)^{\mathrm{T}}\nu^0\right] - \theta_\beta^0 \\
&\overset{(*)}{=} \left[I_L - \left(\tau^2 DA + I_L\right)^{-1}\right]\left[\theta_\beta^0 + A^{-1}\left(\tilde{M}\right)^{\mathrm{T}}\left(\boldsymbol{\eta}^0\right)^{\mathrm{T}}\nu^0\right] - \theta_\beta^0 \\
&= A^{-1}\left(\tilde{M}\right)^{\mathrm{T}}\left(\boldsymbol{\eta}^0\right)^{\mathrm{T}}\nu^0 - \left(\tau^2 DA + I_L\right)^{-1}\left[\theta_\beta^0 - A^{-1}\left(\tilde{M}\right)^{\mathrm{T}}\left(\boldsymbol{\eta}^0\right)^{\mathrm{T}}\nu^0\right]
\end{aligned}$$

Note that $(*)$ uses the Identity $(I+A)^{-1} = I - A(I+A)^{-1} = I - (A^{-1}+I)^{-1}$, and the notation $\tau^{-2} = \frac{\sigma_Y^2}{\sigma_\beta^2}$.

Proof of Part (i)

Now we can see that if $\left(\boldsymbol{\eta}^0\right)^{\mathrm{T}}\nu^0 = \mathbf{0}$, i.e. when the unmeasured confounder effect does

not exist, the bias of $\hat{\theta}_\beta^F$ becomes

$$\text{bias}(\hat{\theta}_\beta^B) = -\left(\tau^2 DA + I_L\right)^{-1} \theta_\beta^0$$

The range of $\text{bias}(\hat{\theta}_\beta^B)$ is controlled by the smallest and largest eigen-values of $(\tau^2 DA + I_L)^{-1}$, scaled up to a rotation of θ_β^0. Note that $\sigma_{\min}(D)\sigma_{\min}(A) \leq \sigma_{\min}(DA) \leq \sigma_{\max}(DA) \leq \sigma_{\max}(D)\sigma_{\max}(A)$. With the assumption $\lambda_L > c_\lambda n^{-1+a_\lambda}$, use $h \gtrsim g$ to denote the inequality $h > c_g g$ up to a positive constant c_g that does not contain any rate of n. We can see that $\sigma_{\min}(D)\sigma_{\min}(A) \gtrsim n^{a_\lambda} \to \infty$ as $n \to \infty$, and $\sigma_{\max}(D)\sigma_{\max}(A) \lesssim n \to \infty$, hence $\text{bias}(\hat{\theta}_\beta^B) \to \mathbf{0}$ as $n \to \infty$.

Proof of Part (ii)

Similarly, when $\left(\boldsymbol{\eta}^0\right)^{\mathrm{T}} \nu^0 \neq \mathbf{0}$,

$$\begin{aligned}
\text{bias}(\hat{\theta}_\beta^B) &= \left[\tau^{-2} D^{-1} + A\right]^{-1} \left(\tilde{M}\right)^{\mathrm{T}} \left(\boldsymbol{\eta}^0\right)^{\mathrm{T}} \nu^0 - \left(\tau^2 DA + I_L\right)^{-1} \theta_\beta^0 \\
\text{bias}(\hat{\theta}_\beta^F) &= \left[\tau^{-2} D^{-1} + A\right]^{-1} \left(\tilde{M}\right)^{\mathrm{T}} \left\{\left(\boldsymbol{\eta}^0\right)^{\mathrm{T}} \nu^0 - \left(\hat{\boldsymbol{\eta}}\right)^{\mathrm{T}} \hat{\nu}\right\} - \left(\tau^2 DA + I_L\right)^{-1} \theta_\beta^0
\end{aligned}$$

As we've shown that $\left(\tau^2 DA + I_L\right)^{-1} \theta_\beta^0 = o_p(1)$, we focus on the first term in both bias expressions.

For the image mediator M_i, the mediator model (4.1) assumes that $M_i(s_j) = \mathbb{E}\{M_i(s_j)\} + \sigma_M Z_{i,j}$, where $Z_{i,j}$ are the independent standard normal variables. Under the orthonormal decomposition, $\tilde{M}_{i,l} = \mu_{i,l} + \sigma_M Z_{i,l}$ where the $Z_{i,l}$ are still independent standard normal under orthonormal transformation, and $\mu_{i,l}$ is a constant mean term that determines mean structure of $\tilde{M}_{i,l}$. Hence we can write $\tilde{M} = \boldsymbol{\mu} + \sigma_M \mathbf{Z} \in \mathbb{R}^{n \times L}$.

$$\begin{aligned}
&\left[\tau^{-2} D^{-1} + A\right]^{-1} \left(\tilde{M}\right)^{\mathrm{T}} \left(\boldsymbol{\eta}^0\right)^{\mathrm{T}} \nu^0 = \\
&\left\{\frac{1}{n}\left[\tau^{-2} D^{-1} + \left(\boldsymbol{\mu} + \sigma_M \mathbf{Z}\right)^{\mathrm{T}} \left(\boldsymbol{\mu} + \sigma_M \mathbf{Z}\right)\right]\right\}^{-1} \left\{\frac{1}{n}\left(\boldsymbol{\mu} + \sigma_M \mathbf{Z}\right)^{\mathrm{T}} \left(\boldsymbol{\eta}^0\right)^{\mathrm{T}} \nu^0\right\}
\end{aligned}$$

The denominator can be broken down into 4 parts,

$$\begin{aligned}
&\frac{1}{n}\left[\tau^{-2} D^{-1} + \left(\boldsymbol{\mu} + \sigma_M \mathbf{Z}\right)^{\mathrm{T}} \left(\boldsymbol{\mu} + \sigma_M \mathbf{Z}\right)\right] \\
=&\frac{1}{n}\tau^{-2} D^{-1} + \frac{1}{n}\sum_{i=1}^n \mu_i \mu_i^{\mathrm{T}} + \frac{\sigma_M^2}{n}\sum_{i=1}^n Z_i Z_i^{\mathrm{T}} + \frac{\sigma_M}{n}\sum_{i=1}^n \left(\mu_i Z_i^{\mathrm{T}} + Z_i \mu_i^{\mathrm{T}}\right)
\end{aligned}$$

The first term $\frac{1}{n}\tau^{-2} D^{-1} = o_p(1)$ since $1/n/\lambda_L \lesssim n^{-a_\lambda} \to 0$.

By Assumption 5, the second term has a constant limit, where $\mu_{i,l} =$

$\int_{\mathcal{S}} \{E_i(s)\psi_l(s)\}\,\lambda(\mathrm{d}s)$. We denote $H = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mu_i\mu_i^{\mathrm{T}} \in \mathbb{R}^{L\times L}$, where the (l,l')-th element in H is $H_{l,l'}$, a finite constant introduced in Assumption 5.

The third term has the limit $\frac{\sigma_M^2}{n}\sum_{i=1}^{n} Z_i Z_i^{\mathrm{T}} \xrightarrow{p} \sigma_M^2 I_L$ due to the i.i.d. normality of $\mathbf{Z}$. The last term is also $o_p(1)$, because the (l,l')-th term is a normal variable with mean 0 and variance $\frac{1}{n^2}\sum_{i=1}^{n}\left(\mu_{i,l}^2+\mu_{i,l'}^2\right)$ for $l \neq l'$, and $\frac{4}{n^2}\sum_{i=1}^{n}\mu_{i,l}^2$ for $l = l'$. Because $(H)_{l,l} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\mu_{i,l}^2\right)$ is a constant, $(H)_{l,l}/n \to 0$.

Hence the denominator is equivalent to $o_p(1) + H + \sigma_M^2 I_L$.

To analyze the numerator and simplify the notations, we use $U^0 := \left(\boldsymbol{\eta}^0\right)^{\mathrm{T}}\nu^0 \in \mathbb{R}^n$, similarly, $\hat{U} := \left(\hat{\boldsymbol{\eta}}\right)^{\mathrm{T}}\hat{\nu}$ to denote the unmeasured confounder term. The numerator can be expressed as

$$\frac{1}{n}\left(\boldsymbol{\mu}+\sigma_M\mathbf{Z}\right)^{\mathrm{T}}\left(\boldsymbol{\eta}^0\right)^{\mathrm{T}}\nu^0 = \left(\frac{1}{n}\sum_{i=1}^{n}\left\{\mu_{i,l}+\sigma_M Z_{i,l}\right\}U_i^0\right)_{l=1}^{L}$$

Note that $\frac{\sigma_M}{n}\sum_{i=1}^{n} Z_{i,l}U_i^0 \xrightarrow{p} \mathbf{0}$ for all l since $\sum_{i=1}^{n}\left(U_i^0\right)^2/n$ is finite (assumption made in part (ii) of Proposition 4).

With the Assumption 5 that $h^0 = \lim_{n\to\infty}\sum_{i=1}^{n}\mu_i U_i^0$ is a finite vector in $\mathbb{R}^L$, we can draw the conclusion that $\mathrm{bias}(\hat{\theta}_\beta^B) \xrightarrow{p} (H+\sigma_M^2 I_L)^{-1}h^0$.

Similarly, if we define $\hat{h} = \lim_{n\to\infty}\sum_{i=1}^{n}\mu_i\hat{U}_i$, the bias of θ_β under the joint model (4.4) becomes $\mathrm{bias}(\hat{\theta}_\beta^F) \xrightarrow{p} (H+\sigma_M^2 I_L)^{-1}\left(h^0-\hat{h}\right)$.

□

C.3 Two-stage Algorithm: update of ν in the outcome model

Let $\hat{\boldsymbol{\eta}} \in \mathbb{R}^{p\times n}$ be the estimator of $\boldsymbol{\eta}$ obtained using the mediator model. We present details for updating ν and δ_ν in Algorithm 1. Updating ν requires fast linear regression using SVD on η, which can be split into two cases, one with the nonzero element in ν greater than n, and the other smaller than n.

C.4 Additional Simulation and Real Data Analysis Results

Below is a visualization of the MSE and bias of $\beta(s)$ over 100 replications, under all six simulation cases.

Algorithm 1. Two-stage Algorithm

for Iterations $t = 1, 2, ...$ **do**
  update θ_β according to the posterior derivations.
  Section 1: Update ν using SVD on the design matrix.
  Let $\delta_1 = \{s_j : \nu(s_j) \neq 0\}$ and $\delta_0 = \{s_j : \nu(s_j) = 0\}$. Let $|\delta_1|$ be the length of δ_1, and $|\delta_0|$ be the length of δ_0.
  Denote $\hat{\boldsymbol{\eta}}_1^{\mathrm{T}} = \left(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\right)_{[:,j],j\in\delta_1} \in \mathbb{R}^{n\times|\delta_1|}$, and do an SVD on $\hat{\boldsymbol{\eta}}_1^{\mathrm{T}} = UDV^{\mathrm{T}}$.
  Let $\mathbf{Y}_\nu = \mathbf{Y} - \gamma\mathbf{X} - \mathbf{C}^{\mathrm{T}}\zeta - \mathbf{M}^{\mathrm{T}}\beta \in \mathbb{R}^n$ be the residual without the $\boldsymbol{\eta}^{\mathrm{T}}\nu$ term, and let $\mathbf{Y}_\nu^* = U^{\mathrm{T}}\mathbf{Y}_\nu$.
  if $|\delta_1| > n$ **then**
    Let $\tau^2 = \sigma_\nu^2/\sigma_Y^2$.
    Sample $\alpha_1 \sim \mathbf{N}_{|\delta_1|}(0, \sigma_\nu^2\mathbf{I}_{|\delta_1|})$, and sample $\alpha_2 \sim \mathbf{N}_n(0, \sigma_Y^2\mathbf{I}_n)$.
    Set $\nu^* = \alpha_1 + \tau^2 VD(1+\tau^2D^2)^{-1}(\mathbf{Y}_\nu^* - DV^{\mathrm{T}}\alpha_1 - \alpha_2)$.
    Set $\nu_{[j],j\in\delta_1} = \nu^*$.
  else
    Sample $\nu^* \sim \mathbf{N}_{|\delta_1|}(E_1, V_1)$, where

$$V_1 = \left(\sigma_Y^{-2}D^2 + \sigma_\nu^{-2}\mathbf{I}_{|\delta_1|}\right)^{-1}, \quad E_1 = V_1\left(\sigma_Y^{-2}D\mathbf{Y}_\nu^*\right).$$

    Set $\nu_{[j],j\in\delta_1} = V\nu^*$.
  end if
  Sample $\nu^0 \sim \mathbf{N}_{|\delta_0|}(0, \sigma_\nu^2\mathbf{I}_{|\delta_0|})$, and let $\nu_{[j],j\in\delta_0} = \nu^0$.
  Save ν as the t-th sample $\nu^{(t)}$.
  Section 2: Update δ_ν sequentially.
  Let p_δ be the hyper-parameter for the Bernoulli prior on δ_ν. Here we set $p_\delta = 0.5$. Compute the residual vector as $R = \mathbf{Y}_\nu - \hat{\boldsymbol{\eta}}^{\mathrm{T}}(\nu * \delta_\nu) \in \mathbb{R}^n$.
  for location $j = 1, \ldots, p$ **do**
    if $\delta_{\nu,j} = 1$ **then**
      $R_1 = R$,
      $R_0 = R + \left(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\right)_j * \nu_j$;
    else
      $R_1 = R - \left(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\right)_j * \nu_j$,
      $R_0 = R$;
    end if
    $\log l_1 = -0.5/\sigma_Y^2 * \|R_1\|_2^2$, $\log l_0 = -0.5/\sigma_Y^2 * \|R_0\|_2^2$.
    $p_1 = \exp^{\log l_1 - \log l_0}$,
    $p_1 = p_1 * p_\delta/(1-p_\delta)$,
    $p_0 = 1/(p_1+1)$, $p_1 = 1 - p_0$.
    Sample $U_j \sim \mathrm{Unif}[0,1]$, $\delta_\nu = 1$ if $U_j < p_1$ and set $R = R_1$, otherwise $\delta_\nu = 0, R = R_0$.
  end for
  Save δ_ν as the t-th sample $\delta_\nu^{(t)}$.
  Update the rest of parameters $\gamma, \zeta, \sigma_Y^2, \sigma_\gamma^2, \sigma_\zeta^2, \sigma_\nu^2$ using standard Gibbs Sampler.
end for
return the MCMC chains of $\theta_\beta, \nu, \delta_\nu, \gamma, \zeta, \sigma_Y^2, \sigma_\gamma^2, \sigma_\zeta^2, \sigma_\nu^2$.

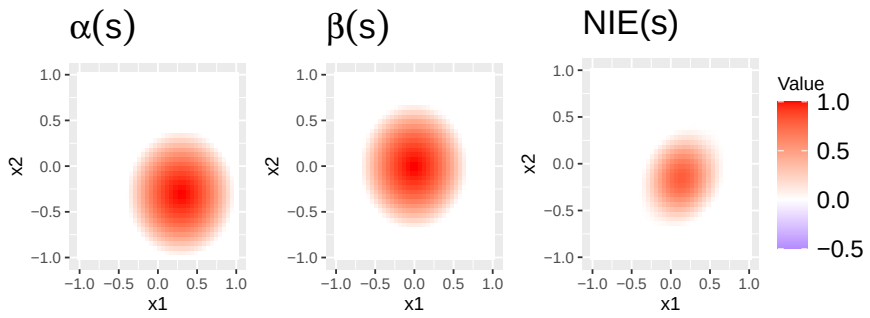


Figure C.1: True signal for $\alpha(s), \beta(s)$, and spatially-varying NIE $\mathcal{E}(s)$.

Table C.1: Simulation result of the scalar Natural Direct Effect, averaged over 100 replications. Each column represent one method. The smallest MSE of $\mathcal{E}$ is bolded in each case.

	BIMA	**BASMU**		**BIMA**	**BASMU**
Case 1	dense ν		Case 4	sparse ν, $n = 600$	
Bias	-0.44	-1.71	Bias	-8.16	-2.42
Var	0.04	0.01	Var	0.07	0.03
MSE	0.23	2.94	MSE	66.71	5.88
Case 2	sparse ν		Case 5	dense ν, $\sigma_\eta = 1$	
Bias	-5.14	-2.24	Bias	-5.90	-2.29
Var	0.06	0.02	Var	0.04	0.02
MSE	26.53	5.05	MSE	34.85	5.25
Case 3	all 0 ν		Case 6	dense ν, $\sigma_M = 4$	
Bias	0.00	-0.26	Bias	-0.51	-1.70
Var	0.00	2.23	Var	0.03	0.01
MSE	0.00	2.28	MSE	0.30	2.91

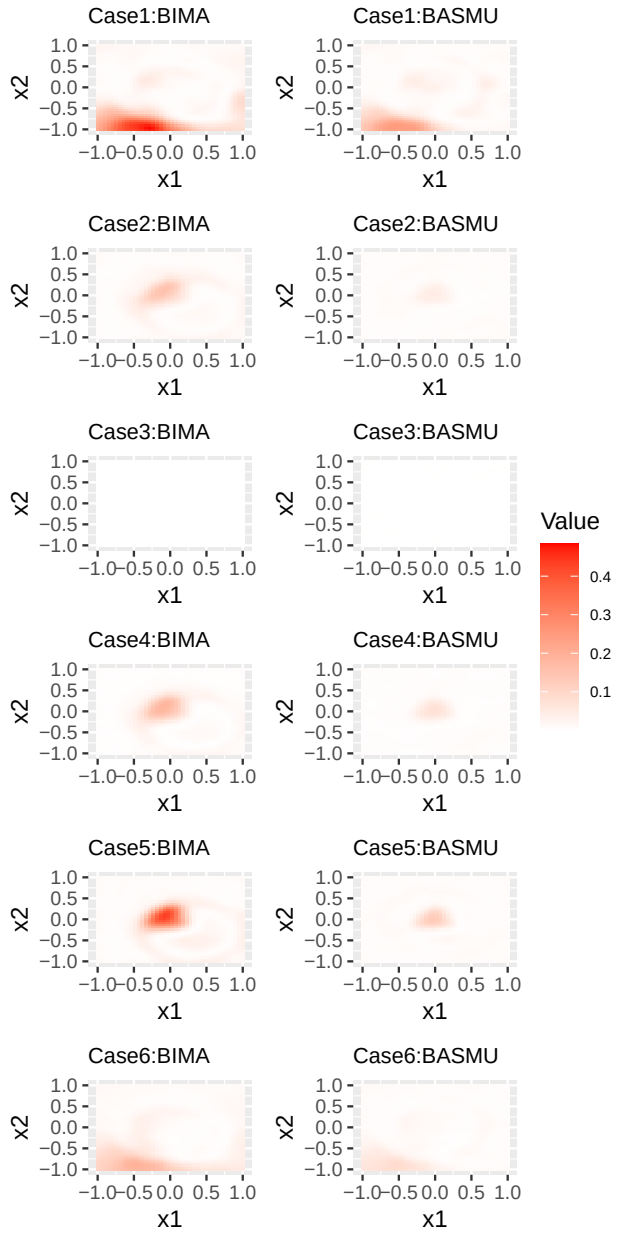


Figure C.2: MSE based on 100 replications for $\beta(s)$ over different spatial locations s, under all simulation cases. The color bar ranges from 0 to 0.48, from white to red.

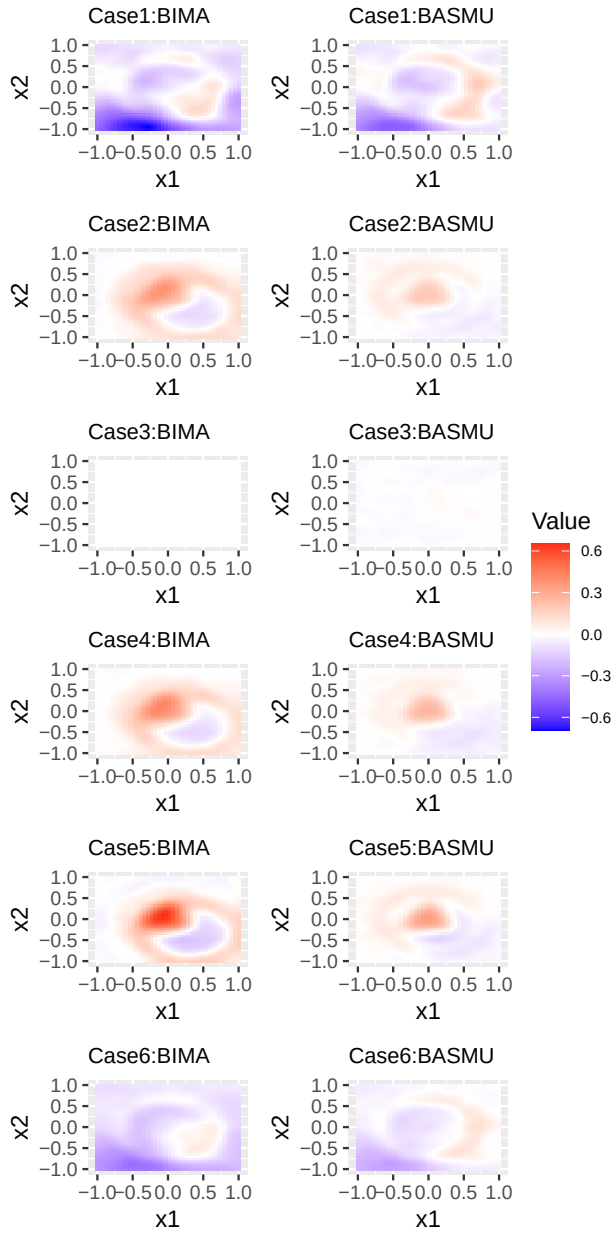
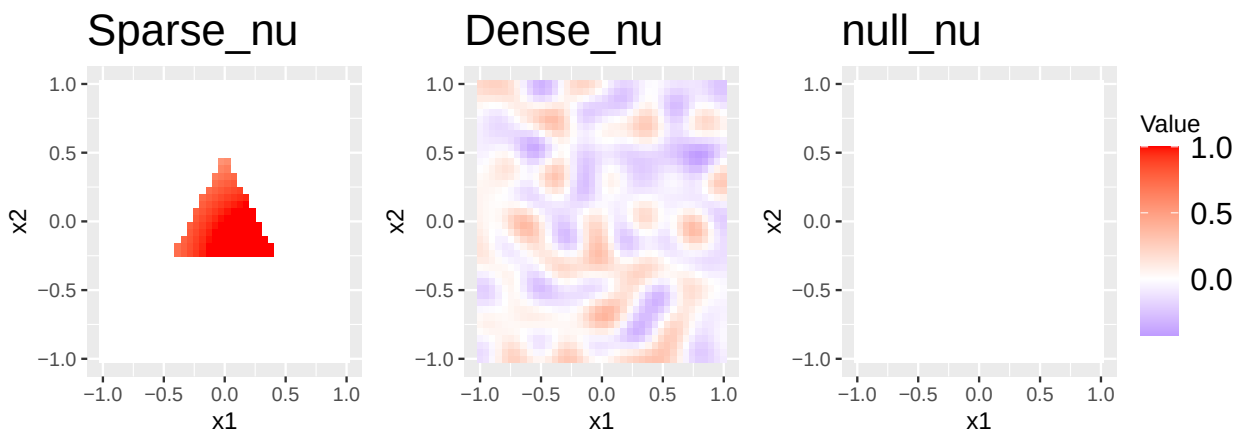
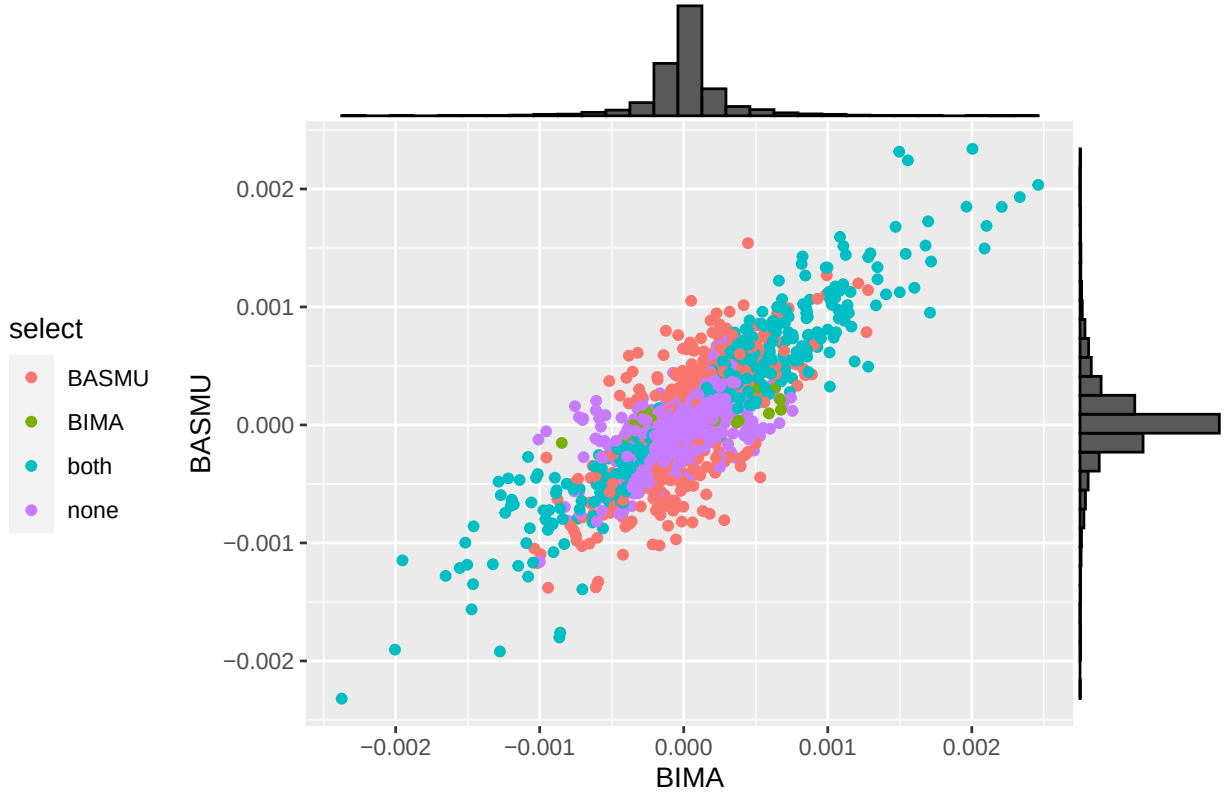


Figure C.3: Bias based on 100 replications for $\beta(s)$ over different spatial locations s, under all simulation cases. The color bar ranges from -0.7 to 0.65, from blue (negative) to white (0) to red (positive).



(a) The true signal pattern for ν, from left to right: sparse ν, dense ν, all 0 ν.



(b) Scatter plot of TIE $\mathcal{E}(s_j)$ comparison of BIMA and BASMU result. Each point is one voxel location. The x-axis is the value of $\mathcal{E}(s_j)$ estimated by BIMA, and the y-axis is estimated by BASMU. The selection is color-coded, with the legend from top to bottom: selected only by BIMA/BASMU, selected by both methods, not selected by either method.

Figure C.4: Additional simulation and real data plots.

Bibliography

[1] Artin Armagan, David B Dunson, Jaeyong Lee, Waheed U Bajwa, and Nate Strawn. Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):1011–1018, 2013.

[2] Sunil Arya, David Mount, Samuel E. Kemp, and Gregory Jefferis. *RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric*, 2019. R package version 2.6.1.

[3] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055 – 2085, 2015.

[4] 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.

[5] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.*, 51(6):1173–1182, December 1986.

[6] Andrew Barron, Mark J Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

[7] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[8] Claudia Blaiotta, M Jorge Cardoso, and John Ashburner. Variational inference for medical image segmentation. *Computer Vision and Image Understanding*, 151:14–28, 2016.

[9] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.*, 112(518):859–877, April 2017.

[10] 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.

[11] Oliver Y Chén, Ciprian Crainiceanu, Elizabeth L Ogburn, Brian S Caffo, Tor D Wager, and Martin A Lindquist. High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2):121–136, 2018.

[12] Tingan Chen, Abhishek Mandal, Hongtu Zhu, and Rongjie Liu. Imaging genetic based mediation analysis for human cognition. *Frontiers in neuroscience*, 16:824069, 2022.

[13] Nidhan Choudhuri, Subhashis Ghosal, and Anindya Roy. Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, 99(468):1050–1059, 2004.

[14] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: extending omitted variable bias. *J. R. Stat. Soc. Series B Stat. Methodol.*, 82(1):39–67, February 2020.

[15] AO Cohen, MI Conley, DV Dellarco, and BJ Casey. The impact of emotional cues on short-term and long-term memory during adolescence. *Proceedings of the Society for Neuroscience. San Diego, CA. November*, 2016.

[16] Yulai Cong, Bo Chen, and Mingyuan Zhou. Fast simulation of hyperplane-truncated multivariate normal distributions. *Bayesian Analysis*, 12(4):1017–1037, 2017.

[17] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, Marilyn S Albert, and Ronald J Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, July 2006.

[18] Peng Ding and Tyler J Vanderweele. Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Biometrika*, 103(2):483–490, 2016.

[19] Daniele Durante and Tommaso Rigon. Conditionally conjugate Mean-Field variational bayes for logistic models. *SSO Schweiz. Monatsschr. Zahnheilkd.*, 34(3):472–485, August 2019.

[20] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[21] Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[22] Dirk Eddelbuettel and Conrad Sanderson. Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, March 2014.

[23] Rami M El-Baba and Mark P Schury. Neuroanatomy, frontal cortex. 2020.

[24] Xiangnan Feng, Tengfei Li, Xinyuan Song, and Hongtu Zhu. Bayesian scalar on image regression with nonignorable nonresponse. *J. Am. Stat. Assoc.*, 115(532):1574–1597, 2020.

[25] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[26] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[27] Alan E Gelfand and Penelope Vounatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–25, January 2003.

[28] Subhashis Ghosal, Jayanta K Ghosh, and RV Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

[29] Subhashis Ghosal and Anindya Roy. Posterior consistency of gaussian process prior for nonparametric binary regression. *Ann. Statist.*, 34(5):2413–2429, 10 2006.

[30] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

[31] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *J. R. Stat. Soc. Series B Stat. Methodol.*, 73(2):123–214, March 2011.

[32] Jeff Goldsmith, Lei Huang, and Ciprian M Crainiceanu. Smooth Scalar-on-Image regression via spatial bayesian variable selection. *J. Comput. Graph. Stat.*, 23(1):46–64, January 2014.

[33] Xu Guo, Runze Li, Jingyuan Liu, and Mudong Zeng. High-dimensional mediation analysis for selecting dna methylation loci mediating childhood trauma and cortisol stress reactivity. *Journal of the American Statistical Association*, (just-accepted):1–32, 2022.

[34] Danella M Hafeman and Sharon Schwartz. Opening the black box: a motivation for the assessment of mediation. *Int. J. Epidemiol.*, 38(3):838–845, June 2009.

[35] M D Hoffman, D M Blei, C Wang, and J Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 2013.

[36] Lei Huang, Jeff Goldsmith, Philip T Reiss, Daniel S Reich, and Ciprian M Crainiceanu. Bayesian scalar-on-image regression with application to association between intracranial DTI and cognitive outcomes. *Neuroimage*, 83:210–223, December 2013.

[37] Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychol. Methods*, 15(4):309–334, December 2010.

[38] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *SSO Schweiz. Monatsschr. Zahnheilkd.*, 25(1):51–71, February 2010.

[39] Shu Jiang and Graham A Colditz. Causal mediation analysis using high-dimensional image mediator bounded in irregular domain with an application to breast cancer. *Biometrics*, February 2023.

[40] Ying Jin, Zhimei Ren, and Zhengyuan Zhou. Sensitivity analysis under the f-sensitivity models: a distributional robustness perspective. *arXiv preprint arXiv:2203.04373*, 2022.

[41] Enrico Kaden, Alfred Anwander, and Thomas R Knösche. Variational inference of the fiber orientation density using diffusion mr imaging. *Neuroimage*, 42(4):1366–1380, 2008.

[42] Jian Kang. *BayesGPfit: Fast Bayesian Gaussian Process Regression Fitting*, 2022. R package version 1.1.0.

[43] Jian Kang, Brian J Reich, and Ana-Maria Staicu. Scalar-on-image regression via the soft-thresholded gaussian process. *Biometrika*, 105(1):165–184, 2018.

[44] Brian Knutson, Andrew Westdorp, Erica Kaiser, and Daniel Hommer. Fmri visualization of brain activity during a monetary incentive delay task. *Neuroimage*, 12(1):20–27, 2000.

[45] Prachi H Kulkarni, SN Merchant, and Suyash P Awate. Mixed-dictionary models and variational inference in task fmri for shorter scans and better image quality. *Medical Image Analysis*, 78:102392, 2022.

[46] Changwoo Lee, Zhao Tang Luo, and Huiyan Sang. T-loho: A bayesian regularization model for structured sparsity and smoothness on graphs. *Advances in Neural Information Processing Systems*, 34:598–609, 2021.

[47] 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.

[48] Fan Li, Tingting Zhang, Quanli Wang, Marlen Z Gonzalez, Erin L Maresh, and James A Coan. Spatial bayesian variable selection and grouping for high-dimensional scalar-on-image regression. 2015.

[49] Xinyi Li, Li Wang, Huixia Judy Wang, and Alzheimer's Disease Neuroimaging Initiative. Sparse learning and structure identification for ultrahigh-dimensional image-on-scalar regression. *Journal of the American Statistical Association*, 116(536):1994–2008, 2021.

[50] Martin A Lindquist. The statistical analysis of fMRI data. *SSO Schweiz. Monatsschr. Zahnheilkd.*, 23(4):439–464, November 2008.

[51] Martin A Lindquist. Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, 107(500):1297–1309, 2012.

[52] Gordon D Logan. Spatial attention and the apprehension of spatial relations. *Journal of experimental psychology: Human perception and performance*, 20(5):1015, 1994.

[53] Chengwen Luo, Botao Fa, Yuting Yan, Yang Wang, Yiwang Zhou, Yue Zhang, and Zhangsheng Yu. High-dimensional mediation analysis in survival models. *PLoS computational biology*, 16(4):e1007768, 2020.

[54] David MacKinnon. Contrasts in multiple mediator models. *Contrasts In Multiple Mediator Models*, pages 141–160, 01 2000.

[55] David P MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2012.

[56] David P MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2012.

[57] David P Mackinnon and James H Dwyer. Estimating mediated effects in prevention studies. *Eval. Rev.*, 17(2):144–158, April 1993.

[58] David P MacKinnon and Amanda J Fairchild. Current directions in mediation analysis. *Curr. Dir. Psychol. Sci.*, 18(1):16–20, February 2009.

[59] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. Mediation analysis. *Annu. Rev. Psychol.*, 58:593–614, 2007.

[60] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

[61] Tanmay Nath, Brian Caffo, Tor Wager, and Martin A Lindquist. A machine learning based approach towards high-dimensional mediation analysis. *Neuroimage*, 268:119843, March 2023.

[62] Judea Pearl. Causal inference in statistics: An overview. *ssu*, 3(none):96–146, January 2009.

[63] James O Ramsay and Bernard W Silverman. *Fitting differential equations to functional data: Principal differential analysis*. Springer, 2005.

[64] Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

[65] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning series. MIT Press, London, England, November 2005.

[66] Philip T Reiss, Lei Huang, and Maarten Mennes. Fast function-on-scalar regression with penalized basis expansions. *The international journal of biostatistics*, 6(1), 2010.

[67] Alexander Rix, Mike Kleinsasser, and Yanyi Song. *bama: High Dimensional Bayesian Mediation Analysis*, 2021. R package version 1.2.

[68] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[69] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[70] Arkaprava Roy and Zhou Lan. Double soft-thresholded model for multi-group scalar on vector-valued image regression. June 2022.

[71] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[72] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[73] Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

[74] Lorraine Schwartz. On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

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

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

[77] KA Smitha, K Akhil Raja, KM Arun, PG Rajesh, Bejoy Thomas, TR Kapilamoorthy, and Chandrasekharan Kesavadas. Resting state fmri: A review on methods in resting state connectivity analysis and resting state networks. *The neuroradiology journal*, 30(4):305–317, 2017.

[78] Yanyi Song, Xiang Zhou, Jian Kang, Max T Aung, Min Zhang, Wei Zhao, Belinda L Needham, Sharon LR Kardia, Yongmei Liu, John D Meeker, et al. Bayesian hierarchical models for high-dimensional mediation analysis with coordinated selection of correlated mediators. *arXiv preprint arXiv:2009.11409*, 2020.

[79] Yanyi Song, Xiang Zhou, Jian Kang, Max T Aung, Min Zhang, Wei Zhao, Belinda L Needham, Sharon LR Kardia, Yongmei Liu, John D Meeker, et al. Bayesian sparse mediation analysis with targeted penalization of natural indirect effects. *arXiv preprint arXiv:2008.06366*, 2020.

[80] Yanyi Song, Xiang Zhou, Min Zhang, Wei Zhao, Yongmei Liu, Sharon LR Kardia, Ana V Diez Roux, Belinda L Needham, Jennifer A Smith, and Bhramar Mukherjee. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, 76(3):700–710, 2020.

[81] Chandra Sripada, Mike Angstadt, Saige Rutherford, Aman Taxali, and Kerby Shedden. Toward a "treadmill test" for cognition: Improved prediction of general cognitive ability from the task activated brain. *Human brain mapping*, 41(12):3186–3197, 2020.

[82] Linda SL Tan and David J Nott. Gaussian variational approximation with sparse precision matrices. *Statistics and Computing*, 28:259–275, 2018.

[83] Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Ann. Stat.*, 40(3):1816–1845, June 2012.

[84] Aad van der Vaart and Harry van Zanten. Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 06 2011.

[85] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[86] T J VanderWeele and S Vansteelandt. Mediation analysis with multiple mediators. *Epidemiol. Method.*, 2(1):95–115, January 2014.

[87] Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115, 2014.

[88] Tyler J VanderWeele. Mediation analysis: A practitioner's guide. *Annu. Rev. Public Health*, 37:17–32, 2016.

[89] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[90] Mikkel Wallentin, Andreas Roepstorff, Rebecca Glover, and Neil Burgess. Parallel memory systems for talking about location and age in precuneus, caudate and broca's region. *Neuroimage*, 32(4):1850–1864, 2006.

[91] Chong Wang and David M Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 2013.

[92] Xiao Wang, Hongtu Zhu, and Alzheimer's Disease Neuroimaging Initiative. Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112(519):1156–1168, 2017.

[93] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[94] Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[95] T Xifara, C Sherlock, S Livingstone, S Byrne, and M Girolami. Langevin diffusions and the metropolis-adjusted langevin algorithm. *Stat. Probab. Lett.*, 91:14–19, August 2014.

[96] Yuliang Xu, Timothy D. Johnson, Thomas E. Nichols, and Jian Kang. Scalable bayesian image-on-scalar regression for population-scale neuroimaging data analysis, 2024.

[97] Yuliang Xu and Jian Kang. Bayesian image mediation analysis. *arXiv preprint arXiv:2310.16284*, 2023.

[98] Shan Yu, Guannan Wang, Li Wang, and Lijian Yang. Multivariate spline estimation and inference for image-on-scalar regression. *Stat. Sin.*, 2021.

[99] Ying Yuan and David P MacKinnon. Bayesian mediation analysis. *Psychological methods*, 14(4):301, 2009.

[100] Zijian Zeng, Meng Li, and Marina Vannucci. Bayesian image-on-scalar regression with a spatial global-local spike-and-slab prior. *Bayesian Analysis*, 1(1):1–26, 2022.

[101] Anru R Zhang and Yuchen Zhou. On the non-asymptotic and sharp lower tail bounds of random variables. *Stat*, 9(1):e314, 2020.

[102] Daiwei Zhang, Lexin Li, Chandra Sripada, and Jian Kang. Image response regression via deep neural networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad073, 07 2023.

[103] Haixiang Zhang, Yinan Zheng, Zhou Zhang, Tao Gao, Brian Joyce, Grace Yoon, Wei Zhang, Joel Schwartz, Allan Just, Elena Colicino, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154, 2016.

[104] Mingrui Zhang and Peng Ding. Interpretable sensitivity analysis for the Baron-Kenny approach to mediation with unmeasured confounding. May 2022.

[105] Yi Zhao, Lexin Li, and Alzheimer's Disease Neuroimaging Initiative. Multimodal data integration via mediation analysis with high-dimensional exposures and mediators. *Hum. Brain Mapp.*, 43(8):2519–2533, June 2022.

[106] Yi Zhao, Martin A Lindquist, and Brian S Caffo. Sparse principal component based high-dimensional mediation analysis. *Computational statistics & data analysis*, 142:106835, 2020.

[107] Yi Zhao and Xi Luo. Granger mediation analysis of multiple time series with an application to functional magnetic resonance imaging. *Biometrics*, 75(3):788–798, 2019.

[108] Yi Zhao and Xi Luo. Pathway lasso: pathway estimation and selection with high-dimensional mediators. *Statistics and Its Interface*, 15(1):39–50, 2022.

[109] Yi Zhao and Xi Luo. Multilevel mediation analysis with structured unmeasured mediator-outcome confounding. *Comput. Stat. Data Anal.*, 179:107623, March 2023.

[110] Hongtu Zhu, Jianqing Fan, and Linglong Kong. Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109(507):1084–1098, 2014.

[111] Hongtu Zhu, Tengfei Li, and Bingxin Zhao. Statistical learning methods for neuroimaging data analysis with applications. *arXiv preprint arXiv:2210.09217*, 2022.

[112] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.