Statistical Inference on Large-scale and Complex Data via Gaussian Process

by

Moyan Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2023

Doctoral Committee:

Professor Jian Kang, Co-Chair
Assistant Professor Raed Al Kontar, Co-Chair
Associate Professor Shi Cong
Assistant Professor Zhenke Wu

Moyan Li

moyanli@umich.edu

ORCID iD: 0009-0002-7392-0038

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ALGORITHMS

ALGORITHM

ABSTRACT

Recent technological advancements have generated vast amounts of complex data in various fields, including biomedical sciences such as neuroimages, electronic health records, and electroencephalogram (EEG) signals obtained from brain computer interface (BCI) systems. However, the analysis of such data presents significant challenges due to its high dimensionality, spatial or temporal resolution, correlation structure, and heterogeneity. Gaussian Processes (GPs) have emerged as a flexible tool for Bayesian nonparametrics and machine learning, enabling the modeling of functional and dependent data over time and space. The nonparametric flexibility and high interpretability of GPs have led to their success in numerous applications. Nevertheless, existing GP models and methods are inadequate in addressing new questions that arise in analyzing large-scale and complex data. This dissertation aims to fill this gap by developing novel GP-based models and proposing efficient posterior computation algorithms using GP priors for the statistical analysis of large-scale and complex data, with a focus on biomedical applications such as brain imaging and EEG-BCI data analysis.

The first chapter of this dissertation addresses the issue of negative transfer in the multi-output Gaussian process (MGP). While the MGP assumes that outputs share commonalities, negative transfer can occur when this assumption is not met, resulting in reduced performance compared to learning outputs independently or in subsets. To avoid negative transfer in MGP models, we first define the concept and derive necessary conditions for avoiding it. Our analysis shows that, under the convolution construction, having a sufficient number of latent functions Q is the key factor in avoiding negative transfer, irrespective of the kernel or inference procedure used. However, increasing Q leads to a higher number of parameters that need to be estimated. To tackle this challenge, we propose two latent structures that can scale to large datasets, prevent negative transfer, and allow the use of any kernel or sparse approximations. We also demonstrate that our model supports regularization, facilitating the automatic selection of related outputs. We evaluate our proposed model on the Parkinson dataset, where it outperforms the original MGP model in predicting disease symptom scores.

In the second chapter, we incorporate GP into the Bayesian framework and focus on the brain image analysis using GP prior. Specifically, we propose a Bayesian nonparametric spatially varying correlation model, which is to address the question of estimation and inference of spatial regions where two imaging modalities are significantly correlated. We build our model based on the thresholded correlation Gaussian process (TCGP), which ensures piecewise smoothness, sparsity, as well as jump discontinuity of spatially varying correlations, and works well even when the number of subjects is limited or the signal-to-noise ratio is small. We study the identifiability of our model, establish the large support property, and derive the posterior consistency and selection consistency. Moreover, we derive a highly efficient Gibbs sampler algorithm and its variant to compute the posterior distribution. We illustrate the method with both simulations and an analysis of functional magnetic resonance imaging data from the Human Connectome Project.

The third chapter focuses on the electroencephalogram (EEG)-based brain computer interface (BCI) analysis, where the goal is to infer the participant's intended character on a 6×6 virtual keyboard using the EEG signal. We developed a Bayesian time-varying classification model with signal interaction via the relaxed thresholded Gaussian Process priors (SI-RTGP), which leads to an enhanced prediction and interpretation. To the best of our knowledge, we are among the first to explicitly consider the effect of the signal interaction across different channels for predicting the stimulus type outcomes. We extend the thresholded GP prior in the second chapter to the relaxed thresholded GP prior, which is more flexible and is able to model both the sparse and the non-sparse patterns by varying the "relaxing" parameter. Moreover, it provides a more computationally efficient way to conduct MCMC sampling compared to other thresholded GP priors. The proposed SI-RTGP model is applied to the P300 speller study conducted by the University of Michigan direct brain interface (UMDBI) laboratory, which achieves improved classification accuracy on multiple subjects. Additionally, the model can identify a number of scientifically meaningful channels and channel pairs, providing valuable insights for future BCI research.

CHAPTER 1

Introduction

1.1 Motivation and challenges

Biomedical research faces a myriad of challenges, including the need for robust data analysis techniques to handle the growing complexity and volume of data generated by modern medical studies. As researchers strive to develop innovative solutions for an aging population, increasing prevalence of disabilities, and rapidly evolving technology, the demand for advanced statistical methods becomes paramount. This dissertation aims to explore the development of novel statistical methods for the analysis of complex data in medical studies. The study will focus on the potential of these advanced methods to identify patterns and transform intricate data into actionable knowledge for precision medicine and informed decision-making. The investigation will encompass specific applications of these statistical techniques in various aspects of biomedical research, such as neuroimaging analysis and electronic health records, to demonstrate their value in improving patient care and fostering innovation within the field.

Neuroimaging has been playing pivotal roles in clinical diagnosis and basic biomedical research in the past decades. In this context, complex data encompasses not only large amounts of data but also complex data structures and various kinds of modalities. The most widely used imaging modalities are magnetic resonance imaging (MRI) (Abd-Ellah et al. 2019), computerized tomography (CT) (Withers et al. 2021), positron emission tomography (PET) (Jiang et al. 2019), and single-photon emission computed tomography (SPECT) (Verger et al. 2021). Recently, more and more research has focused on using multimodal imaging data obtained separately, from different subjects, and/or from different clinical or research sites (Kelberman et al. 2020, Niu et al. 2020). This practice offers the advantages of large and diverse datasets. However, it also comes with challenges of sophisticated models, complicated data normalization that includes correction of errors and variations imbedded in data from different institutions (Calhoun and Sui 2016, Tulay et al. 2019). Moreover, it is very difficulty to appropriately process data across different domains with high quality, while controlling for potential bias introduced during the preprocessing stage. It requires the whole scientific community to work closely to test all major preprocessing tools by using well-designed synthetic and real datasets in terms of reproducibility, generalizability, and reliability (Zhu et al. 2022). Additionally, it remains unclear how to appropriately and efficiently analyze neuroimaging related data sets with multiple Vs (e.g., Volume, Velocity, Variety and Veracity), while ensuring algorithmic fairness. Therefore, the efficient analysis and processing of large-scale and complex data and the development of high-performance computing tools are critical for modern neuroscientific studies.

Another example of large-scale and complex data in biomedical research is from the electroencephalogram (EEG)-based brain computer interface (BCI) speller system (Won et al. 2022), which is a device that enables a person to "type" words by EEG signal patterns in the brain activity in response to external stimuli without using a physical keyboard. It has been used to assist people with severe neuromuscular disabilities, such as amyotrophic lateral sclerosis (ALS), with regular communication (Wolpaw et al. 2018). Different from neuroimaging data which has high spatial resolution, EEG signal has the characteristic of high temporal resolution, hence downsampling is usually needed when analyzing EEG data. Another challenge rises from the low signal-to-noise ratio of EEG signals, so to achieve a decent spelling accuracy, users have to repeat the experiment many times to collect enough data. Moreover, when users spend too much time calibrating this BCI, they may experience variations in attention including fatigue and boredom. Such variations can lead to ignored, misperceived, or delayed brain response activity that may further reduce spelling efficiency.

Due to its high spatial/temporal resolution, complex data structure, low signal to noise ratio and human variability, modeling and analyzing this type of data is one of the most important and challenging work in biomedical research. In this dissertation, we develop novel statistical methods to address those issues and aim to provide insights for cognitive neuroscience research.

1.2 Existing solutions

Given the increasing volume of data, significant efforts are being made to enhance traditional computational and analytical data models. Recent modeling approaches primarily aim to achieve two goals: maximizing learning accuracy with minimal computational cost and processing large datasets quickly and efficiently.

The first type of widely used methods relies on dimension reduction techniques such as support vector machines (SVMs) (Gao et al. 2022, Sethi et al. 2022), independent component analysis (ICA) (Wu et al. 2022a, Boonyakitanont et al. 2022), and principal component analysis (PCA) (Mini et al. 2021, Cimmino et al. 2021). These methods are widely used to extract vital features from high-dimensional, complex datasets for subsequent analysis, such as prediction or classification. For example, Xu et al. (2004) and Kaper et al. (2004) developed ICA- and SVM-based algorithms, respectively, for BCI applications, while Du et al. (2020) proposed an automated, adaptive ICA-based pipeline for identifying reproducible fMRI markers of brain disorders. However, these approaches of reducing data size through dimension reduction may sacrifice some information, making subsequent analyses more challenging, particularly when sample sizes are small or signal-to-noise ratios are already low.

The second type of method is deep learning, which is a subfield of machine learning and recently showed a remarkable performance in various area. Convolutional neural networks (CNNs) are widely used in brain image classification and segmentation. For example, Sultan et al. (2019) proposed a model based on CNN to classify different brain tumor types. Khan et al. (2020) introduce the CNN approach along with data augmentation to categorize brain MRI scan images into cancerous and non-cancerous. More recently, several neural network methods have been proposed in the BCI context for EEG-based classification tasks. For example, multi-task autoencoder-based models such as Ditthapron et al. (2019), compact CNNs like EEGNet (Lawhern et al. 2018), and weighted ensemble strategies such as Kshirsagar and Londhe (2019) have shown promising results. However, deep learning methods are typically regarded as black boxes, lacking interpretability and making it challenging to gain insight into the underlying neural mechanisms despite their high prediction or classification accuracy.

Finally, Bayesian methods are commonly used to analyze large-scale and complex data as they offer statistical inference capabilities that allow us to simulate the entire posterior distribution, compute posterior inclusion probabilities, and quantify uncertainty for selection or prediction. For instance, Marroquín et al. (2002) proposed an efficient Bayesian method for automatic segmentation of brain MRI, while Jayachitra and Prasanth (2021) used a weighted Gaussian Naïve Bayes classifier for brain stroke classification, and Kia et al. (2020) introduced a hierarchical Bayesian regression model for probabilistic modeling of batch effects in neuroimaging data. Researchers have also explored the use of Bayesian methods in EEG-based BCI classification. For example, Barthélemy et al. (2023) proposed a novel Bayesian accumulation of Riemannian probabilities, which is an end-to-end pipeline for P300 BCI classification, and Ma et al. (2022) made the first attempt to study the probability distribution of multi-trial EEG signals using a Bayesian generative model, providing a useful tool to simulate EEG signals in P300 BCI and a novel probabilistic classifier. Despite the success of these methods, challenges still remains in prior specification and computational efficiency. Although several new MCMC sampling algorithms have been developed recently (e.g., Ahn et al. 2012, Chen et al. 2014, Nishimura et al. 2020), they usually converge slowly and require multiple tuning parameters. Furthermore, specifying a proper prior is crucial, particularly for datasets with complex structures and small sample sizes.

1.3 Gaussian Process

Gaussian Process (GP) is a flexible and powerful tool for modeling complex and high-dimensional data, and has been widely used in machine learning, statistics, and signal processing. In Bayesian inference, GP can be adopted for prior specifications of functional parameters in the model. It has flexibility to incorporate prior knowledge into the model and more accurately quantify uncertainty of model inference. This makes GP a valuable tool for many applications, such as regression, classification, optimization, and control.

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution, which is completely specified by its mean function and covariance function. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a d-dimensional input vector and $f(\boldsymbol{x})$ be the corresponding output. We define mean function $m(\boldsymbol{x})$ and the covariance function $k\left(\boldsymbol{x}, \boldsymbol{x}^{\prime}\right)$ of a real process $f(\boldsymbol{x})$ as

$$\begin{aligned} m(\boldsymbol{x}) &= \mathbb{E}[f(\boldsymbol{x})] \\ k\left(\boldsymbol{x}, \boldsymbol{x}^{\prime}\right) &= \mathbb{E}\left[(f(\boldsymbol{x})-m(\boldsymbol{x}))\left(f\left(\boldsymbol{x}^{\prime}\right)-m\left(\boldsymbol{x}^{\prime}\right)\right)\right] \end{aligned}$$

and will write the Gaussian process as

$$f(\boldsymbol{x}) \sim \mathrm{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}^{\prime})). \tag{1.1}$$

The choice of mean and covariance functions determines the properties of the GP. The mean function describes the overall trend of the function, while the covariance function describes the smoothness and correlation between different input points. Commonly used covariance functions include the squared exponential (Stein 1999), Matérn (Matérn 2013), and periodic kernels (Smola and Schölkopf 1998), which can be combined and modified to suit different applications.

Given n training points $\boldsymbol{X} = \left\{\boldsymbol{x}_i \in \mathbb{R}^d\right\}_{i=1}^n$ and their observations $\boldsymbol{y} = \left\{y_i = y\left(\boldsymbol{x}_i\right) \in \mathbb{R}\right\}_{i=1}^n$, where $y\left(\boldsymbol{x}_i\right) = f\left(\boldsymbol{x}_i\right) + \varepsilon$ with the iid noise $\varepsilon \sim \mathrm{N}\left(0, \sigma_{\varepsilon}^2\right)$, GP seeks to infer the latent function $f: \mathbb{R}^d \mapsto \mathbb{R}$ in the function space $\mathrm{GP}\left(m(\boldsymbol{x}), k\left(\boldsymbol{x}, \boldsymbol{x}^{\prime}\right)\right)$ defined by the mean $m(\cdot)$ and the kernel $k(\cdot, \cdot)$. Let $\boldsymbol{K}_{nn} = k(\boldsymbol{X}, \boldsymbol{X})$ and $\boldsymbol{K}_{nn}^{\varepsilon} = \boldsymbol{K}_{nn} + \sigma_{\varepsilon}^2 \boldsymbol{I}_n$, then the model evidence (marginal likelihood) can be represented as $pr(\boldsymbol{y} \mid \boldsymbol{\theta}) = \int pr(\boldsymbol{y} \mid \boldsymbol{f}) pr(\boldsymbol{f}) d\boldsymbol{f} = \mathrm{N}\left(\boldsymbol{y} \mid \mathbf{0}, \boldsymbol{K}_{nn}^{\varepsilon}\right)$ where $\boldsymbol{\theta}$ comprises the hyperparameters which could be inferred by maximizing

$$\log pr(\boldsymbol{y}) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \log \left|\boldsymbol{K}_{nn}^{\varepsilon}\right| - \frac{1}{2} \boldsymbol{y}^{\top}\left(\boldsymbol{K}_{nn}^{\varepsilon}\right)^{-1} \boldsymbol{y}. \tag{1.2}$$

Let $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$ represents the training data, the predictive distribution at a test point $\boldsymbol{x}_*$, i.e.

$pr\left(f_* \mid \mathcal{D}, \boldsymbol{x}_*\right) = \mathrm{N}\left(f_* \mid m\left(\boldsymbol{x}_*\right), \sigma_*^2\left(\boldsymbol{x}_*\right)\right)$ has the mean and variance respectively expressed as

$$\begin{aligned} m\left(\boldsymbol{x}_*\right) &= \boldsymbol{k}_{*n}\left(\boldsymbol{K}_{nn}^{\varepsilon}\right)^{-1}\boldsymbol{y}, \\ \sigma^2\left(\boldsymbol{x}_*\right) &= k_{**} - \boldsymbol{k}_{*n}\left(\boldsymbol{K}_{nn}^{\varepsilon}\right)^{-1}\boldsymbol{k}_{n*}, \end{aligned} \tag{1.3}$$

where $\boldsymbol{k}_{*n} = k\left(\boldsymbol{x}_*, \boldsymbol{X}\right)$ and $k_{**} = k\left(\boldsymbol{x}_*, \boldsymbol{x}_*\right)$. For y_*, we need to consider the noise such that $pr\left(y_* \mid \mathcal{D}, \boldsymbol{x}_*\right) = \mathrm{N}\left(y_* \mid m\left(\boldsymbol{x}_*\right), \sigma_*^2\left(\boldsymbol{x}_*\right) + \sigma_\varepsilon^2\right)$.

Recently, GP models have demonstrated success in various domains. For example, Futoma et al. (2017) developed a GP-based model that can help detect and improve treatment of sepsis, and Li et al. (2021) proposed a deep Bayesian GP to estimate uncertainty in electronic health records. GPs also have wide applications in modeling spatial and temporal dependencies and are widely used in modeling image and time series data. For instance, Wu et al. (2022b) and Kang et al. (2018) both use GP priors to model neuroimaging data, while Ma et al. (2022) uses a GP prior to perform Bayesian inferences on neural activity in EEG-based brain-computer interfaces.

In conclusion, GP is a powerful tool for modeling complex systems and making predictions with uncertainty quantification. Its ability to incorporate prior knowledge and adjust to new data in a Bayesian framework makes it a popular choice in many fields, including engineering, computer science, finance, and healthcare. With its versatility and potential for various applications in modeling and prediction, Gaussian Process is poised to continue to be a useful tool for solving complex problems in the future.

1.3.1 Multi-output Gaussian Process (MGP)

A Multi-output Gaussian Process (MGP) is an extension of the standard GP that can model multiple outputs simultaneously. In a MGP, the outputs are modeled as a collection of correlated GPs, where the covariance structure between outputs is specified by a cross-covariance function. The cross-covariance function describes the correlation between the output of two different GPs at any two input points. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a d-dimensional input vector, a MGP can be formally defined as follows:

$$\boldsymbol{f}(\boldsymbol{x}) \sim \mathrm{MGP}(\boldsymbol{m}(\boldsymbol{x}), \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')), \tag{1.4}$$

where $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \dots, f_n(\boldsymbol{x})] \in \mathbb{R}^n$ is the vector of outputs at input vector $\boldsymbol{x}$ and $f_i : \mathbb{R}^d \mapsto \mathbb{R}$, $i = 1, \dots, n$ are n GPs. $\boldsymbol{m}(\boldsymbol{x}) = [m_1(\boldsymbol{x}), m_2(\boldsymbol{x}), \dots, m_n(\boldsymbol{x})] \in \mathbb{R}^n$ is the vector of mean functions, and $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \begin{bmatrix} K_{11} & \dots & K_{1n} \\ \vdots & \ddots & \vdots \\ K_{n1} & \dots & K_{nn} \end{bmatrix}$ is the block matrix of covariance functions that describe the correlations between the outputs at input vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$. Specifically, the covariance between f_i at input $\boldsymbol{x}$ and output f_j at input $\boldsymbol{x}_0$ is defined as $K_{ij}(\boldsymbol{x}, \boldsymbol{x}_0) = \mathrm{cov}(f_i(\boldsymbol{x}), f_j(\boldsymbol{x}_0))$.

The advantages of MGP lies in its ability to integratively analyze multiple outputs in order to leverage their commonalities and share information among outputs, hence improve predictive and learning accuracy. MGPs have been widely used in various applications, including sensor fusion and spatial-temporal modeling of the health care data. For example, Kia and Marquand (2018) introduce a scalable MGP regression by modeling both spatial and across-sample variances, which provides higher sensitivity in novelty detection scenarios, Kia et al. (2018) proposed a scalable MGP tensor regression for normative modeling of structured variation in neuroimaging data.

Overall, the MGP provides a flexible and powerful framework for modeling multiple related functions, and has been successfully applied in various domains, such as computer vision, robotics, and bioinformatics.

1.3.2 Thresholded Gaussian Process (TGP)

Given a Gaussian Process $f(x)$ defined on $\mathbb{R}$, we use $T_s(f(x),\omega)$ to represent the soft thresholded GP. $T_s(\cdot,\omega)$ is the soft threshoding function.

$$T_s(x,\omega)=\begin{cases} 0, & |x|\leqslant\omega \\ \operatorname{sgn}(x)(|x|-\omega), & |x|>\omega \end{cases}$$

where $\operatorname{sgn}(x)=1$ if $x>0$, $\operatorname{sgn}(x)=-1$ if $x<0$, and $\operatorname{sgn}(0)=0$. The thresholding parameter $\omega>0$ determines the degree of sparsity. On the other hand, we use $T_h(f(x),\omega)$ to represent the hard thresholded GP, where

$$T_h(x,\omega)=\begin{cases} 0, & |x|\leqslant\omega \\ x, & |x|>\omega \end{cases}$$

TGP has numerous applications in modeling neuroimaging and brain activity temporal data, where prior knowledge indicates that the signal should vary smoothly over the brain region or throughout the time. For instance, Kang et al. (2018) utilized the soft thresholded GP as a prior for the scalar-on-image regression problem, enabling a large prior support over the class of piecewise-smooth, sparse, and continuous spatially varying regression coefficient functions. Wu et al. (2022b) introduced a Bayesian model incorporating the hard thresholded GP for separating latent brain networks and detecting activated brain activation.

1.3.3 Gaussian Process Approximation

Although GP models have demonstrated success in various domains, their main limitation is the $O(n^3)$ computation and $O(n^2)$ storage for n training points (Rasmussen 2004). In order to improve the scalability of standard GP for large-scale data, several approximation methods of GP have been extensively presented and studied in recent years.

The first type of model uses global approximations, which achieve the sparsity of the full kernel matrix $\boldsymbol{K}_{nn}$ through (i) using a subset of the training data (subset-of-data) (Keerthi and Chu 2005, Lawrence et al. 2002); (ii) removing the entries of $\boldsymbol{K}_{nn}$ with low correlations (sparse kernels) (Melkumyan and Ramos 2009); and (iii) employing a low-rank representation (sparse approximations). Subset-of-data (SoD) is the simplest strategy to approximate the full GP by using a subset $\mathcal{D}_{\text{sod}}$ of the training data $\mathcal{D}$. Hence, the SoD retains the standard GP inference at lower time complexity of $O\left(m^3\right)$, since it operates on $\boldsymbol{K}_{mm}$ which only comprises $m(m \ll n)$ data points. A recent theoretical work (Hayashi et al. 2020) analyzes the error bounds for the prediction and generalization of SoD through a graphon-based framework, indicating a better speed-accuracy trade-off in comparison to other approximations reviewed below when n is sufficiently large. Sparse kernels (Melkumyan and Ramos 2009) attempt to directly achieve a sparse representation $\tilde{\boldsymbol{K}}_{nn}$ of $\boldsymbol{K}_{nn}$ via the particularly designed compactly supported (CS) kernel, which imposes $k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = 0$ when $\left|\boldsymbol{x}_i - \boldsymbol{x}_j\right|$ exceeds a certain threshold. Therefore, only the non-zero elements in $\tilde{\boldsymbol{K}}_{nn}$ are involved in the calculation. As a result, the training complexity of the GP using CS kernel scales as $O\left(\alpha n^3\right)$ with $0 < \alpha < 1$.

In this dissertation, we will focus on using one of the sparse approximations approaches. We first introduce Mercer's theorem here, which is a continuous analog of the singular-value or eigenvalue decomposition of a symmetric positive definite matrix. One of its main applications is to find convenient ways to express stochastic processes, via the Karhunen-Loeve expansion (Xiu 2010).

Suppose $\kappa(s,t)$ is a symmetric, continuous, and non-negative definite kernel function on $[a,b] \times [a,b]$. Mercer's theorem asserts that there is an orthonormal set of eigenfunctions $\psi_l(t)$ and eigenvalues λ_l such that

$$\kappa(s,t) = \sum_{l=1}^{\infty} \lambda_l \psi_l(s) \psi_l(t), \tag{1.5}$$

where the values and functions satisfy the integral eigenvalue equation $\lambda_l \psi_l(s) = \int_a^b \kappa(s,t)\psi_l(t)$. Now suppose that $f(t)$ is a stochastic process for t in some interval $[a,b]$. The process is characterized by its mean, $m(t)$, and its covariance, $\kappa(s,t)$. Using Mercer's theorem on κ, we can express the process by the K-L expansion

$$f(t) = m(t) + \sum_{l=1}^{\infty} \sqrt{\lambda_l} Z_l \psi_l(t), \tag{1.6}$$

where λ_j and ψ_j are Mercer eigenmodes for κ, and the Z_j are uncorrelated and of unit variance. K-L expansion is a generalization of the singular value decomposition of a matrix, which can be written as a sum of outer products of vectors. In practice, we truncate the above expansions by focusing on the leading L eigenvalues and eigenfunctions to approximate the original GP, where

L can be determined following the usual practice of principal components analysis that retains a certain percentage of total variation. This reduces the problem to finding the first L eigenvalues and eigenfunctions in Eq.(1.5). The truncation point L should depend on the required level of accuracy in the reconstruction of the covariance function.

Another type of methods focus on local approximations, which use localized experts to improve the scalability of GP. Notable examples include (i) Naive-local-experts (Gramacy and Apley 2015, Park and Huang 2016, Datta et al. 2016) (ii) Mixture-of-experts, which devotes to combining the local and diverse experts owning individual hyperparameters for improving the overall accuracy and reliability (Yuksel et al. 2012, Masoudnia and Ebrahimpour 2014) (iii) Product-of-experts (Cao and Fleet 2014, Hinton 2002). Different from the MoE which employs a weighted sum of several probability distributions (experts) via an “or” operation, the product-of-experts (PoE) multiplies these probability distributions.

1.4 Our Contributions

In this dissertation, we present several novel statistical methods based on Gaussian Process and make important contributions to modeling, theory, computation and applications.

For statistical modeling and theory, we develop several new models based on MGP and TGP. In the first project, we introduce two novel latent structures: the pairwise model and the arrowhead model. These structures address the negative transfer problem in MGP and allow for regularization penalties on hyper-parameters, which facilitates selection of related and unrelated outputs. In the second project, we propose a Bayesian nonparametric spatially varying correlation model for conducting multimodal correlation analysis. We propose a new Bayesian nonparametric prior, the thresholded correlation Gaussian process (TCGP), which targets second-order correlations between two modalities. TCGP is a nontrivial extension of TGP or threshold prior, which has been adopted in prior constructions for modeling sparse regressions or spatially varying functions, i.e., either thresholding Gaussian random variables (Nakajima and West 2013, Ni et al. 2019, Cai et al. 2020), or thresholding GP (Kang et al. 2018, Wu et al. 2022b). However, none of those priors are readily applicable for Bayesian analysis of multimodal correlation analysis as in our setting. Moreover, we are among the first to study the theoretical properties of Bayesian analysis of spatially varying correlations. Particularly, we prove the model identifiability and establish the posterior consistency and the selection consistency based upon the foundational work of Choi (2005), Ghosal and Roy (2006), Tokdar and Ghosh (2007). In the third project, we extend the thresholded GP to the relaxed thresholded GP and propose a Bayesian time-varying regression model with signal interactions via relaxed thresholded GP priors (SI-RTGP). The proposed relaxed thresholded GP prior encompasses a large class of temporal varying functions that are piecewise smooth and sparse, which enables the feature selection during the Bayesian MCMC sampling. Moreover, compared to the previous thresholded GP prior (Kang et al. 2018, Cai et al. 2020), the relaxed thresholded GP prior is more flexible and is able to model both the sparse and the non-sparse patterns by varying the "relaxing" parameter. These contributions provide innovative solutions for complex problems and demonstrate the potential of our proposed frameworks in advancing the field of Gaussian Process modeling.

This dissertation reserach also has led to the development of computational tools and new algorithms that can effectively analyze large-scale and complex datasets. One major contribution is the relaxation models we proposed for the MGP structure, which can scale to handle arbitrarily large datasets. By distributing the MGP into a group of bivariate GPs that are independently built, predictions can be obtained by combining the predictions from each bivariate GP. This approach allows for parallelization, enabling each submodel to be estimated with a limited number of parameters. Additionally, we have developed an efficient sampling algorithm for posterior computation in Bayesian models with thresholding-type priors. Most existing solutions resort to gradient based MCMC algorithms (Roberts and Rosenthal 1998, Girolami and Calderhead 2011), where a smooth approximation of the thresholding function is required to get the analytically tractable first derivative (Cai et al. 2020, Wu et al. 2022b). There have also been recent advances in developing new sampling algorithms (e.g., Ahn et al. 2012, Chen et al. 2014, Nishimura et al. 2020). However, these algorithms usually converge relatively slowly, and require multiple tuning parameters. Our approach is based on the full conditional distributions, resulting in a highly efficient Gibbs sampler algorithm that outperforms other methods in terms of convergence speed and tuning parameters.

The proposed work has wide-ranging applications across different domains. One promising area is multimodal neuroimaging analysis, where the goal is to investigate the association between two imaging modalities and identify brain regions where such an association is statistically significant. As an illustrative example, we applied our model to study the association between resting-state fMRI and memory task-related fMRI in the Human Connectome Project. Our findings identified several brain regions where resting-state and task-related brain activities are strongly associated. This type of analysis is particularly useful given the growing interest in predicting task-related brain activations from resting-state fMRI data in recent years (Tavor et al. 2016, Jones et al. 2017, Cohen et al. 2020). In our study, we found that the angular gyrus exhibited the highest positive mean correlation, which is consistent with existing research on this region's role in cognitive processes related to language, number processing and memory retrieval (Farrer et al. 2008, Seghier 2013). We also identified strong positive correlations in the middle temporal gyrus and superior parietal gyrus. The former has been linked to numerous cognitive processes (Acheson and Hagoort 2013), while the latter is critically involved in information manipulation in working memory (Koenigs et al. 2009). Furthermore, our analysis identified two regions in the lingual gyrus with strong negative correlation, which is believed to play an important role in visual memory and word processing (Leshikar et al. 2012). These findings offer useful insights into brain activities during rest and working memory tasks and demonstrate the potential of our proposed model in studying complex brain networks. Another application is analyzing brain-computer interface (BCI) data. We applied our model to EEG data collected from the P300 speller study conducted by the University of Michigan Direct Brain Interface Laboratory. The experiment aimed to infer a participant's intended character on a virtual keyboard using the EEG signal. By leveraging the signal interactions across channels and performing feature selection using relaxed thresholded prior, our proposed method improved prediction accuracy compared to other machine learning methods. We identified several significant channels and channel pairs that contribute most to predicting the stimulus type outcomes, confirming and extending existing findings (Krusienski et al. 2008, McCann et al. 2015). Overall, our proposed model demonstrates its potential for various applications in studying brain structure and activity and offer insight for future neuroscientific study.

CHAPTER 2

On Negative Transfer and Structure of Latent Functions in Multi-output Gaussian Processes

2.1 Introduction

The multi-output, also referred to as multivariate/vector-valued/multitask, Gaussian process (GP) Williams and Rasmussen (2006) draws it root from transfer learning, specifically multitask learning. The goal is to integratively analyze multiple outputs in order to leverage their commonalities and hence improve predictive and learning accuracy. Indeed, the multi-output GP (MGP) has seen many success stories in the last decade (Journel and Huijbregts 1976, Jones and Johnson 2009, Gramacy and Apley 2015, Sung et al. 2020). This success can be largely attributed to the convolution construction which provided the capability to account for heterogeneity and non-trivial commonalities in the outputs. The convolution process (CP) is based on the idea that a GP, $f_i(\boldsymbol{x}) : \mathbb{R}^{\mathcal{D}} \to \mathbb{R}$ can be constructed by convolving a latent Gaussian process $X(\boldsymbol{x})$ with a smoothing kernel $K_i(\boldsymbol{x})$. This construction, first proposed by Ver Hoef and Barry (Ver Hoef and Barry 1998) and Higdon (Higdon 2002), is equivalent to stimulating a linear filter characterized by the impulse response $K_i(\boldsymbol{x})$. The only restriction is a stable filter, i.e., $\int |K_i(\boldsymbol{u})| d\boldsymbol{u} < \infty$. Given the CP construction, if we share multiple latent functions $X_q(\boldsymbol{x})$ across $f_i(\boldsymbol{x}), i \in \{1, ..., N\}$, then all N outputs can be expressed as jointly distributed GP, i.e., an MGP (Álvarez and Lawrence 2011). This is shown in eq. (2.1).

$$f_i(\boldsymbol{x}) = \sum_{q=1}^{Q} K_{qi}(\boldsymbol{x}) \star X_q(\boldsymbol{x}), \tag{2.1}$$

where $\star$ denotes a convolution. The key feature in eq. (2.1) is that it allows information to be shared through different kernels which enables great flexibility in describing the data. Many models used to build cross correlations across outputs including the large class of separable covariances and the linear model of corregonialization were shown to be special cases of the convolution construction Alvarez et al. (2012), Fricker et al. (2013). Since then, work on MGP has mainly focused on two trends: (i) Efficient inference procedures that address the computational complexity (a challenge inherited from the GP) (Nychka et al. 2015, Gramacy and Apley 2015, Damianou et al. 2016, Gramacy 2016, Gramacy and Haaland 2016). This literature has mainly focused on variational inference which laid the theoretical foundation for the commonly used class of inducing point/kernel approximation (Burt et al. 2019, Snelson and Ghahramani 2006, Sung et al. 2016, Yang et al. 2020). Interestingly, variational inference also reduced overfitting and helped generalization due to its regularizing impact (Zhao and Sun 2016, Nguyen et al. 2014, Moreno-Muñoz et al. 2018). A recent push on utilizing distributed computations for exact GP has also shown promise in this area (Wang et al. 2019, Nguyen et al. 2019, Zhang et al. 2021). (ii) Building expressive kernels (Wilson 2014, Zhang and Apley 2016, Parra and Tobar 2017, Chen et al. 2019, Ulrich et al. 2015) that often can represent certain unique features of the data studied. Recent literature have focused on spectral kernels, despite the fact that convolved covariance based on the exponential, Gaussian or Matérn kernels are still very common in applications. Relating to this, many studies have provided elegant formulations of qualitative features in a GP based on an MGP treatment with separable covariances (Li et al. 2020a, Zhou et al. 2011, Qian et al. 2008). This set of literature models outputs with different qualitative features as separate GPs and assumes that correlation over the continuous input is separable from correlation across qualitative features.

However, a key question is yet to be answered. The MGP is based on the assumption that outputs share commonalities, but what happens if this assumption does not hold? would **negative transfer** occur? which in turn leads to *forced correlation and decreased performance relative to learning each output independently* (Caruana 1997). This question is especially relevant when using the CP, which implicitly implies that outputs have heterogeneous, possibly unique, features. For instance, following recent literature, would an expressive kernel and an efficient inference procedure for finding good kernel parameter estimates automatically avoid spurious correlations? Or say in an extreme case where all outputs share no commonalities, would the MGP automatically collapse into independent GPs? Indeed, recent literature Mak et al. (2018), Li et al. (2020a) has recognized that an MGP is not necessarily better than an individual GP.

In this article we shed light on the aforementioned questions. Specifically, we first define negative transfer in the context of an MGP. We then show that addressing the challenge above is mainly dependent on having a sufficient number of latent functions, i.e., Q in eq. (2.1). However, even when N is relatively small, a small increase in Q would cause the number of parameters to be estimated to skyrocket. This renders estimation in such a non-convex and highly nonlinear setting impractical, which explains why current literature including the above cited papers only use $1 \leq Q \leq 4$. To this end, we investigate *easy-to-implement relaxation models* on the MGP structure that scale to arbitrarily large datasets, can avoid negative transfer and allow any kernel or sparse approximations to be used within. A key feature of our models is that they allow regularization penalties on the hyper-parameters which can provide selection of related/unrelated outputs.

We organize the remaining paper as follows. Sec.2.2 provides some preliminaries. Sec.2.3 defines negative transfer and provides conditions to avoid it in the MGP. In Sec.2.4 we provide scalable relaxation models on the MGP structure, we then explore regularization schemes that help generalization and automatic selection in the relaxation models. A proof of concept and illustration on real-data is given in Sec.2.5. Finally, we conclude our paper in Sec.2.6. Technical details are given in the supplementary materials.

2.2 Preliminaries

Consider the set of N noisy output functions $\boldsymbol{y}(\boldsymbol{x}) = [y_1(\boldsymbol{x}), \cdots, y_N(\boldsymbol{x})]^\top$ and let $\mathcal{I} = \{1, \cdots, N\}$ be the corresponding index set.

$$\begin{bmatrix} y_1(\boldsymbol{x}) \\ y_2(\boldsymbol{x}) \\ \vdots \\ y_N(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ \vdots \\ f_N(\boldsymbol{x}) \end{bmatrix} + \begin{bmatrix} \epsilon_1(\boldsymbol{x}) \\ \epsilon_2(\boldsymbol{x}) \\ \vdots \\ \epsilon_N(\boldsymbol{x}) \end{bmatrix} = F(\boldsymbol{x}) + E(\boldsymbol{x}),$$

where $F : \mathbb{R}^D \to \mathbb{R}^N$ is zero mean multivariate process with covariance

$$\mathrm{cov}^f_{ij}(\boldsymbol{x}, \boldsymbol{x}') := \mathrm{cov}\big(f_i(\boldsymbol{x}), f_j(\boldsymbol{x}')\big)$$

for $i, j \in \mathcal{I}$ and $\epsilon_i(\boldsymbol{x}) \sim \mathcal{N}(0, \sigma_i^2)$ represents additive noise. For the ith output the observed data is denoted as $\mathcal{D}_i = \{(\boldsymbol{y}_i, \boldsymbol{X}_i)\}$, where $\boldsymbol{y}_i = [y_i^1, \cdots, y_i^{p_i}]^\top$, $y_i^c := y_i(\boldsymbol{x}_{ic})$, $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{ip_i}]^\top$ and p_i represents the number of observations for output i. Now let $P = \sum p_i$ and $D_{\mathcal{I}} = \{D_1, \cdots, D_N\}$, then the predictive distribution for output i at $\boldsymbol{x}_0$ is given as

$$pr(y_i(\boldsymbol{x}_0)|D_{\mathcal{I}}) = \mathcal{N}\big(\boldsymbol{C}^\top_{\boldsymbol{f},f_i^0}(\boldsymbol{C}_{\boldsymbol{f},\boldsymbol{f}} + \Sigma)^{-1}\boldsymbol{y},\ C_{f_i^0,f_i^0} + \sigma_i^2 - \boldsymbol{C}^\top_{\boldsymbol{f},f_i^0}(\boldsymbol{C}_{\boldsymbol{f},\boldsymbol{f}} + \Sigma)^{-1}\boldsymbol{C}_{\boldsymbol{f},f_i^0}\big) \quad (2.2)$$

where $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_N^\top]^\top$ corresponds to the latent function values $\boldsymbol{f} = [\boldsymbol{f}_1^\top, ..., \boldsymbol{f}_N^\top]^\top$, $\boldsymbol{C}_{\boldsymbol{f},\boldsymbol{f}} \in \mathbb{R}^{P\times P}$ is the covariance matrix from the operator $\mathrm{cov}^f_{ij}(\boldsymbol{x}, \boldsymbol{x}')$ and $\boldsymbol{\Sigma} = \mathrm{diag}[\sigma_1^2\boldsymbol{I}_{p_1}, ..., \sigma_N^2\boldsymbol{I}_{p_N}]$ is a block diagonal matrix with $\boldsymbol{I}$ as the identity matrix.

As shown in (2.2), information transfer is facilitated via $\mathrm{cov}^f_{ij}(\boldsymbol{x}, \boldsymbol{x}')$. Now, under the CP in (2.1) and assuming independent latent function X_q with $\mathrm{cov}(X_i(\boldsymbol{u}), X_i(\boldsymbol{u}')) = \delta(\boldsymbol{u} - \boldsymbol{u}') = \delta_{\boldsymbol{u}\boldsymbol{u}'}$ (δ is the Dirac delta function) we then have

$$\mathrm{cov}^f_{ij}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{q=1}^{Q} \int_{-\infty}^{\infty} K_{q_i}(\boldsymbol{u})K_{q_j}(\boldsymbol{u} - \boldsymbol{d})d\boldsymbol{u}. \quad (2.3)$$

where $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}' \in \mathbb{R}^D$ denotes a convolution. Here we note that a more general case can be used where X_i is a GP generated from a CP, i.e., $\text{cov}(X_i(\boldsymbol{u}), X_i(\boldsymbol{u}')) = \int K_{X_i}(\boldsymbol{u})K_{X_i}(\boldsymbol{u} - \boldsymbol{d})d\boldsymbol{u}$. In the appendix we show that the following results also hold under such a case.

2.3 Negative Transfer: Definition and Conditions

2.3.1 Definition of negative transfer

Similar to multi-task learning, negative transfer draws its roots from transfer learning Pan and Yang (2009). A widely accepted description of negative transfer is stated as *"transferring knowledge from the source can have a negative impact on the target learner"*. In an MGP, negative transfer could be defined similarly: the integrative analysis of all outputs can have negative impact on the performance of the model compared with separate modeling of each output or a subset of them.

Definition 1. *Consider an* MGP *with N possible outputs, and assume y_i represents the target output. Let the index set of all outputs $\mathcal{I}$ comprise of M non-empty disjoint subsets $\mathcal{I} = \{\mathcal{I}_1 \cup \cdots \mathcal{I}_m \cup \cdots \mathcal{I}_M\}$. Then, we can define the information transfer metric (IT) of the i^{th} output y_i, $i \in \mathcal{I}_m$ as follows:*

$$IT_i(D_{\mathcal{I}_m}) = R_i[\text{MGP}(D_{\mathcal{I}})] - R_i[\text{MGP}(D_{\mathcal{I}_m})],$$

where $\text{MGP}(D_u)$ *is an* MGP *using data D_u, $R_i[\text{MGP}(D_u)] = E\big[\mathcal{L}\big(y_i(\boldsymbol{x}), y_{i,true}(\boldsymbol{x})\big)\big|D_u\big]$ defines the expected risk using some loss function $\mathcal{L}$ and $y_i(\boldsymbol{x})$ denotes the predicted random variable in (2.2). Here the expectation is taken over the data distribution $(\boldsymbol{x}, y_i(\boldsymbol{x})) \sim \mathcal{P}_i$. We say negative transfer occurs for output y_i if $IT_i(D_{\mathcal{I}_m})$ is positive.*

Definition 1 implies that negative transfer happens for output y_i when using $D_{\mathcal{I}}$ leads to worse accuracy compared to using a subset of the data or just an individual GP. Therefore, one can provide a model flexible enough to avoid negative transfer if there exists an MGP such that $\forall\, \boldsymbol{x}_0 \in \mathbb{R}^D$

$$pr(y_i(\boldsymbol{x}_0)|D_{\mathcal{I}}) = pr(y_i(\boldsymbol{x}_0)|D_{\mathcal{I}_m}), \forall\, i \in \mathcal{I}_m \subseteq \mathcal{I} \tag{2.4}$$

One can think of $\mathcal{I}_m$ as the index for the subset of outputs that share commonalities with y_i ($\mathcal{I}_m$ here includes i). For instance, if y_i shares no commonalities with any other output ($\mathcal{I}_m = i$) then the MGP should be able to have $pr(y_i(\boldsymbol{x}_0)|D_{\mathcal{I}}) = pr(y_i(\boldsymbol{x}_0)|D_i)$ for all $\boldsymbol{x}_0$, i.e., the conditional predictive distribution in (2.2) for output i is independent of all other outputs. In other words, we need an MGP that is able to collapse into independent GPs or an MGP with only related outputs.

Building a model that can achieve (2.4) is a challenging task under the non-separable covariance structure (See Remark 4). However, a sufficient condition to achieve (2.4) is to ensure $\forall\, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^D$,

$$\text{cov}^f_{ij}(\boldsymbol{x}, \boldsymbol{x}') = 0, \forall\, i \in \mathcal{I}_m \text{ and } j \in \mathcal{I}_{/\mathcal{I}_m} \tag{2.5}$$

In the following section we study the necessary condition for the MGP to achieve (2.5). Specifically we show that if M is known (we know how many distinct/unrelated subgroups of outputs exist) then the necessary condition to achieve (2.5) is $Q \geq M$. However the fact that M is not known beforehand implies that we need $Q \geq N$.

2.3.2 Conditions to avoid negative transfer

We first provide a lemma based on the CP covariance in (2.3) needed to establish our result.

Lemma 2. *Given two outputs, y_1 and y_2, modeled using one latent function X_1, i.e. $y_i = K_{1i}(\boldsymbol{x}) \star X_1(\boldsymbol{x}) + \epsilon_i(\boldsymbol{x})$ for $i = 1, 2$. Assume, the kernels $K_{1i}(\boldsymbol{x}) \in L^1(\mathbb{R}^D)$, $i \in \{1, 2\}$, satisfy one of the following conditions:*

- *$K_{1i}(\boldsymbol{x}) = \alpha_{1i} k_{1i}(\boldsymbol{x})$, $\alpha_{1i} \in \mathbb{R}$ and $k_{1i}(\boldsymbol{x}) > 0 \, \forall \, \boldsymbol{x} \in \mathbb{R}^{\mathcal{D}}$. Typical cases include squared exponential, Matern, quadratic kernel, periodic and local periodic.*
- *k_{1i} has the form $\sum_u a_u^2 exp\big(\boldsymbol{x}^T \boldsymbol{B}_u \boldsymbol{x}\big) cos(2\pi {\boldsymbol{c}_u}^T \boldsymbol{x})$ with parameters $(a_u, \boldsymbol{B}_u, \boldsymbol{c}_u)$. Typical cases include the Spectral, generalized spectral, MOCSM Chen et al. (2019), CSM Ulrich et al. (2015) and SMD Chen et al. (2018) kernels.*

Then for $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^D$, $cov_{12}^f(\boldsymbol{x}, \boldsymbol{x}') = \int_{-\infty}^{\infty} K_{11}(\boldsymbol{u}) K_{12}(\boldsymbol{u} - \boldsymbol{d}) d\boldsymbol{u} = 0$ if and only if at least one of K_{11} and K_{12} is identically equal to zero.

The technical details for Lemma 22 are given in Appendix A. Clearly when using one latent function X_1 if one of the kernels is identically zero then the MGP is invalid ($\mathrm{cov}_{uu}^f(\boldsymbol{x}, \boldsymbol{x}') = 0 \, \forall \, \boldsymbol{x}, u \in \{1, 2\}$). On the other hand, if we use $Q \geq 2$ latent functions, then the model has enough flexibility to construct f_i from different latent functions, i.e. $f_1 = K_1 \star X_1$ and $f_2 = K_2 \star X_2$. In this case, $\mathrm{cov}_{12}^f = 0, \, \forall \, \boldsymbol{x}, \boldsymbol{x}'$. Hence, Lemma 22 implies that only if $Q \geq 2$ we can achieve $\mathrm{cov}_{12}^f(\boldsymbol{x}, \boldsymbol{x}') = 0 \, \forall \, \boldsymbol{x}, \boldsymbol{x}'$. In Lemma 22 the assumption that kernels belong to the L^1 space is also needed for a stable CP construction in (2.1). Here we note that despite the fact that the conditions presented satisfy most (if not all) of the kernels currently used in the CP, in the appendix we also provide some simple means based on the injectivty of the Fourier transform to check the conditions. We now give the main theorem for the necessary condition to avoid negative transfer.

Theorem 3. *Given $K_{qi}(\boldsymbol{x}) \in L^1(\mathbb{R}^D)$ that satisfies the conditions in Lemma 22, and*

$$y_i(\boldsymbol{x}) = f_i(\boldsymbol{x}) + \epsilon_i(\boldsymbol{x}) = \sum_{q=1}^{Q} K_{qi}(\boldsymbol{x}) \star X_q(\boldsymbol{x}) + \epsilon_i(\boldsymbol{x}), \tag{2.6}$$

then there exists an MGP, constructed using a CP, that can achieve $cov_{ij}^f(\boldsymbol{x}, \boldsymbol{x}') = 0, \forall \, i \in \mathcal{I}_m$ and $j \in \mathcal{I}_{/\mathcal{I}_m}$ if and only if we have $Q \geq M$ latent functions.

Proof. We use an induction argument to establish the proof.

In Lemma 1, we have shown that if we model two outputs using one latent function, then $\mathrm{cov}_{12}^{f}(\boldsymbol{x}, \boldsymbol{x}') \neq 0$ for any $\boldsymbol{x}, \boldsymbol{x'} \in \mathbb{R}^d$. On the other hand, if we use $Q(Q \geq 2)$ latent functions, then the model has enough flexibility to construct f_i, $i = 1, 2$ from different latent functions, i.e. $f_1 = K_1 \star X_1$ and $f_2 = K_2 \star X_2$, where $X_1 \neq X_2$. In this case, $\mathrm{cov}_{12}^{f}(\boldsymbol{x}, \boldsymbol{x}') = 0$ for any $\boldsymbol{x}$ and $\boldsymbol{x'}$. i.e. we have proved that when $M = 2$, the model could achieve

$$\mathrm{cov}_{ij}^{f}(\boldsymbol{x}, \boldsymbol{x}') = 0, \forall\, i \in \mathcal{I}_m \text{ and } j \in \mathcal{I}_{/\mathcal{I}_m}, m = 1, 2$$

if and only if the number of latent function $Q \geq 2$.

Then we use induction: Assume the conclusion holds for $M = K - 1$: Consider a MGP with N outputs $y_1, y_2 \cdots, y_N$. Let the index set of all outputs $\mathcal{I}$ comprise of $K - 1$ non-empty disjoint subsets $\mathcal{I} = \{\mathcal{I}_1, \cdots, \mathcal{I}_{K-1}\}$. If for any $i, j \in \{1, 2, \cdots, K - 1\}$ and $\forall\ \boldsymbol{x}, \boldsymbol{x'} \in \mathbb{R}^D$, $\mathrm{cov}(f_{i_s}(\boldsymbol{x}), f_{j_t}(\boldsymbol{x'})) = 0$, then we need at least $K - 1$ latent functions, where $i_s \in \mathcal{I}_i$ and $j_t \in \mathcal{I}_j$. Now consider $M = K$. We could separate this problem into two steps: first, we want $K - 1$ disjoint subsets of $y_1, \cdots, y_N$ to be uncorrelated. Denote the index of these $K - 1$ subsets as $\mathcal{I}_1, \mathcal{I}_2 \cdots, \mathcal{I}_{K-1}$. Follow the assumption in the induction, we at least need $K - 1$ latent functions $\{X_1, X_2, \cdots, X_{K-1}\}$, i.e. any output f_{i_s} in $\mathcal{I}_i$, $i = 1, 2, \cdots, K - 1$ is constructed from the convolution of X_i and a smooth kernel: $f_{is} = K_{i_s} \star X_i$, $i = 1, 2 \cdots, K - 1$. Then, we want the outputs with index $\mathcal{I}_K = \mathcal{I} \backslash \{\mathcal{I}_1 \cup \mathcal{I}_2 \cdots \cup \mathcal{I}_{K-1}\}$ to be uncorrelated with the outputs in the previous $K - 1$ subsets. If we still use $K - 1$ latent functions, then the outputs y_i, $i \in \mathcal{I}_K$ has to be constructed using the latent functions in $\{X_1, X_2, \cdots, X_{K-1}\}$. Then, similar to the case when we have 2 outputs, there must exist a subset $\mathcal{I}_{i_0}$, $i_0 \in \{1, 2, \cdots, K - 1\}$, such that y_i, $i \in \mathcal{I}_{i_0}$ has non-zero covariance function with the outputs in $\mathcal{I}_K$, i.e. these two subsets are correlated. On the other hand, if we use K latent functions, then the model has capability to construct the outputs in $\mathcal{I}_i$, $i = 1, 2, \cdots, K$ from different latent functions, i.e. any output f_{i_s}, $i_s \in \mathcal{I}_i$ can be constructed as $f_{i_s} = K_{i_s} \star X_i$, $i = 1, 2, \cdots, K$. In this case, $\mathrm{cov}(f_{i_s}(\boldsymbol{x}), f_{j_t}(\boldsymbol{x'})) = 0$ for any $\boldsymbol{x}$ and $\boldsymbol{x'}$, where f_{i_s} and f_{j_t} are respectively arbitrary outputs with index in $\mathcal{I}_i$ and $\mathcal{I}_j$.

Therefore, we have proved that when $\mathcal{I} = \{\mathcal{I}_1 \cup \mathcal{I}_2 \cdots \cup \mathcal{I}_M\}$, where $\mathcal{I}_1, \cdots, \mathcal{I}_M$ are M non-empty disjoint subset of $\mathcal{I}$, $\mathrm{cov}(f_i(\boldsymbol{x}), f_j(\boldsymbol{x'})) = 0$ for any $\boldsymbol{x}$ and $\boldsymbol{x'}$ if and only if the number of latent functions $Q \geq M$, where $i \in \mathcal{I}_m$ and $j \in \mathcal{I}/\mathcal{I}_m$. That is to say, the model could achieve

$$\mathrm{cov}_{ij}^{f}(\boldsymbol{x}, \boldsymbol{x}') = 0, \forall\, i \in \mathcal{I}_m \text{ and } j \in \mathcal{I}_{/\mathcal{I}_m}$$

if and only if $Q \geq M$.

□

It is crucial to note here that in reality we do not know M, i.e., we do not know how many distinct subgroups that are uncorrelated exist. Thus, in order to guarantee that negative transfer can be avoided we need $Q \geq N$. This also implies that the model is flexible enough to collapse to N independent GPs and hence predict each output independently. We note that $Q = N$ only gives the model enough flexibility to collapse into N independent $\mathrm{GP}s$ when necessary, while the model can still borrow strength from other outputs to improve the prediction, i.e., the model can choose whether to have information sharing or not by itself.

Remark 4. *Equation (2.5) is a sufficient condition for independent predictions (2.4). However, for non-separable constructions, achieving (2.4) without (2.5) is a very challenging task. Basically, if $\Omega = C_{\boldsymbol{f},\boldsymbol{f}}^{-1}$, then we need (see Uhler (2017))*

$$\Omega_{c_i,c_j} = 0 \quad \textit{IFF} \quad \det\left(C_{[P]\backslash c_i,[P]\backslash c_j}\right) = 0 \quad \forall\, i \in \mathcal{I}_m,\; j \in \mathcal{I}_{/\mathcal{I}_m}$$

for independent predictions (2.4). Here we use c_i to denote the input entry for data from output i. Designing a kernel that can achieve this is very challenging. This points to an interesting research direction: defining kernels specifically intended to minimize negative transfer without the need for $Q \geq M$ which is necessary for (2.5) in the convolution construction. Additive kernels that cancel out the covariance can be an interesting structure to investigate.

Remark 5. *It is no surprise that when $Q \geq M$, negative transfer can be avoided. Basically, each correlated group (or individual output) has flexibility to be modeled via a separate latent function. The interesting results are necessity. Under most mature kernels used for* MGP*'s, negative transfer cannot be avoided without $Q \geq M$.*

2.3.3 Induced Challenges

Despite the many works in the previous decade on reducing the computational complexity of both the GP and MGP, the results in Sec.2.3.2 induce another key challenge for MGP: the high dimensional parameter space. This challenge is inherited from the CP construction which provides different covariance parameters (via the kernels) to different outputs levels. For instance, assume any kernel $K_{qi}(\boldsymbol{x})$ has ω parameters to be estimated then using the CP, this implies estimating $QN\omega + N$ parameters, where the added N parameters are for $\epsilon_i(\boldsymbol{x})$. Following our results, a model that can avoid negative transfer thus needs at least $N^2\omega + N$. Note here that ω also increases with $\mathcal{D}$, i.e., the dimension of $\boldsymbol{x}$. Obtaining good estimates in such a high dimensional space is an impractical task specifically under a non-convex and highly nonlinear objective, be it the exact Gaussian likelihood or its variational bound. Indeed, it is crucial to note that computational complexity and parameter space are two separate challenges and the many papers that tackle the former still suffer from the latter challenge. We conjecture that for this reason, most (if not all) MGP literature (including all aforementioned cited papers) have used $1 \leq Q \leq 4$. To address this challenge, in Sec. 2.4 we provide relaxation models that can significantly reduce the parameters space and scale to arbitrarily large datasets by parallelization. Further, our proposed models allow any sparse approximation to be plugged in. This in turns allows utilization of the many advances in reducing the computational complexity (inducing point, state space approximation, etc..).

2.4 Relaxation models

We investigate two relaxation models: the arrowhead and pairwise models. Without loss of generality, we focus on predicting y_1 using the other $N-1$ outputs. We use $\mathcal{I}_{/1}$ to index all outputs except y_1.

2.4.1 Arrowhead Model

The idea of an arrowhead model originates from the arrowhead matrix. While still using N latent functions, we can assume all outputs y_i, $i \in \mathcal{I}_{/1}$, are independent and only share information with y_1, the output of interest. This implies $\mathrm{cov}^f_{ij}(\boldsymbol{x}, \boldsymbol{x}') = 0, \forall\, i, j \in \mathcal{I}_{/1}$. The structure and covariance matrix are highlighted in Fig. 2.1(a) and (2.7) respectively. As shown in the figure, y_1 possesses unique features encoded in X_1 and shared features with other outputs encoded in $X_i, i \in \mathcal{I}_{/1}$.

$$\boldsymbol{C}^{P\times P}_{\boldsymbol{f},\boldsymbol{f}} = \left(\begin{array}{c|ccc} \boldsymbol{C}_{\boldsymbol{f}_1,\boldsymbol{f}_1} & \boldsymbol{C}_{\boldsymbol{f}_1,\boldsymbol{f}_2} & \cdots & \boldsymbol{C}_{\boldsymbol{f}_1,\boldsymbol{f}_N} \\ \hline \boldsymbol{C}_{\boldsymbol{f}_2,\boldsymbol{f}_1} & \boldsymbol{C}_{\boldsymbol{f}_2,\boldsymbol{f}_2} & \cdots & \boldsymbol{0}_{p_2\times p_N} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}_{\boldsymbol{f}_N,\boldsymbol{f}_1} & \boldsymbol{0}_{p_N\times p_2} & \cdots & \boldsymbol{C}_{\boldsymbol{f}_N,\boldsymbol{f}_N} \end{array}\right) \tag{2.7}$$

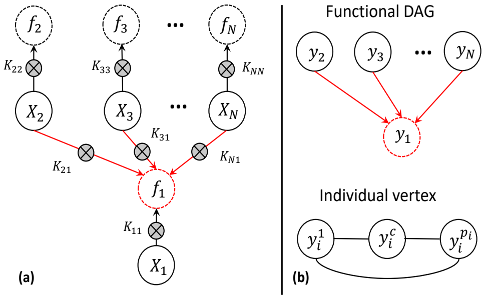


Figure 2.1: Arrowhead Model

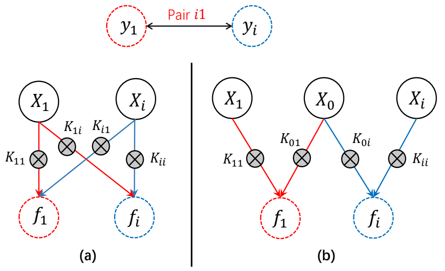


Figure 2.2: Pairwise Model

The arrowhead structure in fact poses many unique advantages: (1) Linear increase of parameter space dimension with N: The number of parameters to be estimated is reduced to $(2N-1)\omega + N$. (2) provides enough flexibility to achieve (2.5) and hence avoid negative transfer. For instance, if $\text{cov}^f_{1i}(\boldsymbol{x}, \boldsymbol{x}') = 0 \, \forall \boldsymbol{x}$ and $i \in \mathcal{I}_{/1}$ then outputs y_1 and y_i are predicted independently. (3) can be parallelized, where independent arrowhead models are build to predict each output. (4) One nice interpretation of the arrowhead structure is through a Gaussian directed acyclic graphical model (DAG) with vertices $V = \{\boldsymbol{y}_i : i \in I\}$. Unlike typical DAGs, each vertex in this graph is in itself a fully connected undirected Gaussian graphical model, i.e. a functional response. This is shown in Fig.2.1(b). Based on this, the full likelihood factorizes over parent nodes. To see this, let $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$ denote the likelihood of the dataset, where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_f^\top, \boldsymbol{\sigma}^\top\}^\top$, such that $\boldsymbol{\theta}_f$ and $\boldsymbol{\sigma}$ are kernel and noise parameters. Then, $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \mathcal{L}^{1|i \in \mathcal{I}_{/1}}(\boldsymbol{\theta}; \boldsymbol{y}_1 | \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N) \prod_{i \in \mathcal{I}_{/1}} \mathcal{L}^i(\boldsymbol{\theta}; \boldsymbol{y}_i)$. This reduces the complexity of exact inference to $O(Np^3)$ assuming $p_i = p \, \forall i$, i.e, complexity of N independent GPs. This complexity is similar to the well known inducing point sparse approximation in Alvarez and Lawrence (2009), however without the assumption of conditional independence given discrete observations from the latent functions. Despite reduced complexity, the main advantage is the reduction in the parameter space. Here it is crucial to note that any sparse approximation, be it an inducing point/variational approximation, a state space approximation, a matrix tapering approach or just a faster matrix inversion/determinant calculation scheme, can be plugged in into this structure.

Remark 6. *To predict all N outputs the arrowhead model should be learned N times. Yet, it is critical to highlight that a key benefit of the arrowhead model is optimization in a much lower dimensional space. In a full* MGP *the number of parameters to be estimated is $O(N^2)$. However, in the arrowhead model then the number of parameters to be estimated reduces to $O(N)$ when predicting one output. For predicting all N outputs, this will rise up to $O(N^2)$. In this case, although the parameters to be estimated are both $O(N^2)$, the optimizer in full MGP and arrowhead approach operate in different parameter space dimensions. In the* MGP *model, we optimize over $O(N^2)$ parameters together, which may lead to a highly complex objective function. Instead, in the arrowhead model, we parallel run N models where each model only has $O(N)$ parameters. This is the key reason why in our numerical studies to follow, the arrowhead approach is much faster to fit and converge compared to the full* MGP *model even when predicting all N output.*

2.4.2 Pairwise Model

Despite the linear increase in parameter space in the arrowhead model, when N is extremely large, model estimation can still be prohibitive. To this end, we investigate distributing the MGP into a group of bivariate GPs which are independently built. Predictions are then obtained through combining predictions from each bivariate GP. We here note that pairwise modeling of longitudinal data is not new and was first investigated by Fieuws and Verbeke (2006) in the context of linear models. Pairwise models where then used to scale GPs while allowing qualitative features to be added Li et al. (2018), Kontar et al. (2020). However, the ability of pairwise models to provide $Q \geq M$ while reducing complexity has not been highlighted and poses great value beyond scalability. As previously mentioned we focus on predicting output 1 through borrowing strengths from the other $N-1$ outputs. Fig. 2.2(a) illustrates the pairwise submodel between y_1 and $y_i, i \in \mathcal{I}_{/1}$, where two latent functions are used to avoid negative transfer. Note here that the structure in Fig. 2.2(b) is proposed for efficient regularization and is discussed later in Sec. 2.4.3. The key advantage of the pairwise structure is that: (i) it can scale to an arbitrarily large N by parallelization where each submodel is estimated with a limited number of parameters $(4\omega + 2)$ and with complexity of $O(2p^3)$ (assuming exact inference with no approximations and $p_i = p \, \forall i$). (ii) It can avoid negative transfer as each sub-model satisfies $Q \geq M$. (iii) After building the $N-1$ sub-models, combining predictions boils down to combining $N-1$ predictive distributions $pr(y_1(\boldsymbol{x}_0)|D_1, D_i)_{i \in \mathcal{I}_{/1}}$ in (2.2). This can be readily done using the rich literature on product of experts (PoE) and Bayesian committee machines Deisenroth and Ng (2015), Moore and Russell (2015), Tresp (2000). For instance, in the PoE model, each expert is weighted by the inverse covariance, therefore experts which are uncertain about their predictions are automatically weighted less than experts that are certain about their predictions. As a result, we obtain the final prediction as follows,

$$\hat{y}_1(\boldsymbol{x}_0) = \frac{\sum\limits_{i=1}^{N-1} \hat{y}_{1i}(\boldsymbol{x}_0)/\mathcal{V}_i(\boldsymbol{x}_0)}{\sum\limits_{i=1}^{N-1} 1/\mathcal{V}_i(\boldsymbol{x}_0)} \tag{2.8}$$

where $\hat{y}_{1i}(\boldsymbol{x}_0)$ represents the mean prediction of $y_1(\boldsymbol{x}_0)$ in the i_{th} bivariate submodel for $i = 1, 2, \cdots, N-1$ and $\mathcal{V}_i(\boldsymbol{x}_0)$ is the corresponding variance (see 2.2). The key idea across such approaches, in our context, is that sub-models that are uncertain about their predictions of $y_1(\boldsymbol{x}_0)$ (i.e., have larger predictive variance) will get less weight.

Remark 7. *Our two proposed models share a common theme which aims at guaranteeing a sparse precision matrix from a sparse covariance. This is the key of* GP*s as they model the covariance yet predictions are based on the precision matrix. For instance, assume that y_i and y_j should be predicted independently, then in the arrowhead matrix, the model allows the inverse for* ***any*** *input/output pair to be zero if their covariance $\boldsymbol{C}_{\boldsymbol{f}_i, \boldsymbol{f}_j}$ is zero. However, in a regular* MGP *they will still be able to talk to each other via other outputs (i.e. the precision matrix will not be sparse). Similarly for the pairwise model a zero covariance between the outputs will directly lead to zero precision.*

Remark 8. *The arrowhead model is designed for predicting y_i using $y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N$.*

If we need to get the predictions of $y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N$, we need to parallel run separate arrrowhead models. That being said, the arrowhead model enjoys multiple key properties, (1) it provides enough flexibility to achieve Eq. (3.2) and hence avoid negative transfer, (2) it leads only to a linear increase in the parameter space with N, (3) it reduces the complexity of exact inference to $O(Np^3)$ assuming $p_i = p\ \forall i$, i.e, complexity of N independent GP*s. However, the pairwise model is readily applied to predicting N outputs. The only thing we need to do is to parallel run all possible bivariate* MGP *models and then directly get the predictions for any output by combining the predictions using product of experts (PoE). Both pairwise and arrowhead models require a fraction of time to that of the* MGP *with N latent functions.*

2.4.3 Encouraging sparsity via regularization

Besides the fact that these two models can avoid negative transfer, an interesting feature is that we can add regularization that helps reduce negative transfer and is capable of automatic variable selection. Here variable selection implies selection of which functions should be predicted independently or not. Let $\ell(\boldsymbol{\theta}; \boldsymbol{y}) = -\log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{1}{2}\langle \boldsymbol{Y}, (\boldsymbol{C_{f,f}} + \boldsymbol{\Sigma})^{-1}\rangle + \frac{1}{2}\log|\boldsymbol{C_{f,f}} + \boldsymbol{\Sigma}|$ where $\langle \boldsymbol{A}, \boldsymbol{A}'\rangle = \mathrm{trace}(\boldsymbol{A}\boldsymbol{A}')$ and $\boldsymbol{Y} = \boldsymbol{y}\boldsymbol{y}^\top$. A penalized version of $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ is defined as

$$\ell_{\mathbb{P}}(\boldsymbol{\theta}; \boldsymbol{y}, \lambda) = \ell(\boldsymbol{\theta}; \boldsymbol{y}) + \mathbb{P}_\lambda(\boldsymbol{\theta}_0)\ , \tag{2.9}$$

where $\mathbb{P}_\lambda(|\boldsymbol{\theta}_0|)$ is a penalty function and $\boldsymbol{\theta}_0 \subseteq \boldsymbol{\theta}$. Possible well-known choices include the ridge penalty $\mathbb{P}_\lambda(|\theta_i|) = \lambda\theta_i^2$, L^1 penalty $\mathbb{P}_\lambda(|\theta_i|) = \lambda|\theta_i|$, bridge penalty $\mathbb{P}_\lambda(|\theta_i|) = \lambda|\theta_i|^{0<\cdot<1}$, and SCAD penalty which includes two tuning parameters (λ and γ) $\mathbb{P}_\lambda(|\theta_i|) = \lambda|\theta_i|$ if $|\theta_i| \leq \lambda$, $(\theta_i^2 - 2\gamma\lambda|\theta_i| + \lambda^2)/(2\gamma - 2)$ if $\lambda < |\theta_i| \leq \gamma\lambda$, $\lambda^2(\gamma+1)/2$ if $|\theta_i| > \gamma\lambda$ Fan and Li (2001). The tuning parameters can be estimated using cross validation Friedman et al. (2001).

In the arrowhead model, one can directly observe that for $K_{qi}(\boldsymbol{x}) = \alpha_{qi}k_{qi}(\boldsymbol{x})$, then $\boldsymbol{\theta}_0 = \{\alpha_{q1}\}_{q=2}^N$. Therefore when $\alpha_{q1} \to 0 \implies \mathrm{cov}_{1i}^f(\boldsymbol{x}, \boldsymbol{x}') = \int_{-\infty}^{\infty} K_{i1}(\boldsymbol{u})K_{ii}(\boldsymbol{u} - \boldsymbol{d})d\boldsymbol{u} \to 0$ and hence outputs y_1 and $y_{q=i}$ will be predicted independently. Thus any shrinkage penalty will encourage the arrowhead model to limit information sharing across unrelated output. Another advantage besides automatic shrinkage is *functional variable selection* where the sparse elements in $\{\alpha_{q1}\}_{q=1}^N$ would identify which outputs are related to y_1.

Similar to the arrowhead model, the pairwise approach also facilitates regularization and automatic variable selection. For the structure illustrated in Fig. 2.2(a), we have $\mathrm{cov}_{1i}^f(\boldsymbol{x}, \boldsymbol{x}') = \int_{-\infty}^{\infty} K_{11}(\boldsymbol{u})K_{1i}(\boldsymbol{u} - \boldsymbol{d})d\boldsymbol{u} + \int_{-\infty}^{\infty} K_{i1}(\boldsymbol{u})K_{ii}(\boldsymbol{u} - \boldsymbol{d})d\boldsymbol{u}$. Therefore to encourage sparsity, a group penalty $\mathbb{P}_\lambda^G$ on K_{i1} and K_{1i} is needed.

$$\ell_{\mathbb{P}}(\boldsymbol{\theta}_{1i}; \boldsymbol{y}_1, \boldsymbol{y}_i, \lambda) = \ell(\boldsymbol{\theta}_{1i}; \boldsymbol{y}_1, \boldsymbol{y}_i) + \mathbb{P}_\lambda^G(\alpha_{1i}, \alpha_{i1}). \tag{2.10}$$

One well-known option for $\mathbb{P}_\lambda^G$ is the group Lasso $\mathbb{P}_\lambda^G = \sqrt{2}\lambda||(\alpha_{1i}, \alpha_{i1})^T||_2$. Alternatively, one can utilize the structure in Fig. 2.2(b) and instead of penalizing the kernels, one can regularize the shared latent function X_0. For instance, one can augment the covariance of X_0 with a parameter α_0 such that $\mathrm{cov}(X_0(\boldsymbol{u}), X_0(\boldsymbol{u}')) = \alpha_0\delta(\boldsymbol{u} - \boldsymbol{u}')$. Then, $\ell_{\mathbb{P}}(\boldsymbol{\theta}_{1i}; \boldsymbol{y}_1, \boldsymbol{y}_i, \lambda) = \ell(\boldsymbol{\theta}_{1i}; \boldsymbol{y}_1, \boldsymbol{y}_i) + \mathbb{P}_\lambda(\alpha_0)$. It can be directly verified that as $\alpha_0 \to 0$ outputs y_1 and y_i are predicted independently.

Remark 9. *In Section 2.5, we demonstrate the empirical benefits of the relaxation models and having $Q \geq M$ latent functions. However, it is important to note that having sufficient latent functions does not automatically imply that the* MGP *will perform well. It only says that the model is flexible enough (spans the hypothesis space) to avoid negative transfer. Model estimation still may suffer from getting stuck at critical points with bad generalization. Unfortunately, clearly understanding the predictive capability of our model, albeit the effect of Q, is a tough question. It requires tight generalization bounds, possibly as a function of Q. At this stage, only few papers were able to get generalization bounds in* GP*s particularly for simple kernels Wang et al. (2020). Also these bounds turn out to be loose in moderate to high dimensions. For these reasons, understanding how Q affects prediction is a challenge. Hopefully a big crack to this problem will happen in either deep learning (great progress is happening here - see Dziugaite and Roy (2017)) or kernel methods soon, and hence opening up this challenge.*

2.5 Simulations and Case Studies

Since negative transfer is a subject yet to be explored in MGP, *we dedicate most of this section towards a proof of concept* for the: (1) impact of negative transfer, (2) need for sufficient latent functions as shown in theorem 3, (3) advantageous properties of the proposed relaxations.

2.5.1 Illustration of Negative Transfer

2.5.1.1 Convolved Squared exponential Kernel

In this setting, we aim to illustrate theorem 3 using the well-known convolved squared exponential kernel in Álvarez and Lawrence (2011). We generate outputs y_1, y_2 and y_3 from

$$\begin{aligned} y_1(x) &= 5 \cdot \sin(3x/2) + 3 + \epsilon_1(x) \\ y_2(x) &= 5 \cdot \sin(x) - 3 + \epsilon_2(x) \\ y_3(x) &= x^2/10 - 5 + \epsilon_3(x) \end{aligned}$$

where $x \in \mathbb{R}$ is evenly spaced in $[0, 10]$, $p_1 = p_2 = p_3 = 20$ and $\sigma_1 = \sigma_2 = \sigma_3 = 0.05$. All parameters are estimated by optimizing the log-likelihood function, i.e. $\ell(\boldsymbol{\theta}; \boldsymbol{y}) = -\log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) =$

$\frac{1}{2}\langle \boldsymbol{Y}, (\boldsymbol{C}_{\boldsymbol{f},\boldsymbol{f}} + \boldsymbol{\Sigma})^{-1}\rangle + \frac{1}{2}\log|\boldsymbol{C}_{\boldsymbol{f},\boldsymbol{f}} + \boldsymbol{\Sigma}|$. In Table 2.1 we report the means squared error (MSE), averaged over the 3 outputs, on $p = 70$ uniformly spaced points in $[0, 10]$ when $Q = 1, 2, 3$ and 4. Table 2.1 provides many interesting insights. Indeed from the function specifications, it is clear that they have very different shape and length scales (i.e., frequency and amplitude). As a result, when using one or two latent functions negative transfer leads to large predictive errors. It is also noticeable that the result of using $Q = 4$ does not have much difference with that $Q = 3$. This confirms our theorem which implies that with at least N latent functions an MGP is capable of avoiding negative transfer.

Table 2.1: Predictive error with varying Q

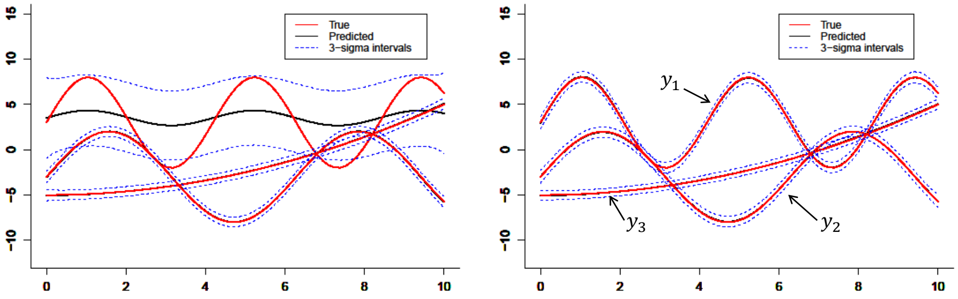
Q	1	2	3	4
MSE	25.183	11.464	0.00159	0.00157

2.5.1.2 Spectral Kernel

The immediate follow up question is what if we use the recently proposed, more flexible class of spectral kernels. The aim is to illustrate that as shown in Lemma 22, avoiding negative transfer is mainly independent of what kind of kernel we use, i.e. even if we use a more flexible kernel. We use the same data with that in setting 2.5.1.1. The covariance function is given as:

$$\mathrm{cov}_{ij}^{f}(x, x') = \sum_{q=1}^{Q} \frac{a_{qi}a_{qj}}{2}\sqrt{\frac{\pi}{\sigma_{qi}^2 + \sigma_{qj}^2}}H(d)$$

where $H(d) = e^{A_1(d)}\cos(\theta_1 d) + e^{A_2(d)}\cos(\theta_2 d)$. Formulation of $A_i(d)$ and θ_i, $i = 1, 2$ are given in Appendix C. This covariance is the result of a convolution across spectral kernels.



(a) one latent function

(b) three latent functions

Figure 2.3: Illustration of predictions using a spectral kernel

The predictive results for the three outputs are illustrated in Fig. 2.3. The results confirm that even with a flexible kernel, negative transfer will detrimentally affect model performance without enough latent functions. Indeed in Fig. 2.3(a) one can observe that $y_1(x)$ and $y_3(x)$ have larger length scales and hence are smoother. As a results when $Q = 1$, output $y_1(x)$ is forced to have a larger length scale in lieu of the two other outputs. This however, can be avoided with a sufficient number of latent functions as shown in Fig. 2.3(b).

2.5.2 Role of Regularization

Still under the setting of Sec 2.5.1.1, we try to verify theorem 3 and the impact of $\mathbb{P}_\lambda$ on automatic selection of related output. We use the pairwise model described in Fig. 2.2(b) where two bivariate submodels are used to predict y_1: (y_1, y_2) and (y_1, y_3). The covariance function of y_1 and y_i ($i = 2, 3$) constructed using Fig. 2.2(b) are given as

$$\begin{aligned}
\mathrm{cov}^f_{11}(d) &= \alpha^2_{11} \exp\{-\frac{d^2}{4 \cdot l^2_{11}}\} + \alpha^2_{01} \exp\{-\frac{d^2}{4 \cdot l^2_{01}}\} + \sigma^2 \\
\mathrm{cov}^f_{ii}(d) &= \alpha^2_{ii} \exp\{-\frac{d^2}{4 \cdot l^2_{ii}}\} + \alpha^2_{0i} \exp\{-\frac{d^2}{4 \cdot l^2_{0i}}\} + \sigma^2 \\
\mathrm{cov}^f_{1i}(d) &= \alpha_{01}\alpha_{0i}\sqrt{\frac{2|l_{01}l_{0i}|}{l^2_{01} + l^2_{0i}}} \exp\{-\frac{1}{2}\frac{(d-\mu)^2}{l^2_{01} + l^2_{0i}}\}
\end{aligned} \tag{2.11}$$

We applied the pairwise model with a regularization term respectively to the data. For the penalty we use $\mathbb{P}_\lambda(\boldsymbol{\alpha}_0) = \lambda|\alpha_{01} \cdot \alpha_{0i}|$ where $\boldsymbol{\alpha}_0 = (\alpha_{01}, \alpha_{0i})^T$ and $i = 2, 3$. Table 2.2 shows the estimated parameters. Note that here we use the structure described in Fig. 2.2(b), hence as long as one of α_{01} and α_{0i} is penalized to be 0, then the negative transfer between y_1 and y_i could be avoided, for $i = 2, 3$.

Table 2.2: Estimated parameters for the regularized pairwise model

pair	α_{01}	α_{0i}	l_{01}	l_{0i}
(y_1, y_2)	**8.27 e-7**	1.91	6.61	-1.21
(y_1, y_3)	**-3.27 e-6**	0.93	2.37	1.77

One can directly observe from Table 2.2 that when adding regularization on $\boldsymbol{\theta}_0$, α_{i1} is shrunk to nearly zero in both submodels. This implies that $\mathrm{cov}^f_{1i}(d) \approx 0 \; \forall x$ and $i \in \{2, 3\}$ and hence y_1 is predicted independently. This not only confirms that regularization can limit information sharing but also illustrates that in the proposed MGP models, one can automatically perform variable selection (cluster the outputs that ought to be predicted independently). A user might then choose to perform a separate MGP on the selected subsets. To the best of our knowledge this is the first model that can achieve simultaneous estimation and functional selection for non-separable and dependent GPs.

2.5.3 Illustration with Subsets of Correlated Outputs

2.5.3.1 Low Dimensional Setting

We then study the case when subsets of outputs are correlated. We first perform inference in a low dimensional regime to compare with the full MGP that does not face the challenge of large complexity and extremely high dimensional parameter space. We generate outputs $y_i^{(j)}(x) = f_i^{(j)}(x) + \epsilon_i^{(j)}(x)$ from $f_i^{(1)}(x) = x^2/(0.8 \cdot (1-x))$ for $i \in \{1,2\}$; $f_i^{(2)}(x) = x/(1-x)$ for $i \in \{3,4\}$; $f_i^{(3)}(x) = 2 \cdot x^2$ for $i \in \{5,6\}$; $f_i^{(4)}(x) = x^3$ for $i \in \{7,8\}$.

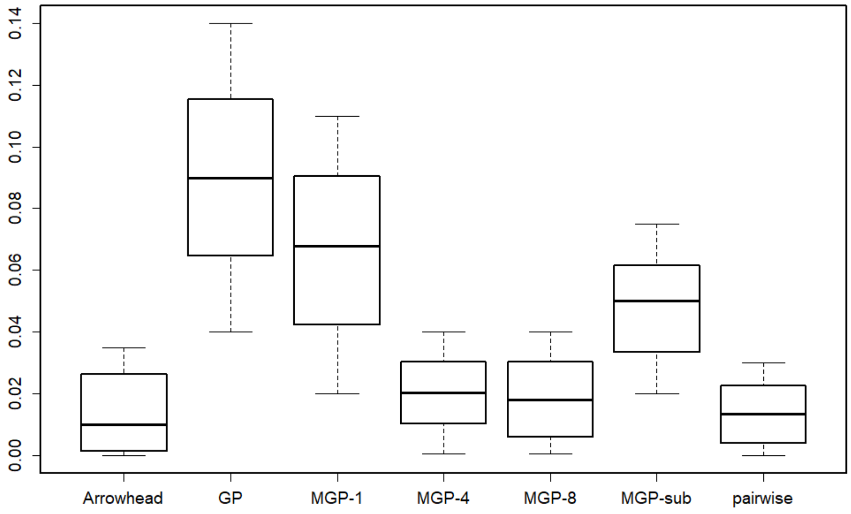


Figure 2.4: Prediction comparison for $y_1(x)$

Here we focus on predicting output y_1, using the following models: (1) $\mathrm{MGP}-Q$ where $Q = 1, 4$ and 8 respectively, (2) Pairwise model where predictions are combined using the robust product of experts in Deisenroth and Ng (2015), (3) Arrowhead model, (4) A univariate GP on y_1, (5) A bivariate GP with outputs y_1 and y_2 (i.e. outputs in $y_i^{(1)}(x)$) denoted as $\mathrm{MGP}-sub$.

The main difference between $\mathrm{MGP}-8$ and GP lies in whether there is information sharing among the outputs or not. GP represent the result when the model is only trained using the data from y_1. However, $\mathrm{MGP}-8$ represents the result when we model $y_1, y_2, \cdots, y_8$ together as a MGP model using $Q = 8$ latent functions.

We use $p = 7$, $\sigma_i = 0.1$ for $i = 1, 2$ and $\sigma_i = 0.01$ for $i = 3, 4 \cdots, 8$. The convolved squared exponential kernel in Alvarez and Lawrence (2009) is used. Results for the mean squared error (MSE) over $p = 30$ uniformly spaced points in $[0, 0.8]$ are given in Fig. 2.4 where the experiment is replicated 30 times. Also Tukey's multiple comparison test is done and only significant results are reported in the discussion below. The first result to observe is that $\mathrm{MGP} - sub$ outperformed $\mathrm{MGP} - 1$ which confirms that negative transfer occurred since when outputs from $y_i^{(1)}$ are analyzed separately they produce better predictive results. However the key observation is that the pairwise and arrowhead models outperform $\mathrm{MGP} - 1$. *This is because, when learning an output from* $y_i^{(1)}$*, both pairwise and arrowhead models can leverage the correlation with other outputs and still avoid negative transfer evidenced through* $\mathrm{MGP} - 1$. Also both proposed latent structures had comparable performance with $\mathrm{MGP} - 8$, which confirms their capability to provide competitive predictive results with lower number of parameters and computational complexity. Another interesting result is that $\mathrm{MGP} - 4$ and $\mathrm{MGP} - 8$ have similar performance. Indeed, this is expected based on theorem 3, where if we have M distinct subsets we only need $Q = M$ to avoid negative transfer. However in reality M is not given in advance; which is why N latent functions are needed. To this point, we highlight the remark below.

Remark 10. *Estimating M beforehand, i.e., defining how to cluster or what aspects of the functions to cluster upon, is very hard. As shown in the simulation above, even if we cluster the functions based on their generating latent function,* MGP*-sub performs worse than our approach. Recall,* MGP*-sub is equivalent to a pre-processing step where only outputs from the same function are modeled together. This is because, outputs from different functions have common knowledge to share.*

For instance, assume we have $sin(\cdot)$ and $cos(\cdot)$ functions. A preprocessing clustering approach will separate those two functions. Yet, an MGP *will see the commonalities in both the length and shape scale of the $sin(\cdot)$ and $cos(\cdot)$ functions and hence borrow strength between them. As such, clustering should be done with the fact that an* MGP *will be used afterward, in mind. Theoretically, clustering should be done based on an* MGP *and then a separate* MGP *is fitted on the clusters. This is what our approach does, yet without doing it in two steps to allow error propagation.*

2.5.3.2 Moderate Dimensional Setting

In this setting we aim to compare our proposed structures when the number of parameters is significantly increased. Specifically $N = 20$ and $N = 50$ outputs are used. For the $N = 20$ setting, outputs are generated from a GP with zero mean and $\mathrm{cov}_{ii}^{y}(x, x') = \alpha_i^2 \exp((x - x')^2/2 \cdot l_i^2) + \sigma_i^2 \delta(x, x')$ under; $\alpha_i = 4, l_i = 1, \sigma_i = 0.005\,\mathrm{for}\, i = 1, \cdots, 5;\ \alpha_i = 1, l_i = 4, \sigma_i = 0.0001\,\mathrm{for}\, i = 6, \cdots, 12;\ \alpha_i = 4, l_i = 1, \sigma_i = 0.001\,\mathrm{for}\, i = 13, \cdots, 20$.

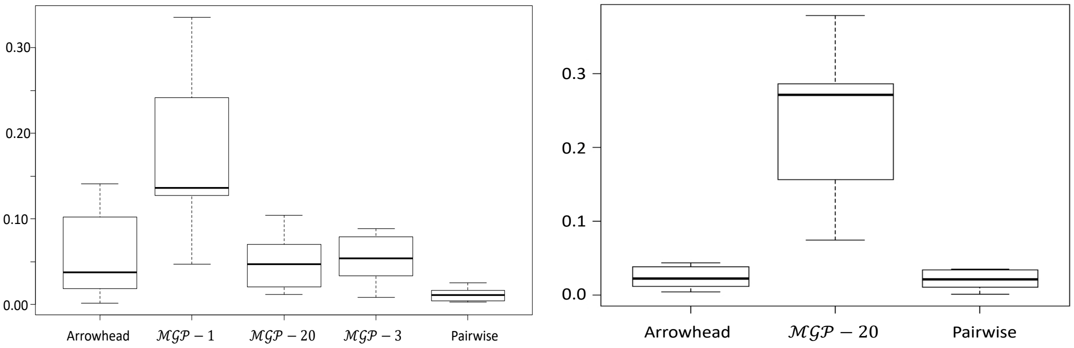
For $N = 50$ setting, we generate 50 outputs from a GP with mean zero and $\mathrm{cov}_{ii}^{y}(x, x') =$

$\alpha_i^2 \exp \frac{(x-x')^2}{2\cdot l_i^2} + \sigma_i^2 \delta\left(x, x'\right)$ under the following setting

$$\begin{aligned}
&\alpha_i = 4, l_i = 1, \sigma_i = 0.001 \quad \text{for} \quad i = 1, 2, \cdots, 9\\
&\alpha_i = 1, l_i = 4, \sigma_i = 0.0001 \quad \text{for} \quad i = 10, \cdots, 19\\
&\alpha_i = 1, l_i = 8, \sigma_i = 0.0001 \quad \text{for} \quad i = 20, \cdots, 29\\
&\alpha_i = 8, l_i = 1, \sigma_i = 0.001 \quad \text{for} \quad i = 30, \cdots, 39\\
&\alpha_i = 3, l_i = 1, \sigma_i = 0.005 \quad \text{for} \quad i = 40, \cdots, 50
\end{aligned}$$

We have 15 points evenly spaced in $[0, 3]$ and we randomly choose 8 points from them as training data and the left 7 points are testing points. For the full MGP model in this setting, we use 20 latent functions to construct the model.

Similar to the setting in Sec. 2.5.3.1 we test on y_1 under 30 replications. The results for $N = 20$ are shown in Fig. 2.5a. From the result we can see that $\mathrm{MGP} - 3$, $\mathrm{MGP} - 20$ and the arrowhead model yield similar results (also confirmed via Tukey's test). This once again confirms that the arrowhead model has competitive performance and that with enough latent functions one can avoid negative transfer. Yet the interesting result is that the pairwise model showed much better performance. This is intuitively understandable as in each pair the number of estimated parameters is very small and thus one can except better estimators compared to the competing models. This fact is further illustrated through the results of $N = 50$ shown in Fig. 2.5b.



(a) Moderately Large Parameter Space

(b) Large Parameter Space

Figure 2.5: Predictive error on moderate and large parameter space

In Fig. 2.5b for the MGP we use $Q = 20$ thus we have $QN\omega + N = 2050$ parameters to estimate. The results show that with $N = 50$ there is a huge decrease in predictive performance. This result is expected as it is extremely challenging to obtain good parameter estimates specifically for a GP likelihood function which is known to be highly non-linear with many local critical points with bad generalization power. Indeed, similar decrease in performance in a high dimensional parameter space has been reported in Li and Zhou (2016) and Li et al. (2018). Here both the arrowhead and pairwise models offer a solution that, not only scales with any N, but also can lead to better performance. We note that on average $\mathrm{MGP}-20$ with $N = 50$ took ≈ 24 hours to estimate despite the computational complexity being relatively small with $P = 400$. While the arrowhead model took ≈ 30 minutes and ≈ 30 seconds for the pairwise model. In practice its very common to have $N >> 50$. This indeed exacerbates the challenge above and further highlights the needs for the proposed relaxation models.

2.5.3.3 Large Dimensional Setting

In this setting we aim to show that the proposed pairwise model can be applied to a very large N by parallelization. We generate 1000 outputs from a GP with mean zero and $\mathrm{cov}_{ii}^{y}\left(x, x'\right) = \alpha_i^2 \exp \frac{(x-x')^2}{2 \cdot l_i^2} + \sigma_i^2 \delta\left(x, x'\right)$ with different values of α_i, l_i and σ_i. Specifically, we generate 1000 outputs from 10 different groups, where $\alpha = \{4, 1, 1, 8, 3, 10, 1, 5, 20, 3\}$, $l = \{1, 4, 8, 1, 2, 1, 2, 4, 1, 3\}$ for each group. We generate 40 points evenly spaced in $[0, 3]$ and we randomly choose 10 points from them for training and the remaining 30 points are for testing. Under this setting, the original MGP model has $QN\omega + N = 30000$ parameters to be estimated. However, the pairwise model can be easily applied in solving this problem by inferring the bivariate GPs in parallel. The following result is based on the structure in Fig. 2.2(a). We replicate the simulation 100 times and show the mean MSE and standard deviation for the prediction of $y_1, \cdots, y_{1000}$ using different penalty methods. We also compare with using independent GPs. The results are shown in Table 2.3. Besides the significant improvement in performance compared to individual modeling, it is critical to note that the pairwise model took less than 10 minutes to obtain the predictive results for all 1000 outputs in a single simulation run.

Table 2.3: Mean MSE and std for the predictions of $y_1, \cdots, y_{1000}$ using different penalties based on the pairwise model and 1000 individual GPs. The result is based on 100 replications.

model (penalty)	pairwise (ridge)	pairwise (bridge)	pairwise (SCAD)	GP
MSE	0.38	0.32	0.33	1.59
std	0.11	0.09	0.08	0.33

2.5.4 Case Studies

2.5.4.1 Exchange Rate Data

We perform a case study on the pacific exchange rate service (`http://fx.sauder.ubc.ca/data.html`). Our goal is to predict the foreign exchange rate compared to the United States dollar currency. This dataset is specifically chosen as research has shown that only subsets of exchange rates are correlated Reboredo et al. (2014). Here we provide MSE and Negative Loglikelihood (NLL) using a two phase cross-validation on two key rates during the 157 weeks (each week is a datapoint) from 2017 to 2020. Once again the results confirm the need for the proposed relaxation models as parameter estimates tend to deteriorate as the parameter dimension increases. Fig. 2.6 provides illustrations of the prediction results of KRW/USD and MXN/USD using pairwise model and arrowhead model.

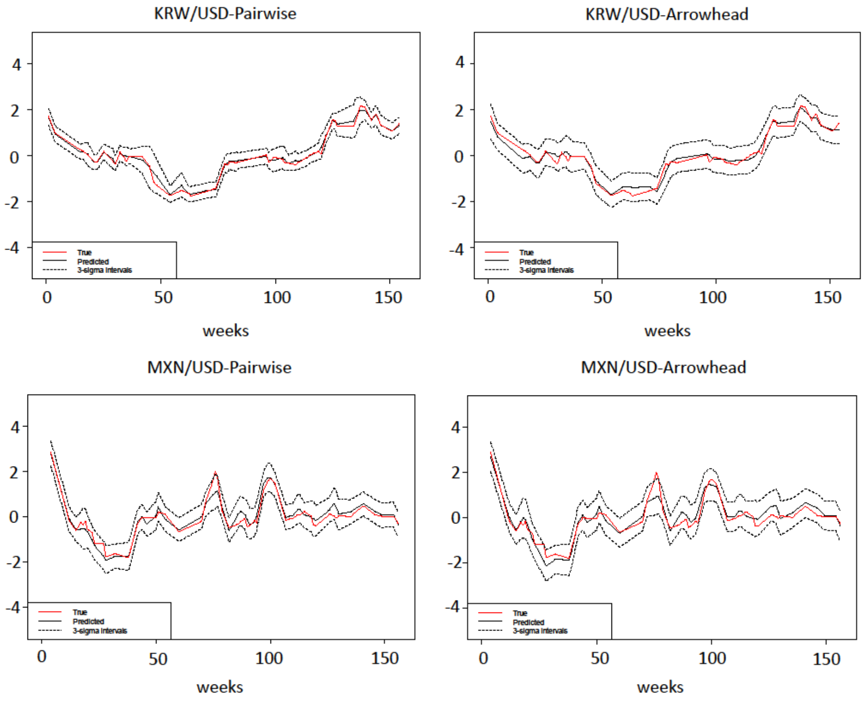


Figure 2.6: illustrations on exchange rate data

Table 2.4: Predictive Error of MXN/USD and KRW/USD

model	pairwise	arrowhead	MGP−10
(MSE, NLL) of MXN/USD	(0.040, 70)	(0.031, 102)	(0.217, 747)
(MSE, NLL) of KRW/USD	(0.015, 53)	(0.035, 66)	(0.322, 719)

2.5.4.2 Parkinson Data

We use the Parkinson data set to predict the disease symptom score (motor UPDRS and total UPDRS) of Parkinson patients at different times. The data set is available on `http://`

archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring. At each time, we randomly choose 10 patients from the data set to model a MGP model with 10 outputs and randomly split 60% data of each patient as training sets and 40% as testing sets. Our goal is to predict the motor UPDRS and total UPDRS of the 10th patient in each round. We run our model for 70 times. Fig. 2.7 shows the predictive error using different models.

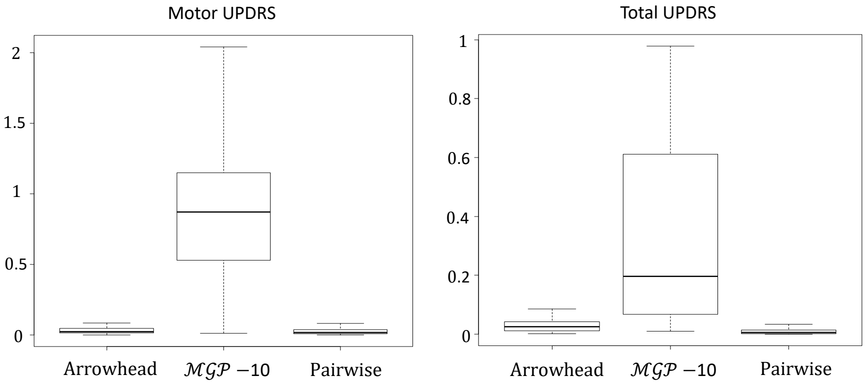


Figure 2.7: Parkinson data set

2.6 Conclusion

This chapter addresses the key challenge of constructing an MGP that can borrow strength across outputs without forcing correlation. We show that this is achieved by having a sufficient number of latent functions regardless of the kernel used. We then propose two latent structures that can avoid negative transfer and maintain estimation in a low-dimensional parameter space. A key feature of our structures is that they allow functional variable selection via regularization. Further analysis into the use of such latent structures and other dependent GP models for selection in functional data settings or probabilistic graphical models can be an interesting topic to explore.

CHAPTER 3

Bayesian Inference of Spatially Varying Correlations via the Thresholded Correlation Gaussian Process

3.1 Introduction

Multimodal neuroimaging is now prevailing in neuroscience research, where different types of brain images are collected for a common set of subjects. Common imaging modalities include anatomic magnetic resonance imaging (MRI), resting-state or task-based functional MRI (fMRI), diffusion tensor imaging (DTI), positron emission tomography (PET), among many others. Multimodal neuroimaging analysis aggregates such diverse but often complementary information, consolidates knowledge across different modalities, and produces improved understanding of neurological development or disorders (Uludağ and Roebroeck 2014). Multimodal analysis is receiving increasing attention in numerous other scientific applications as well, e.g., the multi-omics studies (Richardson et al. 2016).

A central question in multimodal neuroimaging analysis is to understand the association between two imaging modalities and to identify brain regions where such an association is statistically significant. This question is of great scientific interest. For instance, Zhu et al. (2014a) surveyed and showed joint analysis of fMRI and DTI reveals important interplays between brain functions and structures. Cavaliere et al. (2018) showed fMRI and PET together improve the characterization of patients with consciousness disorder. Li et al. (2019) jointly analyzed two PET modalities with different nuclear tracers, and identified brain regions where the tau protein and glucose metabolism are strongly correlated to facilitate the understanding of Alzheimer's disease pathology. Harrewijn et al. (2020) studied resting-state and task-based fMRI, and found that functional connectivities during the rest and the dot-probe task are positively correlated, which conforms to and further extends the current studies of human cognitive behaviors.

In this article, we propose a Bayesian nonparametric spatially varying correlation model to address the question of estimation and inference of spatial regions where two imaging modalities are significantly correlated. We build our model based on the thresholded correlation Gaussian process, which ensures piecewise smoothness, sparsity, as well as jump discontinuity of spatially varying correlations, and works well even when the number of subjects is limited or the signal-to-noise ratio is small. We study the identifiability of our model, establish the large support property, and derive the posterior consistency and selection consistency. We derive the full conditional distributions, propose a Gibbs sampling algorithm that is highly efficient, and propose a hybrid mini-batch Markov chain Monte Carlo (MCMC) to further improve the computational efficiency. We apply our proposed method to jointly analyze the resting-state and working memory task-based fMRIs from a study of the Human Connectome Project (HCP), and identify a number of scientifically meaningful brain regions that offer useful insights for cognitive neuroscience research.

Our proposal is related to but also substantially different from the existing literature on multimodal correlation analysis as well as Bayesian modeling and inference.

For multimodal correlation analysis, there are, broadly speaking, three categories of solutions. The first category is voxel-wise analysis, which estimates the correlation at each voxel separately, then conducts massive voxel-wise significance tests with false discovery control. This approach is computationally easy to implement, but it does not incorporate any spatial or scientific knowledge into statistical inference. Besides, the number of voxels, and thus the number of tests, is huge, whereas the number of subjects in most studies is limited. As a result, voxel-wise analysis often suffers from a particularly low detection power. Although the random field theory has been suggested for multiple testing correction so to improve voxel-wise analysis (Worsley et al. 2004), it does not fully address the low power issue, and is also not directly applicable in our problem due to the complex structure of spatially varying correlations. The second category is region-wise analysis, which first summarizes, usually by averaging, the imaging signals within each brain region defined by some pre-specified brain atlas, then carries the correlation analysis at the region level. Although region-wise analysis generally enjoys a better power than voxel-wise analysis, it is sensitive to the choice of brain atlas. More importantly, the voxels within the same region may not always share the same correlation patterns. Averaging the signals by regions may weaken or cancel out significant correlations. The third category merges voxel-wise and region-wise analysis. In particular, Li et al. (2019) adapted the spatially varying coefficient model, which is widely used in neuroimaging analysis but generally for a different purpose (e.g., Zhu et al. 2014b, Li et al. 2017, 2020b), to the problem of multimodal correlation analysis. They proposed a multi-step procedure, which first fits a spatially varying coefficient model and obtains a smoothed correlation estimate at the voxel level, then applies a graph clustering algorithm to partition the brain into regions with homogeneous correlations, and finally carries out a likelihood ratio test at the region level to identify the regions where two imaging modalities are significantly correlated. However, this procedure involves multiple tuning parameters, and the testing results may be sensitive to their choices. In addition, due to multiple steps of estimation, it is difficult to establish the theoretical guarantees for the final inference method.

For Bayesian modeling and inference, our proposal also makes a number of useful contributions. First of all, we propose a new Bayesian nonparametric prior, i.e, the thresholded correlation Gaussian process, for spatially varying correlation coefficients that are sparse and piecewise smooth over the space. It is constructed under a Bayesian hierarchical model, by thresholding a Gaussian process of the variances for another two correlated Gaussian processes. Our model targets the second-order correlations between two modalities. Relatedly, Bhattacharya and Dunson (2011) proposed a multiplicative Gamma process shrinkage prior with latent factors to model high-dimensional covariance matrices. Nevertheless, their method places the sparsity on the individual latent factors, whereas we need to impose the sparsity at the voxel level, and the sparsity on the latent factors does not lead to the sparsity on the voxels. In addition, our model hinges on the idea of thresholding a Gaussian process. A similar strategy has been adopted in prior constructions for modeling sparse regressions or spatially varying functions, i.e., either thresholding Gaussian random variables (Nakajima and West 2013, Ni et al. 2019, Cai et al. 2020), or thresholding Gaussian processes (Kang et al. 2018, Wu et al. 2022b). However, none of those priors are readily applicable for Bayesian analysis of spatially varying correlations as in our setting.

Second, we contribute to posterior computations for Bayesian models with thresholding type priors. Most existing solutions resort to gradient based MCMC algorithms (Roberts and Rosenthal 1998, Girolami and Calderhead 2011), where a smooth approximation of the thresholding function is required to get the analytically tractable first derivative (Cai et al. 2020, Wu et al. 2022b). There have also been recent advances in developing new sampling algorithms (e.g., Ahn et al. 2012, Chen et al. 2014, Nishimura et al. 2020). However, these algorithms usually converge relatively slowly, and require multiple tuning parameters. By contrast, instead of using a gradient-based MCMC, we successfully derive the full conditional distributions, and propose a Gibbs sampler algorithm that is highly efficient. Besides, the proposed posterior computation algorithm is fairly general, and can be applied to other Bayesian models with thresholding priors as well.

Finally, we are among the first to study the theoretical properties of Bayesian analysis of spatially varying correlations. Particularly, we show that the proposed thresholded correlation Gaussian process has a large prior support on a wide class of sparse, piecewise smooth, and spatially varying correlation functions. We establish the posterior consistency based upon the foundational work of Choi (2005), Ghosal and Roy (2006), Tokdar and Ghosh (2007). However, it is far from a simple extension, as it involves a two-level Bayesian hierarchical model, multiple Gaussian processes, as well as some thresholding functions. To address these challenges, we propose an equivalent model representation for the transformed data, where the spatially varying correlation coefficients become model parameters that specify the mean of the transformed data. This equivalent formulation substantially simplifies the theoretical analysis in the original model. In light of the sparsity, we further establish the selection consistency of activation regions with nonzero correlation coefficients.

The rest of the article is organized as follows. We develop our spatially varying correlation model in Section 3.2. We derive the theoretical properties in Section 3.3, and the Gibbs sampling algorithm in Section 3.4. We carry out the simulations in Section 3.5, and analyze the fMRI data in Section 3.6. We relegate all technical proofs to the Supplementary Material.

3.2 Spatially Varying Correlation Model

In this section, we first propose our Bayesian spatially varying correlation model and the correlation Gaussian process. We then present an equivalent model formulation.

3.2.1 Nonparametric model and correlation Gaussian process

Suppose the observed data consist of n subjects, each with two imaging modalities. Suppose these two imaging modalities are well aligned in a d-dimensional compact spatial space $\mathcal{B} \subset \mathbb{R}^d$, which is generally true for multimodal neuroimaging. Suppose each image consists of measurements at m voxel locations $\mathcal{B}_m = \{v_1, \ldots, v_m\} \subseteq \mathcal{B}$, and we often use $v, v' \in \mathcal{B}$ to denote some generic voxel locations in $\mathcal{B}$. Let $Y_{1,i}(v)$ and $Y_{2,i}(v)$ denote the two imaging measures at location v, for subject $i = 1, \ldots, n$. We consider the following model:

$$Y_{k,i}(v) = \mu_{k,i}(v) + \varepsilon_{k,i}(v), \quad \varepsilon_{k,i}(v) \sim \mathrm{N}\big(0, \tau_k^2(v)\big), \quad \text{for } \ k = 1, 2, \tag{3.1}$$

where $\mu_{k,i}(v)$ are the spatially varying functions that represent the expected values of $Y_{k,i}(v)$, $\varepsilon_{k,i}(v)$ are the random noises that are mutually independent over k, i, v, and follow a normal distribution $\mathrm{N}(\cdot, \cdot)$ with mean zero and variance $\tau_k^2(v)$, $k = 1, 2$.

We next propose a novel prior model for $\mu_{1,i}(v)$ and $\mu_{2,i}(v)$.i.e.,

$$\begin{aligned} &\mu_{1,i}(v) = \eta_{+,i}(v) + \eta_{-,i}(v), \quad \mu_{2,i}(v) = \eta_{+,i}(v) - \eta_{-,i}(v), \quad \text{for } \ i = 1, \ldots, n, \\ &\eta_{+,i} \sim \mathrm{GP}(0, \kappa_+), \qquad \eta_{-,i} \sim \mathrm{GP}(0, \kappa_-), \end{aligned} \tag{3.2}$$

where $\eta_{+,i}$ and $\eta_{-,i}$ are two independent Gaussian processes $\mathrm{GP}(\cdot, \cdot)$, with mean zero and covariance kernel $\kappa_+(v, v')$ and $\kappa_-(v, v')$, $v, v' \in \mathcal{B}$, which capture the positive and negative correlations between the two modalities, respectively. We assume κ_+ and κ_- are of the form,

$$\kappa_+(v, v') = \sigma_+(v)\sigma_+(v')\kappa(v, v'), \qquad \kappa_-(v, v') = \sigma_-(v)\sigma_-(v')\kappa(v, v'), \tag{3.3}$$

where $\sigma_+^2(v)$ and $\sigma_-^2(v)$ are the spatially varying variance functions for $\eta_{+,i}(v)$ and $\eta_{-,i}(v)$, respectively, and $\kappa(\cdot, \cdot)$ is a stationary correlation kernel function. There are various choices for the kernel function $\kappa(\cdot,\cdot)$; for instance, we employ a Matern kernel in our implementation,

$$\kappa(v,v') = \frac{2^{1-\gamma_1}}{\Gamma(\gamma_1)} \left(\sqrt{2\gamma_1} \frac{\|v-v'\|}{\gamma_2} \right)^{\gamma_1} B_{\gamma_1} \left(\sqrt{2\gamma_1} \frac{\|v-v'\|}{\gamma_2} \right), \tag{3.4}$$

where $\Gamma(\cdot)$ is the gamma function, $B_{\gamma_1}(\cdot)$ is the modified Bessel function of the second kind, and γ_1 and γ_2 are two positive hyperparameters that can be determined by the Bayes factor. To impose the sparsity as well as to ensure the identifiability, we require that $\sigma_+^2(v)\sigma_-^2(v) = 0$. In other words, only one of the two terms $\sigma_+^2(v)$ and $\sigma_-^2(v)$ is nonzero.

Finally, we impose that the variance functions are of the form,

$$\begin{aligned} &\sigma_+(v) = G_\omega\{\xi(v)\}, \quad \sigma_-(v) = G_\omega\{-\xi(v)\}, \\ &\xi(v) \sim \mathrm{GP}(0,\kappa), \end{aligned} \tag{3.5}$$

where $G_\omega(x) = xI(x > \omega)$ is a thresholding function with the thresholding parameter $\omega \geq 0$ and $I(\cdot)$ the indicator function, $\xi(v)$ is a spatially varying function that determines both $\sigma_+(v)$ and $\sigma_-(v)$ through $G_\omega(x)(\cdot)$. As a prior specification, we assume $\xi(v)$ follows another Gaussian process with mean zero and correlation kernel $\kappa(\cdot,\cdot)$, and $\kappa(\cdot,\cdot)$ is the same as that in (3.3). Note that the construction in (3.5) ensures $\sigma_+^2(v)$ and $\sigma_-^2(v)$ are uniquely determined by $\xi(v)$, and $\sigma_+^2(v)\sigma_-^2(v) = 0$.

Following the prior specifications (3.2) to (3.5), and integrating out $\mu_{1,i}(v)$ and $\mu_{2,i}(v)$ in (3.1), we obtain the spatially varying correlation function between $Y_{1,i}(v)$ and $Y_{2,i}(v)$ of the form,

$$\begin{aligned} \rho(v) &= \mathrm{Corr}\Big\{Y_{1,i}(v), Y_{2,i}(v) \mid \xi(v), \tau_1^2(v), \tau_2^2(v)\Big\} \\ &= \frac{G_\omega^2\{\xi(v)\} - G_\omega^2\{-\xi(v)\}}{\sqrt{G_\omega^2\{\xi(v)\} + G_\omega^2\{-\xi(v)\} + \tau_1^2(v)}\sqrt{G_\omega^2\{\xi(v)\} + G_\omega^2\{-\xi(v)\} + \tau_2^2(v)}}. \end{aligned} \tag{3.6}$$

We say that $\rho(v)$ in (3.6) follows a *thresholded correlation Gaussian Process*, as formally defined below.

Definition 11. *Given any nonzero spatially varying variance functions $\tau_1^2(v)$ and $\tau_2^2(v)$, and the thresholding parameter $\omega \geq 0$, suppose $\xi(v) \sim GP(0,\kappa)$, then $\rho(v)$ in* (3.6) *follows a thresholded correlation Gaussian process, denoted as $\rho \sim TCGP(\omega, \kappa, \tau_1^2, \tau_2^2)$.*

Under this construction, with probability one, a correlation Gaussian process is between -1 and 1, and enjoys both piecewise smoothness and sparsity.

Our proposed model enjoys several benefits. It encompasses a large class of spatially varying functions that are piecewise smooth, sparse, and jump discontinuous, the features that we commonly encounter in neuroimaging data (Zhu et al. 2014b). Moreover, instead of specifying a voxel-wise prior, our thresholded correlation Gaussian process incorporates the spatial information of the image, leading to potentially more accurate detection.

3.2.2 Equivalent model representation

To facilitate both the theoretical investigation and posterior computation, we next derive an equivalent representations of model (3.1) under the prior specifications (3.2) to (3.5).

We first note that, from (3.5) and (3.6), $\sigma_+(v)$ and $\sigma_-(v)$ can be uniquely determined by $\rho(v)$, in that, given $\tau_1^2(v)$ and $\tau_2^2(v)$,

$$\begin{aligned} &\sigma_+(v) = G_\omega\{\xi(v)\} = s\left\{\rho(v); \tau_1^2(v), \tau_2^2(v)\right\}, \\ &\sigma_-(v) = G_\omega\{-\xi(v)\} = s\left\{-\rho(v); \tau_1^2(v), \tau_2^2(v)\right\}, \quad \text{where} \\ &s(x; t_1, t_2) = \sqrt{\frac{2t_1t_2}{\sqrt{(t_1-t_2)^2 + 4x^{-2}t_1t_2} - (t_1+t_2)}} I(x>0), \text{ for any } x \in [-1,1], t_1, t_2 > 0. \end{aligned} \tag{3.7}$$

We next consider a transformation of the observed images $\{Y_{1,i}(v), Y_{2,i}(v)\}$, the average $Y_{+,i}(v) = \{Y_{1,i}(v) + Y_{2,i}(v)\}/2$, and the contrast $Y_{-,i}(v) = \{Y_{1,i}(v) - Y_{2,i}(v)\}/2$. Denote

$$E_{+,i}(v) = \frac{\eta_{+,i}(v)}{\sigma_+(v)}, \qquad E_{-,i}(v) = \frac{\eta_{-,i}(v)}{\sigma_-(v)}. \tag{3.8}$$

By (3.7), model (3.1) is equivalent to

$$\begin{aligned} Y_{+,i}(v) &= s\left\{\rho(v); \tau_1^2(v), \tau_2^2(v)\right\} E_{+,i}(v) + \varepsilon_{+,i}(v), \\ Y_{-,i}(v) &= s\left\{-\rho(v); \tau_1^2(v), \tau_2^2(v)\right\} E_{-,i}(v) + \varepsilon_{-,i}(v), \end{aligned} \tag{3.9}$$

where $\varepsilon_{+,i}(v)$ and $\varepsilon_{-,i}(v)$ are random noises that are independent over i, v, and follow a normal distribution with mean zero and variance $\{\tau_1^2(v) + \tau_2^2(v)\}/4$. The covariance between $\varepsilon_{+,i}(v)$ and $\varepsilon_{-,i}(v)$ is $\{\tau_1^2(v) - \tau_2^2(v)\}/\{\tau_1^2(v) + \tau_2^2(v)\}$.

Following the prior specifications (3.2) to (3.5), we have the equivalent prior specifications for $E_{+,i}(v)$, $E_{-,i}(v)$, and $\rho(v)$ as,

$$E_{+,i} \sim \mathrm{GP}(0,\kappa), \qquad E_{-,i} \sim \mathrm{GP}(0,\kappa), \qquad \rho \mid \tau_1^2, \tau_2^2 \sim \mathrm{TCGP}(\omega, \kappa, \tau_1^2, \tau_2^2), \tag{3.10}$$

where $\kappa(\cdot,\cdot)$ is the correlation kernel as specified in both (3.3) and (3.5), and in our modeling process, we use the Matérn kernel as specified in (3.4) for $\kappa(\cdot,\cdot)$.

Figure 3.1 gives a graphical illustration of our nonparametric Bayesian spatially varying correlation model. In our subsequent theoretical and numerical analysis, we focus on the equivalent transformed model.

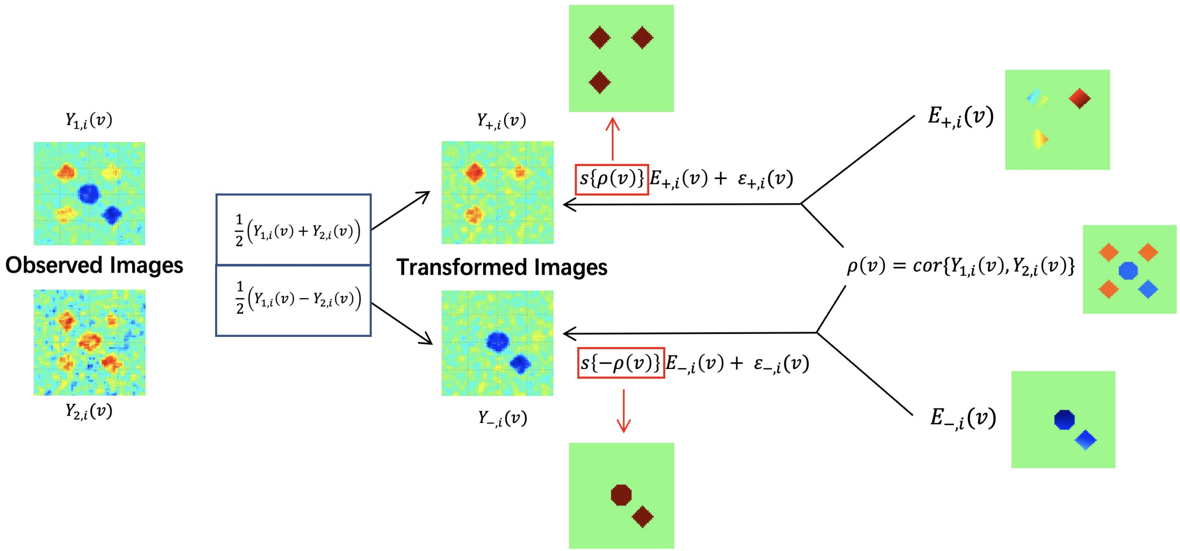

$Y_{1,i}(v)$
Observed Images
$Y_{2,i}(v)$
$\frac{1}{2}(Y_{1,i}(v)+Y_{2,i}(v))$
$\frac{1}{2}(Y_{1,i}(v)-Y_{2,i}(v))$
$Y_{+,i}(v)$
Transformed Images
$Y_{-,i}(v)$
$s\{\rho(v)\}E_{+,i}(v)+\varepsilon_{+,i}(v)$
$s\{-\rho(v)\}E_{-,i}(v)+\varepsilon_{-,i}(v)$
$E_{+,i}(v)$
$\rho(v)=cor\{Y_{1,i}(v),Y_{2,i}(v)\}$
$E_{-,i}(v)$

Figure 3.1: Graphical illustration of the proposed Bayesian spatially varying correlation model. The transformed image $Y_{\pm,i}(v)$ are modeled based on (3.9).

3.3 Theory

In this section, we study the model identifiability, derive the large support property, and establish the posterior and selection consistency.

3.3.1 Notations and definitions

We begin with some notations and definitions. For any vector $v = (v_1, \dots, v_d)^\intercal \in \mathbb{R}^d$, let $\|v\|_p = \left(\sum_{l=1}^d |v_l|^p\right)^{1/p}$ denote the L_p-norm, $p \geq 1$, and $\|v\|_\infty = \max_{l=1}^d |v_l|$ the supremum norm. For any real function f on the region $\mathcal{B}$, let $\|f\|_p = \left\{\int_{\mathcal{B}} |f(v)|^p \, \mathrm{d}v\right\}^{1/p}$ denote the L_p-norm, $p \geq 1$, and $\|f\|_\infty = \sup_{v\in\mathcal{B}} |f(v)|$ be the supremum norm. Suppose $\mathcal{B}$ is a compact convex set. Recall there are n subjects, and m spatial locations for each image. Denote $Y_\pm = \{Y_{\pm,1}^\intercal, \dots, Y_{\pm,n}^\intercal\}^\intercal$, where $Y_{\pm,i} = \{Y_{\pm,i}(v_1), \dots, Y_{\pm,i}(v_m)\}^\intercal$. Furthermore, denote our parameter of interest as $\theta(\cdot) = \left\{\rho(\cdot), E_+^\intercal(\cdot), E_-^\intercal(\cdot)\right\}^\intercal$, where $E_\pm(\cdot) = \{E_{\pm,1}(\cdot), \dots, E_{\pm,n}(\cdot)\}^\intercal$, and the true parameter $\theta_0(\cdot) = \left\{\rho_0(\cdot), E_{+,0}^\intercal(\cdot), E_{-,0}^\intercal(\cdot)\right\}^\intercal$.

Definition 12. *Define $\mathcal{C}^q(\mathcal{B})$ as a set of differentiable functions of order q defined on $\mathcal{B}$, such that a function $f \in \mathcal{C}^q(\mathcal{B})$ has the partial derivative,*

$$D^b f(v) = \frac{\partial^{\|b\|_1} f}{v_1^{b_1} \dots v_d^{b_d}}(v) = \sum_{\|a\|_1+\|b\|_1\leq q} \frac{D^{b+a} f(u)}{a!}(v-u)^a + R_q(v,u),$$

where $b = (b_1, \dots, b_d)^\intercal \in \mathbb{Z}_+^d$, $a \in \mathbb{Z}_+^d$, $\mathbb{Z}_+$ denotes the set of non-negative integers, $u \in \mathbb{R}^d$, and the *remainder $R_q(v,u)$ satisfies the following properties: (i) Given any point v_0 of $\mathcal{B}$ and any constant $\epsilon > 0$, there is a constant $\delta > 0$, such that if v and u are any two points of $\mathcal{B}$ with $\|v - v_0\|_1 < \delta$ and $\|u - v_0\|_1 < \delta$, then $|R_q(v,u)| \leq \|v-u\|_1^{q-\|b\|_1}\epsilon$; (ii) If $\|D^b f\|_\infty \leq C < \infty$ for some constant C, then $|R_q(v,u)| \leq (C\|v-u\|_1^{q+1})/(q+1)!$.*

Definition 13. *Define $\Theta_\rho = \big\{\rho(v) \in (-1,1) : v \in \mathcal{B}\big\}$ as a collection of spatially varying correlation functions that satisfy the following properties: (i) There exist two disjoint non-empty open sets $\mathcal{R}_{-1}$ and $\mathcal{R}_1$ with $\overline{\mathcal{R}}_1 \cap \overline{\mathcal{R}}_{-1} = \emptyset$, such that $\rho(v)$ is smooth over $\overline{\mathcal{R}}_{-1} \cup \overline{\mathcal{R}}_1$, i.e., $\rho(v)I\left[v \in \overline{\mathcal{R}}_{-1} \cup \overline{\mathcal{R}}_1\right] \in \mathcal{C}^\alpha\left(\overline{\mathcal{R}}_{-1} \cup \overline{\mathcal{R}}_1\right)$, with $\alpha = \lceil d/2 \rceil + 1$, the least integer greater than or equal to $d/2$; (ii) $\rho(v) = 0$ for $v \in \mathcal{R}_0$, $\rho(v) > 0$ for $v \in \mathcal{R}_1$, and $\rho(v) < 0$ for $v \in \mathcal{R}_{-1}$, where $\mathcal{R}_0 = \mathcal{B} - (\mathcal{R}_{-1} \cup \mathcal{R}_1)$ and $\mathcal{R}_0 - (\partial\mathcal{R}_1 \cup \partial\mathcal{R}_{-1}) \neq \emptyset$; (iii) $\rho(v)$ is a discontinuous function and is bounded away from zero for any $v \notin \mathcal{R}_0$, i.e., $\gamma = \inf_{v \notin \mathcal{R}_0} |\rho(v)| > 0$.*

Definition 14. *Define $\Theta_E = \{E(v) \in \mathbb{R}^n : \|E(v)\|_2^2 = C_v\}$ for some constant $C_v < \infty$.*

In summary, Θ_ρ is the collection of all piecewise smooth, sparse, and jump discontinuous correlation functions $\rho(v)$ defined on $\mathcal{B}$, where γ in Definition 13 represents the minimum nonzero effect size of the correlation functions that have discontinuity jumps, and Θ_E is the collection of the spatially varying functions $E(v)$ that satisfy some second moment constraints.

3.3.2 Model identifiability and large support

We first show that model (3.9) is identifiable, then show that the prior specification in (3.10) has a large support. We begin with a regularity condition.

Assumption 14.1. *The true correlation function ρ_0 is piecewise smooth, sparse, and jump discontinuous, in that $\rho_0 \in \Theta_\rho$. In addition, the true functions $E_{+,0}$ and $E_{-,0}$ have constant second moments with respect to the location v, i.e., $E_{+,0} \in \Theta_E$ and $E_{-,0} \in \Theta_E$.*

Assumption 14.1 essentially specifies the class of true functions that we target. Denote $\mathcal{V}(\rho) = \{v : \rho(v) \neq 0\}$, and $\mathcal{V}(\rho') = \{v : \rho'(v) \neq 0\}$. The next proposition shows that model (3.9) is identifiable. Specifically, $\rho(v)$ is identifiable for all $v \in \mathcal{B}$, and $E_+(v), E_-(v)$ are identifiable for $v \in \mathcal{V}(\rho) \cup \mathcal{V}(\rho')$. The identifiability of $E_+(v)$ and $E_-(v)$ is constrained on $\mathcal{V}(\rho) \cup \mathcal{V}(\rho')$ because when $\rho(v) = 0$, $s\{\rho(v)\} = 0$ in model (3.9).

Proposition 15. *(Identifiability) Suppose Assumption 14.1 holds. Then model* (3.9) *is identifiable. That is, if the probability distributions of $\{Y_+, Y_-\}$ under $\theta = \left\{\rho, E_+^\intercal, E_-^\intercal\right\}^\intercal$ and $\theta' = \left\{\rho', {E'_+}^\intercal, {E'_-}^\intercal\right\}^\intercal$ are equal, then we have $\rho = \rho'$ for $v \in \mathcal{B}_m$, and $\theta = \theta'$ for $v \in \mathcal{V}(\rho) \cup \mathcal{V}(\rho')$.*

To ensure the large-support property, we introduce another condition on the correlation kernel function $\kappa(\cdot,\cdot)$. The same condition was imposed in Ghosal and Roy (2006) as well.

Assumption 15.1. *The correlation kernel $\kappa(\cdot,\cdot)$ satisfies that, for any $v \in \mathcal{B}$, $\kappa(v,\cdot)$ has continuous partial derivatives up to order $2\alpha+2$, where $\alpha = \lceil d/2 \rceil + 1$. In addition, suppose $\kappa(v,v') = \prod_{l=1}^{d} \kappa_l(v_l - v_l'; \nu_l)$, for any $v = (v_1,\ldots,v_d)^\intercal$, and $v' = (v_1',\ldots,v_d')^\intercal \in [0,1]^d$, where $\kappa_l(\cdot;\nu_l)$ is a continuous, nowhere zero, symmetric density function on $\mathbb{R}$ with parameter $\nu_l \in \mathbb{R}^+$, for $l = 1,\ldots,d$.*

The next theorem shows that our prior specification in (3.10) is desirable, in that it has support over a large class of sparse, piecewise smooth and jump discontinuous spatially varying correlation functions. That is, there is a positive probability that $\theta = \left\{\rho, E_+^\intercal, E_-^\intercal\right\}^\intercal$ concentrates on an arbitrarily small neighborhood of any true parameter in the parameter space $\Theta = \Theta_\rho \times \Theta_E \times \Theta_E$.

Theorem 16. *(Large Support) Suppose Assumptions 14.1 and 15.1 hold. Under the prior specification in (3.10), for any $\epsilon > 0$, $\Pi\big(\|\theta - \theta_0\|_\infty < \epsilon\big) > 0$, where $\Pi(\cdot)$ denotes a probability measure on the Borel set of Θ.*

3.3.3 Posterior consistency

Next, we establish the posterior consistency, then the selection consistency.

Assumption 16.1. *There exist constants $d/(2\alpha) < \nu_0 < 1$, $C_0 > 0, C_1 > 0$, and $N \geq 1$, with $\alpha = \lceil d/2 \rceil + 1$, such that $C_0 n^d \leq m \leq C_1 n^{2\alpha\nu_0}$ for all $n > N$.*

Assumption 16.1 imposes that the number of spatial locations m should be of the polynomial order of the sample size n. The lower bound indicates that m needs to be sufficiently large to ensure that the posterior distribution of the spatially varying coefficient function concentrates around the true value. The upper bound ensures that a sufficient amount of information is collected across subjects to identify the population level true parameters.

The next theorem shows that, under the proposed prior, the posterior distribution of θ concentrates in an arbitrarily small neighborhood of the true parameter θ_0, when the number of subjects n and the number of spatial locations m are sufficiently large.

Theorem 17. *(Posterior Consistency) Suppose Assumptions 14.1, 15.1 and 16.1 hold. Under model (3.9) and the prior specification in (3.10), for any $\epsilon > 0$, as $m \to \infty$ and $n \to \infty$,*

$$\Pi\Big(\{\theta \in \Theta : \|\theta - \theta_0\|_1 < \epsilon\} \,\Big|\, Y_+, Y_-\Big) \to 1 \;\; \textit{in} \;\; \mathbb{P}_{\theta_0}^{(m,n)}\textit{-probability},$$

where $\mathbb{P}_{\theta_0}^{(m,n)}$ denotes the distribution of $\{Y_+, Y_-\}$ given the true parameter θ_0, and $\Pi(\cdot \mid Y_+, Y_-)$ denotes the posterior probability measure on the Borel set of Θ given data $\{Y_+, Y_-\}$.

The next theorem shows that, with probability tending to one, our method can identify the true activation regions that have positive correlations, negative correlations, and no correlations, respectively, when both n and m tend to infinity.

Theorem 18. *Suppose the same conditions in Theorem 17 hold. Then, as $m \to \infty$ and $n \to \infty$,*

$$\Pi\Big(\operatorname{sgn}\{\rho(v)\} = \operatorname{sgn}\{\rho_0(v)\}, v \in \mathcal{B} \,\Big|\, Y_+, Y_-\Big) \to 1 \ \textit{ in } \ \mathbb{P}_{\theta_0}^{(m,n)}\textit{-probability},$$

where $\operatorname{sgn}(x) = 1$ *if* $x > 0$, $\operatorname{sgn}(x) = -1$ *if* $x < 0$ *and* $\operatorname{sgn}(0) = 0$.

3.4 Posterior Computation

In this section, we first adopt the Karhunen-Loève expansion to simplify the model to a finite number of parameters. We next derive the full conditional distributions of the model parameters, and develop an efficient Gibbs sampling algorithm. We also propose a hybrid mini-batch MCMC to further improve the computational efficiency.

3.4.1 Karhunen-Loève approximation

Model (3.9) involves three Gaussian processes, for $E_{+,i}(v)$, $E_{-,i}(v)$, and $\xi(v)$, respectively, and all hinge on the infinite dimensional correlation kernel function $\kappa(\cdot,\cdot)$. We first adopt the usual strategy of Karhunen-Loève expansion to simplify the model to a finite number of parameters. Specifically, consider the spectral decomposition of the kernel function,

$$\kappa\left(v, v'\right) = \sum_{l=1}^{\infty} \lambda_l \psi_l(v) \psi_l\left(v'\right),$$

where $\{\lambda_l\}_{l=1}^{\infty}$ are the eigenvalues in descending order, and $\{\psi_l(v)\}_{l=1}^{\infty}$ are the corresponding orthonormal eigenfunctions. By Mercer's Theorem (Mercer 1909), we can represent the Gaussian processes in our model by the Karhunen-Loève (KL) expansion,

$$E_{+,i}(v) = \sum_{l=1}^{\infty} e_{i,l,+} \psi_l(v), \quad E_{-,i}(v) = \sum_{l=1}^{\infty} e_{i,l,-} \psi_l(v), \quad \xi(v) = \sum_{l=1}^{\infty} c_l \psi_l(v).$$

where $c_l, e_{i,l,\pm}$ are Karhunen-Loève coefficients. We further truncate the above expansions by focusing on the leading L eigenvalues and eigenfunctions, where L can be determined following the usual practice of principal components analysis that retains a certain percentage of total variation.

Based on the Karhunen-Loève truncation, model (3.1) becomes,

$$\begin{aligned} Y_{+,i}(v) &= G_\omega\left\{\sum_{l=1}^{L} c_l\psi_l(v)\right\}\left\{\sum_{l=1}^{L} e_{i,l,+}\psi_l(v)\right\}+\varepsilon_{+,i}(v),\\ Y_{-,i}(v) &= G_\omega\left\{-\sum_{l=1}^{L} c_l\psi_l(v)\right\}\left\{\sum_{l=1}^{L} e_{i,l,-}\psi_l(v)\right\}+\varepsilon_{-,i}(v). \end{aligned} \tag{3.11}$$

Recall that $\mathcal{B}_m=\{v_1,\ldots,v_m\}$ denotes the set of locations where the imaging data are observed, and let $Y=\{Y_{1,i}(v),Y_{2,i}(v),i=1,\ldots,n,v\in\mathcal{B}_m\}$ denote the imaging data observed at the set of voxels in $\mathcal{B}_m$. Then all the parameters in our model include:

$$\tilde{\Theta}=\Big\{\{c_l\}_{l=1}^{L},\ \big\{\{e_{i,l,+}\}_{l=1}^{L},\{e_{i,l,-}\}_{l=1}^{L}\big\}_{i=1}^{n},\ \{\tau_1^2(v),\tau_2^2(v)\}_{v\in\mathcal{B}_m},\ \omega\Big\}. \tag{3.12}$$

We specify their prior distributions as,

$$c_l\sim\mathrm{N}(0,\lambda_l),\ \ e_{i,l,\pm}\sim\mathrm{N}(0,\lambda_l),\ \ \tau_1^2(v),\tau_2^2(v)\sim\mathrm{IG}(a_\tau,b_\tau),\ \ \omega\sim\mathrm{U}(a_\omega,b_\omega), \tag{3.13}$$

That is, we impose a normal distribution for the Karhunen-Loève coefficients $c_l, e_{i,l,\pm}$, where λ_l is the eigenvalue of the kernel $\kappa(v,v')$ as specified above. We impose an inverse Gamma prior for the variance terms $\tau_1^2(v)$, $\tau_2^2(v)$, with shape a_τ and scale b_τ, and we choose some small values for a_τ, b_τ, so that this prior is non-informative. We also impose a uniform prior for the thresholding parameter ω, with range from a_ω to b_ω, and we choose a_ω, b_ω based on the quantiles of $|\xi(v)|_{v\in\mathcal{B}_m}$. It is also possible to consider other types of prior for ω, e.g., an exponential distribution. Note that the conditional prior for ω allows it to be adaptively learnt in a fully Bayesian way in our Gibbs sampling. This is different from the gradient based MCMC methods, which require a smooth approximation of the thresholding function.

3.4.2 Gibbs sampling

We first present a general result that is useful for deriving the full conditional distributions of some of our key parameters. We note that this result is both new and general, and can be applied to deriving the Gibbs sampler for other types of models involving Gaussian process.

Proposition 19. *Consider a random variable θ, and two sets of functions $f_p(\theta)=a_{1p}\theta^2+a_{2p}\theta+a_{3p}$, and $h_k(\theta)=b_{1k}\theta^2+b_{2k}\theta+b_{3k}$, where $a_{1j},a_{2j},a_{3j},b_{1k},b_{2k},b_{3k}$ are some coefficients, $p=$*

$1, \ldots, P, k = 1, \ldots, K$. *Suppose the density of* θ *is proportional to*

$$\exp\left\{\sum_{p=1}^{P} f_p(\theta) I(\theta > L_p) + \sum_{k=1}^{K} h_k(\theta) I(\theta < U_k)\right\}, \tag{3.14}$$

where U_k, L_p *are some thresholding coefficients,* $p = 1, \ldots, P, k = 1, \ldots, K$. *Then,*

(i) If at least one of $\{a_{1p}, \ldots, a_{1P}, b_{1k}, \cdots, b_{1K}\}$ *is not equal to 0, then* θ *follows a mixture of truncated normal distributions.*

(ii) If $a_{1p} = b_{1k} = a_{2p} = b_{2k} = 0$ *for all* p, k*, and at least one of* $\{a_{3p}, \ldots, a_{3P}, b_{3k}, \ldots, b_{3K}\}$ *is not equal to 0, then* θ *follows a mixture of uniform distributions.*

(iii) If $a_{1p} = b_{1k} = 0$ *for all* p, k*, and at least one of* $\{a_{2p}, \cdots, a_{2P}, b_{2k}, \cdots, b_{2K}\}$ *is not equal to 0, then* θ *follows a mixture of exponential distributions.*

We next derive the full conditional distributions of our model parameter $\tilde{\Theta}$ in (3.12). Specifically, we first derive the full conditionals of $\{c_l\}_{l=1}^{L}$ and ω, both of which are based on Proposition 19. We then derive the full conditionals of $\{e_{i,l,\pm}\}_{l=1,i=1}^{L,n}$ and $\{\tau_1^2(v), \tau_2^2(v)\}_{v \in \mathcal{B}_m}$, both of which have closed forms thanks to their conjugate priors. Let $\tilde{\Theta}_{\setminus\theta}$ denote the set of parameters in $\tilde{\Theta}$ but without θ.

The full conditional of c_l is a mixture of truncated normal distributions, as we show in Section B.3.2 of the Supplementary Material. This is because the density of c_l is of the form,

$$\pi(c_l \mid Y, \tilde{\Theta}_{\setminus c_l}) \propto \exp\left(\sum_{\substack{j=1 \\ \psi_l(v_j)>0}}^{m} \left[g_+(c_l; v_j) I\{c_l > T_+(v_j)\} + g_-(c_l; v_j) I\{c_l < T_-(v_j)\}\right]\right.$$
$$\left. + \sum_{\substack{j=1 \\ \psi_l(v_j)<0}}^{m} \left[g_+(c_l; v_j) I\{c_l < T_+(v_j)\} + g_-(c_l; v_j) I\{c_l > T_-(v_j)\}\right]\right).$$

Given the location v_j, $g_\pm(c_l; v_j)$ are two quadratic functions of c_l, and $T_\pm(v_j)$ are two scalars, whose detailed forms are given in Section B.3.2. If we set $f_p(\theta) = g_+(c_l; v_j)$, $h_k(\theta) = g_-(c_l; v_j)$ for those locations v_j satisfying $\psi_1(v_j) > 0$, and set $f_p(\theta) = g_-(c_l; v_j)$, $h_k(\theta) = g_+(c_l; v_j)$ otherwise, then the density of c_l satisfies the condition of Proposition 19(i), and thus it follows a mixture of truncated normal distributions.

The full conditional of ω is a mixture of uniform distributions, as we show in Section B.3.3 of

Algorithm 1 Gibbs sampling for TCGP

```
input: the observed imaging data $Y = \{Y_{1,i}(v), Y_{2,i}(v), i = 1, \ldots, n, v \in \mathcal{B}_m\}$,
       the kernel function $\kappa(\cdot,\cdot)$, the Karhunen-Loève truncation number $L$,
       the prior hyperparameters $a_\tau, b_\tau, a_\omega, b_\omega$.
output: the posterior samples of $\tilde{\Theta} = \{\{c_l\}_{l=1}^L, \{e_{i,l,\pm}\}_{l=1,i=1}^{L,n}, \{\tau_1^2(v), \tau_2^2(v)\}_{v\in\mathcal{B}_m}, \omega\}$.
initialize $\tilde{\Theta}$: sample $\tilde{\Theta}$ from the prior distribution.
for $t = 1, \cdots, T$ do
    parallel sample ${\tau_k}^2(v)$ from the inverse Gamma distribution, $v \in \mathcal{B}_m, k = 1, 2$.
    for $l = 1, \ldots, L$ do
        sample $c_l$ from the mixture of truncated normal distributions.
        sample $\omega$ from the mixture of uniform distributions.
        sample $e_{i,l,\pm}$ from the normal distribution, $i = 1, \ldots, n$.
    end for
end for
```

the Supplementary Material. This is because the density of ω is of the form,

$$\omega \mid Y, \tilde{\Theta}_{\backslash\omega} \sim \exp\left[\sum_{\substack{j=1 \\ a_\omega < \xi(v_j) < b_\omega}}^{m} C_+(v_j) I\{\omega < \xi(v_j)\} + \sum_{\substack{j=1 \\ a_\omega < -\xi(v_j) < b_\omega}}^{m} C_-(v_j) I\{\omega < -\xi(v_j)\} \right].$$

Given the location v_j, $\xi(v_j) = \sum_{l=1}^L c_l \psi_l(v_j)$, $C_\pm(v_j)$ are two scalars, whose detailed forms are given in Section B.3.3. If we set $h_k(\theta) = C_+(v_j)$ for those v_j satisfying $a_\omega < \xi(v_j) < b_\omega$, and set $h_k(\theta) = C_-(v_j)$ for those locations satisfying $a_\omega < -\xi(v_j) < b_\omega$, then the density of ω satisfies the condition of Proposition 19(ii), and thus it follows a mixture of uniform distributions. We make two additional remarks. First, we specify the prior of ω as $\mathrm{U}(a_\omega, b_\omega)$, where we choose a_ω, b_ω to have a non-informative prior. In practice, we may adopt the empirical Bayes idea, by running the Gibbs sampling once with a non-informative prior first, then using the quantile values of the sorted $\{|\xi(v)|\}_{v\in\mathcal{B}_m}$ to refine the range of the uniform distribution for ω. This can further improve the convergence behavior of the algorithm. Second, if we specify the prior of ω as an exponential distribution, then we may apply Proposition 19(iii) to obtain the full conditional of ω.

The full conditionals of $e_{i,l,\pm}$ $(i = 1, \ldots, n; l = 1, \ldots, L)$ is a normal distribution, thanks to the conjugate prior. Its derivation is given in Section B.3.4.

The full conditional of $\tau_k^2(v)$ $(k = 1, 2; v \in \mathcal{B}_m)$ is an inverse Gamma distribution, again thanks to the conjugate prior. Its derivation is given in Section B.3.5.

We summarize the Gibbs sampling for the thresholded correlation Gaussian process in Algorithm 1.

3.4.3 Hybrid mini-batch MCMC

The proposed Gibbs sampler is computationally efficient in general. Meanwhile, the complexity of computing the full conditional of c_l is $O(m^2)$, where m is the total number of voxels. When m is large, this step can be expensive. We next propose a hybrid mini-batch MCMC, with two key components, to further improve the computational efficiency.

The first component is to develop an adaptive proposal function in Gibbs sampling. We note that the Gibbs sampler can be viewed as a special case of Metropolis–Hastings, in which the newly proposed state is always accepted with probability one, and the proposal function in Metropolis–Hastings corresponds to the full conditional distribution in the Gibbs sampler. There have been some recent progress developing scalable MCMC methods (Li and Wong 2017, Wu et al. 2022c). However, those algorithms mainly focus on how to more efficiently evaluate the ratio of the likelihood function at each iteration, instead of focusing on the proposal function. Moreover, their aims are not to perform Bayesian inference from the exact posterior, but rather to exploit the tempered posterior with an efficient MCMC sampler to obtain a better solution from the global optimization.

We propose an adaptive proposal function, by subsampling voxel locations, instead of individual subjects. More specifically, let $\mathcal{B}_{m_s} \subset \mathcal{B}_m$ denote a random subset of all the observed locations $\mathcal{B}_m$, Y_{m_s} the corresponding imaging data observed at those voxels in $\mathcal{B}_{m_s}$, and $m_s = |\mathcal{B}_{m_s}|$ the cardinality of $\mathcal{B}_{m_s}$. Recognizing that the Gibbs sampler is a special case of Metropolis–Hastings, the proposal function for the parameter $\theta \in \tilde{\Theta}$ is $P\{\theta|\tilde{\Theta}_{\setminus\theta}, Y\}$, which is the full conditional distribution of θ. Instead of using the entire imaging data Y to derive the full conditional distribution of θ, we propose to use a mini-batch of data Y_{m_s} to obtain the proposal function $P\{\theta|\tilde{\Theta}_{\setminus\theta}, Y_{m_s}\}$. The acceptance ratio of θ is,

$$\phi(\theta', \theta) = \min\left[1, \frac{\Pi_{v \notin \mathcal{B}_{m_s}} P\{Y(v)|\theta', \tilde{\Theta}_{\setminus\theta}\}}{\Pi_{v \notin \mathcal{B}_{m_s}} P\{Y(v)|\theta, \tilde{\Theta}_{\setminus\theta}\}}\right],$$

whose derivation is given in Section B.3.6 of the Supplementary Material. In this case, the computational complexity of sampling c_l is reduced from $O(m^2)$ to $O(m_s^2)$.

The second component is to consider a hybrid version of mini-batch. This is because, when keeping using the mini-batch of voxels during the whole sampling process, the Markov chain may converge to local modes, and may also converge slowly. To overcome these issues, we propose to use the full dataset after, say, every T_0 iterations of using the mini-batch data.

We summarize the hybrid mini-batch MCMC procedure in Algorithm 2. In our implementation, we set $m_s = m/16$ and $T_0 = 20$, which leads to a good empirical performance. We also carry out a sensitivity analysis in Section B.4.2 of the Supplementary Material, and find that the result is not

Algorithm 2 Hybrid mini-batch MCMC for TCGP.

```
input: the observed imaging data $Y = \{Y_{1,i}(v), Y_{2,i}(v), i = 1, \ldots, n, v \in \mathcal{B}_m\}$,
       the kernel function $\kappa(\cdot, \cdot)$, the Karhunen-Loève truncation number $L$,
       the prior hyperparameters $a_\tau, b_\tau, a_\omega, b_\omega$.
output: the posterior samples of $\tilde{\Theta} = \{\{c_l\}_{l=1}^{L}, \{e_{i,l,\pm}\}_{l=1,i=1}^{L,n}, \{\tau_1^2(v), \tau_2^2(v)\}_{v \in \mathcal{B}_m}, \omega\}$.
initialize $\tilde{\Theta}$: sample $\tilde{\Theta}$ from the prior distribution.
for $t = 1, \cdots, T$ do
    parallel sample ${\tau_k}^2(v)$ from the inverse Gamma distribution, for all $v \in \mathcal{B}_m, k = 1, 2$.
    random sample $m_s$ locations from $\mathcal{B}_m$ and form $\mathcal{B}_{m_s}$ and $Y_{m_s}$.
    for $l = 1, \cdots, L$ do
        if $t \mod T_0 = 0$ then
            sample $c_l$ from the mixture of truncated normal distributions based on $Y$.
            sample $\omega$ from the mixture of uniform distributions based on $Y$.
        else
            sample $c_l^{(t)}$ from the mixture of truncated normal distributions based on $Y_{m_s}$.
            accept $c_l^{(t)}$ with probability $\phi(c_l^{(t)}, c_l^{(t-1)})$.
            sample $\omega^{(t)}$ from the mixture of uniform distributions based on $Y_{m_s}$.
            accept $\omega^{(t)}$ with probability $\phi(\omega^{(t)}, \omega^{(t-1)})$.
        end if
        parallel sample $e_{i,l,\pm}$ from the normal distribution, $i = 1, \ldots, n$.
    end for
end for
```

sensitive to m_s and T_0, as long as they are in a reasonable range.

3.5 Simulations

In this section, we carry out two simulation studies, one for a 2D example and the other a 3D example, to investigate the empirical performance of the proposed method.

3.5.1 2D image simulation

We simulate the data from model (3.1), with the sample size $n = 50$, and the image resolution $m = 64 \times 64$. We simulate the mean $\mu_{k,i}$ from (3.2) and (3.3), $k = 1, 2$, with $\kappa(v, v') = \exp -0.1(v^2 + v'^2) - 10(v - v')^2$, $\sigma_+^2(v) = \zeta_+ \sum_{j=1}^{3} I(\|v - u_{+,j}\|_1 < 0.1)$, where $u_{+,1} = (0.3, 0.7)$, $u_{+,2} = (0.7, 0.7)$, $u_{+,3} = (0.3, 0.3)$, and $\sigma_-^2(v) = \zeta_- \{I(\|v - u_{-,1}\|_1 < 0.1) + I(\|v - u_{-,2}\|_2 < 0.1)\}$, where $u_{-,1} = (0.5, 0.5)$, $u_{-,2} = (0.7, 0.3)$. Here (ζ_+, ζ_-) controls the signal strength, and we consider two settings, with $(\zeta_+, \zeta_-) = (0.15, 0.25)$ for a weak signal, and $(\zeta_+, \zeta_-) = (0.75, 0.85)$ for a strong signal. We simulate the noise $\varepsilon_{k,i}$ from the normal distribution with mean zero and variance $\tau_k^2(v)$, and simulate $\log(\tau_k^2(v))$ from a Gaussian process with mean zero and correlation kernel $\kappa(v, v')$, $k = 1, 2$.

Table 3.1: Results of 2D image simulations. Reported are the average sensitivity, specificity, and FDR, with standard error in the parenthesis, based on 100 data replications. Six methods are compared: the voxel-wise analysis, the region-wise analysis, the integrated method with two thresholding values, 0.95 and 0.90, and the proposed Bayesian method (TCGP) with the Gibbs sampler and the hybrid mini-batch MCMC.

Signal	Method	Positive Correlation			Negative Correlation		
		Sensitivity	Specificity	FDR	Sensitivity	Specificity	FDR
Weak	Voxel-wise	0.000 (0.000)	1.000 (0.000)	0.020 (0.010)	0.000 (0.001)	1.000 (0.001)	0.010 (0.001)
	Region-wise	0.238 (0.001)	0.953 (0.002)	0.447 (0.002)	0.473 (0.002)	0.956 (0.003)	0.629 (0.004)
	Integrated (0.95)	0.612 (0.001)	0.994 (0.000)	0.134 (0.010)	0.844 (0.003)	0.993 (0.000)	0.131 (0.003)
	Integrated (0.90)	0.821 (0.001)	0.971 (0.000)	0.341 (0.010)	0.963 (0.003)	0.966 (0.000)	0.398 (0.006)
	TCGP (Gibbs)	0.855 (0.003)	0.996 (0.001)	0.057 (0.008)	0.997 (0.002)	0.993 (0.001)	0.108 (0.005)
	TCGP (Hybrid)	0.851 (0.006)	0.993 (0.001)	0.092 (0.010)	0.993 (0.002)	0.992 (0.001)	0.126 (0.005)
Strong	Voxel-wise	0.062 (0.002)	1.000 (0.000)	0.000 (0.014)	0.091 (0.002)	1.000 (0.000)	0.000 (0.006)
	Region-wise	0.741 (0.002)	0.852 (0.003)	0.747 (0.004)	0.479 (0.002)	0.950 (0.002)	0.645 (0.003)
	Integrated (0.95)	0.773 (0.001)	0.998 (0.000)	0.036 (0.002)	0.933 (0.002)	0.996 (0.000)	0.067 (0.001)
	Integrated (0.90)	0.996 (0.020)	0.959 (0.000)	0.378 (0.017)	0.999 (0.020)	0.953 (0.000)	0.468 (0.001)
	TCGP (Gibbs)	0.976 (0.002)	0.999 (0.000)	0.015 (0.004)	1.000 (0.001)	0.999 (0.000)	0.018 (0.001)
	TCGP (Hybrid)	0.960 (0.003)	0.997 (0.001)	0.049 (0.005)	0.990 (0.001)	0.999 (0.000)	0.023 (0.002)

To apply the proposed method, we employ the Matérn kernel in (3.4) in our data analysis. We set the prior hyperparameters $a_\tau = b_\tau = 0.001$ to obtain a non-informative prior, and choose a_ω and b_ω adaptively as the minimum and the maximum of $|\xi(v)|_{v\in\mathcal{B}_m}$, respectively, from each iteration. We run the Gibbs sampler for 1000 iterations, with the first 200 iterations as the burn-in. We also run the hybrid mini-batch MCMC for 1200 iterations, with the first 400 iterations as the burn-in. We claim a voxel having a nonzero correlation by simply thresholding the posterior inclusion probability at 0.5, an approach commonly used in Bayesian analysis. We also compare with a number of alternative solutions, including the voxel-wise analysis, the region-wise analysis, and the integrated analysis method of Li et al. (2019) with two different thresholding values, 0.90 and 0.95, following the analysis in Li et al. (2019). We evaluate the performance of each method by the sensitivity, specificity, and false discovery rate (FDR).

Table 3.1 reports the results averaged over 100 data replications, and Figure 3.2 visualizes the result for one data replication. We see that our proposed method clearly outperforms the alternative solutions. In particular, the voxel-wise analysis suffers from a low detection power, the region-wise analysis yields a high false discovery rate, and the integrated method of Li et al. (2019) is sensitive to the thresholding parameter. With the 90% threshold, the integrated method enjoys a better sensitivity and specificity, but yields a larger FDR, whereas with the 95% threshold, it can well control the FDR, is not as powerful. In addition, the proposed Bayesian method is also capable of statistical inference, in that we can simulate the entire posterior distribution, compute the posterior inclusion probability, and quantify the uncertainty for the spatially varying correlation. Figure 3.3 shows the

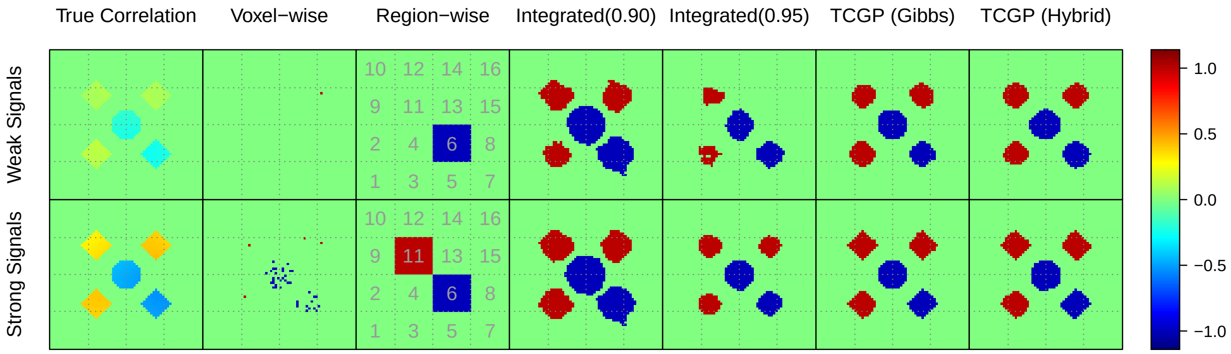


Figure 3.2: Results of 2D image simulations. The first row is for a weak signal and the second row a strong signal. The panels from left to right show the true correlation map, the significantly positively (red) and negatively (blue) correlated regions selected by different methods.

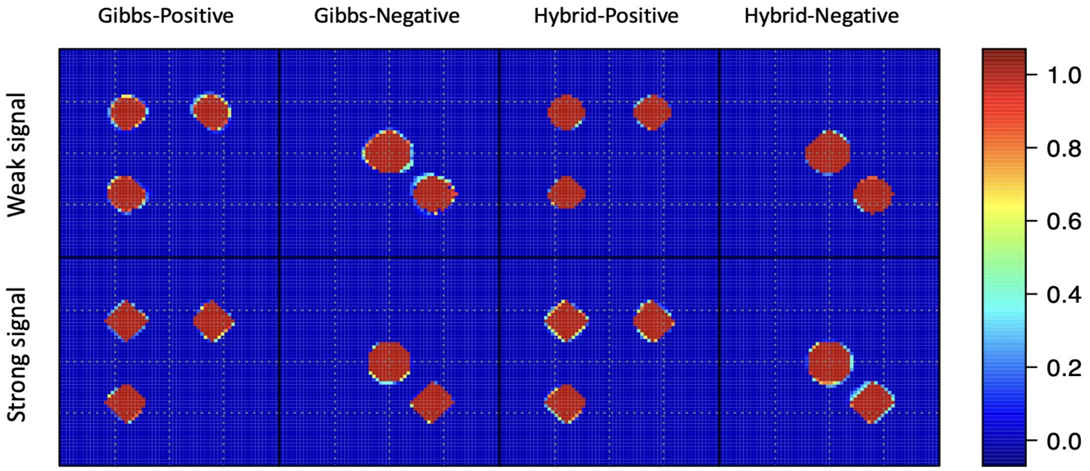


Figure 3.3: Results of 2D image simulations. The posterior inclusion probability map of the positive and negative spatially-varying correlations using the Gibbs sampler and the hybrid mini-batch MCMC.

probability map of the identified positively and negatively correlated regions, which are close to the truth. Finally, we briefly remark on the computational time of the two Gibbs samplers. On a laptop with 2 cores, 3.1GHz clock speed and 8GB memory, the Gibbs sampler algorithm took about 90 minutes for one data replication, while the hybrid algorithm took about 30 minutes, with the mean acceptance ratio around 0.3.

Table 3.2: Simulation results of the 3D image example. Reported are the average sensitivity, specificity, and FDR, with standard error in the parenthesis, based on 100 data replications. Six methods are compared: the voxel-wise analysis, the region-wise analysis, the integrated method with two thresholding values, 0.95 and 0.90, and the proposed Bayesian method with the Gibbs sampler and the hybrid mini-batch MCMC.

Signal	Method	Positive Correlation			Negative Correlation		
		Sensitivity	Specificity	FDR	Sensitivity	Specificity	FDR
Weak	Voxel-wise	0.084 (0.003)	0.999 (0.005)	0.001 (0.006)	0.101 (0.005)	0.999 (0.000)	0.002 (0.003)
	Region-wise	0.357 (0.001)	0.865 (0.002)	0.573 (0.011)	0.472 (0.002)	0.891 (0.003)	0.451 (0.004)
	Integrated(0.95)	0.489 (0.002)	0.981 (0.001)	0.160 (0.010)	0.582 (0.001)	0.952 (0.005)	0.100 (0.001)
	Integrated(0.90)	0.663 (0.008)	0.959 (0.001)	0.230 (0.009)	0.731 (0.004)	0.946 (0.005)	0.150 (0.001)
	TCGP (Gibbs)	0.891 (0.005)	0.984 (0.001)	0.071 (0.007)	0.883 (0.002)	0.977 (0.004)	0.073 (0.001)
	TCGP (Hybrid)	0.878 (0.002)	0.977 (0.002)	0.083 (0.009)	0.869 (0.005)	0.965 (0.003)	0.089 (0.002)
Strong	Voxel-wise	0.210 (0.005)	0.999 (0.002)	0.001 (0.001)	0.237 (0.004)	0.999 (0.000)	0.002 (0.001)
	Region-wise	0.638 (0.003)	0.765 (0.001)	0.587 (0.010)	0.627 (0.006)	0.824 (0.005)	0.532 (0.003)
	Integrated(0.95)	0.553 (0.005)	0.992 (0.000)	0.066 (0.005)	0.882 (0.005)	0.970 (0.000)	0.101 (0.002)
	Integrated(0.90)	0.746 (0.010)	0.974 (0.003)	0.144 (0.007)	0.933 (0.010)	0.955 (0.001)	0.133 (0.000)
	TCGP (Gibbs)	0.933 (0.002)	0.986 (0.001)	0.062 (0.005)	0.929 (0.002)	0.987 (0.001)	0.061 (0.002)
	TCGP (Hybrid)	0.920 (0.003)	0.974 (0.001)	0.075 (0.004)	0.918 (0.005)	0.967 (0.001)	0.085 (0.001)

3.5.2 3D image simulation

We next consider a $d = 3$ example that mimics the HCP data analyzed in Section 3.6. More specifically, we obtain the posterior means of $c_l, e_{i,l,\pm}, \omega$ from our HCP data analysis, then generate Y_+, Y_- following model (3.11). We continue to employ the kernel function $\kappa(v, v') = \exp -0.1(v^2 + v'^2) - 10(v - v')^2$ for data generation. We simulate the noise $\varepsilon_{k,i}$ from the normal distribution with mean zero and variance $\tau_k^2(v)$, and simulate $\log(\tau_k^2(v))$ from a Gaussian process with mean zero and correlation kernel $\zeta_k \kappa(v, v')$, $k = 1, 2$. We consider two noise levels, or equivalently the signal strengths, with $\zeta = 10$ for a weak signal, and $\zeta = 1$ for a strong signal. We follow the HCP data and set the sample size $n = 904$ and the image resolution $91 \times 109 \times 91$ with $m = 117,293$ voxels in the brain region. Table 3.2 reports the results averaged over 100 data replications, and Figure 3.4 visualizes the result for one data replication. We observe essentially the same patterns as in the 2D example. Besides, in Figure 3.4, we only show the result for the positively correlated regions, as the result for the negative correlated regions is very similar.

3.6 Analysis of HCP Data

In this section, we further illustrate our method with an fMRI dataset from the Human Connectome Project. Our specific goal is to study the association between the resting-state fMRI and the memory task-related fMRI, and identify brain regions where the resting-state and task-related brain activities are strongly associated. This type of analysis is useful, as there has been increasing interest in recent years to predict task-related brain activations from resting-state fMRI (Tavor et al.

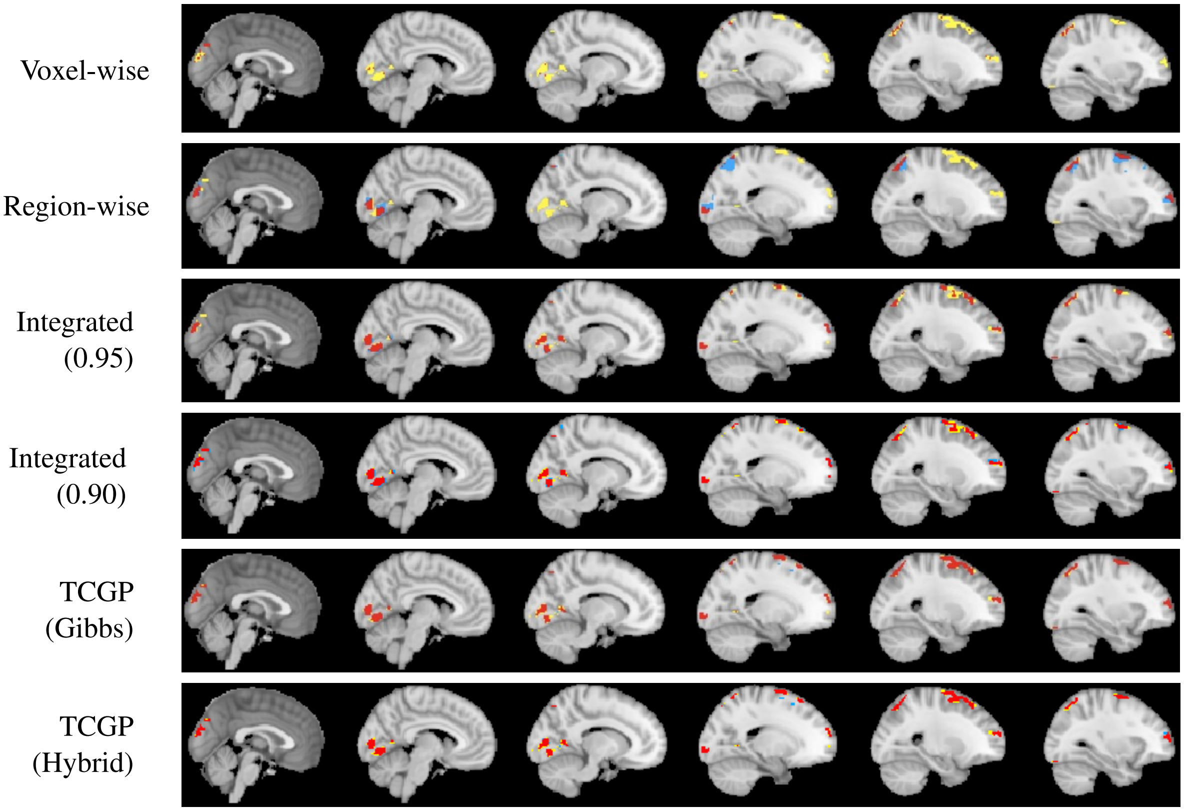


Figure 3.4: Simulation results of the 3D image example. The 2D slices of positively correlated regions identified by the voxel-wise analysis, the region-wise analysis, the integrated method with two thresholding values, 0.95 and 0.90, and the proposed Bayesian method with the Gibbs sampler and the hybrid mini-batch MCMC. The red, yellow and blue regions represent the true positive, the false negative, and the false positive regions, respectively.

2016, Jones et al. 2017, Cohen et al. 2020). It also reveals numerous brain regions and offers useful insights to understand brain activities during rest and working memory tasks.

The dataset we analyze consists of $n = 904$ subjects with both resting-state and task fMRI scans. We preprocess both types of images following the usual pipelines. In particular, the preprocessing of resting-state fMRI includes correction for distortions and head motion, removal of slowest temporal drifts and structured non-neuronal artifact (Smith et al. 2013). The resulting data is the 3D fractional amplitude of low frequency fluctuation (fALFF) image, which quantifies the amplitude of the low frequency oscillations in fMRI signals to reflect the local brain activities at resting state. The preprocessing of task fMRI includes gradient unwarping, motion correction, distortion correction, and grand-mean intensity normalization (Barch et al. 2013). The resulting data is the 3D volumetric image. In addition, we regress out potential confounding variables of age and sex using the image-on-scalar approach of Zhu et al. (2014b). Finally, we register and align both images to the standard MNI space (Mazziotta et al. 2001), and the resulting image resolution is $91 \times 109 \times 91$ with $m = 117,293$ voxels in the brain region.

Table 3.3: Results of HCP data analysis. Reported are the activation regions containing more than 100 voxels that are declared having a nonzero correlation.

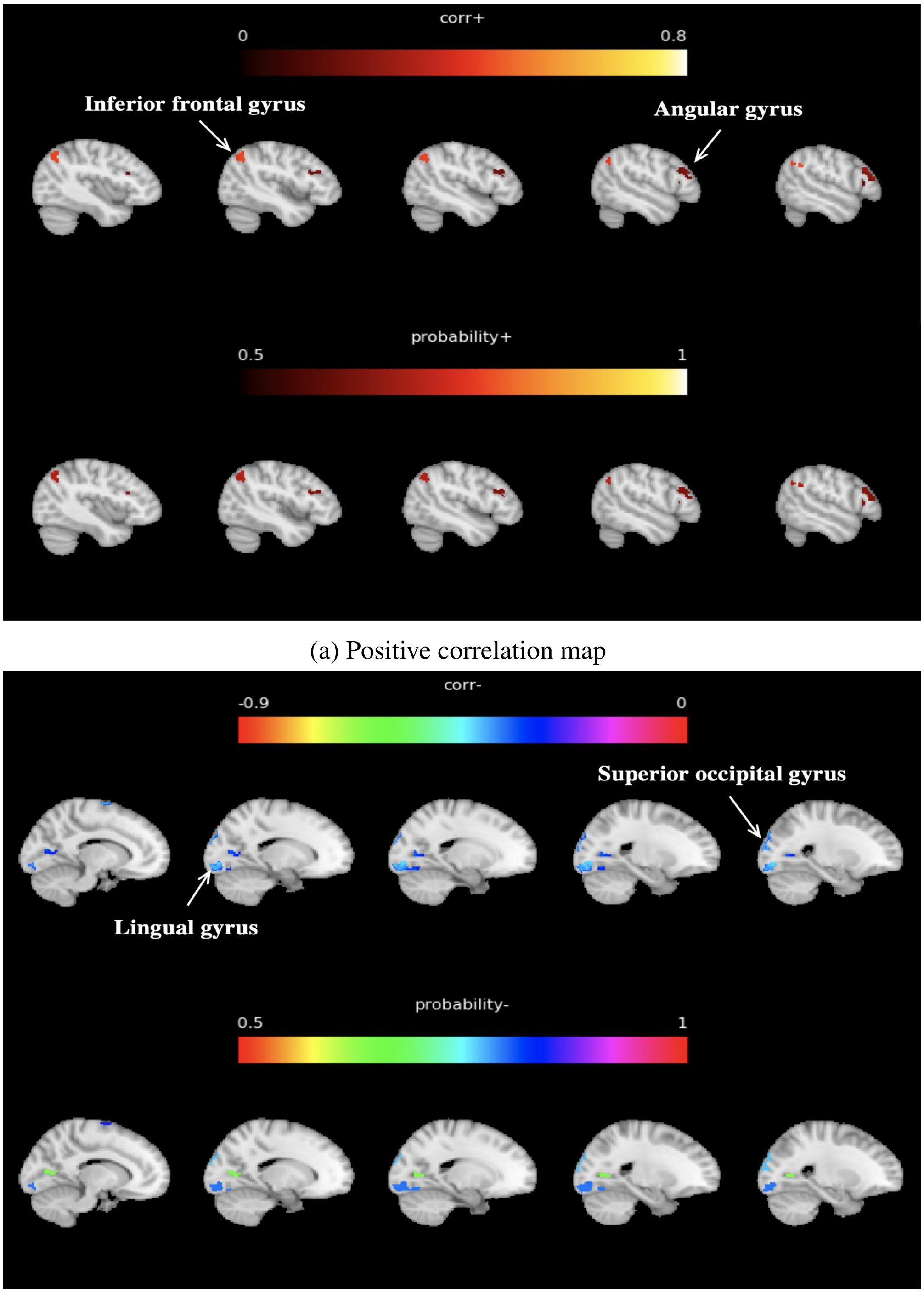
Regions with positive correlations

AAL Regions	Cluster Size	Activation Center	Mean	Std.	PIP
Precentral-L	385	(45.4, 6.75, 42.95)	0.39	0.07	0.79
Frontal-Sup-R	141	(-25.6, 60.5, 19.2)	0.30	0.05	0.64
Frontal-Sup-R	329	(-26.4, 8.0, 65.2)	0.35	0.08	0.62
Frontal-Mid-L	643	(33.7, 32.9, 42.2)	0.35	0.04	0.59
Frontal-Inf-Tri-R	218	(-51.9, 28.0, 22.6)	0.21	0.06	0.64
Calcarine-R	200	(-13.8, -87.4, 3.64)	0.37	0.05	0.69
Cuneus-L	120	(-0.4, -87.4, 22.6)	0.33	0.04	0.65
Lingual-R	144	(-10.4, -75.3, -4.5)	0.35	0.05	0.62
Parietal-Sup-L	187	(20.0, -67.4, 53.7)	0.35	0.05	0.53
Parietal-Sup-L	108	(26.0, -52.8, 62.4)	0.40	0.06	0.58
Parietal-Sup-R	165	(-29.2, -20.9, 68.3)	0.30	0.08	0.76
Parietal-Inf-L	253	(47.2, -46.1, 49.7)	0.40	0.05	0.59
Angular-R	209	(-46.9, -60.2, 44.7)	0.43	0.03	0.70
Temporal-Sup-L	331	(54.3, -31.8, 18.0)	0.40	0.05	0.82
Temporal-Mid-L	104	(63.1, -25.7, 1.38)	0.41	0.07	0.56

Regions with negative correlations

AAL Regions	Cluster Size	Activation Center	Mean	Std.	PIP
Precentral-L	115	(28.6, -23.1, 65.4)	-0.44	0.03	0.90
Precentral-R	183	(-54.4, 8.0, 36.0)	-0.40	0.08	0.59
Frontal-Mid-L	191	(28.2, 52.2, 12.7)	-0.39	0.06	0.78
Rolandic-Oper-L	186	(-45.6, -14.5, 15.9)	-0.36	0.14	0.58
Supp-Motor-Area-L	120	(1.1, -7.9, 66.1)	-0.36	0.05	0.71
Supp-Motor-Area-R	143	(-6.9, -13.3, 69.5)	-0.38	0.07	0.85
Calcarine-R	183	(-15.4, -68.8, 10.5)	-0.32	0.06	0.65
Lingual-L	292	(10.5, -75.0, -5.5)	-0.28	0.04	0.79
Lingual-R	286	(-21.2, -86.3, -9.0)	-0.38	0.05	0.80
Occipital-Sup-L	111	(16.0, -89.8, 25.0)	-0.40	0.06	0.80
Occipital-Sup-R	147	(-25.2, -89.9, 26.2)	-0.37	0.04	0.77
Occipital-Inf-R	122	(-38.8, -81.7, -3.2)	-0.44	0.05	0.56
SupraMarginal-L	191	(58.8, -25.7, 30.8)	-0.37	0.04	0.83
SupraMarginal-R	121	(-58.2, -36.9, 28.3)	-0.38	0.03	0.98
Paracentral-Lobule-R	147	(-5.7,-30.5, 70.4)	-0.33	0.04	0.63
Temporal-Mid-L	109	(58.6, -36.3, 7.7)	-0.43	0.05	0.55

To apply the proposed methods to this data, we employ the Matérn kernel in (3.4) in our data analysis. We set the prior hyperparameters $a_\tau = 0.001$, $b_\tau = 0.001$, and choose a_ω and b_ω as the 75% quantile and 100% quantile of $\{|\xi(v)|\}_{v\in\mathcal{B}}$, respectively. The choice of a_ω is based on the belief that at most 25% voxels have non-zero correlations. We perform sensitivity analysis on the small changes of prior specfiication for the threshold parameters in Section B.4.3 of the Supplementary Material. We run the Gibbs sampler for 1000 iterations, with the first 200 iterations as the burn-in. We also run the hybrid mini-batch MCMC for 1200 iterations, with the first 400 iterations as the



(a) Positive correlation map

(b) Negative correlation map

Figure 3.5: Results of HCP data analysis. The sagittal slices of activation regions with significant correlations. Panel (a) shows the slices of positive correlation map and the associated inclusion probability map. Panel (b) shows the slices of negative correlation map and the associated posterior inclusion probability map.

burn-in. We again claim a voxel having a nonzero correlation by simply thresholding the posterior inclusion probability at 0.5.

Table 3.3 summarizes the estimated activation regions with strong positive or negative correlations. Here we only report those regions containing more than 100 voxels that are declared having a nonzero correlation. We also map those estimated regions to the automatic anatomical labeling (AAL) brain atlas, and report where each estimated activation region is located, the cluster size, the activation center coordinates, the mean and the standard deviation of the correlation in a specific cluster, and the posterior inclusion probability. We make the following observations. We identify a region in angular gyrus that has the highest positive mean correlation. This finding agrees with the literature, as intensive research has shown that angular gyrus is involved in cognitive processes related to language, number processing, spatial cognition, memory retrieval, and attention (Farrer et al. 2008, Seghier 2013). We also identify a region with strong positive correlations in middle temporal gyrus and superior parietal gyrus. The former region is connected with numerous cognitive processes including recognition of known faces, audio-visual emotional recognition, and accessing word meaning while reading (Acheson and Hagoort 2013), and the latter is critically involved in information manipulation in working memory (Koenigs et al. 2009). In addition, we identify two regions in lingual gyrus with strong negative correlation, while lingual gyrus is believed to play an important role in visual memory and word processing (Leshikar et al. 2012). Figure 3.5 shows the identified activation regions with significant correlations.

CHAPTER 4

Bayesian Time-varying Classification with Signal Interactions via the Relaxed Thresholded Gaussian Process

4.1 Introduction

Brain-computer interfaces (BCIs) are “speech translators” between the human brain and computers. These signal processing systems acquire brain signals through noninvasive electroencephalography (EEG) with the help of scalp electrodes, usually in the 10-20 configuration. Among various EEG signals, the event-related potentials (ERPs) refer to a set of brain signals that are generated in response to some external stimuli such as visual, auditory, or somatosensory (Farwell and Donchin 1988). Its popularity is aided by its tremendous application possibilities in movement and communication assistance, especially for rehabilitation of the disabled. In the health care sector, it can assist partially paralyzed people such as those who have diseases like Amyotrophic Lateral Sclerosis (ALS) and stroke. Although it was originally developed for people with disabilities (Pfurtscheller et al. 2008), the prevailing spectrum of use has been expanded to competent users (Van Erp et al. 2012), neuro-rehabilitation (Kwak et al. 2015), etc.

To better illustrate the framework, we briefly introduce the motivating dataset following the experimental protocol by (Thompson et al. 2014). The EEG data is collected from the P300 speller study conducted by the University of Michigan direct brain interface (UMDBI) laboratory. The goal of the experiment is to infer the participant’s intended character on a 6 × 6 virtual keyboard using the electroencephalography (EEG) signal. Before the experiment, the subjects were told that a specified character in the visual stimulator was the target character. During the experiment, the subjects were asked to keep an eye on the target character position in the visual stimulator, while any row or column in the visual stimulator flashed randomly. When the target character’s row or column was flashing, a positive potential (P300 ERP) related to the event could be detected in the subject’s scalp; if not, the detected EEG data were non-P300 event-related potentials. Each event was either a row stimulus or a column stimulus. Rows and columns of the keyboard flash in a random order, and it looped through all rows and columns every consecutive 12 stimuli, called a sequence. Thus, each sequence always had two events (stimuli) that were supposed to elicit P300 ERPs (one row and one

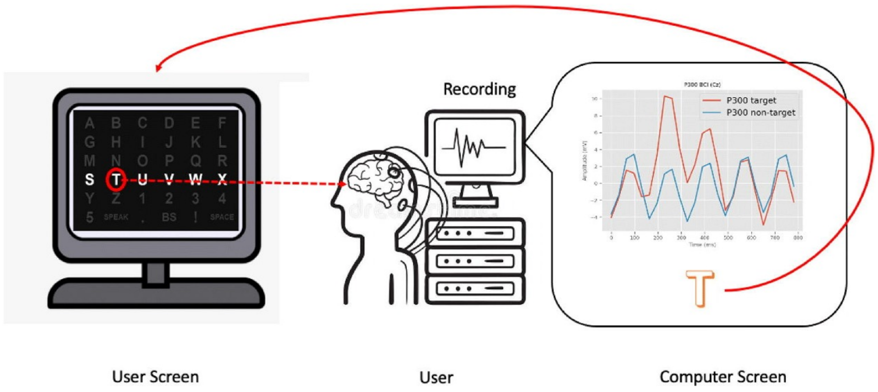


Figure 4.1: An illustration of the conventional procedure of the P300 ERP-BCI operation. The P300 ERP-BCI design presents a sequence of events on a virtual screen to the user. The user focuses on a specific character and responds to different events with different brain signals (P300 or non-P300). These brain signals are recorded by the EEG machine. Classifiers are then constructed to analyze EEG signals in a fixed time response window after each event to make a binary decision whether a P300 ERP response is produced. Finally, the binary classification results are converted into character-level probabilities, and the character with the highest probability is shown on the screen.

column) out of every 12 events. In particular, Figure 4.1 shows 36 characters in a 6×6 grid with the fourth row being highlighted.

The most-studied, and foundation computational neuroscience challenge in ERP BCIs is the classification of the brain activity after each stimulus as either a target or non-target response. Correct binary classification of the brain activity after each stimulus enables identification of the stimulus groups of interest, and at their intersection, the target key. The original work (Farwell and Donchin 1988) developed four classification methods, stepwise linear discriminant analysis (SWLDA), peak picking, area, and covariance, with the best performance achieved by SWLDA with 95% accuracy. Subsequent research aimed to improve P300 Speller performance using various techniques, including independent component analysis (Xu et al. 2004) and support vector machine (SVM) (Kaper et al. 2004). Philip and George (2020) did a comprehensive review of the articles related to classification methods, finding that most of the methods fall into three categories: ensemble learning, SVM, and discriminant analysis. The paper highlighted the superiority of ensemble learning methods as they take the advantages of different classifiers while accurately classifying the imbalanced P300 dataset. Researchers have also explored the use of Bayesian methods in BCI classification. For example, Zhang et al. (2015) introduced a sparse Bayesian method by exploiting Laplace priors for EEG classification. Barthélemy et al. (2023) proposed a novel Bayesian accumulation of Riemannian probabilities, which is an end-to-end pipeline for P300 BCI classification. Ma et al. (2022) made the first attempt to study the probability distribution of multi-trial EEG signals using a Bayesian generative model, which provides a useful tool to simulate

EEG signals in P300 BCI and a novel probabilistic classifier.

Despite the success of these methods, they mostly rely on EEG signals on each channel but ignore the possible functional relationships between different signals recorded from distinct brain areas. Several existing studies have demonstrated that brain functions require the integration of distributed brain areas (Tononi and Edelman 1998, Friston et al. 1997). In the BCI context, Kabbara et al. (2016) also shows a clear difference between functional networks obtained in the case of target and non-targets visual stimuli. Hence, the signal interactions among channels involved in the brain network is an important feature in BCI studies and their association to a stimulus type outcome offers another potential to contribute to the predictive mechanism.

For modeling signal interactions, two-step methods are commonly used by first identifying main effects and then refitting the model with both main effects and their interaction effects (Hao et al. 2018, Wang et al. 2021). However, this approach may not be suitable for EEG analysis since the selection of the interaction effect is not based on the existence of a main effect. Recent works under Bayesian paradigms have attempted to place both main and interaction terms in one inference system via hierarchical shrinkage priors (Griffin and Brown 2017), but developing specific priors to capture the structure of EEG data is necessary for this approach to be effective. More recently, several neural network methods have been proposed for EEG-based classification tasks. For example, multi-task autoencoder-based models such as Ditthapron et al. (2019), compact CNNs like EEGNet Lawhern et al. (2018), and weighted ensemble strategies such as Kshirsagar and Londhe (2019) have shown promising results. These methods have been able to implicitly capture the interaction effects among different channels and their association to a stimulus type outcome. However, they often require task-specific knowledge to design the network architecture, and the amount of data used to train these networks varied significantly across studies. Furthermore, these methods may suffer from the black-box nature of neural networks, which limits their interpretability. In other words, it may not be possible to identify which channel pairs are most contributing to the stimulus type output (i.e., target or non-target classification). This makes it challenging to gain insight into the underlying neural mechanisms and limits the generalizability of the models. Therefore, there is a need to develop methods that can incorporate the interaction effect among EEG channels while maintaining interpretability and generalizability of the model.

In this paper, we developed a Bayesian time-varying classification model with signal interactions via relaxed thresholded Gaussian Process priors (SI-RTGP). We remove the linearity constraint among EEG signal predictors by including signal interaction effects across different channels for an enhanced prediction and interpretation. To the best of our knowledge, we are among the first to explicitly incorporate the signal interaction across different channels into the EEG prediction model. We also propose a relaxed thresholded Gaussian process prior to capture the association between EEG signal and the stimulus type outcomes, which has several advantages compared to the existing work. First, the proposed relaxed thresholded Gaussian Process (RTGP) prior encompasses a large class of temporal varying functions that are piecewise smooth and sparse, which enables the feature selection during the Bayesian MCMC sampling. Second, compared to the previous thresholded Gaussian Process (GP) prior such as the soft (Kang et al. 2018) and the hard tresholded GP (Cai et al. 2020), the relaxed thresholded GP prior is more flexible and is able to model both the sparse and the non-sparse patterns by varying the "relaxing" parameter. Also, it provides a computationally efficient way to conduct MCMC sampling while other thresholded GP priors may raise a challenge in posterior computation when handling large-scale dataset. We applied the proposed model to the P300 speller study conducted by the University of Michigan direct brain interface (UMDBI) laboratory. The proposed SI-RTGP model can improve the classification accuracy on several subjects and can identify a number of scientifically meaningful channels and channel pairs that offer useful insights for BCI research.

Our contribution are of several folds. First, we incorporate the signal interaction into the prediction model to improve the prediction accuracy. Our model jointly identifies both main and interaction effects within a single Bayesian inference framework, avoiding potential misspecification from a two-step approaches. Compared with neural network model that implicitly add interaction effect, the proposed model is more interpretable and can explicitly identify the important channel pairs. Second, we introduce a new selection prior model based on relaxed thresholded Gaussian Process (RTGP) priors, which takes into account the temporal correlation of the EEG signal on each channel through Gaussian Process and ensures the piecewise smoothness and sparsity of the effect of EEG signal. Moreover, the proposed RTGP prior is more flexible than other thresholded Gaussian Process priors and ensures an efficient posterior sampling process by introducing a "relaxing" parameter.

4.2 Method

4.2.1 Bayesian time-varying classification model with signal interactions

Our model focuses on the multi-channel EEG data for one participant. Suppose a total of R target characters are typed for BCI calibration in the training data. For each character $r(r = 1, ..., R)$, the BCI generates S sequences of $J(J = 12)$ stimuli consisting of six row stimuli, denoted as $1, \cdots, 6$ and six column stimuli, denoted as $7, \cdots, 12$ on the 6×6 keyboard in a random order. Let $I = \{(r, s, j) \mid r = 1, \ldots, R; s = 1, \ldots, S; j = 1, \ldots, J\}$. Hence, there are $n = |I| = R \times S \times J$ flashes in total. Let $i \in I$ index the flash. In practice, people call a random presentation 6 rows and 6 columns a sequence, and run multiple sequences for a single character. Multiple scores for each row and column are averaged to increase the accuracy.

During the BCI calibration stage, each participant was asked to wear an EEG cap with $K(K =$

$16)$ channels corresponding to different regions on the brain surface. Let T be the number of time points of the EEG signals we collect for each stimulus. Let $X_{ki}(t)$ be the observed EEG signal intensity of the i_{th} stimulus from channel k at time t. $\mathbf{X}_{ki} = \{X_{ki}(t)\}_{t=1}^{T} \in \mathbb{R}^T$ and $\mathbf{X}_i = \left(\mathbf{X}_{1i}^\top, \mathbf{X}_{2i}^\top, \cdots, \mathbf{X}_{Ki}^\top\right)^\top \in \mathbb{R}^p$, where $p = K \times T$. We use $X_{0i}(k_1, k_2)$ to represent the signal interaction between the EEG signal on channel k_1 and that on channel k_2, where $1 \leq k_1 < k_2 \leq K$. In this paper, the signal interaction $X_{0i}(k_1, k_2)$ is defined as the Z-transformed correlation between $\mathbf{X}_{k_1 i}$ and $\mathbf{X}_{k_2 i}$. Let $\mathbf{X}_{0i} = \{X_{0i}(k_1, k_2)\}_{1 \leq k_1 < k_2 \leq K} \in \mathbb{R}^{p_0}$, where $p_0 = \binom{K}{2}$. Finally, let $Y_i \in \{0, 1\}$ be the stimulus type outcomes of thr i_{th} flash. A graphical illustration is shown in Figure 4.2.

To build a model that uses the EEG signal and signal interactions to predict the stimulus type outcome, we propose a Bayesian time-varying classification model with signal interactions as follows:

$$
\begin{aligned}
\Pr(Y_i = 1 \mid \mathbf{X}_i, \mathbf{X}_{0i}) &= \Phi(\mu_i), \\
\mu_i = \frac{1}{p} \sum_{k=1}^{K} \left(\sum_{t=1}^{T} \beta_k(t) X_{ki}(t) \right) &+ \frac{1}{p_0} \sum_{k_1 < k_2} \beta_0(k_1, k_2) X_{0i}(k_1, k_2),
\end{aligned} \tag{4.1}
$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $\beta_k(t)$ is a time-varying coefficient function on channel k and $\beta_0(k_1, k_2)$ quantifies the effect of the signal interaction between the EEG signal on channel k_1 and that on channel k_2. The role of the rescaling factor $1/p$ and $1/p_0$ are to rescale the total effects of massive predictors such that they are bounded away from infinity with large probability, when K and T are very large. Model (4.1) uses the signal on each channel and signal interactions across channels as predictors to model stimulus type outcomes.

The model is trained using the data collected from the BCI calibration stage. Then for the BCI working stage, given an unknown character $r^\star$, the binary classification probability of each flash $\hat{p}_i, i \in I_{r^\star}$ can be obtained using the trained model (4.1), where $I_{r^\star} = \{(r^\star, s, j) \mid s = 1, \ldots, S; j = 1, \ldots, J\}$. The binary classification are then converted into character-level probabilities by averaging over all the sequences of $r^\star$. Specifically, let $I_{r^\star, j} = \{(r^\star, s, j) \mid s = 1, \ldots, S\}$ be the set of index of all the sequences of the jth flash for the unknown character $r^\star$. Then $p_{r^\star, j} = \frac{1}{S} \sum\limits_{i \in I_{r^\star, j}} \hat{p}_i$, $j = 1, \ldots, 12$ represents the character-level probabilities. The predicted character is then located in the $\underset{j=1,\ldots,6}{\arg\max}\, p_{r^\star, j}$ row and $\underset{j=7,\ldots,12}{\arg\max}\, p_{r^\star, j}$ column on the keyboard. We give an illustration of this process in Figure 4.3.

Based on the following prior construction, our model can perform channel selection and explicitly identify important channel pairs.

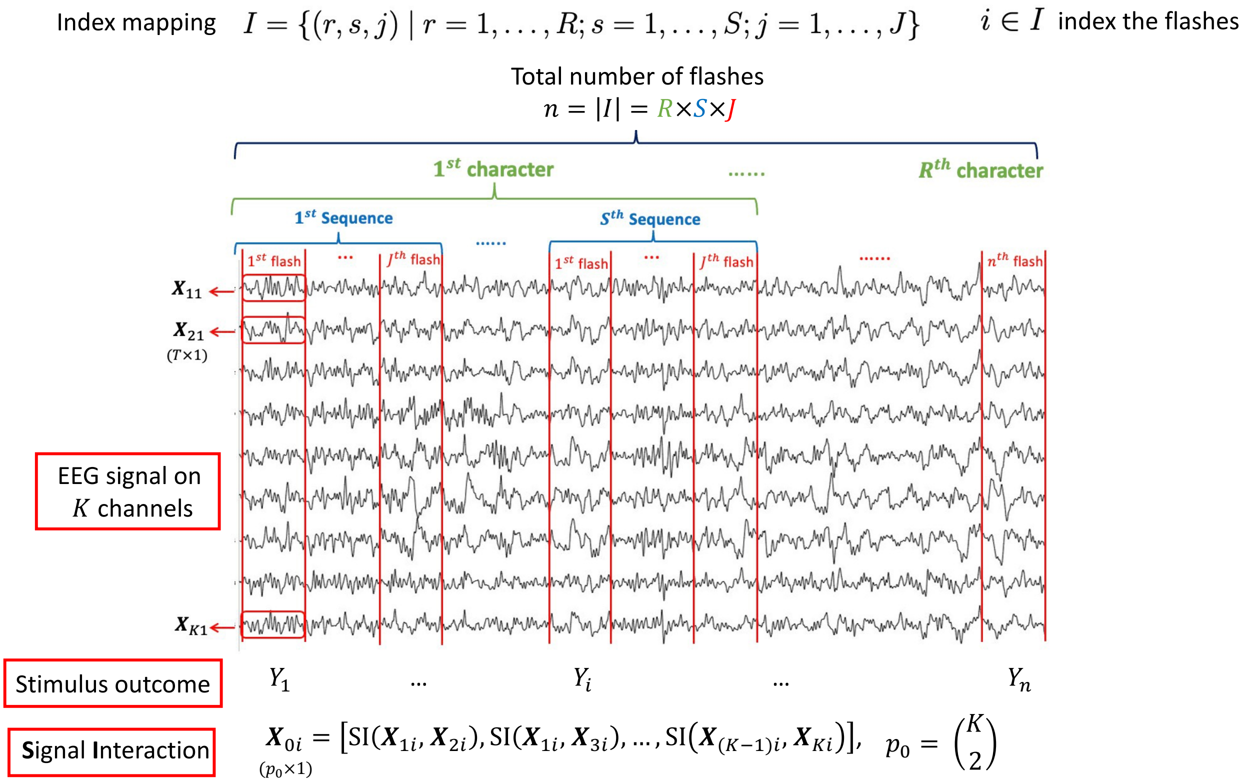

Index mapping $I = \{(r,s,j) \mid r = 1,\ldots,R; s = 1,\ldots,S; j = 1,\ldots,J\}$
$i \in I$ index the flashes
Total number of flashes
$n = |I| = R \times S \times J$
1^{st} character
......
R^{th} character
1^{st} Sequence
......
S^{th} Sequence
1^{st} flash
...
J^{th} flash
1^{st} flash
...
J^{th} flash
......
n^{th} flash
$\boldsymbol{X}_{11}$
$\boldsymbol{X}_{21}$
$(T\times 1)$
$\boldsymbol{X}_{K1}$
EEG signal on K channels
Stimulus outcome
Y_1
...
Y_i
...
Y_n
Signal Interaction
$\boldsymbol{X}_{0i} = [\mathrm{SI}(\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}), \mathrm{SI}(\boldsymbol{X}_{1i}, \boldsymbol{X}_{3i}), \ldots, \mathrm{SI}(\boldsymbol{X}_{(K-1)i}, \boldsymbol{X}_{Ki})]$, $p_0 = \binom{K}{2}$
$(p_0 \times 1)$

Figure 4.2: Illustration of the data preparation step for the SI-RTGP model. $\mathbf{X}_{ki}$ represents the EEG signal on channel k of the i_{th} flash. For the i_{th} flash, the signal interaction between channel k_1 and channel k_2, i.e. $\mathrm{SI}(X_{k_1 i}, X_{k_2 i})$, is measured by the z-transformed correlation between EEG signals on channel k_1 and that on channel k_2.

4.2.2 Relaxed thresholded Gaussian Process prior

Thresholded Gaussian Process prior can capture the sparsity and temporal dependency of the associations between stimulus type and the EEG signal predictors. However, existing thresholded GP priors such as soft-thresholded Gaussian Process in Wu et al. (2022b) and multiscale thresholded Gaussian process Shi and Kang (2015) raise a challenge in posterior computation when handling large-scale data. In order to incorporate with feature selection in Gaussian process with computation feasibility, we propose a relaxed thresholded Gaussian Process (RTGP) prior as follows.

Definition 20. *Given the kernel κ, the thresholding parameter $\omega \geq 0$ and the relaxing parameter $\xi > 0$, suppose $f(x) \sim GP(0, \kappa)$ and $\tilde{f}(x) \sim N(f(x), \xi^2)$. Let $g(x) = f(x) I(|\tilde{f}(x)| > \omega) \triangleq T_r(f, \omega, \xi^2)$, then $g(x)$ follows a relaxed thresholded Gaussian Process, denoted as $g(x) \sim RTGP(\kappa, \omega, \xi^2)$.*

Here, $I(\cdot)$ is the indicator function and T_r represents the relaxed thresholding function. The introduction of $\tilde{f}(x)$ allows for the full conditional distribution of $f(x)$ to have a conjugate, nice

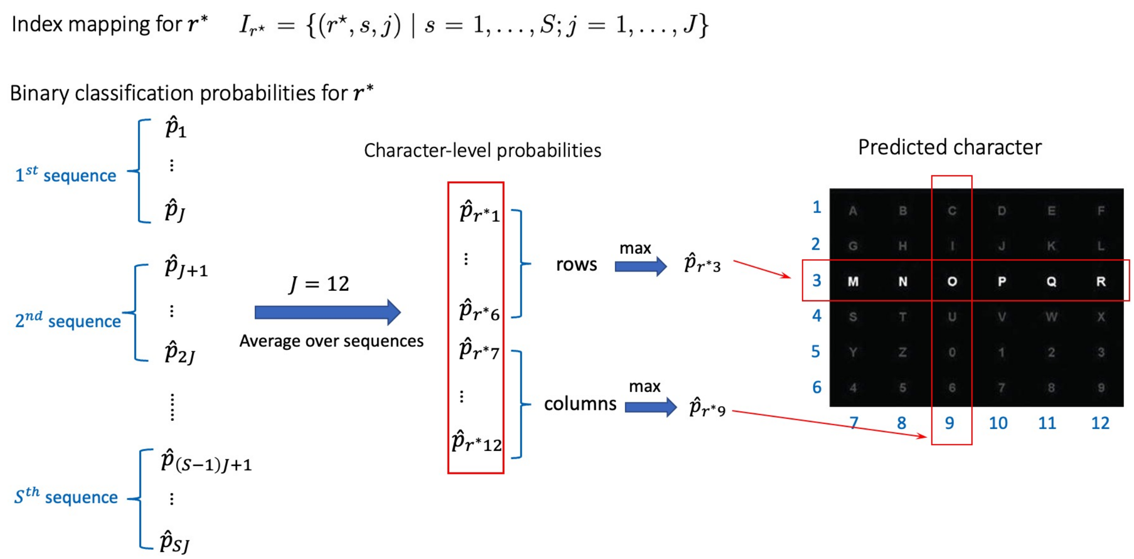


Figure 4.3: The conversion from binary classification probability to character-level probabilities. For an unknown character $r^\star$, character-level probabilities is obtained by averaging the binary classification probabilities over all the sequences. The corresponding row and column are determined by the maximum value of probabilities.

explicit closed-form, ensuring an efficient posterior sampling process. ξ^2 represents the variance of $\tilde{f}(x)$, which serves as a "relaxing" parameter to control the independent white noise added to $f(x)$. A smaller value of ξ^2 indicates more strict constraint on preserving the mean structure of $f(x)$. On the other hand, a larger value of ξ^2 provides more flexibility. To illustrate, we compare different kinds of thresholded functions in Figure 4.4, where $f(x) \sim \mathrm{GP}(0, \kappa)$, $T_s(\cdot, 0.5)$ represent the soft thresholding function, i.e. $T_s(f(x), 0.5) = 0$, if $|f(x)| < 0.5$, otherwise $T_s(f(x), 0.5) = \mathrm{sgn}\{f(x)\}(|f(x)| - \lambda)$. $T_h(\cdot, 0.5)$ denote the hard thresholding function, i.e. $T_h(f(x), 0.5) = 0$, if $|f(x)| < 0.5$, otherwise $T_h(f(x), 0.5) = f(x)$. Both of the soft and hard thresholding functions on GP can impose sparsity and piece-wise smoothness. While the hard thresholded GP has the jump discontinuity property, the soft thresholded GP is continuous. The second row in Figure 4.4 shows the plot of the proposed relaxed thresholded GP with different value of ξ. When $\xi = 0.01$, $T_r\{f(x), 0.5, 0.01\}$ almost converge to the hard thresholded GP. On the other hand, when $\xi = 0.1$, $T_r\{f(x), 0.5, 0.1\}$ is a continuous function with sparsity, which is similar to the structure of $T_s\{f(x), 0.5\}$. If we keep increasing ξ to 1, then $T_r\{f(x), 0.5, 1\}$ can recover $f(x)$ with some probability. When modeling the EEG data, we do not know what the true curve looks like. For example, if the true curve around $x = 100$ does not have sparsity, the relaxed thresholded GP has the flexibility to recover the pattern by choosing a reasonable ξ. However, the soft and hard thresholded GP both will impose sparsity with probability 1. The following proposition shows the relationship between the RTGP and other TGPs.

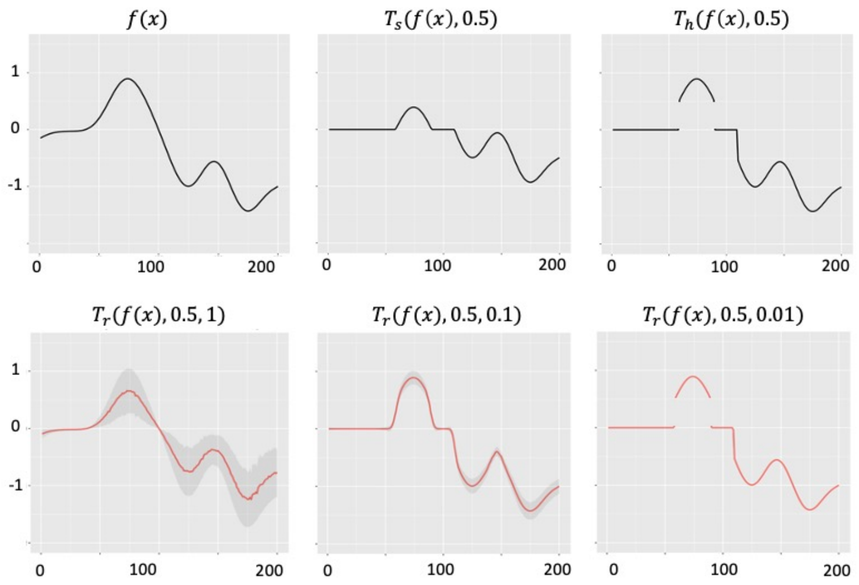


Figure 4.4: Illustration of different thresholded Gaussian Process prior. $T_s(\cdot, 0.5)$ and $T_h(\cdot, 0.5)$ represents the soft and the hard thresholding function thresholded at 0.5. $T_r(\cdot, 0.5, \xi^2)$ represents the proposed relaxed thresholding function with different value of relaxing parameter ξ^2.

Proposition 21. *Given the thresholding parameter $\omega > 0$, let $T_r(\theta, \omega, \xi^2) = \theta \cdot I(|\tilde{\theta}| > \omega)$, $T_h(\theta, \omega) = \theta \cdot I(|\theta| > \omega)$ and $T_s(\theta, \omega) = \mathrm{sgn}(\theta)(|\theta| - \omega) \cdot I(|\theta| > \omega)$ where $\theta \sim \mathcal{P}_\theta(\theta)$, then for any $\epsilon > 0$, there exist ξ^2, such that*

$$\begin{aligned} \Pr\Big(|T_r(\theta, \omega, \xi^2) - T_h(\theta, \omega)| < \epsilon\Big) > 0, \\ \Pr\Big(|T_r(\theta, \omega, \xi^2) - \theta| < \epsilon\Big) > 0, \\ \Pr\Big(|T_r(\theta, \omega, \xi^2) - T_s(\theta^\star, \omega)| < \epsilon\Big) > 0, \end{aligned} \tag{4.2}$$

where $\theta^\star = \theta + \omega$ when $\theta > 0$ and $\theta^\star = \theta - \omega$ when $\theta < 0$. Furthermore, $\lim_{\xi^2 \to 0} T_r(\theta, \omega, \xi^2) = T_h(\theta, \omega)$ and $\lim_{\xi^2 \to \infty} T_r(\theta, \omega, \xi^2) = \theta$.

Proposition 21 gives a mathematical illustration of Figure 4.4, which shows that relaxed thresholded function has certain probability to reduce to soft or hard thresholded function and the flexibility is controlled by the relaxing parameter ξ^2. The proof is shown in the supplementary material.

Given a stationary kernel κ, we assign $\beta_k(t) \sim \mathrm{RTGP}(\kappa, \omega_1, \xi^2)$. Specifically,

$$\begin{aligned} \beta_k(t) &= E_k(t) I(|\tilde{E}_k(t)| > \omega_1), \\ E_k(t) &\sim \mathrm{GP}(0, \kappa), \\ \tilde{E}_k(t) &\sim \mathrm{N}(E_k(t), \xi^2), \end{aligned} \tag{4.3}$$

where $\{E_k(\cdot)\}_{k=1}^{K}$ follow independent Gaussian Process $\mathrm{GP}(\cdot,\cdot)$ with mean 0 and stationary covariance kernel $\kappa(\cdot,\cdot)$. There are various choices for the kernel function $\kappa(\cdot,\cdot)$; for instance, we use the modified form of squared exponential (SE) covariance kernel, defined as:

$$\kappa\left(x, x^{\prime}\right)=\exp \left\{-\alpha\left(|x|+\left|x^{\prime}\right|^2\right)-\rho\left|x-x^{\prime}\right|^2\right\} \tag{4.4}$$

for $\alpha > 0, \rho > 0$, where $|\cdot|$ is the L2-norm. Here α is the decay parameter which controls the decay rate of $\mathrm{Var}\{E_k(\cdot)\}$ compared to $\mathrm{Var}\{E_k(0)\}$. The parameter ρ is the smoothing parameter, a smaller value of ρ corresponds to a smoother GP. The relaxed thresholded GP prior for $\beta_k(t)$ can incorporate the temporal information of the EEG signal, leading to potentially more accurate prediction. Moreover, it encompasses a large class of temporal varying functions that are piecewise smooth and sparse, which enable the channel selection during the Bayesian MCMC sampling.

Similarly, to quantify the effect of signal interaction across channels, we assign $\beta_0(k_1, k_2) \sim \mathrm{RTGP}(\kappa_I, \omega_2, \xi), k_1 < k_2$, where κ_I represents the identity kernel. Specifically,

$$\begin{aligned}
\beta_0(k_1,k_2) &= \eta(k_1,k_2) I(|\tilde{\eta}(k_1,k_2)| > \omega_2),\\
\eta(k_1,k_2) &\sim \mathrm{N}(0,\sigma_\eta^2),\\
\tilde{\eta}(k_1,k_2) &\sim \mathrm{N}(\eta(k_1,k_2),\sigma_\eta^2),\\
\sigma_\eta^2 &\sim \mathrm{IG}(a,b).
\end{aligned} \tag{4.5}$$

We choose to use the identity kernel here to assume the independency across channel pairs. Different kernel can be applied when other prior information is available.

Combining Model (4.1) and the prior specification in Eq. (4.3) and Eq. (4.5), we name the proposed model as Bayesian time-varying classification model with signal interactions via relaxed thresholded Gaussian Process prior (SI-RTGP). Our proposed model enjoys several benefits. First, it explicitly incorporate the signal interaction into the Bayesian probit regression model, leading to potential improvement in prediction accuracy. Second, we use RTGP prior to capture the characteristics of the EEG signal, signal interaction and their effects on the stimulus type outcomes, which is more flexible than other thresholded GP priors and can lead to a more efficient posterior computation. Third, our model can perform variable selection and explicitly identify important channels and channel pairs when predicting stimulus type outcomes, which provides more interpretable results in EEG data analysis.

4.3 Posterior Computation

In this section, we introduce the details of the posterior computation of the proposed SI-RTGP model. We use the modified form of squared exponential (SE) covariance kernel, defined in Eq.(4.4).

We first adopt the usual strategy of Karhunen-Loève expansion to simplify the model to a finite number of parameters. Specifically, consider the spectral decomposition of the kernel function,

$$\kappa\left(x, x'\right)=\sum_{l=1}^{\infty} \lambda_l \psi_l(x) \psi_l\left(x'\right),$$

where $\{\lambda_l\}_{l=1}^{\infty}$ are the eigenvalues in descending order, and $\{\psi_l(x)\}_{l=1}^{\infty}$ are the corresponding orthonormal eigenfunctions. By Mercer's Theorem, we can represent the Gaussian processes in our model as $E_k(t)=\sum_{l=1}^{\infty} e_{kl}\psi_l(t)$. where e_{kl} are Karhunen-Loève coefficients. We further truncate the above expansions by focusing on the leading L eigenvalues and eigenfunctions, where L can be determined following the usual practice of principal components analysis that retains a certain percentage of total variation. Based on the Karhunen-Loève truncation, the proposed SI-RTGP model becomes:

$$\begin{gathered}
\Pr(Y_i=1 \mid \mathbf{X}_i, \mathbf{X}_{0i})=\Phi(\mu_i), \\
\mu_i=\frac{1}{p} \sum_{k=1}^{K}\left(\sum_{t=1}^{T} \beta_k(t) X_{ki}(t)\right)+\frac{1}{p_0} \sum_{k_1<k_2} \beta_0(k_1, k_2) X_{0i}(k_1, k_2); \\
\beta_k(t)=E_k(t) I(|\tilde{E}_k(t)|>\omega_1); \quad E_k(t)=\sum_{l=1}^{L} e_{kl} \psi_l(t); \quad \tilde{E}_k(t) \sim \mathrm{N}(E_k(t), \xi^2); \\
\beta_0(k_1, k_2)=\eta(k_1, k_2) I(|\tilde{\eta}(k_1, k_2)|>\omega_2); \quad \eta(k_1, k_2) \sim \mathrm{N}(0, \sigma_\eta^2); \quad \tilde{\eta}(k_1, k_2) \sim \mathrm{N}(\eta(k_1, k_2), \xi^2).
\end{gathered} \tag{4.6}$$

Then all the parameters in our model include:

$$\Theta=\{\{\{e_{kl}\}_{l=1}^{L}, \{\tilde{E}_k(t)\}_{t=1}^{T}\}_{k=1}^{K}, \{\eta(k_1, k_2), \tilde{\eta}(k_1, k_2)\}_{k_1<k_2}, \sigma_\eta^2, \xi^2, \omega_1, \omega_2\}.$$

The priors are set as $e_{kl} \sim \mathrm{N}(0, \sigma_e^2\lambda_l)$, $\eta(k_1, k_2) \sim N(0, \sigma_\eta^2)$ and $\sigma_\eta^2 \sim \mathrm{IG}(a_\eta, b_\eta)$. We set σ_e^2 and σ_η^2 to be large values and $a=b=0.001$ so that the priors becomes non-informative prior. Following the idea of simulated annealing, we give ξ^2 a starting value and an ending value, where ξ^2 gradually decrease during the MCMC sampling process. In practice, we first set $\xi^2=1$ in the first 200 iterations during sampling and then gradually decrease the value of ξ until $\xi^2=0.0001$. As for the thresholding parameters, we set $\omega_1=\omega_2=0$ in the first 200 steps. Then we assign an adaptive discrete prior to ω_1 and ω_2, i.e. $P(\omega_1=\gamma_{1z})=1/Z$ and $P(\omega_2=\gamma_{2z})=1/Z$, $z=1,\cdots,Z$, where $\{\gamma_{1z}\}_{z=1}^{Z}$ and $\{\gamma_{2z}\}_{z=1}^{Z}$ are Z evenly spaced number between $a_\omega=0.25$ quantile and $b_\omega=0.9$ quantile of $\{|\tilde{E}_k(t)|\}_{k=1,t=1}^{K,T}$ and $\{|\tilde{\eta}(k_1, k_2)|\}_{k_1<k_2}$ respectively. We will discuss the details of the prior of ω_1 and ω_2 in Section 4.4. Given the prior specification, the full conditionals of $\{e_{k,l}\}_{k=1,l=1}^{K,L}$, $\{\tilde{E}_k(t)\}_{t=1,k=1}^{T,K}$, $\{\eta(k_1, k_2)\}_{k_1<k_2}$ and $\{\tilde{\eta}(k_1, k_2)\}_{k_1<k_2}$ are normal distributions,

thanks to the conjugate prior. The full conditionals of ω_1 and ω_2 are discrete distribution, again thanks to the conjugate prior. The detailed derivation and the sampling procedure is provided in the supplementary material.

4.4 Analysis of EEG-BCI Data

In this section, we apply the proposed method to the analysis of EEG data. We show that the proposed model can improve the prediction accuracy by leveraging the signal interactions across different channels and perform variable selection using relaxed thresholded Gaussian Process prior.

4.4.1 Dataset and preprocessing

The EEG data is collected from the P300 speller study conducted by the University of Michigan direct brain interface (UMDBI) laboratory. The goal of the experiment is to infer the participant's intended character on a 6×6 virtual keyboard using the EEG signal.

For the training session, each participant was asked to wear an EEG cap with $K = 16$ channels corresponding to different regions on the brain surface and sit approximately 0.8 m from a 17-inch monitor with the BCI display. Figure 4.5 shows the spatial distribution of channels. Channels marked with red were used for recording and analysis purposes. The abbreviated names were F3, Fz, F4, T7, C3, Cz, C4, T8, CP3, CP4, P3, Pz, P4, PO7, PO8, and Oz (Thompson et al. 2014). We defined the $J = 12$ stimuli flashing all rows and columns as a sequence and defined multiple sequences as a super-sequence. In our P300 ERP-BCI design, a super-sequence corresponded to the EEG signals associated with the given target character. During the training session, each super-sequence included $S = 15$ sequences, and a total of 19 super-sequences were collected, corresponding to 19 target charaters, "THE_QUICK_BROWN_FOX". The length of each super-sequence was about 29,000ms with the sampling rate of 256 Hz. We first apply a notch filter at 60 Hz to remove the power line noise and a band-pass filter between 0.5 Hz and 6 Hz to all 16 channels. We then extract 800 milliseconds of EEG signal after each flash. With this processing, each flash is followed by 205 time points of $K = 16$ dimensional EEG signal.

4.4.2 SI-RTGP modeling

To construct $\mathbf{X}_{ki}$, we use the first 500 milliseconds of EEG signal for each flash, i.e., $T = 128$, because the EEG signal collected in the following 300 milliseconds usually contains a lot of noise. We then calculate $X_{0i}(k_1, k_2)$, which is measured by the z-transformed correlation between the EEG signal on channel k_1 and that on channel k_2. Under this setting, we have $\{\mathbf{X}_i\}_{i=1}^n \in \mathbb{R}^{n \times p}$, $\{\mathbf{X}_{0i}\}_{i=1}^n \in \mathbb{R}^{n \times p_0}$, where $n = R \times S \times J = 19 \times 15 \times 12$, $p = K \times T = 16 \times 128$, and $p_0 = \binom{K}{2} = 120$.

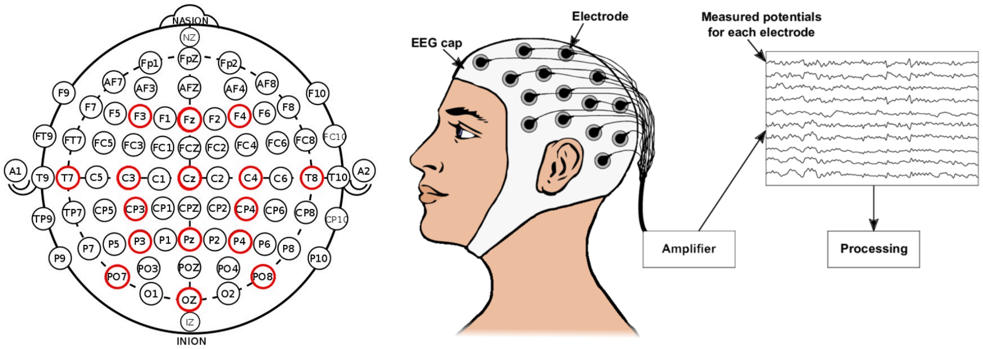


Figure 4.5: A figure from *Wikimedia Commons* by Brylie Christopher Oxley / CC0, 2017, demonstrating a 64-channel EEG locations using the International 10–20 standard. Channels marked with red were used in our ERP-BCI design.

We then apply the proposed model to EEG data analysis and perform Gibbs sampling for 2000 steps with the first 400 steps as burn in. We set $\alpha = 0.01$, $\rho = 50$ in Eq.(4.4), which is selected based on Bayes factor. We assign adaptive priors for ω_1 and ω_2. Specifically, in the first 200 steps, $\omega_1 = \omega_2 = 0$, then we assign ω_1 a discrete distribution with $P(\omega_1 = C_{1,m}) = 1/10$, $m = 1, \ldots, 10$. The value of $\{C_{1m}\}_{m=1}^{10}$ also adaptively change during MCMC sampling. In each iteration, let $C_{1,1}, \ldots, C_{1,10}$ be 10 evenly spaced number between the 0.25 and 0.90 quantile of $\{|\tilde{E}_k(t)|\}_{k=1;t=1}^{K;T}$. Similarly, the prior of ω_2 is also a discrete distribution with $P(\omega_1 = C_{2,m}) = 1/10$, $m = 1, \ldots, 10$, where $\{C_{2,m}\}_{m=1}^{10}$ are 10 evenly spaced number between the 0.25 and 0.90 quantile of $\{|\tilde{\eta}_{(}k_1, k_2)|\}_{k=1;k_1<k_2}^{K}$. We claim a node (main effect) or a edge (interaction effect) should be selected by simply thresholding the posterior inclusion probability at 0.5, an approach commonly used in Bayesian analysis. The character-level prediction accuracy is listed in Table 4.1, where we compare the proposed model with random forest (RF), support vector machine (SVM), deep neural network (DNN), SWLDA (Krusienski et al. 2008), SMGP (Ma et al. 2022) and EEGnet (Lawhern et al. 2018). The architecture of DNN is chosen such that it has the comparable number of parameters with the proposed SI-RTGP.

Since each super sequence consists of 15 sequences during the BCI training session, we apply the proposed model using $1, 2, \ldots, 15$ sequences respectively to obtain the selection results and then calculate selection reproducibility based on the frequency of each feature is selected. The results are shown in Figure 4.6. We also visualize the signal interaction effect (positive or negative) of those channel pairs with high reproducibility value in Figure 4.7.

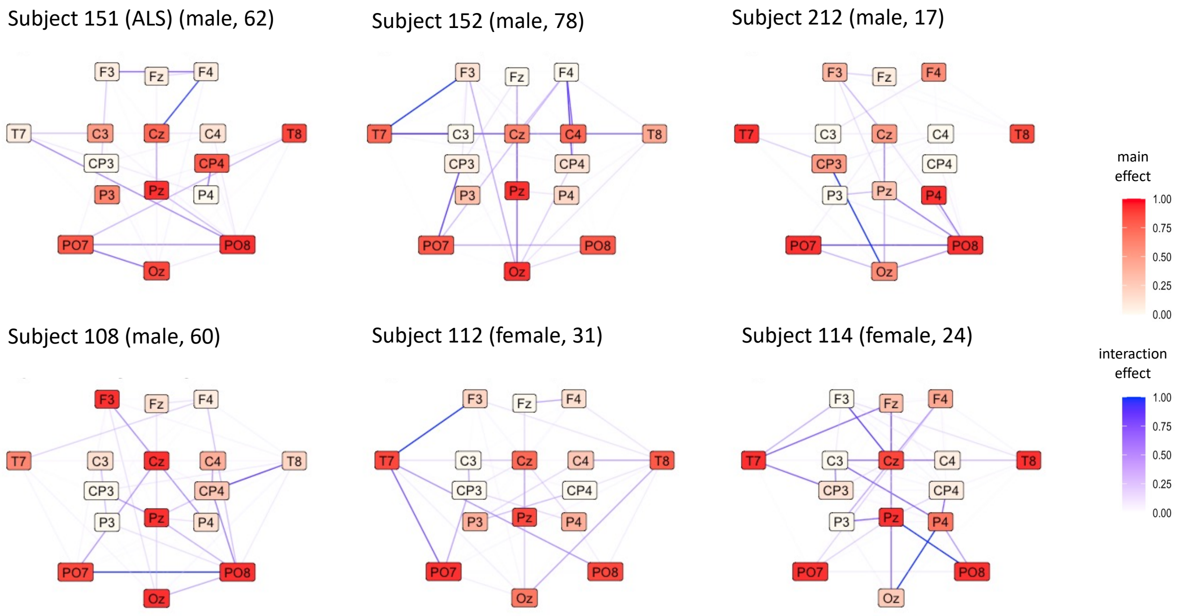
Table 4.1: Character-level prediction accuracy on test data for six subjects. The subjects' gender and age are provided. Subject 151 among these six subjects has been diagnosed with ALS.

	108	112	114	151(ALS)	152	212
	(male, 60)	(female, 31)	(female, 24)	(male, 62)	(male, 78)	(male, 17)
SI-RTGP	0.932	0.566	0.965	0.931	0.634	0.686
EEGnet	0.930	0.515	0.965	0.925	0.587	0.623
SWLDA	0.889	0.509	0.944	0.922	0.596	0.620
SMGP	0.877	0.511	0.935	0.880	0.601	0.609
DNN	0.711	0.305	0.935	0.850	0.399	0.461
RF	0.650	0.421	0.877	0.605	0.519	0.445
SVM	0.671	0.275	0.735	0.601	0.355	0.311

4.4.3 Results interpretation

Based on these results, we made the following observations. First, by leveraging the signal interactions across channels and performing feature selection using relaxed thresholded prior, the proposed methods can improve the prediction accuracy compared to other machine learning methods, which confirms that the signal interactions among channels play important role in predicting stimulus type outcomes. We use two selected channel pairs Fz-T7 and Pz-PO8 of subject 114 as an illustrative example. The mean of the EEG signal across all target or non-target flash on channel Fz and T7 is shown in Figure 4.8a, and the scatter plot is shown in Figure 4.8b, from which we can observe that the EEG signal on channel Fz and T7 have higher correlation on those target flashes. On the other hand, the EEG signal on channel Pz and PO8 have higher correlation on those non-target flashes as shown in Figure 4.9. That is consistent to the results in Figure 4.7 where the signal interaction effect of Fz-T7 is positive while that of Pz-PO8 is negative. That is to say, the correlation structure among channels are different between target and non-target flashes. Therefore, incorporating signal interaction, i.e., signal correlation information, into the prediction model can help the model differentiate target and non-target flashes more accurately, resulting in improved prediction accuracy.

Second, we identify important main effect (nodes) and interaction effect (edges) as shown in Figure 4.6 and 4.7. In terms of the main effects, PO7, PO8, Oz, Pz and Cz are important across all subjects. This result is reasonable given that the brain regions corresponding to PO7 and PO8 are involved in visual object recognition and receive processed visual information. The occipital lobe, corresponding to Oz, is primarily responsible for visual processing. The brain regions related to Pz are called the precuneus, which is involved in memory tasks, such as when people look at images and try to respond based on what they have remembered. Similar findings have been reported by Krusienski et al. Krusienski et al. (2008) and McCann et al. McCann et al. (2015). It is also worth mentioning that channel CP4 is selected with a high reproducibility value for subject 151, who has



Subject 151 (ALS) (male, 62)

Subject 152 (male, 78)

Subject 212 (male, 17)

Subject 108 (male, 60)

Subject 112 (female, 31)

Subject 114 (female, 24)

Figure 4.6: The reproducibility results of the six subjects. The subjects' gender and age are provided. Specifically, we use different number of sequences to train the model, leading to several feature selection results. The reproducibility is measured based on the frequency with which each feature was selected.

been diagnosed with amyotrophic lateral sclerosis (ALS), whereas CP4 was not as important for other healthy subjects. Regarding the signal interaction effect, the channel pairs Cz-Pz and Pz-Oz are relatively important across all subjects and the interaction between PO7 and PO8 is generally important for males. Furthermore, Cz is a hub region, indicating that it has a strong interaction effect with different nodes across subjects.

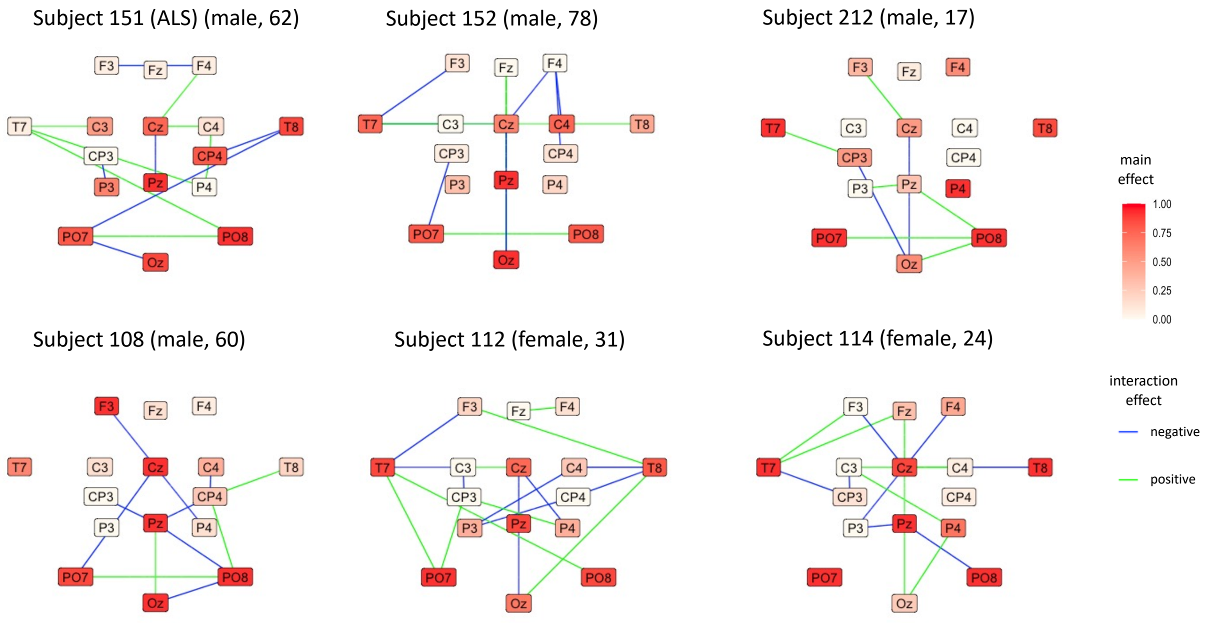
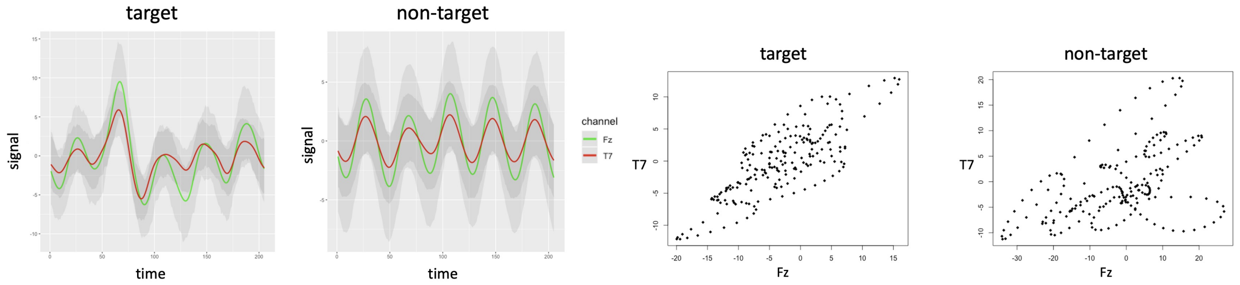


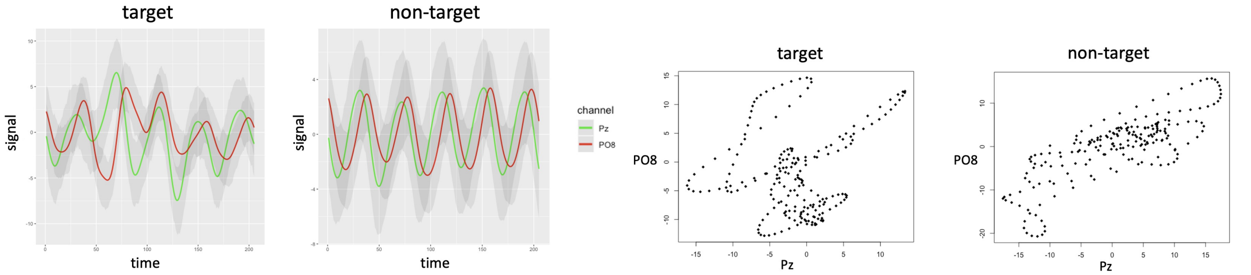
Figure 4.7: The signal interaction effect of the six subjects. The subjects' gender and age are provided. The green edges represent the positive effect ($\beta_0(k_1, k_2) > 0$) while the blue edges represent the negative effect ($\beta_0(k_1, k_2) < 0$) of the selected channel pairs when predicting stimulus type outcomes.



(a) EEG signal on channel Fz and T7

(b) Scatter plot of EEG signal on channel Fz and T7

Figure 4.8: EEG Signal comparison on channel Fz and T7. Figure 4.8a shows the overall EEG signal across all target and non-target stimulus and Figure 4.8b gives the scalar plot of EEG signal on one of the target or non-target stimulus



(a) EEG signal on channel Pz and PO8

(b) Scatter plot of EEG signal on channel Pz and PO8

Figure 4.9: EEG Signal comparison on channel Pz and PO8. Figure 4.8a shows the overall EEG signal across all target and non-target stimulus and Figure 4.8b gives the scalar plot of EEG signal on one of the target or non-target stimulus

CHAPTER 5

Conclusion and Future Work

5.1 Summary

In this dissertation, we developed several models and algorithm based on Gaussian Process, which address the challenges of analyzing large-scale and complex data in biomedical research. Conclusions drawn from this dissertation work are summarized below.

Novel GP-based models with scalability and generalizability We present several contributions to the field of GP-based methods for analyzing complex datasets. First, we have addressed the negative transfer problem in the context of an MGP and proposed two novel latent structures that can avoid negative transfer and maintain estimation in a low-dimensional parameter space. Our approach is highly scalable and allows for functional variable selection through regularization, making it a valuable tool for many real-world applications. Second, we have introduced a Bayesian framework based on two levels of hierarchical Gaussian process priors to perform multimodal correlation analysis in neuroimaging data. The proposed thresholded correlation Gaussian process prior is highly flexible and can accommodate a wide range of spatially varying functions, including those that are piecewise smooth, sparse, and jump discontinuous, which are commonly encountered in neuroimaging data. This model is quite general and can be applied to various neuroimaging correlation analysis problems. Finally, we have proposed a novel Bayesian classification model with signal interactions using the relaxed thresholded Gaussian process prior, which enables us to model the effect of signal interactions and perform variable selection during the Bayesian MCMC sampling. Our proposed approach represents a significant advancement in the field of GP-based methods and has the potential to be further applied to future neuroscience studies.

Algorithm with computational efficiency Throughout this dissertation, we have proposed several efficient algorithms based on GP. In the first project, we introduce a pairwise structure that can be parallelized to scale to an arbitrarily large N. Each sub-model is estimated with a limited number of parameters, resulting in a significant reduction in computation time. For instance, in an MGP model with $N = 50$ outputs, the original MGP model took approximately 24 hours to estimate, while the pairwise model only took ≈ 30 seconds. In the second project, we propose an efficient

MCMC sampling algorithm based on thresholded GP. Compared to the existing gradient-based MCMC algorithm, our Gibbs sampler is highly efficient, and does not require any tuning parameters. A hybrid mini-batch MCMC is also introduced to further improve computational efficiency. Finally, in the third project, we extended the thresholded GP prior to the relaxed thresholded GP, leading to a more efficient posterior computation in analyzing large-scale and complex data. Our proposed algorithms have demonstrated significant improvements in computational efficiency and can be applied to other Bayesian model with thresholding priors.

Applications in neuroimaging studies Our proposed model has been successfully applied to various real-world data analyses. One notable application is in multimodal neuroimaging analysis in the Human Connectome Project. Our goal is to study the association between resting-state fMRI and memory task-related fMRI to identify brain regions where these two brain activities are strongly associated. This type of analysis provides valuable insights into predicting task-related brain activations from resting-state fMRI, which is a useful research area that provides a "task-free" method for mapping brain functions in patients who are unable to perform tasks. We also apply our proposed model to brain-computer interface data, with the goal of inferring the participant's intended character on a 6×6 virtual keyboard using the EEG signal. Our proposed model improves the character-level prediction accuracy and identifies several important channels and channel pairs that contribute most to predicting stimulus type outcomes. These applications demonstrate the effectiveness of our proposed model and its potential for use in a wide range of fields.

5.2 Future work

Based on the findings of this dissertation, there are several potential avenues for future research.

One such direction would be to explore the use of variational inference (VI) as an alternative to Markov chain Monte Carlo (MCMC) sampling. Although MCMC sampling is asymptotically exact, it can be computationally expensive, particularly when there is no closed form of full conditional distributions. VI, on the other hand, is a machine learning technique that approximates probability densities through optimization and has been successfully applied in various domains. Unlike MCMC, VI assumes a model, which introduces some bias but also reduces the variance. Although it is generally less accurate than MCMC, it is much faster and thus better suited for large-scale statistical problems. Recently, some researchers have proposed hybrid methods that combine the advantages of both MCMC and VI, such as contrastive VI and auxiliary-likelihood MCMC. These hybrid methods have shown promising results in various applications (Ruiz and Titsias 2019, Habib and Barber 2019). By leveraging these approaches, it may be possible to reduce the computational burden and improve the efficiency of inference for the problem addressed in this dissertation, and further research in this direction could yield valuable insights and applications.

Another possible direction would be to study the correlation structure among multiple imaging modalities. In this dissertation, we perform spatially-varying correlation analysis between two modalities. We could further apply the proposed method to study the correlation structure among multiple modalities. In this case, the pairwise structure introduced in the first project can be adopted to reduce the computational complexity. By studying the correlation structure among multiple modalities, it may be possible to uncover new insights into the complex interactions and functional relationships between different brain regions. This could lead to improved diagnostic and treatment approaches for neurological and psychiatric disorders. Additionally, incorporating more modalities may provide a more complete picture of the brain, potentially enabling researchers to better understand the mechanisms underlying cognitive processes, such as learning and memory, and their disruptions in neurological and psychiatric disorders.

Finally, it is worth exploring the possibility of incorporating spatial dependency when modeling the EEG signals. While some work has already incorporated temporal dependency in ERP-based BCIs (Koçanaoğullari et al. 2018, Ma et al. 2022), there have been few attempts to account for spatial information. Our work has considered the interaction between EEG signals across different channels, but the prior we imposed on the effect of signal interactions is spatially independent. Given that several studies have shown that brain functions require the integration of distributed brain areas (Tononi and Edelman 1998, Friston et al. 1997), a promising direction would be to incorporate prior knowledge based on the brain network and design a more informative kernel or other prior structure that can capture the spatial dependency of EEG signals. Such an approach could lead to improved prediction accuracy of BCI systems by better capturing the underlying neural mechanisms involved in the cognitive task.

APPENDIX A

Supplementary Material of Chapter 1

A.1 Proof of Lemma 3.2

Consider the $\mathcal{MGP}$ model with 2 outputs y_1 and y_2, modeled with one latent function X_1, where $f_i(\boldsymbol{x}) = K_{1i}(\boldsymbol{x}) \star X_1(\boldsymbol{x})$ and X_1 is Dirac Delta function. We will show that for any $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{R}^{\mathcal{D}}$, $\text{cov}^f_{12}(\boldsymbol{x}, \boldsymbol{x}')$ = 0 if and only if at least one of K_{11}, K_{12} is identically equal to zero. The sufficiency is obvious, then we prove necessity.

First assume K_{11} and K_{12} satisfy the first condition, i.e. $\exists\, \alpha_{11}, \alpha_{12} \in \mathcal{R}$ such that $K_{11} = \alpha_{11}k_{11}$ and $K_{12} = \alpha_{12}k_{12}$, where $k_{11} > 0$ and $k_{12} > 0$. Gaussian, Matern, rational quadratic, periodic and locally periodic kernels are typical examples for this case. For any two inputs points $\boldsymbol{x}$ and $\boldsymbol{x'}$, denote $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}'$. Then

$$\begin{aligned}\text{cov}^f_{12}(\boldsymbol{x}, \boldsymbol{x}') &= \int_{-\infty}^{+\infty} K_{11}(\boldsymbol{u})K_{12}(\boldsymbol{u} - \boldsymbol{d})d\boldsymbol{u} \\ &= \alpha_{11}\alpha_{12}\int_{-\infty}^{+\infty} k_{11}(\boldsymbol{u})k_{12}(\boldsymbol{u} - \boldsymbol{d})d\boldsymbol{u}\end{aligned}$$

Since $k_{11} > 0$ and $k_{21} > 0$, $\int_{-\infty}^{+\infty} k_{11}(\boldsymbol{u})k_{12}(\boldsymbol{u} - \boldsymbol{d})d\boldsymbol{u} \neq 0$ for $\forall \boldsymbol{d} \in \mathcal{R}^D$. Thus, $\text{cov}^f_{12}(\boldsymbol{x}, \boldsymbol{x}') = 0$ if and only if at least one of α_{1i}, $i = 1, 2$ is equal to zero, i.e. at least one of K_{1i}, $i = 1, 2$ is identically equal to 0.

Then consider the second case when k_{1i} has the form $\sum_u a_u^2 \text{exp}\big(\boldsymbol{x}^T \boldsymbol{B}_u \boldsymbol{x}\big)\text{cos}(2\pi \boldsymbol{c}_u \boldsymbol{x}^T)$ with parameters $(a_u, \boldsymbol{B}_u, \boldsymbol{c}_u)$. Here for simplicity, we only prove for one dimension case when $x, x' \in \mathcal{R}$ and the proof for general case is similar. We rewrite k_{1i}, $i = 1, 2$ as

$$k_{1i} = \sum_{q=1}^{Q_i} a_{iq}^2 \exp\{-\sigma_{iq}^2 d^2\} \cos(\mu_{iq} d)$$

. Since now $k_{1i}(d) = k_{1i}(-d)$,

$$\begin{aligned}\text{cov}_{12}^{f}(x, x') &= \int_{-\infty}^{+\infty} K_{11}(u)K_{12}(u-d)du \\ &= \int_{-\infty}^{+\infty} K_{11}(u)K_{12}(d-u)du \\ &= K_{11} \star K_{12}(d) \\ &= \mathcal{F}^{-1}(\mathcal{F}(K_{11}) \cdot \mathcal{F}(K_{12}))(d)\end{aligned}$$

where $\mathcal{F}$ is the Fourier operator and $\star$ denote the convolution operator. The last equality is derived using the conclusion of Convolution Theorem. Hence $\text{cov}_{12}^{f}(x, x') = 0$ if and only if

$$\mathcal{F}(K_{11}) \cdot \mathcal{F}(K_{12})(\xi) = \sum_{k=1}^{Q_1}\sum_{j=1}^{Q_2} \frac{a_{1k}^2 a_{2j}^2}{4} M_{1k}M_{2j} = 0$$

where

$$\begin{aligned}M_{1k} &= \sqrt{\frac{\pi}{\sigma_{1k}^2}}(\exp(\frac{(\mu_{1k} - 2\pi\xi)^2}{-4\sigma_{1k}^2}) + \exp(\frac{(\mu_{1k} + 2\pi\xi)^2}{-4\sigma_{1k}^2})) \\ M_{2j} &= \sqrt{\frac{\pi}{\sigma_{2j}^2}}(\exp(\frac{(\mu_{2j} - 2\pi\xi)^2}{-4\sigma_{2j}^2}) + \exp(\frac{(\mu_{2j} + 2\pi\xi)^2}{-4\sigma_{2j}^2}))\end{aligned}$$

Note that if the kernel does not satisfy this two condition, we can still check if $\mathcal{F}^{-1}(\mathcal{F}(K_{11}) \cdot \mathcal{F}(K_{12}))(d) = 0$, i.e. whether $(\mathcal{F}(K_{11}) \cdot \mathcal{F}(K_{12}))(\xi) = 0$ or not using Fourier Transform. Since $M_{1k}, M_{2j} > 0$, we have $a_{1k}^2 a_{2j}^2 = 0$ for any $k \in \{1, 2, \cdots Q_1\}$ and $j \in \{1, 2, \cdots Q_2\}$. Therefore, we reach the conclusion either $a_{1k} = 0, \forall k$ or $a_{2j} = 0, \forall j$, i.e. at least one of $K_{1i}, i = 1, 2$ is identically equal to 0.

For general case when X_1 is a $\mathcal{GP}$ constructed from $\mathcal{CP}$, i.e.

$$\text{cov}(X_1(\boldsymbol{u}), X_1(\boldsymbol{u}')) = \int_{-\infty}^{+\infty} K_{X_1}(\boldsymbol{v})K_{X_1}(\boldsymbol{v} - \boldsymbol{d})d\boldsymbol{v}$$

where $\boldsymbol{d} = \boldsymbol{u} - \boldsymbol{u}'$. Consider the first case when there $\exists\ \alpha_{11}, \alpha_{12}, \alpha_1 \in \mathcal{R}$ such that $K_{11} = \alpha_{11}k_{11}$, $K_{12} = \alpha_{12}k_{12}$ and $K_{X_1} = \alpha_1 k_{X_1}$, where $k_{11} > 0$, $k_{12} > 0$ and $k_{X_1} > 0$.

$$\begin{aligned}\text{cov}_{12}^{f}(\boldsymbol{x}, \boldsymbol{x}') &= \int_{-\infty}^{+\infty} K_{11}(\boldsymbol{x} - \boldsymbol{z}) \int_{-\infty}^{+\infty} K_{12}(\boldsymbol{x}' - \boldsymbol{z}') \cdot \int_{-\infty}^{+\infty} K_{X_1}(\boldsymbol{v})K_{X_1}(\boldsymbol{v} - \boldsymbol{d})d\boldsymbol{v}d\boldsymbol{z}'d\boldsymbol{z} \\ &= \alpha_{11}\alpha_{12}\alpha_1 \int_{-\infty}^{+\infty} k_{11}(\boldsymbol{x}-\boldsymbol{z}) \int_{-\infty}^{+\infty} k_{12}(\boldsymbol{x}'-\boldsymbol{z}') \cdot \int_{-\infty}^{+\infty} k_{X_1}(\boldsymbol{v})k_{X_1}(\boldsymbol{v}-\boldsymbol{d})d\boldsymbol{v}d\boldsymbol{z}'d\boldsymbol{z}\end{aligned}$$

Similar as the argument when X_1 is Dirac Delta function, we have $\text{cov}_{12}^{f}(\boldsymbol{x}, \boldsymbol{x}') = 0$ if and only if one of K_{11}, K_{12} and K_{X_1} is identically equal to 0.

Now consider the case when K_{11}, K_{12} and K_{X_1} satisfy the second condition, then

$$\begin{aligned}
&\operatorname{cov}_{12}^{f}(\boldsymbol{x}, \boldsymbol{x}') \\
&= K_{11} \star K_{12} \star K_{X_1} \star K_{X_1}(\boldsymbol{d}) \\
&= \mathcal{F}^{-1}(\mathcal{F}(K_{11} \star K_{12}) \cdot \mathcal{F}(K_{X_1} \star K_{X_1}))(\boldsymbol{d}) \\
&= \mathcal{F}^{-1}(\mathcal{F}(K_{11}) \cdot \mathcal{F}(K_{12}) \cdot \mathcal{F}(K_{X_1}) \cdot \mathcal{F}(K_{X_1}))(\boldsymbol{d})
\end{aligned}$$

Hence $\operatorname{cov}_{12}^{f}(\boldsymbol{x}, \boldsymbol{x}') = 0$ if and only if $\mathcal{F}(K_{11}) \cdot \mathcal{F}(K_{12}) \cdot \mathcal{F}(K_{X_1}) \cdot \mathcal{F}(K_{X_1}) = 0.$ Similar as the proof when X_1 is Dirac Delta function, we reach the conclusion that $\operatorname{cov}_{12}^{f}(\boldsymbol{x}, \boldsymbol{x}') = 0$ if and only if one of K_{11}, K_{12} and K_{X_1} is identically equal to 0.

A.2 Illustrative example of Theorem 3.3

Here we present a simple example when $N = 2$ and $Q = 2$ to illustrate Theorem 2. We have proved that in this case, the model could achieve $\operatorname{cov}_{12}^{f}(\boldsymbol{x}, \boldsymbol{x'}) = 0,\ \forall \boldsymbol{x}, \boldsymbol{x'} \in \mathcal{R}^D$. For any new input point $\boldsymbol{x}^\star$, the integrative analysis of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ leads to the prediction:

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{y}_1^\star \\ \boldsymbol{y}_2^\star \end{pmatrix} &= \begin{pmatrix} C_{11}^\star & C_{12}^\star \\ C_{21}^\star & C_{22}^\star \end{pmatrix} \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \\
&= \begin{pmatrix} C_{11}^\star & 0 \\ 0 & C_{22}^\star \end{pmatrix} \begin{pmatrix} C_{11} & 0 \\ 0 & C_{22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \\
&= \begin{pmatrix} C_{11}^\star C_{11}^{-1} \boldsymbol{y}_1 \\ C_{22}^\star C_{22}^{-1} \boldsymbol{y}_2 \end{pmatrix}
\end{aligned}$$

where $\boldsymbol{y}_i = f_i(\boldsymbol{x}) + \epsilon_i(\boldsymbol{x})$ and $\boldsymbol{y}_i^\star = f_i(\boldsymbol{x}^\star) + \epsilon_i(\boldsymbol{x})$, $i = 1, 2$; $C_{ij} = \operatorname{cov}_{12}^{f}(\boldsymbol{x}, \boldsymbol{x'})$, $C_{ij}^\star = \operatorname{cov}_{12}^{f}(\boldsymbol{x}^\star, \boldsymbol{x'})$ The prediction result is exactly the same with that when we model y_1 and y_2 independently, which means that the model has capability to make the $\mathcal{MGP}$ model collapse into two 2 independent $\mathcal{GP}s$, hence we achieve our goal of avoiding negative transfer between $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$.

A.3 Formula of covariance functions using spectral kernels

Consider the $\mathcal{MGP}$ model with two outputs and one latent function X_1

$$f_i(x) = K_{1i}(x) \star X_1(x), \quad i = 1, 2, \quad x \in \mathcal{R} \tag{A.1}$$

where

$$K_{1i}(d) = \sum_{q=1}^{Q_i} a_{qi} \cdot \exp\{-\sigma_{qi}^2 d^2\} \cdot \cos(\mu_{qi} d)$$

From the proof of Lemma 1, we know that

$$\begin{aligned}
\operatorname{cov}_{12}^{f}(x,x') &= \int_{-\infty}^{+\infty} K_{11}(u)K_{12}(u-d)du \\
&= \int_{-\infty}^{+\infty} K_{11}(u)K_{12}(d-u)du \\
&= K_{11} \star K_{12}(d) \\
&= \mathcal{F}^{-1}\left(\mathcal{F}\left(K_{11}\right)\cdot\mathcal{F}\left(K_{12}\right)\right)(d)
\end{aligned}$$

Compute the Fourier Transform of K_{11} and K_{12} respectively:

$$\begin{aligned}
\mathcal{F}(K_{11})(\xi) &= \sum_{q=1}^{Q_1}\frac{a_{q1}}{2}\sqrt{\frac{\pi}{\sigma_{q1}^2}}(\exp\{-\frac{(\mu_{q1}-2\pi\xi)^2}{4\sigma_{q1}^2}\}+\exp\{-\frac{(\mu_{q1}+2\pi\xi)^2}{4\sigma_{q1}^2}\}) \\
&= \sum_{q=1}^{Q_1}\frac{a_{q1}}{2}M_{q1}(\xi)
\end{aligned}$$

$$\begin{aligned}
\mathcal{F}(K_{12})(\xi) &= \sum_{q=1}^{Q_2}\frac{a_{q2}}{2}\sqrt{\frac{\pi}{\sigma_{q2}^2}}(\exp\{-\frac{(\mu_{q2}-2\pi\xi)^2}{4\sigma_{q2}^2}\}+\exp\{-\frac{(\mu_{q2}+2\pi\xi)^2}{4\sigma_{q2}^2}\}) \\
&= \sum_{q=1}^{Q_2}\frac{a_{q2}}{2}M_{q2}(\xi)
\end{aligned}$$

Thus

$$\mathcal{F}(K_1)(\xi)\cdot\mathcal{F}(K_2)(\xi) = \sum_{s=1}^{Q_1}\sum_{t=1}^{Q_2}\frac{a_{s1}a_{t2}}{4}M_{s1}M_{t2}$$

We hence get the covariance function between f_1 and f_2:

$$\operatorname{cov}_{12}^{f}(x,x') = \mathcal{F}^{-1}\left(\mathcal{F}\left(K_{11}\right)\cdot\mathcal{F}\left(K_{12}\right)\right)(d) = \sum_{s=1}^{Q_1}\sum_{t=1}^{Q_2}\frac{a_{s1}a_{t2}}{2}\sqrt{\frac{\pi}{\sigma_{s1}^2+\sigma_{t2}^2}}H(d)$$

where

$$\begin{aligned}
H(d) &= (e^{A_1(d)}\cos(\theta_1 d)+e^{A_2(d)}\cos(\theta_2 d)) \\
A_1(d) &= \frac{-(\mu_{s1}-\mu_{t2})^2-4\sigma_{s1}^2\sigma_{t2}^2\pi^2d^2}{4(\sigma_{s1}^2+\sigma_{t2}^2)} \\
A_2(d) &= \frac{-(\mu_{s1}+\mu_{t2})^2-4\sigma_{s1}^2\sigma_{t2}^2\pi^2d^2}{4(\sigma_{s1}^2+\sigma_{t2}^2)} \\
\theta_1 &= \frac{\mu_{s1}\sigma_{t2}^2+\mu_{t2}\sigma_{s1}^2}{\sigma_{s1}^2+\sigma_{t2}^2} \\
\theta_2 &= \frac{\mu_{s1}\sigma_{t2}^2-\mu_{t2}\sigma_{s1}^2}{\sigma_{s1}^2+\sigma_{t2}^2}
\end{aligned}$$

Note that in our simulation, we use the kernel with $Q_1 = Q_2 = 1$ and the number of latent functions is Q. Thus the covariance function between f_i and f_j becomes

$$\mathrm{cov}_{ij}^{f}(x, x') = \sum_{k=1}^{Q} \frac{a_{qi}a_{qj}}{2} \sqrt{\frac{\pi}{\sigma_{qi}^2 + \sigma_{qj}^2}} H(d)$$

APPENDIX B

Supplementary Material of Chapter 2

In this supplement, we first present the proofs of all the theoretical results in the paper, along with a number of useful lemmas. We next derive the full conditional distributions of the model parameters, and present some additional numerical results.

B.1 Proofs

B.1.1 Proof of Proposition 15

Given $\tau_1^2(v)$ and $\tau_2^2(v)$, if $\pi(Y_{+,i}(v), Y_{-,i}(v) \mid \theta) = \pi\left(Y_{+,i}(v), Y_{-,i}(v) \mid \theta'\right)$, for any $i = 1, \ldots, n$, $v \in \mathcal{B}_m$, and since $\{Y_{+,i}(v), Y_{-,i}(v)\}$ follows a bivariate normal distribution, we have that $\mu_{+,i}(v) = \mu'_{+,i}(v)$, and $\mu_{-,i}(v) = \mu'_{-,i}(v)$, i.e., $s\{\rho(v)\}E_{+,i}(v) = s\{\rho'(v)\}E'_{+,i}(v)$, and $s\{-\rho(v)\}E_{-,i}(v) = s\{-\rho'(v)\}E'_{-,i}(v)$, for any $i = 1, \ldots, n$, $v \in \mathcal{B}_m$.

Furthermore, we have that,

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \left[s\{\rho(v)\}E_{+,i}(v) - s\{\rho'(v)\}E'_{+,i}(v)\right]^2 \\
&= \sum_{i=1}^{n} \left[s\{\rho(v)\}^2 E_{+,i}(v)^2 - 2s\{\rho(v)\}s\{\rho'(v)\}E_{+,i}(v)E'_{+,i}(v) + s\{\rho'(v)\}^2 E'_{+,i}(v)^2\right] \\
&= \left[s\{\rho(v)\} - s\{\rho'(v)\}\right]^2 \sum_{i=1}^{n} E^2_{+,i}(v) + s\{\rho'(v)\}s\{\rho(v)\} \sum_{i=1}^{n} \{E_{+,i}(v) - E'_{+,i}(v)\}^2 \\
&\quad + s\{\rho'(v)\}\left[s\{\rho(v)\} - s\{\rho'(v)\}\right] \sum_{i=1}^{n} \left\{E_{+,i}(v)^2 - E'_{+,i}(v)^2\right\}
\end{aligned}$$

By Definition 14, we have $\sum_{i=1}^{n} E_{+,i}(v)^2 = \sum_{i=1}^{n} E'_{+,i}(v)^2$.

When $v \in \mathcal{V}(\rho) \cup \mathcal{V}(\rho')$, we have $s\{\rho(v)\} \geq 0$, $s\{\rho'(v)\} \geq 0$, and at least one of $s\{\rho(v)\}$ and $s\{\rho'(v)\}$ is not equal to 0. Therefore, $s\{\rho(v)\} = s\{\rho'(v)\}$, and $E_{+,i}(v) = E'_{+,i}(v)$, for any $i = 1, \ldots, n,\ v \in \mathcal{B}_m$. On the other hand, if $v \notin \mathcal{V}(\rho) \cup \mathcal{V}(\rho')$, then $s\{\rho(v)\} = s\{\rho'(v)\} = 0$. Similarly, we have $E_{-,i}(v) = E'_{-,i}(v) = 0$, for any $i = 1, \ldots, n$, $v \in \mathcal{B}_m$.

Since $s(\cdot)$ is a monotonic function, we have $\rho(v) = \rho'(v)$ for all $v \in \mathcal{B}_m$. This completes the proof of Proposition 15. □

B.1.2 Proof of Theorem 16

By Lemma 22, we have $\rho(v) = T_\omega\{\xi(v)\} = H[R_\omega\{\xi(v)\}]$, where $H(t) = t^2/(t^2+1)$ when $\xi(v) > \omega$, $H(t) = -t^2/(t^2+1)$ when $\xi(v) < -\omega$, and $H(t) = 0$ otherwise, and $R_\omega(x) = G_\omega(x) - G_\omega(-x)$ is the hard thresholded function. Therefore, we have that,

$$\begin{aligned}\Pi\left(\|\rho-\rho_0\|_\infty < \varepsilon\right) &= \Pi\left(\|H[R_\omega\{\xi(v)\}] - H[R_\omega\{\xi_0(v)\}]\| < \epsilon\right)\\ &\geq \Pi\left(\|R_\omega\{\xi(v)\} - R_\omega\{\xi_0(v)\}\| < \epsilon\right),\end{aligned}$$

by the Lipschitz continuity of $H(\cdot)$. Given the assumptions for $\rho_0(v)$, we have that $\xi(v)$ is bounded away from 0 for $v \notin \mathcal{R}_0$. Henceforth,

$$\begin{aligned}&\Pi(\|R_\omega(\xi(v)) - R_\omega(\xi_0(v))\| < \epsilon)\\ &\geq \Pi\left(\sup_{v\notin\mathcal{R}_0}|\xi(v)-\xi_0(v)| < \epsilon, \inf_{v\notin\mathcal{R}_0}|\xi(v)| > \omega, \sup_{v\in\mathcal{R}_0}|\xi(v)| \leq \omega\right).\end{aligned} \tag{B.1}$$

Without loss of generality, we only consider $0 < \epsilon < \omega - \omega_0$, where $\omega_0 = \inf_{v\notin\mathcal{R}_0}|\rho(v)|$. Note that for all $v \notin \mathcal{R}_0$, $|\xi(v) - \xi_0(v)| < \epsilon$ and $|\xi_0(v)| \geq \omega_0$, which implies that $|\xi(v)| \geq \omega_0 - \epsilon > \omega$. Then (B.1) is equivalent to

$$\Pi(\|\rho(v) - \rho_0(v)\| < \epsilon) \geq \Pi\left(\sup_{v\notin\mathcal{R}_0}|\xi(v)-\xi_0(v)| < \epsilon, \sup_{v\in\mathcal{R}_0}|\xi(v)| \leq \omega\right).$$

Let $\psi_l(v)$ and λ_l be the normalized eigenfunctions and eigenvalues of the kernel function $\kappa(\cdot,\cdot)$. The KL expansions of $\xi(v)$ and $\xi_0(v)$ are $\xi(v) = \sum_{l=1}^\infty c_l\psi_l(v)$, $\xi_0(v) = \sum_{l=1}^\infty c_{l0}\psi_l(v)$.

For $v \notin \mathcal{R}_0$, we have that,

$$\sup_{v\notin\mathcal{R}_0}|\xi(v)-\xi_0(v)| \leq \sup_{v\notin\mathcal{R}_0}|\xi_L(v)-\xi_L^0(v)| + \sup_{v\notin\mathcal{R}_0}|\xi(v)-\xi_L(v)| + \sup_{v\notin\mathcal{R}_0}|\xi_L^0(v)-\xi_0(v)|.$$

Since the RKHS of $\kappa(\cdot,\cdot)$ is the space of the continuous functions on $\mathcal{R}$, $\xi(v)$ is uniformly continuous on $\mathcal{B}\backslash\mathcal{R}_0$ with probability 1. Then by Theorem 3.1.2 of Adler and Taylor (2009), $\lim_{L\to\infty}\sup_{v\notin\mathcal{R}_0}|\xi(v)-\xi_L(v)| = 0$ with probability 1. By the uniform convergence of the series $\sum_{l=1}^L c_{l0}\psi_l(v)$ to $\xi_0(v)$ on $\mathcal{B}\backslash\mathcal{R}_0$, as $L\to\infty$, we have $\lim_{L\to\infty}\sup_{v\notin\mathcal{R}_0}|\xi_0(v)-\xi_L^0(v)| = 0$. Then we can find a finite integer L', such that, for all $L > L'$, $\sup_{v\notin\mathcal{R}_0}|\xi(v)-\xi_L(v)| < \epsilon/3$ with probability 1, and $\sup_{v\notin\mathcal{R}_0}|\xi_0(v)-\xi_L^0(v)| < \epsilon/3$. Since $\psi_l(v), l = 1,\ldots,L$, are all continuous functions in $\mathcal{R}$, we have $\max_{1\leq l\leq L}\|\psi_l(v)\|_\infty < M_{\psi,L}$, for some constant $M_{\psi,L}$. When $|c_l - c_{l0}| < \epsilon/(3LM_{\psi,L})$ for all $l = 1,\ldots,L$, we have $\sup_{v\notin\mathcal{R}_0}|\xi_L(v)-\xi_L^0(v)| \leq \epsilon/3$. Therefore, $|c_l - c_{l0}| < \epsilon/(3LM_{\psi,L})$, $l = 1,\ldots,L$, guarantees that $\sup_{v\notin\mathcal{R}_0}|\xi(v)-\xi_0(v)| \leq \epsilon$ with probability one.

For $v \in \mathcal{R}_0$, we have that,

$$\sup_{v\in\mathcal{R}_0} |\xi(v)| \leq \sup_{v\in\mathcal{R}_0} |\xi(v) - \xi_L(v)| + \sup_{v\in\mathcal{R}_0} |\xi_L(v)|.$$

Similarly, we can find L and $M_{\psi,L}$, such that $|c_l| \leq \omega/(2LM_{\psi,L})$, $l = 1, \ldots, L$, guarantees that $\sup_{v\in\mathcal{R}_0} |\xi(v)| \leq \omega$ with probability 1.

Then we have that,

$$\begin{aligned}\Pi\left(\|\rho - \rho_0\|_\infty < \varepsilon\right) \geq \Pi \Bigg(& \{|c_l - c_{l0}| < \frac{\epsilon}{3LM_{\psi,L}} : L = 1, 2, \ldots, L \; when \; v \notin \mathcal{R}_0\} \\ & \cup \{|c_l| \leq \frac{\omega}{2LM_{\psi,L}} : L = 1, 2, \ldots, L \; when \; v \in \mathcal{R}_0\}\Bigg).\end{aligned}$$

This completes the proof of Theorem 16. □

B.1.3 Proof of Theorem 17

Based on Theorem 16, Lemma 24 shows the positivity of prior neighborhoods. We then construct sieves for $\theta(v)$ as follows:

$$\begin{aligned}\Theta_n = \Bigg\{ & \rho \in \Theta_\rho, E_+, E_- \in \Theta_E : \\ & \|\rho\|_\infty \leq H\left(m^{1/(2d)}\right), \sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |D^\tau \rho(v)| \leq m^{1/(2d)},\ 1 \leq \|\tau\|_1 \leq \alpha \\ & \|E_{+,i}\|_\infty \leq m^{1/(2d)}, \sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |D^\tau E_{+,i}(v)| \leq m^{1/(2d)}, \\ & \|E_{-,i}\|_\infty \leq m^{1/(2d)}, \sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |D^\tau E_{-,i}(v)| \leq m^{1/(2d)},\ \text{for } i = 1, \ldots, n\Bigg\},\end{aligned} \tag{B.2}$$

where α and m are defined in Assumption 16.1.

We can then find an upper bound for the tail probability, and construct the uniform consistent tests in Lemmas 25, 26, 27 and 29. These lemmas verify the three key conditions in Theorem A1 of Choudhuri et al. (2004), which leads to the posterior consistency. That is, by Lemmas 25, 26, 27 and 29, as $n \to \infty$, $m \to \infty$, we have that,

$$\begin{aligned}& \mathbb{E}_{\theta_0}\left(\Psi_n\right) \to 0, \\ & \sup_{\theta\in\mathcal{U}_\epsilon^C\cap\Theta_n} \mathbb{E}_\theta\left(1 - \Psi_n\right) \leq C_0 \exp\left(-C_1 n\right), \\ & \Pi\left(\Theta_n^C\right) \leq K \exp\left(-bm^{1/d}\right) \leq K \exp\left(-C_3 n\right).\end{aligned}$$

where $\mathcal{U}_\epsilon = \{\theta \in \Theta : \|\theta - \theta_0\|_1 < \epsilon\}$ for any $\epsilon > 0$, and Ψ_n is the test statistic defined in (B.10). This completes the proof of Theorem 17. □

B.1.4 Proof of Theorem 18

Let $\mathcal{R}_0 = \{v : \rho_0(v) = 0\}$, $\mathcal{R}_1 = \{v : \rho_0(v) > 0\}$, and $\mathcal{R}_{-1} = \{v : \rho_0(v) < 0\}$. For any $\mathcal{A} \subset \mathcal{B}$ and any integer $k \geq 1$, define

$$\mathcal{F}_k(\mathcal{A}) = \left\{\rho \in \Theta_\rho : \int_{\mathcal{A}} |\rho(v) - \rho_0(v)| \,\mathrm{d}v < \frac{1}{k}\right\}.$$

Then $\mathcal{F}_{k+1}(\mathcal{A}) \subseteq \mathcal{F}_k(\mathcal{A})$ for all k, and $\mathcal{F}_k(\mathcal{B}) \subseteq \mathcal{F}_k(\mathcal{A})$. Consider

$$\mathcal{F}_k\left(\mathcal{R}_0\right) = \left\{\rho \in \Theta_\rho : \int_{\mathcal{R}_0} |\rho(v)|\mathrm{d}v < \frac{1}{k}\right\}.$$

Define $\mathcal{U}^\rho_\epsilon = \{\rho \in \Theta_\rho : \|\rho - \rho_0\|_1 < \epsilon\}$. By Theorem 17 and the fact that $\mathcal{U}^\rho_{1/k} = \mathcal{F}_k(\mathcal{B})$, we have

$$\Pi\left\{\mathcal{F}_k\left(\mathcal{R}_0\right) \mid Y_+, Y_-\right\} \geq \Pi\left(\mathcal{U}^\rho_{1/k} \mid Y_+, Y_-\right) \to 1, \text{ as } n \to \infty.$$

In addition,

$$\{\rho(v) = 0, \text{ for all } v \in \mathcal{R}_0\} = \left\{\int_{\mathcal{R}_0} |\rho(v)|\mathrm{d}v = 0\right\} = \bigcap_{k=1}^{\infty} \mathcal{F}_k\left(\mathcal{R}_0\right).$$

By the monotonic continuity of the probability measure, we have,

$$\Pi\big\{\rho(v) = 0, \text{ for all } v \in \mathcal{R}_0 \mid Y_+, Y_-\big\} = \lim_{k\to\infty} \Pi\big\{\mathcal{F}_k\left(\mathcal{R}_0\right) \mid Y_+, Y_-\big\} = 1, \text{ as } n \to \infty.$$

For any $v_0 \in \mathcal{R}_1$ and any integer $k \geq 1$, there exists $\delta_0 > 0$, such that $|\rho\left(v_1\right) - \rho\left(v_0\right)| < 1/2k$, for any $v_1 \in \mathcal{B}\left(v_0, \delta_0\right) = \{v : \|v_1 - v_0\|_1 < \delta_0\}$. As $\mathcal{R}_1$ is an open set, there exists $\delta_1 > 0$, such that $\mathcal{B}\left(v_0, \delta_1\right) \subseteq \mathcal{R}_1$. Let $\delta = \min\left\{\delta_1, \delta_0\right\} > 0$, we have that,

$$\begin{aligned}
&\left\{\rho\left(v_0\right) > -\frac{1}{k}, \text{ for all } v_0 \in \mathcal{R}_1\right\} \\
&\supseteq \left\{\rho\left(v_0\right) > \rho\left(v_1\right) - \frac{1}{2k} \text{ and } \rho\left(v_1\right) > -\frac{1}{2k}, \text{ for some } v_1 \in \mathcal{B}\left(v_0, \delta\right), \text{ for all } v_0 \in \mathcal{R}_1\right\} \\
&\supseteq \left\{\int_{\mathcal{B}(v_0,\delta)} \rho(v)dv > -\frac{1}{2k}, \text{ for all } v_0 \in \mathcal{R}_1\right\} \\
&\supseteq \left\{\int_{\mathcal{B}(v_0,\delta)} \rho(v)dv > \int_{\mathcal{B}(v_0,\delta)} \rho_0(v)dv - \frac{1}{2k}, \text{ for all } v_0 \in \mathcal{R}_1\right\} \\
&\supseteq \mathcal{F}_{2k}\left[\mathcal{B}\left(v_0, \delta\right)\right] \supseteq \mathcal{U}^\rho_{1/2k}.
\end{aligned}$$

Therefore, $\Pi\left\{\rho\left(v_0\right) > -1/k, \text{ for all } v_0 \in \mathcal{R}_1 \mid Y_+, Y_-\right\} \geq \Pi\left(\mathcal{U}^\rho_{1/2k} \mid Y_+, Y_-\right) \to 1$, as $n \to \infty$. By the monotonic continuity of the probability measure, we have that,

$$\Pi\left\{\rho(v) > 0, \text{ for all } v \in \mathcal{R}_1 \mid Y_+, Y_-\right\} = \lim_{k\to\infty} \Pi\left\{\rho\left(v_0\right) > -\frac{1}{k}, \text{ for all } v_0 \in \mathcal{R}_1 \mid Y_+, Y_-\right\} \to 1,$$

as $n \to \infty$. Similarly, we can obtain that $\Pi\left\{\rho(v) < 0, \text{for all } v \in \mathcal{R}_{-1} \mid Y_+, Y_-\right\} \to 1, n \to \infty$. This completes the proof of Theorem 18. □

B.1.5 Proof of Proposition 19

We prove this proposition by sorting all the thresholding values, and derive the unnormalized density on each interval, respectively. We then obtain the full conditional density function of θ by normalizing the function on each interval as the density function.

We sort $(L_1, \ldots, L_P, U_1, \ldots, U_K)$ in ascending order, which leads to $P + K + 1$ intervals, and denoted them as $I_1, I_2, \ldots, I_{P+K+1}$. For each interval I_i, $i = 1, \ldots, P + K + 1$, the full conditional distribution of θ is proportional to $\exp(-D_i\theta^2 - E_i\theta - F_i)$. We initialize $D_i = E_i = F_i = 0$, then loop through $p = 1, \ldots, P$ and $k = 1, \ldots, K$ to update D_i, E_i and F_i. More specifically, if $I_i \subset [L_p, +\infty)$, we update $D_i = D_i + a_{1p}$, $E_i = E_i + a_{2p}$, and $F_i = F_i + a_{3p}$. If $I_i \subset (-\infty, U_k]$, we update $D_i = D_i + b_{1k}$, $E_i = E_i + b_{2k}$, and $F_i = F_i + b_{3k}$. We consider three specific cases.

- If at least one of $\{a_{1p}, \ldots, a_{1P}, b_{1k}, \ldots, b_{1K}\}$ is not equal to 0, then $D_i \neq 0$, for any $i = 1, \ldots, P + K + 1$. Therefore, when $\theta \in I_i$, the full conditional distribution of θ is $\mathrm{N}\{-E_i/(2D_i), -1/(2D_i)\}$. Incorporating the normalizing constant M_i for each interval, which is independent of θ, the full conditional distribution of θ is the mixture of truncated normal distributions, $\sum_{i=1}^{P+K+1} M_i \cdot \mathrm{TruncatedNormal}_{I_i}\{-E_i/(2D_i), -1/(2D_i)\}$.

- If at least one of $\{a_{2p}, \ldots, a_{2P}, b_{2k}, \ldots, b_{2K}\}$ is not equal to 0 and $a_{1p} = b_{1k} = 0$, for any $p = 1, \ldots, P$ and $k = 1, \ldots, K$, then $D_i = 0$ and $E_i \neq 0$, for any $i = 1, \ldots, P + K + 1$. Therefore, when $\theta \in I_i$, the full conditional distribution of θ is the exponential distribution $\mathrm{Exp}(E_i)$. Incorporating the normalizing constant M_i, the full conditional distribution of θ is $\sum_{i=1}^{P+K+1} M_i \cdot \mathrm{Exponential}_{I_i}(E_i)$.

- If at least one of $\{a_{3p}, \ldots, a_{3P}, b_{3k}, \ldots, b_{3K}\}$ is not equal to 0, $a_{1p} = b_{1k} = a_{2p} = b_{2k} = 0$, for any $p = 1, \ldots, P$ and $k = 1, \ldots, K$, then $D_i = E_i = 0$, and at least one of $F_i \neq 0$, for any $i = 1, \ldots, P + K + 1$. Therefore, when $\theta \in I_i$, the full conditional distribution of θ is proportional to the uniform distribution on $I_i = [u_{1i}, u_{2i}]$. Incorporating the normalizing constant M_i, the full conditional distribution of θ is $\sum_{i=1}^{P+K+1} M_i \cdot \mathrm{U}(u_{1i}, u_{2i})$.

This completes the proof of Proposition 19. □

B.2 Additional Lemmas

Lemma 22. *Rewrite $\rho(v) = T_\omega\{\xi(v); \tau_1^2(v), \tau_2^2(v)\}$ in* (3.6). *Then $T_\omega(\cdot)$ is a piecewise Lipschitz continuous function for any ω.*

Proof: From (3.6), it is straightforward to verify that $\rho(v)$ can be written as

$$\begin{aligned}\rho(v) &= \mathrm{Corr}\{Y_{1,i}(v), Y_{2,i}(v)\} \\ &= \frac{G_\omega^2\{\xi(v)\} - G_\omega^2\{-\xi(v)\}}{\sqrt{G_\omega^2\{\xi(v)\} + G_\omega^2\{-\xi(v)\} + \tau_1^2(v)}\sqrt{G_\omega^2\{\xi(v)\} + G_\omega^2\{-\xi(v)\} + \tau_2^2(v)}} \\ &= \frac{\mathrm{sign}\{\xi(v)\}R_\omega^2\{\xi(v)\}}{\sqrt{R_\omega^2\{\xi(v)\} + \tau_1^2(v)}\sqrt{R_\omega^2\{\xi(v)\} + \tau_2^2(v)}},\end{aligned}$$

where $R_\omega(x) = G_\omega(x) - G_\omega(-x)$. Without loss of generality, suppose $\tau_1^2(v)$ and $\tau_2^2(v)$ are both equal to one. Then $T_\omega(x) = H\{R_\omega(x)\}$, where $H(t) = t^2/(t^2+1)$ when $\xi(v) > \omega$, $H(t) = -t^2/(t^2+1)$ when $\xi(v) < -\omega$, and $H(t) = 0$ otherwise. Since $H(t)$ is continuous and $|H'(t)| \leq 1/(2\omega)$, $H(t)$ is Lipschitz continuous. As $R_\omega(x)$ is the hard thresholding function, which is piecewise Lipschitz continuous function, $T_\omega(x) = H\{R_\omega(x)\}$ is also a piecewise Lipschitz continuous function. This completes the proof of Lemma 22. □

Lemma 23. *Given* $\rho(v) = T_\omega\{\xi(v); \tau_1^2(v), \tau_2^2(v)\}$ *in* (3.6)*, there exist a piecewise Lipschitz continuous function* $s(\cdot)$*, such that* $G_\omega\{\xi(v)\} = s\{\rho(v); \tau_1^2(v), \tau_2^2(v)\}$

Proof: It is straightforward to show that $G_\omega\{\xi(v)\} = s\left\{\rho(v); \tau_1^2(v), \tau_2^2(v)\right\}$, and $G_\omega\{-\xi(v)\} = s\left\{-\rho(v); \tau_1^2(v), \tau_2^2(v)\right\}$, where $s(x; t_1, t_2)$ is as given in (3.7). Therefore, s is a piecewise Lipschitz continuous function. This completes the proof of Lemma 23. □

Lemma 24. *Let* $\Pi_{n,i}(\cdot;\theta)$ *denote the density function of* $Z_{n,i} = (Y_{+,i}, Y_{-,i})$*. Define* $\Lambda_{n,i}(\cdot;\theta_0,\theta) = \log \pi_{n,i}(\cdot;\theta) - \log \pi_{n,i}(\cdot;\theta_0)$*,* $K_{n,i}(\theta_0,\theta) = \mathbb{E}_{\theta_0}\left\{\Lambda_{n,i}\left(Z_{n,i};\theta_0,\theta\right)\right\}$*, and* $V_{n,i}\left(\theta_0,\theta\right) = \mathrm{var}_{\theta_0}$ $\{\Lambda_{n,i}(Z_{n,i};\theta_0,\theta)\}$*. There exists a set* O *with* $\Pi(O) > 0$*, such that, for any* $\epsilon > 0$*,*

$$\liminf_{n\to\infty} \Pi\left[\left\{\theta \in O, n^{-1}\sum_{i=1}^n K_{n,i}\left(\theta_0,\theta\right) < \epsilon\right\}\right] > 0 \textit{ and } n^{-2}\sum_{i=1}^n V_{n,i}\left(\theta_0,\theta\right) \to 0 \textit{ for } \theta \in O.$$

Proof: The density function is of the form,

$$\Pi_{n,i}(Z_{n,i};\theta) = \sum_{v\in\mathcal{B}_m} \frac{1}{2\pi u^2(v)\sqrt{1-r^2(v)}} \cdot \exp\left[-\frac{W_i(v)}{2\{1-r^2(v)\}u^2(v)}\right],$$

where $W_i(v) = \{Y_{+,i}(v) - \mu_{+,i}(v)\}^2 + \{Y_{-,i}(v) - \mu_{-,i}(v)\}^2 + 2r(v)\{Y_{+,i}(v)\mu_{-,i}(v) + Y_{-,i}(v)\mu_{+,i}(v)\}$, $r(v) = \{\tau_1^2(v) - \tau_2^2(v)\}/\{\tau_1^2(v) + \tau_2^2(v)\}$, and $u^2(v) = \{\tau_1^2(v) + \tau_2^2(v)\}/4$. Therefore, we have,

$$\begin{aligned}\Lambda_{n,i}\left(Z_{n,i};\theta_0,\theta\right) &= \log\Pi(Z_{n,i};\theta) - \log\Pi(Z_{n,i};\theta_0) \\ &= \sum_{v\in\mathcal{B}_m}\left[-\frac{1}{2\{1-r^2(v)\}u^2(v)}\right]\Big[\mu_{+,i}^2(v) - \mu_{+,i,0}^2(v) + \mu_{-,i}^2(v) - \mu_{+,i,0}^2(v) \\ &\quad + 2Y_{+,i}(v)\{\mu_{+,i,0}(v) - \mu_{+,i}(v)\} + 2Y_{-,i}(v)\{\mu_{-,i,0}(v) - \mu_{-,i}(v)\}(v) \\ &\quad + 2rY_{+,i}(v)\{\mu_{-,i}(v) - \mu_{-,i,0}(v)\} + 2rY_{-,i}(v)\{\mu_{+,i}(v) - \mu_{+,i,0}(v)\}\Big],\end{aligned}$$

$$\begin{aligned}K_{n,i}\left(\theta_0,\theta\right) &= \mathbb{E}_{\theta_0}\left\{\Lambda_{n,i}\left(Z_{n,i};\theta_0,\theta\right)\right\}\\ &= \sum_{v\in\mathcal{B}_m}\Big(-\frac{1}{2\{1-r^2(v)\}u^2(v)}\Big[\{\mu_{+,i}(v)-\mu_{+,i,0}(v)\}^2+\{\mu_{-,i}(v)-\mu_{-,i,0}(v)\}^2\\ &\quad + 2r(v)\mu_{+,i,0}(v)\mu_{-,i}(v)+2r(v)\mu_{-,i,0}(v)\mu_{+,i}(v)\\ &\quad - 2r(v)\mu_{+,i,0}(v)\mu_{-,i,0}(v)-2r(v)\mu_{-,i,0}(v)\mu_{+,i,0}(v)\Big]\Big).\end{aligned}$$

Given any $\zeta > 0$, let $O(\zeta) = \{\theta : \|\theta-\theta_0\|_\infty < \zeta\}$, with

$$\|\theta-\theta_0\|_\infty = \max_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)}\left\{\|\rho-\rho_0\|_\infty, \max_{1\le i\le n}\|E_{-,i}-E_{-,i,0}\|_\infty, \max_{1\le i\le n}\|E_{+,i}-E_{+,i,0}\|_\infty\right\}.$$

and $\mathcal{V}(\rho) = \{v : \rho(v) \neq 0\}$, $\mathcal{V}(\rho_0) = \{v : \rho_0(v) \neq 0\}$. Then for any $v \in O(\zeta)$,

$$\begin{aligned}&|\mu_{i,+}(v)-\mu_{i,+,0}(v)| \le |s\{\rho(v)\}E_{+,i}(v) - s\{\rho_0(v)\}E_{i,+,0}(v)|\\ &\le |E_{+,i}(v)\left(s\{\rho(v)\}-s\{\rho_0(v)\}\right)| + |s\{\rho_0(v)\}\left(E_{+,i}(v)-E_{i,+,0}(v)\right)| \le K_1\zeta,\end{aligned}$$

where the last inequality is due to the compactness and convexity of $\mathcal{B}_m$, and

$$K_1 = \max_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)}\{E_{+,i}(v), s\{\rho_0(v)\}\} \tag{B.3}$$

Similarly, we have $|\mu_{i,-}(v)-\mu_{i,-,0}(v)| \le K_2\zeta$, for any v, where

$$K_2 = \max_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)}\{E_{-,i}(v), s\{-\rho_0(v)\}\}. \tag{B.4}$$

Therefore, we have that,

$$\begin{aligned}\left|\sum_{i=1}^n K_{n,i}(\theta,\theta_0)\right| &\le \sum_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)}\frac{1}{2\left\{1-r^2(v)\right\}u^2(v)}\Bigg(\sum_{i=1}^n|\mu_{i,+}(v)-\mu_{i,+,0}(v)|^2\\ &\quad + \sum_{i=1}^n|\mu_{i,-}(v)-\mu_{i,-,0}(v)|^2\\ &\quad + 2r(v)M\sum_{i=1}^n|\mu_{i,-}(v)-\mu_{i,-,0}(v)| + 2r(v)M\sum_{i=1}^n|\mu_{i,+}(v)-\mu_{i,+,0}(v)|\Bigg)\\ &\le \sum_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)}\frac{1}{2\{1-r^2(v)\}u^2(v)}\left(nK_1^2\zeta^2+nK_2^2\zeta^2+2|r(v)|Mn(K_1+K_2)\zeta\right)\\ &\le An\zeta^2 + Bn\zeta,\end{aligned}$$

where $M = \max_{v\in\mathcal{V}(\rho)\cup\mathcal{V}_0(\rho_0),\forall i}\{\mu_{+,i,0}(v), \mu_{-,i,0}(v)\}$, $A = (K_1^2+K_2^2)\sum_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)}\frac{1}{2\{1-r^2(v)\}u^2(v)}$, and $B = M(K_1 + K_2) \sum_{v \in \mathcal{V}(\rho) \cup \mathcal{V}(\rho_0)} \frac{|r(v)|}{2\{1 - r^2(v)\}u^2(v)}$. Henceforth, for any $\epsilon > 0$, we obtain that,

$$\liminf_{n\to\infty} \Pi \left[\left\{ \theta \in O, n^{-1} \sum_{i=1}^{n} K_{n,i}(\theta_0, \theta) < \epsilon \right\} \right] > 0.$$

Similarly, we have that,

$$\begin{aligned} V_{n,i}(\theta_0, \theta) = & \sum_{v \in \mathcal{V}(\rho) \cup \mathcal{V}(\rho_0)} \frac{1}{\{1 - r^2(v)\}u^2(v)} \left[\{\mu_{+,i}(v) - \mu_{+,i,0}(v)\}^2 + \{\mu_{-,i}(v) - \mu_{-,i,0}(v)\}^2 \right. \\ & \left. + \{r^3(v) - 3r(v)\}\{\mu_{+,i}(v) - \mu_{+,i,0}(v)\}\{\mu_{-,i}(v) - \mu_{-,i,0}(v)\} \right], \\ |V_{n,i}(\theta_0, \theta)| \leq & \sum_{v \in \mathcal{V}(\rho) \cup \mathcal{V}(\rho_0)} \frac{1}{\{1 - r^2(v)\}u^2(v)} \left(K_1^2 \zeta^2 + K_2^2 \zeta^2 + |r^3(v) - 3r(v)| K_1 K_2 \zeta^2 \right) \leq C\zeta^2, \end{aligned}$$

where $C = (K_1^2 + K_2^2) \sum_{v \in \mathcal{V}(\rho) \cup \mathcal{V}(\rho_0)} \frac{1}{\{1 - r^2(v)\}u^2(v)} + K_1 K_2 \sum_{v \in \mathcal{V}(\rho) \cup \mathcal{V}(\rho_0)} \frac{|r^3(v) - 3r(v)|}{\{1 - r^2(v)\}u^2(v)}$. Henceforth, we obtain that,

$$\left| \sum_{i=1}^{n} V_{n,i}(\theta_0, \theta) \right| \leq nC\zeta^2 \text{ and } \frac{1}{n^2} \sum_{i=1}^{n} V_{i,n}(\theta_0, \theta) \to 0, \text{as } n \to \infty.$$

This completes the proof of Lemma 24. □

Given the sieves we construct in (B.2), we next derive an upper bound for the tail probability, and construct the uniform consistent tests in Lemmas 25, 26, 27 and 29.

Lemma 25. *Suppose $\rho \sim \mathrm{TCGP}(\omega_0, \kappa)$ with $\omega_0 > 0$, the kernel function κ satisfies Assumption 15.1, and $E_{+,i}, E_{-,i} \sim \mathcal{GP}(0, I)$, for $i = 1, \ldots, n$. Then there exist constants K and b, such that $\Pi\left(\Theta_n^C\right) \leq K \exp(-C_3 n)$.*

Proof: Following the same notation as that in the proof of Lemma 22, we have $\rho(v) = T_\omega\{\xi(v)\} = H[R_\omega\{\xi(v)\}]$. Let $\mathcal{R}_1 = \{v : \rho(v) > 0\}$, and $\mathcal{R}_{-1} = \{v : \rho(v) < 0\}$. We have $R_\omega\{\xi(v)\} =$

$\xi(v) > \omega$ when $v \in \mathcal{R}_1$, and $R_\omega\{\xi(v)\} = \xi(v) < -\omega$ when $v \in \mathcal{R}_{-1}$. Then

$$\begin{aligned}\Pi\left(\Theta_n^C\right) \leq\, & \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |H(\xi(v))| > H\left(m^{1/2d}\right)\right\} \qquad \text{(B.5)}\\ & + \sum_{\tau:1\leq\|\tau\|_1\leq\alpha} \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |D^\tau H(\xi(v))| > m^{1/2d}\right\}\\ & + \sum_{i=1}^n \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |E_{+,i}| > m^{1/2d}\right\} + \sum_{i=1}^n \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |E_{-,i}| > m^{1/2d}\right\}\\ & + \sum_{i=1}^n \sum_{\tau:1\leq\|\tau\|_1\leq\alpha} \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |D^\tau E_{+,i}| > m^{1/2d}\right\}\\ & + \sum_{i=1}^n \sum_{\tau:1\leq\|\tau\|_1\leq\alpha} \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |D^\tau E_{-,i}| > m^{1/2d}\right\}. \qquad \text{(B.6)}\end{aligned}$$

Since $H(t)$ is a monotonic function,

$$\begin{aligned}\Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |H(\xi(v))| > H(m^{1/2d})\right\} & \leq \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |\xi(v)| > m^{1/2d}\right\}\\ & \leq K_1 \exp\left(-b_1 m^{1/d}\right) + K_{-1}\exp\left(-b_{-1}m^{1/d}\right),\end{aligned}$$

where the existence of K_1, K_{-1}, b_1, b_{-1} in the second inequality is ensured by Theorem 5 of Ghosal and Roy (2006).

We next consider the second term in (B.5). Since $|H'(t)| \leq 1$ and $|H''(x)| \leq 2$, we have,

$$\begin{aligned}& \sum_{\tau:1\leq\|\tau\|_1\leq\alpha} \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |D^\tau H(\xi(v)-\omega)| > m^{1/2d}\right\}\\ & \leq \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |D^\tau \xi(v)| > m^{1/2d}\right\} + \Pi\left\{\sup_{v\in\mathcal{R}_1\cup\mathcal{R}_{-1}} |2\cdot D^\tau \xi(v)| > m^{1/2d}\right\}\\ & \leq \sum_{\tau:0<\|\tau\|_1\leq\alpha} K_\tau \exp\left(-b_\tau m^{1/d}\right).\end{aligned}$$

Denote the sum of the last four terms in (B.5) as S_E. By Theorem 5 of Ghosal and Roy (2006) again, there exist K_{E_+}, b_{E_+},K_{E_-}, b_{E_-}, K_{E_τ} and b_{E_τ}, such that

$$S_E \leq K_{E_+}\exp(-b_{E_+}m^{1/d}) + K_{E_-}\exp(-b_{E_-}m^{1/d}) + \sum_{\tau:0<\|\tau\|_1\leq\alpha} K_{E_\tau}\exp\left(-b_{E_\tau}m^{1/d}\right).$$

Taking

$$\begin{aligned} K &= K_{-1} + K_1 + K_{E_+} + K_{E_-} + \sum_{\tau:0<\|\tau\|\leq\alpha} K_\tau + \sum_{\tau:0<\|\tau\|\leq\alpha} K_{E_\tau}\\ b &= \min\left\{b_{-1}, b_1, b_{E_+}, b_{E_-}, \min_{1\leq|\tau|\leq\alpha} b_\tau, \min_{1\leq|\tau|\leq\alpha} b_{E_\tau}\right\}\end{aligned} \qquad \text{(B.7)}$$

we have

$$\Pi\left(\Theta_n^C\right) \leq K \exp\left(-bm^{1/d}\right) \leq K \exp\left(-C_3 n\right).$$

This completes the proof of Lemma 25. □

Lemma 26. *Suppose Assumption 14.1 holds. The hypothesis testing problem,*

$$\begin{aligned}
&H_0 : \rho(v) = \rho_0(v), \;\; E_{\pm,i}(v) = E_{\pm,i,0}(v), \quad i = 1, \ldots, n, \; v \in \mathcal{V}(\rho_1) \cup \mathcal{V}(\rho_0),\\
&H_1 : \rho(v) = \rho_1(v), \;\; E_{\pm,i}(v) = E_{\pm,i,1}(v),
\end{aligned}$$

is equivalent to the hypothesis testing problem,

$$\begin{aligned}
&H_0^* : \mu_{\pm,i}(v) = \mu_{\pm,i,0}(v), \quad i = 1, \ldots, n, \; v \in \mathcal{V}(\rho_1) \cup \mathcal{V}(\rho_0),\\
&H_1^* : \mu_{\pm,i}(v) = \mu_{\pm,i,1}(v),
\end{aligned}$$

where $\mathcal{V}(\rho_1) = \{v : \rho_1(v) \neq 0\}$ and $\mathcal{V}(\rho_0) = \{v : \rho_0(v) \neq 0\}$.

Proof: For any $k \in \{0, 1\}$, it is straightforward to see that if H_k holds, then H_k^* also holds. We show that, if H_k^* holds, then H_k also holds. For any $v \in \mathcal{B}_m$,

$$\begin{aligned}
0 &= \sum_{i=1}^n \left[s\{\rho(v)\}E_{+,i}(v) - s\{\rho_k(v)\}E_{+,i,k}(v)\right]^2\\
&= \sum_{i=1}^n \left[s\{\rho(v)\}^2 E_{+,i}(v)^2 - 2s\{\rho(v)\}s\{\rho_k(v)\}E_{+,i,1}(v)E_{+,i,k}(v) + s\{\rho_k(v)\}^2 E_{+,i,0}(v)^2\right]\\
&= \left[s\{\rho(v)\} - s\{\rho_0(v)\}\right]^2 \sum_{i=1}^n E_{+,i,k}^2(v) + s\{\rho_k(v)\}s\{\rho(v)\} \sum_{i=1}^n \left\{E_{+,i}(v) - E_{+,i,k}(v)\right\}^2\\
&\quad + s\{\rho_0(v)\}\left[s\{\rho(v)\} - s\{\rho_0(v)\}\right] \sum_{i=1}^n \left\{E_{+,i}(v)^2 - E_{+,i,k}(v)^2\right\},
\end{aligned}$$

By Definition 14, we have $\sum_{i=1}^n E_{+,i}(v)^2 = \sum_{i=1}^n E_{+,i,0}(v)^2 = \sum_{i=1}^n E_{+,i,1}(v)^2$. When $v \in \mathcal{V}(\rho_1) \cup \mathcal{V}(\rho_0)$, $s\{\rho_0(v)\} \geq 0$, $s\{\rho_1(v)\} \geq 0$, and at least one of $s\{\rho_0(v)\}$ and $s\{\rho_1(v)\}$ is not equal to 0,

$$s\{\rho(v)\} - s\{\rho_k(v)\} = 0, \quad E_{+,i}(v) - E_{+,i,k}(v) = 0, \quad i = 1, \ldots, n.$$

Similarly, we have that $E_{-,i}(v) - E_{-,i,k}(v) = 0$ for any $v \in \mathcal{V}(\rho_1) \cup \mathcal{V}(\rho_0)$, $i = 1, \ldots, n$. Since $s(\cdot)$ is a monotonic function, $\rho(v) = \rho_k(v)$ for any $v \in \mathcal{B}_m$, which ccompletes the proof of Lemma 26. □

Lemma 27. *For the hypothesis testing problem,*

$$\begin{aligned}
&H_0 : \mu_{\pm,i}(v_j) = \mu_{\pm,i,0}(v_j), \quad i = 1, \ldots, n, \; v_j \in \mathcal{V}(\rho_1) \cup \mathcal{V}(\rho_0), \; j = 1, \ldots, m,\\
&H_1 : \mu_{\pm,i}(v_j) = \mu_{\pm,i,1}(v_j),
\end{aligned}$$

construct the testing statistic, $\Psi_n = \Psi_{+n} + \Psi_{-n} - \Psi_{+n}\Psi_{-n}$*, where*

$$\Psi_{\pm n} = \max_{i=1,\ldots,n} \left\{ I\left(\sum_{j=1}^{m} \delta_{\pm,i}(v_j)(Y_{\pm,i}(v_j) - \mu_{\pm,i,0}(v_j)) > 2\left(\frac{m}{C_0}\right)^{\frac{\nu}{d}+\frac{1}{2d}} \right) \right\},$$

$\delta_{\pm,i}(v_j) = 2I\{\mu_{\pm,i,1}(v_j) \geq \mu_{\pm,i,0}(v_j)\} - 1$*,* $\nu_0/2 < \nu < 1/2$*, and* ν_0, d*,* C_0 *are as defined in Assumption* 16.1*. Write* $\mu = \{\mu_{i,\pm}(v_j)\}$*, and* $\mu_k = \{\mu_{i,\pm,k}(v_j)\}$ *for* $k = 0, 1$*. Then, for any* $\epsilon_0 > 0$*, there exist constants* C_0*,* C_1 *and* $i_* \in \{1, \ldots, n\}$*, such that, for any* μ_1 *and* μ_0 *satisfying that* $\sum_{j=1}^{m} |\mu_{+,i_*,1}(v_j) - \mu_{+,i_*,0}(v_j)| > m\epsilon_0$*, or* $\sum_{j=1}^{m} |\mu_{-,i_*,1}(v_j) - \mu_{-,i_*,0}(v_j)| > m\epsilon_0$*, and* μ *satisfying that* $\|\mu - \mu_1\|_\infty < \epsilon_0/4$*, we have* $\mathbb{E}_{\mu_0}(\Psi_n) < C_0 \exp(-2n^{2\nu})$ *and* $\mathbb{E}_{\mu}(\Psi_n) < C_0 \exp(-C_1 n)$*.*

Proof: To bound the type I error, we have $\mathbb{E}_{\mu_0}(\Psi_n) \leq \mathbb{E}_{\mu_0}(\Psi_{+n}) + \mathbb{E}_{\mu_0}(\Psi_{-n})$. By Assumption 16.1, we have $(m/C_0)^{\nu/d} \geq n^\nu$. By the definition of Ψ_{+n}, we have that,

$$\begin{aligned}
\mathbb{E}_{\mu_0}(\Psi_{+n}) &\leq \Pr\left(\sum_{j=1}^{m} \delta_{+,i_*}(v_j)\{Y_{+,i_*}(v_j) - \mu_{+,i_*,0}(v_j)\} > 2\left(\frac{m}{C_0}\right)^{\frac{\nu}{d}+\frac{1}{2d}} \right) \\
&= \Pr\left(\sqrt{\frac{C_0}{m^d}} \sum_{j=1}^{m} \delta_{+,i_*}(v_j)\{Y_{+,i_*}(v_j) - \mu_{+,i_*,0}(v_j)\} > 2\left(\frac{m}{C_0}\right)^{\frac{\nu}{d}} \right) \\
&= 1 - \Phi\left(2\left(\frac{m}{C_0}\right)^{\frac{\nu}{d}} \right) \leq 1 - \Phi\left(2n^\nu\right) \leq \frac{\phi(2n^\nu)}{2n^\nu} = \frac{1}{2\sqrt{2\pi}} \frac{\exp(-2n^{2\nu})}{n^\nu}.
\end{aligned}$$

Similarly, we have that $\mathbb{E}_{\mu_0}(\Psi_{-n}) \leq \frac{1}{2\sqrt{2\pi}} \frac{\exp(-2n^{2\nu})}{n^\nu}$. Therefore,

$$\mathbb{E}_{\mu_0}(\Psi_n) \leq \frac{1}{\sqrt{2\pi}} \frac{\exp(-2n^{2\nu})}{n^\nu}.$$

To bound the type II error, we have that,

$$\mathbb{E}_\mu\left[1 - \Psi_n\right] \leq \min\left\{\mathbb{E}_\mu\left(1 - \Psi_{+n}\right), \mathbb{E}_\mu\left(1 - \Psi_{-n}\right)\right\}.$$

As such, we only need to show that at least one of the type II error probabilities for Ψ_{+n} and Ψ_{-n} is exponentially small. Suppose $\sum_{j=1}^{m} |\mu_{+,i_*,0}(v_j) - \mu_{+,i_*,1}(v_j)| > m\epsilon_0$. Since $\sum_{j=1}^{m} |\mu_{+,i_*}(v_j) -$

$\mu_{+,i_*,1}(v_j)| < m\epsilon_0/4$, we have,

$$
\begin{aligned}
&\mathbb{E}_\mu(1-\Psi_{+n})\\
&\le \Pr\left(\sum_{j=1}^m \delta_{+,i_*}(v_j)\{Y_{+,i_*}(v_j)-\mu_{i_*,+,0}(v_j)\} > 2\left(\frac{m}{C_0}\right)^{\frac{\nu}{d}+\frac{1}{2d}}\right)\\
&= \Pr\left(\sqrt{\frac{C_0}{m^d}}\sum_{j=1}^m \delta_{+,i_*}(v_j)\{Y_{+,i}(v_j)-\mu_{+,i,0}(v_j)\} \le 2\left(\frac{m}{C_0}\right)^{\frac{\nu}{d}}\right)\\
&= \Pr\left(\sqrt{\frac{C_0}{m^d}}\sum_{j=1}^m \delta_{+,i_*}(v_j)\{Y_{+,i}(v_j)-\mu_{+,i}(v_j)\} + \sqrt{\frac{C_0}{m^d}}\sum_{j=1}^m \delta_{+,i_*}(v_j)\{\mu_{+,i}(v_j)-\mu_{+,i,1}(v_j)\}\right.\\
&\quad\left. + \sqrt{\frac{C_0}{m^d}}\sum_{j=1}^m \delta_{+,i_*}(v_j)\{\mu_{+,i,1}(v_j)-\mu_{+,i,0}(v_j)\} < 2(m/C_0)^{\nu/d}\right)\\
&\le \Pr\left(\sqrt{\frac{C_0}{m^d}}\sum_{j=1}^m \delta_{+,i_*}(v_j)\{Y_{+,i}(v_j)-\mu_{+,i}(v_j)\} \le \frac{C_0\epsilon_0 m^{1/2d}}{4} - C_0\epsilon_0 m^{1/2d} + 2(m/C_0)^{\nu/d}\right).
\end{aligned}
$$

Since $\nu < 1/2$, there exists $N > N_0$, such that, for all $n \ge N$, $(m/C_0)^{\nu/d} < C_0 m^{1/2d}\epsilon_0/4$. ByAssumption 16.1, this further implies that,

$$
\begin{aligned}
\mathbb{E}_\mu(1-\Psi_{+n}) &\le \Pr\left(\sqrt{\frac{C_0}{m^d}}\sum_{j=1}^m \delta_{+,i_*}(v_j)\{Y_{+,i_*}(v_j)-\mu_{+,i_*}(v_j)\} \le -\frac{C_0\epsilon_0 m^{1/2d}}{4}\right)\\
&\le \Phi\left(-\frac{C_0\epsilon_0 m^{1/2d}}{4}\right) \le \Phi\left(-\frac{\epsilon_0 n^{1/2}}{4}\right) \le \frac{4}{\epsilon_0(2\pi n)^{1/2}}\exp\left(-\frac{n\epsilon_0^2}{32}\right).
\end{aligned}
$$

Taking $C_0 = \max\left\{2^{-1}(2\pi)^{-1/2}, 4\epsilon_0^{-1}(2\pi)^{-1/2}\right\}$ and $C_1 = \epsilon_0^2/32$ completes the proof of Lemma 27. □

Lemma 28. *Suppose Assumption 14.1, 15.1 and 16.1 hold. For any $\epsilon > 0$, there exist N, i and $\epsilon_0 > 0$, such that, for all $n \ge N$ and all $\theta \in \Theta_n$ that $\|\theta-\theta_0\|_1 > \varepsilon$, we have*

$$
\sum_{j=1}^m |\mu_{\pm,i}(v_j) - \mu_{\pm,i,0}(v_j)| > \epsilon_0 m
$$

.

Proof: We first note that,

$$
\begin{aligned}
\|\theta-\theta_0\|_1 = &\sum_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)} |\rho(v)-\rho_0(v)| + \max_{i=1,\dots,n}\sum_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)} |E_{+,i}(v)-E_{+,i,0}(v)|\\
&+ \max_{i=1,\dots,n}\sum_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)} |E_{-,i}(v)-E_{-,i,0}(v)|
\end{aligned}
\tag{B.8}
$$

Since $\|\theta-\theta_0\|_1 > \epsilon$, at lease one of the three terms in (B.8) is greater than $\epsilon/3$. Without loss of generality, suppose $\max_{i=1,\dots,n}\left\{\sum_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)}|E_{+,i}(v)-E_{+,i,0}(v)|\right\} > \epsilon/3$. Then there exist i, such that

$\sum\limits_{v\in\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)} |E_{+,i}(v) - E_{+,i,0}(v)| > \epsilon/3$. Therefore,

$$\begin{aligned}
&\sum_{j=1}^{m} |\mu_{\pm,i}(v_j) - \mu_{\pm,i,0}(v_j)| = \sum_{j=1}^{m} |s\{\rho(v_j)\}E_{+,i}(v) - s\{\rho_0(v_j)\}E_{+,i,0}(v)| \\
&= \sum_{j=1}^{m} |s\{\rho(v_j)\}\{E_{+,i}(v) - E_{+,i,0}(v)\} + E_{+,i,0}(v)[s\{\rho(v_j)\} - s\{\rho_0(v_j)\}]| \qquad \text{(B.9)} \\
&> \sum_{j=1}^{m} |s\{\rho(v_j)\}|\,|E_{+,i}(v_j) - E_{+,i,0}(v_j)| - \sum_{j=1}^{m} |E_{+,i,0}(v_j)|\,|s(\rho(v_j)) - s(\rho_0(v_j))|
\end{aligned}$$

By Definition 13, there exists $C_\rho > 0$, such that $|s\{\rho(v_j)\}| > C_\rho$ when $v_j \in \mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)$. By the compactness of $\mathcal{V}(\rho)\cup\mathcal{V}(\rho_0)$, there exists C, such that $\max\limits_{j=1,\ldots,m} |E_{+,i,0}(v_j)|\,|s(\rho(v_j)) - s(\rho_0(v_j))| < C$. Therefore,

$$\sum_{j=1}^{m} |\mu_{+,i}(v_j) - \mu_{+,i,0}(v_j)| > C_\rho m\epsilon/3 - mC$$

Taking $\epsilon_0 = C_\rho\epsilon/3 - C$ completes the proof of Lemma 28. □

Lemma 29. *For any $\epsilon^\star > 0$ and $\nu_0 < \nu < \frac{1}{2}$, there exist N, C_0, C_1 and C_2, such that, for all $n > N$ and $\theta \in \Theta_n$, if $\|\theta - \theta_0\|_1 > \epsilon^\star$, a test function Ψ_n can be constructed satisfying that $\mathbb{E}_{\theta_0}(\Psi_n) \le C_0 \exp(-C_2 n^{2\nu})$ and $\mathbb{E}_\theta(1 - \Psi_n) \le C_0\exp(-C_1 n)$, where ν_0 is as defined in Assumption 16.1.*

Proof: Let N_t be the t covering number of Θ_n in the supremum norm. Let $\theta^1, \ldots, \theta^{N_t} \in \Theta_n$ satisfy that, for each $\theta \in \Theta_n$, there exist at least one l such that $\left\|\theta - \theta^l\right\|_\infty < t$. For any $\theta \in \Theta_n$, define

$$\Psi_n = \max_{1\le l\le N_t} \Psi_n\left(\theta_0, \theta^l\right), \qquad \text{(B.10)}$$

where $\Psi_n\left(\theta_0, \theta^l\right)$ is the test statistic constructed in Lemma 27 for the hypothesis testing problem $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta^l$. If $\|\theta - \theta_0\|_1 > \epsilon^\star$, then for θ^l satisfying that $\left\|\theta - \theta^l\right\|_1 < t \le \epsilon^\star/2$, we have $\left\|\theta^l - \theta_0\right\|_1 > \epsilon^\star/2$. By Lemma 28, there exist N_0^*, i and $\epsilon > 0$, such that $\sum_{j=1}^m \left|\mu^l_{+,i}(v_j) - \mu_{+,i,0}(v_j)\right| > \epsilon m$. By Lemma 27, we can choose ϵ_0, such that

$$\mathbb{E}_{\theta_0}\left\{\Psi_n\left(\theta_0, \theta^l\right)\right\} \le C_0\exp\left(-2n^{2\nu}\right), \quad \text{and} \quad \mathbb{E}_\theta\left\{1 - \Psi_n\left(\theta_0, \theta^l\right)\right\} \le C_0\exp(-C_1 n).$$

Furthermore, we have,

$$\begin{aligned}
\mathbb{E}_{\theta_0}(\Psi_n) &\le \sum_{l=1}^{N_t} \Psi_n\left(\theta_0, \theta^l\right) \le C_0 N_t \exp\left(-2n^{2\nu}\right) = C_0\exp\left(\log N_t - 2n^{2\nu}\right) \\
&\le C_0\exp\left\{Cn^{1/(2\alpha)}t^{-d/\alpha} - 2n^{2\nu}\right\} \le C_0\exp\left(Cn^{\nu_0}t^{-d/\alpha} - 2n^{2\nu}\right) \\
&= C_0\exp\left\{-\left(2 - Cn^{\nu_0 - 2\nu}t^{-d/\alpha}\right)n^{2\nu}\right\}.
\end{aligned}$$

When $Ct^{-d/\alpha} < 2$, $\mathbb{E}_{\theta_0}(\Psi_n) \le C_0\exp\left\{-\left(2 - Ct^{-d/\alpha}\right)n^{2\nu}\right\}$. When $Ct^{-d/\alpha} \ge 2$, since $\nu_0 - 2\nu <$

0, there exists $N_1^\star$, such that, for all $n > N_1^*$, $Cn^{\nu_0 - 2\nu} t^{-d/\alpha} < 1$. Then $\mathbb{E}_{\theta_0}(\Psi_n) \leq C_0 \exp\{-n^{2\nu}\}$. In addition,

$$\mathbb{E}_\theta (1 - \Psi_n) = \mathbb{E}_\theta \left[\min_{1 \leq l \leq N_t} \left\{ 1 - \Psi_n \left(\theta_0, \theta^l \right) \right\} \right] \leq \mathbb{E}_\theta \left[\left\{ 1 - \Psi_n \left(\theta_0, \theta^l \right) \right\} \right] \leq C_0 \exp(-C_1 n)$$

Taking $C_2 = \left(2 - Ct^{-d/\alpha}\right) I \left(Ct^{-d/\alpha} < 2\right) + I \left(Ct^{-d/\alpha} \geq 2\right) > 0$, and $N = \max\{N_1^*, N_0^*\}$ completes the proof of Lemma 29. □

B.3 Derivation of full conditional distributions

B.3.1 Full conditional distribution

We first summarize in Algorithm 3 the general procedure of deriving the full conditional distribution of θ using Proposition 19. The main steps are to first rewrite the density of θ in the form of (3.14), where $\{L_p\}_{p=1}^P$, $\{U_k\}_{k=1}^K$, $\{f_p(\theta)\}_{p=1}^P$, $\{h_k(\theta)\}_{k=1}^K$ are the input to Algorithm 3. We then sort $(L_1, \ldots, L_P, U_1, \ldots, U_K)$ in ascending order, which leads to $P + K + 1$ intervals. We next loop through all the intervals, and update the coefficient of $H_i(\theta)$ as shown in Algorithm 3 (line 12). Finally, after obtaining the unnormalized conditional density function of θ on each interval, we derive the full conditional density of θ by incorporating the corresponding normalizing constants.

B.3.2 Full conditional distribution of c_l

Without loss of generality, we only consider c_1 in the following discussion. By model (9) and the Karhunen-Loève expansion, we have $\mu_{\pm,i}(v) = G_\omega\{\pm\xi(v)\} E_{\pm,i}(v)$, $\xi(v) = \sum_{l=1}^L c_l \psi_l(v)$, and $E_{\pm,i}(v) = \sum_{l=1}^L e_{i,l,\pm} \psi_l(v)$. Given $Y_+, Y_-, \tilde{\Theta}_{\setminus c_1}$, the full conditional density of c_1 is,

$$\pi(c_1 \mid Y_+, Y_-, \tilde{\Theta}_{\setminus c_1}) \propto \exp\left(- \sum_{v \in \mathcal{B}_m} \frac{\sum_{i=1}^n W_i(v)}{K(v)} \right) \cdot \exp\left(- \frac{c_1^2}{2\lambda_l} \right), \tag{B.11}$$

where

$$W_i(v) = \{Y_{+,i}(v) - \mu_{+,i}(v)\}^2 + \{Y_{-,i}(v) - \mu_{-,i}(v)\}^2 + 2r(v)\{Y_{+,i}(v)\mu_{-,i}(v) + Y_{-,i}(v)\mu_{+,i}(v)\}$$

, and $K(v) = 2\{1 - r^2(v)\}u^2(v)$, with $r(v) = \{\tau_1^2(v) - \tau_2^2(v)\}/\{\tau_1^2(v) + \tau_2^2(v)\}$ and $u^2(v) = \{\tau_1^2(v) + \tau_2^2(v)\}/4$. Write $T_\pm(v) = \{\pm\lambda_1 - \sum_{l=2}^L c_l \psi_l(v)\}/\{\psi_1(v)\}$. According to the sign of $\psi_1(v)$, we have two different representations of $\sum_{i=1}^n W_i(v)$.

When $\psi_1(v) > 0$,

$$\sum_{i=1}^n W_i(v) = \{A_+(v)c_1^2 + B_+(v)c_1 + C_+(v)\} I\{c_1 > T_+(v)\}$$
$$+ \{A_-(v)c_1^2 + B_-(v)c_1 + C_-(v)\} I\{c_1 < T_-(v)\}.$$

Algorithm 3 Full conditional distribution of θ

```
Input: $\{L_p\}_{p=1}^{P}$, $\{U_k\}_{k=1}^{K}$, $\{f_p(\theta)\}_{p=1}^{P}$, $\{h_k(\theta)\}_{k=1}^{K}$.
Output: the full conditional distribution of $\theta$
Sort $(L_1, \dots, L_P, U_1, \dots, U_K)$ in ascending order, which leads to $P + K + 1$ intervals, denoted as $I_1, I_2, \dots, I_{P+K+1}$.
for interval $I_i$, $i = 1, \dots, P + K + 1$ do
    Initialize $D_i = E_i = F_i = 0$
    for $p = 1, \dots, P$, $k = 1, \dots, K$ do
        if $I_i \subset [L_p, +\infty)$ then
            $D_i = D_i + a_{1p}$, $E_i = E_i + a_{2p}$, $F_i = F_i + a_{3p}$.
        end if
        if $I_i \subset (-\infty, U_k]$ then
            $D_i = D_i + b_{1k}$, $E_i = E_i + b_{2k}$, $F_i = F_i + b_{3k}$.
        end if
    end for
    Write $H_i(\theta) = D_i\theta^2 + E_i\theta + F_i$.
end for
if there exists $i$, such that $D_i \neq 0$ then
    the full conditional distribution of $\theta$ is a mixture of truncated normal distributions.
end if
if $D_i = 0$ for all $i$, and there exists $i$, such that $E_i \neq 0$ then
    the full conditional distribution of $\theta$ is a mixture of truncated exponential distributions.
end if
if $D_i = E_i = 0$ for all $i$, and there exists $i$, such that $F_i \neq 0$ then
    the full conditional distribution of $\theta$ is a mixture of uniform distributions.
end if
```

When $\psi_1(v) < 0$,

$$\sum_{i=1}^{n} W_i(v) = \{A_+(v)c_1^2 + B_+(v)c_1 + C_+(v)\}I\{c_1 < T_+(v)\} \\ + \{A_-(v)c_1^2 + B_-(v)c_1 + C_-(v)\}I\{c_1 > T_-(v)\}.$$

where $A_\pm(v), B_\pm(v), C_\pm(v)$ are all functions of $\tilde{\Theta}_{\backslash c_1}$, and are of the form,

$$A_\pm(v) = \left\{\sum_{i=1}^{n} E_{\pm,,i}(v)^2\right\} \cdot \psi_1^2(v),$$

$$B_\pm(v) = 2\psi_1(v)\left[\left\{\sum_{l=2}^{L} c_l\psi_1(v)\right\}\left\{\sum_{i=1}^{n} E_{\pm,i}(v)^2\right\} \mp \sum_{i=1}^{n}\{Y_{\pm,i}(v) \cdot E_{\pm,i}(v)\} \mp r(v)\sum_{i=1}^{n}\{Y_{\mp,i}(v) \cdot E_{\pm,i}(v)\}\right],$$

$$C_\pm(v) = \left\{\sum_{l=2}^{L} c_l\psi_1(v)\right\}^2 \left\{\sum_{i=1}^{n} E_{\pm,i}(v)^2 \mp \frac{2 \cdot \sum_{i=1}^{n} Y_{\pm,i}(v)E_{\pm,i}(v)}{\sum_{l=2}^{L} c_l\psi_1(v)} \pm \frac{2r(v)\sum_{i=1}^{n} Y_{\mp,i}(v)E_{\pm,i}(v)}{\sum_{l=2}^{L} c_l\psi_1(v)}\right\}.$$

Algorithm 4 Full conditional distribution of c_l

```
Input: $P = K = m$, where $m$ is the number of spatial locations,
    $L_p = \begin{cases} T_+(v_j) & \text{if } \psi_l(v_j) > 0 \\ T_-(v_j) & \text{if } \psi_l(v_j) < 0 \end{cases}, U_k = \begin{cases} T_-(v_j) & \text{if } \psi_l(v_j) > 0 \\ T_+(v_j) & \text{if } \psi_l(v_j) < 0 \end{cases},$
    $f_p(\theta) = \begin{cases} g_+(c_l; v_j) & \text{if } \psi_l(v_j) > 0 \\ g_-(c_l; v_j) & \text{if } \psi_l(v_j) < 0 \end{cases}, h_k(\theta) = \begin{cases} g_-(c_l; v_j) & \text{if } \psi_l(v_j) > 0 \\ g_+(c_l; v_j) & \text{if } \psi_l(v_j) < 0 \end{cases}.$
Output: the full conditional distribution of $c_l$
Follow the procedure in Algorithm 3.
```

Therefore, given Y_+, Y_-, $\tilde{\Theta}_{\backslash c_1}$ and the eigenfunctions $\{\psi_1(v_j)\}_{j=1}^m$ evaluated on $\mathcal{B}_m$, we have,

$$\pi(c_1 \mid Y_+, Y_-, \tilde{\Theta}_{\backslash c_1}) \propto \exp\left(\sum_{\substack{j=1 \\ \psi_1(v_j)>0}}^{m} [g_+(c_1; v_j) I\{c_1 > T_+(v_j)\} + g_-(c_1; v_j) I\{c_1 < T_-(v_j)\}] \right.$$

$$\left. + \sum_{\substack{j=1 \\ \psi_1(v_j)<0}}^{m} [g_+(c_1; v_j) I\{c_1 < T_+(v_j)\} + g_-(c_1; v_j) I\{c_1 > T_-(v_j)\}] \right),$$

where

$$g_\pm(c_1; v_j) = \left\{ -\frac{A_\pm(v_j)}{K(v_j)} - \frac{1}{2\lambda_1^2} \right\} c_1^2 + \frac{B_\pm(v_j)}{K(v_j)} c_1 + \frac{C_\pm(v_j)}{K(v_j)}.$$

By Proposition 1, the full conditional distribution of c_1 is a mixture of truncated normal distributions. We summarize the procedure of obtaining this distribution in Algorithm 4.

B.3.3 Full conditional distribution of ω

Recall that the prior of ω is the uniform distribution on $[a_\omega, b_\omega]$. Then we have,

$$\pi(\omega \mid Y_+, Y_-, \tilde{\Theta}_{\backslash \omega}) \propto \exp\left\{ -\sum_{v \in \mathcal{B}_m} \frac{\sum_{i=1}^n W_i(v)}{K(v)} \right\} \cdot \frac{1}{b_\omega - a_\omega} I(a_\omega \leq \omega \leq b_\omega), \tag{B.12}$$

where $W_i(v)$ is defined as in (B.11). Then,

$$\sum_{i=1}^n W_i(v) = Q_+(v) I\{\omega < \xi(v)\} + Q_-(v) I\{\omega < -\xi(v)\},$$

where

$$Q_\pm(v) = \xi(v)^2 \left\{ \sum_{i=1}^n E_{\pm,i}(v)^2 \right\} \mp 2\xi(v) \left\{ \sum_{i=1}^n Y_{\pm,i}(v) E_{\pm,i}(v) \right\} \pm 2r(v)\xi(v) \left\{ \sum_{i=1}^n Y_{\mp,i}(v) E_{\pm,i}(v) \right\}.$$

Therefore, given Y_+, Y_-, $\tilde{\Theta}_{\backslash \omega}$ and the eigenfunctions $\psi_l(v_j), j = 1, \ldots, m, l = 1, \ldots, L$, evaluated on $\mathcal{B}_m$, we have,

$$\pi(\omega \mid Y_+, Y_-, \tilde{\Theta}_{\setminus\omega}) \propto \exp\left[\sum_{\substack{j=1 \\ a_\omega < \xi(v_j) < b_\omega}}^{m} C_+(v_j) I\{\omega < \xi(v_j)\} + \sum_{\substack{j=1 \\ a_\omega < -\xi(v_j) < b_\omega}}^{m} C_-(v_j) I\{\omega < -\xi(v_j)\}\right],$$

where $C_\pm(v_j) = -\dfrac{Q_\pm(v_j)}{K(v_j)} - \log(b_\omega - a_\omega)$, and we only consider those $\xi(v_j)$ and $-\xi(v_j)$ that are between a_ω and b_ω.

By Proposition 15, the full conditional distribution of ω is a mixture of uniform distributions. We summarize the procedure of obtaining this distribution in Algorithm 5.

B.3.4 Full conditional distribution of $e_{i,l\pm}$

Since $e_{i,l,+}$ only exist in $\mu_{+,i}(v)$, we can rewrite $\mu_{+,i}(v)$ as $\mu_{+,i}(v) = a_{+,i}(v) + b_{+,i}(v)$, where $a_{+,i}(v) = G_\omega\left\{\sum_{l=1}^{L} c_l \psi_l(v)\right\} e_{i,l,+}\psi_l(v) = C_{l,+}(v) \cdot e_{i,l,+}$, and $b_{+,i}(v) = G_\omega\left\{\sum_{l=1}^{L} c_l \psi_l(v)\right\}$ $\sum_{l' \neq l} e_{i,l',+}\psi_{l'}(v)$. Note that $b_{+,i}(v)$ does not depend on $e_{i,l,+}$. Henceforth, we have that,

$$\begin{aligned}
&\{Y_{+,i}(v) - \mu_{+,i}(v)\}^2 = \\
&\quad Y_{+,i}^2(v) + a_{+,i}^2(v) + b_{+,i}^2(v) + 2a_{+,i}(v)b_{+,i}(v) - 2Y_{+,i}(v)a_{+,i}(v) - 2Y_{+,i}(v)b_{+,i}(v), \\
&\{Y_{+,i}(v) - \mu_{+,i}(v)\}\{Y_{-,i}(v) - \mu_{-,i}(v)\} = \\
&\quad Y_{+,i}(v)\{Y_{-,i}(v) - \mu_{-,i}(v)\} - a_{+,i}(v)\{Y_{-,i}(v) - \mu_{-,i}(v)\} - b_{+,i}(v)\{Y_{-,i}(v) - \mu_{-,i}(v)\}.
\end{aligned}$$

Ignoring the terms $\{Y_{-,i}(v) - \mu_{-,i}(v)\}^2$ that do not contain $e_{i,l,+}$, we have,

$$\begin{aligned}
&\pi(e_{i,l+} \mid Y_+, Y_-, \tilde{\Theta}_{\setminus e_{i,l+}}) \\
&\propto \prod_{v \in \mathcal{B}_m} \exp\left(-\frac{a_{+,i}^2(v) + 2a_{+,i}(v)[b_{+,i}(v) - Y_{+,i}(v) - r(v)\{Y_{-,i}(v) - \mu_{-,i}(v)\}]}{2\{1 - r^2(v)\}u^2(v)}\right) \\
&\quad \cdot \exp\left(-\frac{e_{i,l,+}^2}{2\lambda_l}\right) \\
&\propto \exp\left[-\frac{1}{2}\frac{\{e_{i,l,+} - M_{i,l,+}\}^2}{V_{i,l,+}^2}\right].
\end{aligned}$$

Algorithm 5 Full conditional distribution of ω

```
Input: $P = 0$, $K = 2m$,
    $U_k = \begin{cases} \xi(v_j), & \text{if } a_\omega < \xi(v_j) < b_\omega \\ -\xi(v_j) & \text{if } a_\omega < -\xi(v_j) < b_\omega \end{cases}$, $h_k(\theta) = \begin{cases} C_+(v_j), & \text{if } U_k = \xi(v_j) \\ C_-(v_j) & \text{if } U_k = -\xi(v_j) \end{cases}$.
Output: the full conditional distribution of $\omega$
Follow the procedure in Algorithm 3
```

where $M_{i,l,\pm} = \sum\limits_{v\in\mathcal{B}_m} \left[\{\lambda_l m_{i,l,\pm}(v)\}/\{\lambda_l + \sigma^2_{i,l,\pm}(v)\}\right]$, $V^2_{i,l,\pm} = \sum\limits_{v\in\mathcal{B}_m} \left[\lambda_l \sigma^2_{i,l,\pm}(v)/\{\lambda_l + \sigma^2_{i,l,\pm}(v)\}\right]$, with $m_{i,l,\pm}(v) = -\left[\{Y_{\pm,i}(v) - b_{\pm,i}(v)\} - r(v)\cdot\{Y_{\pm,i}(v) - \mu_{\pm,i}(v)\}\right]/C_{l,\pm}(v)$, and $\sigma^2_{i,l,\pm}(v) = \{1 - r^2(v)\}u^2(v)/C^2_{l,\pm}(v)$. Therefore, $e_{i,l\pm}$ follows a normal distribution, i.e.,

$$e_{i,l\pm} \mid Y_+, Y_-, \tilde{\Theta}_{\backslash e_{i,l\pm}} \sim \mathrm{N}(M_{i,l,\pm}, V^2_{i,l,\pm}).$$

B.3.5 Full conditional distribution of $\tau_1^2(v)$ and $\tau_2^2(v)$

For a given $v_0 \in \mathcal{B}_m$, we have,

$$\begin{aligned}
&\pi\left\{\tau_1^2(v_0) \mid Y_+, Y_-, \tilde{\Theta}_{\backslash \tau_1^2(v_0)}\right\} \\
&\propto \prod_{i=1}^n \frac{1}{\sqrt{\tau_1^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{1}{\tau_1^2} + \frac{1}{\tau_2^2}\right)\left\{\tilde{Y}_{+,i}(v_0)^2 + \tilde{Y}_{-,i}(v_0)^2 - 2\frac{\tau_1^2 - \tau_2^2}{\tau_1^2 + \tau_2^2}\tilde{Y}_{+,i}(v_0)\tilde{Y}_{-,i}(v_0)\right\}\right] \cdot \Gamma^{-1}_{\tau_1^2}(a_\tau, b_\tau) \\
&\propto \left\{\frac{1}{\tau_1^2(v_0)}\right\}^{\frac{n}{2}} \exp\left[-\frac{1}{2\tau_1^2(v_0)}\sum_{i=1}^n \{Y_{+,i}(v_0) - \mu_{+,i}(v_0) + Y_{-,i}(v_0) - \mu_{-,i}(v_0)\}^2\right]
\end{aligned}$$

where $\tilde{Y}_{\pm,i}(v_0) = Y_{\pm,i}(v_0) - \mu_{\pm,i}(v_0)$. Therefore, we have,

$$\tau_1^2(v_0) \mid Y_+, Y_-, \tilde{\Theta}_{\backslash \tau_1^2(v_0)} \sim \mathrm{IG}\left(a_\tau + \frac{n}{2}, \frac{\sum_{i=1}^n \{\tilde{Y}_{+,i}(v_0) + \tilde{Y}_{-,i}(v_0)\}^2}{2} + nb_\tau\right).$$

Similarly, we have,

$$\tau_2^2(v_0) \mid Y_+, Y_-, \tilde{\Theta}_{\backslash \tau_2^2(v_0)} \sim \mathrm{IG}\left(a_\tau + \frac{n}{2}, \frac{\sum_{i=1}^n \left\{\tilde{Y}_{+,i}(v_0) - \tilde{Y}_{-,i}(v_0)\right\}^2}{2} + nb_\tau\right).$$

B.3.6 Derivation of hybrid mini-batch MCMC

We derive the acceptance ratio in the hybrid mini-batch MCMC. Let $Y = \{Y_{1i}(v), Y_{2i}(v), i = 1,\dots,n, v \in \mathcal{B}_m\}$, $Y_{m_s} = \{Y_{1i}(v), Y_{2i}(v), i = 1,\dots,n, v \in \mathcal{B}_{m_s},\}$, and $\tilde{\Theta} = \{\theta, \tilde{\Theta}_{\backslash\theta}\}$, where $m_s < m$, and henceforth $\mathcal{B}_{m_s} \subset \mathcal{B}_m$. In the Gibbs sampler, we use the full conditional distribution $P(\theta|Y, \tilde{\Theta}_{\backslash\theta})$ as the proposal function, with the acceptance ratio equal to 1. In the hybrid mini-batch MCMC, we use $P(\theta|Y_{m_s}, \tilde{\Theta}_{\backslash\theta})$ as the proposal function, and the acceptance ratio becomes,

$$\begin{aligned}
\phi(\theta', \theta) &= \min\left\{1, \frac{P(Y|\theta', \tilde{\Theta}_{\backslash\theta})}{P(Y|\theta, \tilde{\Theta}_{\backslash\theta})}\frac{P(\theta|Y_{m_s}, \tilde{\Theta}_{\backslash\theta})}{P(\theta'|Y_{m_s}, \tilde{\Theta}_{\backslash\theta})}\right\} \\
&= \min\left\{1, \frac{\prod_{v\in\mathcal{B}_m} P(Y(v)|\theta', \tilde{\Theta}_{\backslash\theta})}{\prod_{v\in\mathcal{B}_m} P(Y(v)|\theta, \tilde{\Theta}_{\backslash\theta})} \cdot \frac{\prod_{v\in\mathcal{B}_{m_s}} P(Y(v)|\theta, \tilde{\Theta}_{\backslash\theta})p(\theta)}{\prod_{v\in\mathcal{B}_{m_s}} P(Y(v)|\theta', \tilde{\Theta}_{\backslash\theta})p(\theta')}\right\} \\
&= \min\left\{1, \frac{\prod_{v\notin\mathcal{B}_{m_s}} P(Y(v)|\theta', \tilde{\Theta}_{\backslash\theta})}{\prod_{v\notin\mathcal{B}_{m_s}} P(Y(v)|\theta, \tilde{\Theta}_{\backslash\theta})}\right\}.
\end{aligned}$$

B.4 Additional numerical results

B.4.1 Additional simulations

We carry out some additional simulations for the 2D image example, where we vary the sample size $n = \{30, 50, 100\}$ while fixing the image resolution $m = 64 \times 64$, or when we vary $m = \{32 \times 32, 64 \times 64, 100 \times 100\}$ while fixing $n = 50$. Table B.1 reports the results averaged over 100 data replications. We see that our proposed method performs the best across different values of n and m. Meanwhile, it maintains a competitive performance even when n is relatively small or when m is relatively large.

B.4.2 Sensitivity analysis

In our hybrid mini-batch MCMC, we sample a subset of m_s voxels and use the full dataset after every T_0 iterations of using the mini-batch data. We next carry out a sensitivity analysis to study the effect of m_s and T_0. Table B.2 reports the results averaged over 100 data replications. We see that the results are relatively stable for different values of m_s and T_0.

B.4.3 Prior specification for the HCP data analysis

In our HCP data analysis, we set the prior for ω as $\mathrm{U}(a_\omega, b_\omega)$, and we choose a_ω and b_ω as the 75% quantile and 100% quantile of $\{|\xi(v)|\}_{v \in \mathcal{B}}$, respectively. The choice of a_ω is based on the belief that at most 25% voxels have non-zero correlations. Here we vary $a_\omega = \{0.73, 0.75, 0.77\}$, and investigate the corresponding performance of our proposed method. Table B.3 reports the results, which we see that are relatively stable across different choices of a_ω.

Table B.1: The 2D simulation example with the varying sample size n and the varying image resolution m. Reported are the average sensitivity, specificity, and FDR, with standard error in the parenthesis, based on 100 data replications. Six methods are compared: the voxel-wise analysis, the region-wise analysis, the integrated method with two thresholding values, 0.95 and 0.90, and the proposed Bayesian method (TCGP) with the Gibbs sampler and the hybrid mini-batch MCMC.

	Method	Positive Correlation			Negative Correlation		
		Sensitivity	Specificity	FDR	Sensitivity	Specificity	FDR
$n = 30$	Voxel-wise	0.080(0.002)	1.000(0.000)	0.004(0.003)	0.102(0.002)	1.000(0.000)	0.001(0.002)
	Region-wise	0.148(0.005)	0.971(0.002)	0.326(0.003)	0.473(0.006)	0.957(0.003)	0.624(0.003)
	Integrated(0.95)	0.518(0.005)	0.992(0.003)	0.199(0.008)	0.781(0.003)	0.993(0.004)	0.146(0.009)
	Integrated(0.90)	0.855(0.007)	0.960(0.005)	0.378(0.010)	0.871(0.004)	0.937(0.004)	0.392(0.011)
	TCGP (Gibbs)	0.910(0.004)	0.991(0.003)	0.109(0.005)	0.990(0.002)	0.993(0.001)	0.065(0.007)
	TCGP (Hybrid)	0.890(0.005)	0.990(0.003)	0.111(0.008)	0.983(0.004)	0.990(0.002)	0.110(0.008)
$n = 50$	Voxel-wise	0.098(0.002)	1.000(0.000)	0.002(0.001)	0.150(0.002)	1.000(0.000)	0.003(0.001)
	Region-wise	0.438(0.004)	0.953(0.005)	0.547(0.010)	0.573(0.003)	0.956(0.001)	0.629(0.010)
	Integrated(0.95)	0.659(0.003)	0.995(0.002)	0.130(0.008)	0.899(0.005)	0.997(0.001)	0.110(0.009)
	Integrated(0.90)	0.959(0.009)	0.970(0.005)	0.308(0.009)	0.969(0.003)	0.969(0.003)	0.355(0.010)
	TCGP (Gibbs)	0.941(0.004)	0.995(0.002)	0.081(0.005)	0.996(0.002)	0.992(0.001)	0.063(0.005)
	TCGP (Hybrid)	0.931(0.005)	0.993(0.003)	0.092(0.005)	0.993(0.002)	0.992(0.002)	0.086(0.006)
$n = 100$	Voxel-wise	0.102(0.004)	1.000(0.001)	0.002(0.003)	0.198(0.001)	1.000(0.000)	0.003(0.001)
	Region-wise	0.617(0.010)	0.881(0.003)	0.744(0.004)	0.476(0.005)	0.955(0.002)	0.631(0.010)
	Integrated(0.95)	0.714(0.005)	0.998(0.003)	0.099(0.005)	0.898(0.004)	0.997(0.002)	0.099(0.008)
	Integrated(0.90)	0.980(0.010)	0.969(0.010)	0.300(0.010)	0.975(0.003)	0.971(0.003)	0.298(0.011)
	TCGP (Gibbs)	0.953(0.002)	0.997(0.002)	0.041(0.002)	0.999(0.001)	0.997(0.001)	0.033(0.001)
	TCGP (Hybrid)	0.945(0.003)	0.997(0.002)	0.069(0.003)	0.993(0.003)	0.996(0.001)	0.085(0.002)
$m = 32 \times 32$	Voxel-wise	0.017(0.001)	1.000(0.000)	0.005(0.001)	0.040(0.002)	1.000(0.000)	0.004(0.002)
	Region-wise	0.297(0.005)	0.945(0.005)	0.531(0.010)	0.472(0.003)	0.957(0.002)	0.617(0.010)
	Integrated(0.95)	0.620(0.005)	0.989(0.004)	0.138(0.005)	0.852(0.004)	0.989(0.001)	0.198(0.009)
	Integrated(0.90)	0.933(0.010)	0.971(0.006)	0.287(0.008)	0.944(0.004)	0.957(0.005)	0.300(0.011)
	TCGP (Gibbs)	0.931(0.003)	0.993(0.002)	0.083(0.003)	0.991(0.004)	0.992(0.003)	0.065(0.004)
	TCGP (Hybrid)	0.922(0.005)	0.992(0.002)	0.082(0.005)	0.991(0.005)	0.991(0.002)	0.089(0.005)
$m = 64 \times 64$	Voxel-wise	0.098(0.002)	1.000(0.000)	0.002(0.001)	0.150(0.002)	1.000(0.000)	0.003(0.001)
	Region-wise	0.438(0.004)	0.953(0.005)	0.547(0.010)	0.573(0.003)	0.956(0.001)	0.629(0.010)
	Integrated(0.95)	0.659(0.003)	0.995(0.002)	0.130(0.008)	0.899(0.005)	0.997(0.001)	0.110(0.009)
	Integrated(0.90)	0.959(0.009)	0.970(0.005)	0.308(0.009)	0.969(0.003)	0.969(0.003)	0.355(0.010)
	TCGP (Gibbs)	0.941(0.004)	0.995(0.002)	0.081(0.005)	0.996(0.002)	0.992(0.001)	0.063(0.004)
	TCGP (Hybrid)	0.931(0.005)	0.993(0.003)	0.092(0.005)	0.993(0.002)	0.992(0.002)	0.086(0.006)
$m = 100 \times 100$	Voxel-wise	0.005(0.001)	1.000(0.000)	0.004(0.002)	0.011(0.001)	1.000(0.000)	0.000(0.001)
	Region-wise	0.627(0.002)	0.861(0.003)	0.763(0.005)	0.462(0.002)	0.948(0.002)	0.663(0.008)
	Integrated(0.95)	0.843(0.002)	0.998(0.003)	0.039(0.005)	0.952(0.004)	0.997(0.002)	0.052(0.007)
	Integrated(0.90)	0.960(0.003)	0.965(0.005)	0.300(0.012)	0.977(0.002)	0.965(0.004)	0.298(0.011)
	TCGP (Gibbs)	0.971(0.001)	0.999(0.000)	0.029(0.002)	0.997(0.001)	0.998(0.002)	0.031(0.001)
	TCGP (Hybrid)	0.964(0.001)	0.999(0.000)	0.033(0.001)	0.995(0.001)	0.997(0.002)	0.033(0.002)

Table B.2: The sensitivity analysis of the batch size m_s and the number of iterations T_0 for the hybrid mini-batch MCMC. Reported are the average sensitivity, specificity, and FDR, with standard error in the parenthesis, based on 100 data replications.

m_s	T_0	Positive Correlation			Negative Correlation		
		Sensitivity	Specificity	FDR	Sensitivity	Specificity	FDR
$m/32$	20	0.950(0.003)	1.000(0.001)	0.015(0.003)	0.991(0.002)	0.989(0.003)	0.050(0.005)
$m/16$	20	0.953(0.003)	0.996(0.001)	0.061(0.002)	0.991(0.003)	0.997(0.001)	0.049(0.005)
$m/4$	20	0.955(0.002)	0.997(0.001)	0.058(0.002)	0.990(0.001)	0.997(0.001)	0.047(0.003)
$m/16$	50	0.948(0.003)	0.998(0.001)	0.045(0.002)	0.990(0.002)	0.990(0.003)	0.062(0.003)
$m/16$	20	0.953(0.003)	0.996(0.001)	0.061(0.002)	0.991(0.003)	0.997(0.001)	0.049(0.005)
$m/16$	10	0.953(0.001)	0.995(0.001)	0.059(0.003)	0.993(0.002)	0.998(0.001)	0.041(0.004)

Table B.3: Prior specification for the HCP example under different choices of a_ω. Reported are the activation regions containing more than 100 voxels that are declared having a nonzero correlation.

		Lingual-R		
a_ω	cluster size	Activation center	overlap rate	mean correlation
0.73	151	(-10.0, -74.5, -4.0)	0.931	0.35
0.75	144	(-10.4, -75.3, -4.5)	1.000	0.35
0.77	140	(-10.6, -75.8, -5.4)	0.905	0.38
		Angular-R		
a_ω	cluster size	cluster center	overlap rate	mean correlation
0.73	215	(-45.9, -60.1, 45.5)	0.910	0.41
0.75	209	(-46.9, -60.2, 44.7)	1.000	0.43
0.77	200	(-46.0, -59.9, 43.9)	0.911	0.43
		Temporal-Mid-L		
a_ω	cluster size	cluster center	overlap rate	mean correlation
0.73	110	(62.1, -24.9, 1.3)	0.940	0.42
0.75	104	(63.1, -25.7, 1.4)	1.000	0.41
0.77	99	(62.7, -25.5, 1.3)	0.921	0.43
		Precentral-L		
a_ω	cluster size	cluster center	overlap rate	mean correlation
0.73	130	(29.1, -23.0, 64.5)	0.930	-0.41
0.75	115	(28.6, -23.1, 65.4)	1.000	-0.44
0.77	107	(28.8, -23.1, 65.8)	0.931	-0.42
		Occipital-Inf-R		
a_ω	cluster size	cluster center	overlap rate	mean correlation
0.73	130	(-38.1, -81.0, -3.9)	0.910	-0.45
0.75	122	(-38.8, -81.7, -3.2)	1.000	-0.44
0.77	107	(-38.5, -80.0, -4.0)	0.901	-0.43

APPENDIX C

Supplementary Material of Chapter 3

C.1 Derivation of full conditional distributions

For the simplicity of notations, let

$$
\begin{aligned}
S_R &= \frac{1}{KT}\sum_{k=1}^{K}\left(\sum_{t=1}^{T}\beta_k(t)X_{ki}(t)\right) \\
S_0 &= \frac{1}{K_0}\sum_{k_1<k_2}\beta_0(k_1,k_2)X_{0i}(k_1,k_2)
\end{aligned}
$$

C.1.1 Sample e_{kl}

First, consider the full conditional distribution of $e_{k_0l_0}$.

$$\mu_i = S_i + C_i e_{k_0l_0} \tag{C.1}$$

where

$$
\begin{aligned}
C_i &= \frac{1}{KT}\sum_{t=1}^{T}\left(\psi_{l_0}(t)I(|\tilde{E}_{k_0}(t)|>\omega)X_{k_0i}(t)\right) \\
S_i &= S_R + S_0 - C_i e_{k_0l_0}
\end{aligned} \tag{C.2}
$$

Hence

$$
\begin{aligned}
\pi(e_{k_0l_0}|\Theta_{\backslash e_{k_0l_0}}, Y) &\propto \exp\left\{-\frac{\lambda_{l_0}\sum_{i=1}^{n}(Y_i-\mu_i)^2 + e_{k_0l_0}^2}{2\lambda_{l_0}}\right\} \\
&= \exp\left\{-\frac{\lambda_{l_0}\sum_{i=1}^{n}(Y_i-S_i-T_ie_{k_0l_0})^2 + e_{k_0l_0}^2}{2\lambda_{l_0}}\right\} \\
&\propto \exp\left\{-\frac{(\lambda_{l_0}\sum_{i=1}^{n}T_i^2+1)e_{k_0l_0}^2 - 2\lambda_{l_0}(\sum_{i=1}^{n}T_i(Y_i-S_i))e_{k_0l_0}}{2\lambda_{l_0}}\right\}
\end{aligned} \tag{C.3}
$$

Hence the full conditional distribution of $e_{k_0l_0}$ is N $\left(-\frac{2\lambda_{l_0}\sum_{i=1}^{n}T_i(Y_i-S_i)}{\lambda_{l_0}\sum_{i=1}^{n}T_i^2+1}, \frac{\lambda_{l_0}}{\lambda_{l_0}\sum_{i=1}^{n}T_i^2+1}\right)$.

C.1.2 Sample $\tilde{E}_{k_0}(t_0)$

Then considering the full conditional distribution of $\tilde{E}_{k_0}(t_0)$. Rewrite μ_i as a function of $\tilde{E}_{k_0}(t_0)$ as follows.

$$\begin{aligned}\mu_i = {} & E_{k_0}(t_0)I(|\tilde{E}_{k_0}(t_0)| > \omega)X_{k_0 i}(t_0) + \sum_{t \neq t_0} E_{k_0}(t)I(|\tilde{E}_{k_0}(t)| > \omega)X_{k_0 i}(t) \\ & + \sum_{k \neq k_0} \sum_{t=1}^{T} \beta_k(t)X_{ki}(t) + S_0 \\ & = S_i + C_i I(|\tilde{E}_{k_0}(t_0)| > \omega)\end{aligned} \tag{C.4}$$

where

$$\begin{aligned}S_i &= \sum_{t \neq t_0} E_{k_0}(t)I(|\tilde{E}_{k_0}(t)| > \omega)X_{k_0 i}(t) + \sum_{k \neq k_0} \sum_{t=1}^{T} \beta_k(t)X_{ki}(t) + S_0 \\ C_i &= E_{k_0}(t_0)X_{k_0 i}(t_0)\end{aligned} \tag{C.5}$$

Then $(Y_i - \mu_i)^2 = (Y_i - S_i)^2 + \left\{C_i^2 - 2C_i(Y_i - S_i)\right\} I(|\tilde{E}_{k_0}(t_0)| > \omega)$. Hence

$$\begin{aligned}\tilde{E}_{k_0}(t_0)|\Theta_{\setminus \tilde{E}_{k_0}(t_0)}, X, X_0, Y &\propto \exp\left(-\frac{\sum_{i=1}^{n}(Y_i - \mu_i)^2}{2}\right) \cdot \exp\left(-\frac{(\tilde{E}_{k_0}(t_0) - E_{k_0}(t_0))^2}{2\xi^2}\right) \\ &= \exp\left(-\frac{A\tilde{E}^2_{k_0}(t_0) + B\tilde{E}_{k_0}(t_0) + C}{2\xi^2}\right)\end{aligned} \tag{C.6}$$

- when $|\tilde{E}_{k_0}(t_0)| > \omega$, $A = 1$, $B = -2E_{k_0}(t_0)$ and $C = E^2_{k_0}(t_0) + \xi^2 \sum_{i=1}^{n}(Y_i - S_i - C_i)^2$
- when $|\tilde{E}^2_{k_0}(t_0)| < \omega$, $A = 1$, $B = -2E_{k_0}(t_0)$ and $C = E^2_{k_0}(t_0) + \xi^2 \sum_{i=1}^{n}(Y_i - S_i)^2$

Hence

$$\tilde{E}_{k_0}(t_0)|\Theta_{\setminus \tilde{E}_{k_0}(t_0)}, X, Y \propto M_1 \mathrm{N}_{(-\infty,-\omega)}(m, V^2) + M_2 \mathrm{N}_{(-\omega,\omega)}(m, V^2) + M_3 \mathrm{N}_{(\omega,\infty)}(m, V^2) \tag{C.7}$$

where $m = -\dfrac{B}{2A} = E_{k_0}(t_0)$, $V^2 = \dfrac{\xi^2}{A} = \xi^2$ and

$$\begin{aligned}M_1 &= \left\{\Phi\left(\frac{-\omega - m}{V^2}\right) - 0\right\} \exp\left\{-\frac{\xi^2 \sum_{i=1}^{n}(Y_i - S_i - C_i)^2}{2\xi^2}\right\} \\ M_2 &= \left\{\Phi\left(\frac{\omega - m}{V^2}) - \Phi(\frac{-\omega - m}{V^2}\right)\right\} \exp\left\{-\frac{\xi^2 \sum_{i=1}^{n}(Y_i - S_i)^2}{2\xi^2}\right\} \\ M_3 &= \left\{1 - \Phi\left(\frac{\omega - m}{V^2}\right)\right\} \exp\left\{-\frac{\xi^2 \sum_{i=1}^{n}(Y_i - S_i - C_i)^2}{2\xi^2}\right\}\end{aligned} \tag{C.8}$$

C.1.3 Sample $\eta(k_1, k_2)$

$$\mu_i = S_i + C_i \eta(k_1, k_2) \tag{C.9}$$

where $S_i = S_R + S_0 - C_i\eta(k_1,k_2)$ and $C_i = I(|\tilde{\eta}(k_1,k_2)| > \omega)X_{0i}(k_1,k_2)$.

$$\begin{aligned}\pi(\eta(k_1,k_2)|\Theta_{\setminus e_{k_0 l_0}}, X, X_0, Y) &\propto \exp\left\{-\frac{\sum_{i=1}^n (Y_i-\mu_i)^2}{2} - \frac{\eta(k_1,k_2)^2}{2\sigma_\eta^2}\right\} \\ &\propto \exp\left\{-\frac{\sum_{i=1}^n \{(C_i^2+1)\eta(k_1,k_2)^2 - 2C_i(Y_i-S_i)\eta(k_1,k_2)\}}{2\sigma_\eta^2}\right\}\end{aligned} \tag{C.10}$$

Hence the full conditional distribution of $\eta(k_1,k_2)$ is $\mathrm{N}\left(\frac{\sum_{i=1}^n C_i(Y_i-S_i)}{\sum_{i=1}^n C_i^2+1}, \frac{\sigma_\eta^2}{\sum_{i=1}^n C_i^2+1}\right)$

C.1.4 Sample $\tilde{\eta}(k_1,k_2)$

$$\mu_i = S_i + C_i\eta(k_1,k_2) \tag{C.11}$$

where $S_i = S_R + S_0 - C_i I(|\tilde{\eta}(k_1,k_2)| > \omega)$ and $C_i = \eta(k_1,k_2)X_{0i}(k_1,k_2)$.

Hence

$$\begin{aligned}\tilde{\eta}(k_1,k_2)|\Theta_{\setminus\tilde{\eta}(k_1,k_2)}, X, Y &\propto \exp\left(-\frac{\sum_{i=1}^n (Y_i-\mu_i)^2}{2}\right)\cdot\exp\left(-\frac{\{\tilde{\eta}(k_1,k_2)-\eta(k_1,k_2)\}^2}{2\xi^2}\right) \\ &= \exp\left(-\frac{A\tilde{\eta}^2(k_1,k_2)+B\tilde{\eta}(k_1,k_2)+C}{2\xi^2}\right)\end{aligned} \tag{C.12}$$

- when $|\tilde{\eta}(k_1,k_2)| > \lambda$, $A = 1$, $B = -2\eta(k_1,k_2)$ and $C = \eta(k_1,k_2)^2 + \xi^2\sum_{i=1}^n (Y_i - S_i - C_i)^2$
- when $|\tilde{\eta}(k_1,k_2))| < \lambda$, $A = 1$, $B = -2\eta(k_1,k_2)$ and $C = \eta(k_1,k_2)^2 + \xi^2\sum_{i=1}^n (Y_i - S_i)^2$

Hence

$$\tilde{\eta}(k_1,k_2)|\Theta_{\setminus\tilde{\eta}(k_1,k_2)}, X, Y \propto M_1\mathrm{N}_{(-\infty,-\omega)}(m,V^2) + M_2\mathrm{N}_{(-\omega,\omega)}(m,V^2) + M_3\mathrm{N}_{(\omega,\infty)}(m,V^2) \tag{C.13}$$

where $m = -\frac{B}{2A} = \eta(k_1,k_2)$, $V^2 = \frac{\xi^2}{A} = \xi^2$ and

$$\begin{aligned}M_1 &= \left\{\Phi\left(\frac{-\omega-m}{V^2}\right) - 0\right\}\exp\left\{-\frac{\xi^2\sum_{i=1}^n (Y_i-S_i-C_i)^2}{2\xi^2}\right\} \\ M_2 &= \left\{\Phi\left(\frac{\omega-m}{V^2}) - \Phi(\frac{-\omega-m}{V^2}\right)\right\}\exp\left\{-\frac{\xi^2\sum_{i=1}^n (Y_i-S_i)^2}{2\xi^2}\right\} \\ M_3 &= \left\{1-\Phi\left(\frac{\omega-m}{V^2}\right)\right\}\exp\left\{-\frac{\xi^2\sum_{i=1}^n (Y_i-S_i-C_i)^2}{2\xi^2}\right\}\end{aligned} \tag{C.14}$$

Algorithm 6 Gibbs sampling for SI-RTGP

```
input: the stimulus type outcomes $Y$.
        the kernel function $\kappa(\cdot,\cdot)$, the Karhunen-Loève truncation number $L$,
        the prior hyperparameters $a_{\omega_1}, b_{\omega_1}, a_{\omega_2}, b_{\omega_2}$. The total number of iterations $Q$
output: the posterior samples of
        $\Theta = \{\{\{e_{kl}\}_{l=1}^{L}, \{\tilde{E}_k(t)\}_{t=1}^{T}\}_{k=1}^{K}, \{\eta\left(k_1,k_2\right), \tilde{\eta}\left(k_1,k_2\right)\}_{k_1<k_2}, \omega_1, \omega_2\}$.
initialize $\Theta$: sample $\Theta$ from the prior distribution.
for $q = 1, \cdots, Q$ do
    sample $e_{kl}$ from the normal distribution, $l = 1, \ldots, L$, $k = 1, \cdots, K$.
    sample $\tilde{E}_k(t)$ from the normal distribution, $t = 1, \ldots, T$, $k = 1, \cdots, K$.
    for $k_1 = 1, \ldots, K-1$ do
        for $k_2 = k_1+1, \ldots, K$ do
            sample $\eta\left(k_1,k_2\right)$ and $\tilde{\eta}\left(k_1,k_2\right)$ from the normal distribution.
        end for
    end for
    sample $\omega_1$ and $\omega_2$ from the discrete distribution as shown in Eq.(C.15).
    sample $Z_i$ from the normal distribution, $i = 1, \ldots, n$.
end for
```

C.1.5 Sample ω_1 and ω_2

The prior of ω_1 is discrete prior with $P(\omega_1 = \gamma_{1z}) = 1/Z$. Hence the posterior of ω_1 is still the discrete prior with $P(\omega_1 = \gamma_{1z}) = p_z$. where

$$p_z \propto \frac{1}{Z}\exp\left\{-\frac{\sum_{i=1}^{n}(Y_i-\mu_i)^2|_{\omega_1=\gamma_z}}{2}\right\}. \tag{C.15}$$

C.2 Gibbs Sampling

First, we have the equivalent model representation for the probit regression as follow

$$Y_i|\mu_i \sim \text{ Bernoulli}\left\{\Phi(\mu_i)\right\} \tag{C.16}$$

$$\left(Z_i|\{X_{ki}\}_{k=0}^{K}, \{\beta_k\}_{k=0}^{K}\right) \sim \mathrm{N}(\mu_i, 1), \quad i = 1, \ldots, n \tag{C.17}$$

$$\mu_i = \frac{1}{p}\sum_{k=1}^{K}\left(\sum_{t=1}^{T}\beta_k(t)X_{ki}(t)\right) + \frac{1}{p_0}\sum_{k_1<k_2}\beta_0(k_1,k_2)X_{0i}(k_1,k_2) \tag{C.18}$$

where $Z = (Z_1, \cdots, Z_n)$ are latent variables introduced by Albert and Chib (1993) to obtain partial conjugacy. The sampling process is shown in Algorithm 1.

C.3 Proof of proposition 21

Proof. For the simplicity of notation, we use $T_r(\theta)$ to represent $T_r(\theta, \omega, \xi^2)$. The joint density of $T_r(\theta)$ and θ can be written as

$$\mathcal{P}_{(T_r(\theta),\theta)}(x, y) = \mathcal{P}_{T_r(\theta)|\theta=y}(x; y)\mathcal{P}_\theta(y)$$

where

$$\mathcal{P}_{T_r(\theta)|\theta=y}(x; y) = Pr(|\mathrm{N}(y, \xi^2)| > \omega)\mathcal{P}_\theta(x) + Pr(|\mathrm{N}(y, \xi^2)| \leq \omega)\delta_0(x) \tag{C.19}$$

Hence

$$\begin{aligned}\mathcal{P}_{T_r(\theta)}(x) &= \int \mathcal{P}_{(T_r(\theta),\theta)}(x, y)dy \\ &= \int \mathcal{P}_{(T_r(\theta)|\theta=y)}(x, y)\mathcal{P}_\theta(y)dy \\ &= \int \{Pr(|\mathrm{N}(y, \xi^2)| > \omega)\mathcal{P}_\theta(x) + Pr(|\mathrm{N}(y, \xi^2)| \leq \omega)\delta_0(x)\}\mathcal{P}_\theta(y)dy\end{aligned} \tag{C.20}$$

When $\xi^2 \to 0$,

$$\begin{aligned}\lim_{\xi^2\to 0} \mathcal{P}_{T_r(\theta)}(x) &= \int \lim_{\xi^2\to 0} \left[\{Pr(|\mathrm{N}(y, \xi^2)| > \omega)\mathcal{P}_\theta(x) + Pr(|\mathrm{N}(y, \xi^2)| \leq \omega)\delta_0(x)\}\mathcal{P}_\theta(y)dy\right] \\ &= \int [I_{|y|>\omega}\mathcal{P}_\theta(x) + I_{|y|\leq\omega}I(x = 0)]\mathcal{P}_\theta(y)dy \\ &= \mathcal{P}_{T_h(\theta)|\theta=y}\mathcal{P}_\theta(y)dy \\ &= \mathcal{P}_{T_h(\theta)}(x)\end{aligned} \tag{C.21}$$

On the other hand, when $\xi^2 \to \infty$,

$$\lim_{\xi^2\to\infty} \mathcal{P}_{T_r(\theta)}(x) = \int \mathcal{P}_\theta(x)\mathcal{P}_\theta(y)dy = \mathcal{P}_\theta(x) \tag{C.22}$$

Then consider $\Pr\left(|T_r(\theta, \omega, \xi^2) - T_h(\theta, \omega)| < \epsilon\right)$,

$$|T_r(\theta, \omega, \xi^2) - T_h(\theta, \omega)| = |f\{I(|\tilde{\theta}| > \omega) - I(|\theta| > \omega)\}| \tag{C.23}$$

For the case (1) $|\tilde{\theta}| > \omega$ and $|\theta| > \omega$ and case (2) $|\tilde{\theta}| < \omega$ and $|\theta| < \omega$, $|T_r(\theta, \omega, \xi^2) - T_h(\theta, \omega)| = 0$. Hence, we only consider the case (3) when $|\tilde{\theta}| > \omega$ and $|\theta| \leq \omega$ and the case (4) when $|\tilde{\theta}| \leq \omega$ and $|\theta| > \omega$. For case (3), there exist $\delta > 0$, such that $|\tilde{\theta}| \geq \omega + \delta$. Hence, given θ, we only need to choose ξ_1^2 such that $\Phi(-\theta/\xi_1^2) < \epsilon/(2\theta)$, where $\Phi(\cdot)$ represent the CDF of standard normal distribution. Similarly, we choose ξ_2^2 such that $\Phi(-\theta_2/\xi^2) < \epsilon/(2\theta)$. Therefore, $P(|f\{I(|\tilde{\theta}| > \omega) - I(|\theta| > \omega)\}| < \epsilon) > 0$. Similarly, we can prove that $\Pr\left(|T_r(\theta, \omega, \xi^2) - T_s(\theta^\star, \omega)| < \epsilon\right) > 0$. □

C.4 Sensitivity Analysis

In this section, we provide the sensitivity analysis of the thresholding parameters ω_1 and ω_2 for BCI analysis.

Table C.1: Sensitivity Analysis of the prior of ω_1

a_{ω_1}	108	112	114	151	152	212
0.23	0.933	0.563	0.964	0.930	0.633	0.686
0.25	0.932	0.566	0.965	0.931	0.634	0.686
0.27	0.932	0.564	0.965	0.929	0.635	0.687

Table C.2: Sensitivity Analysis of the prior of ω_2

a_{ω_2}	108	112	114	151	152	212
0.23	0.934	0.564	0.966	0.933	0.632	0.684
0.25	0.932	0.566	0.965	0.931	0.634	0.686
0.27	0.930	0.566	0.959	0.930	0.633	0.687

BIBLIOGRAPHY

Mahmoud Khaled Abd-Ellah, Ali Ismail Awad, Ashraf AM Khalaf, and Hesham FA Hamed. A review on brain tumor diagnosis from mri images: Practical implications, key achievements, and lessons learned. *Magnetic resonance imaging*, 61:300–318, 2019.

Philip J Withers, Charles Bouman, Simone Carmignato, Veerle Cnudde, David Grimaldi, Charlotte K Hagen, Eric Maire, Marena Manley, Anton Du Plessis, and Stuart R Stock. X-ray computed tomography. *Nature Reviews Methods Primers*, 1(1):18, 2021.

Wei Jiang, Yamn Chalich, and M Jamal Deen. Sensors for positron emission tomography applications. *Sensors*, 19(22):5019, 2019.

Antoine Verger, Stephan Grimaldi, Maria-Joao Ribeiro, Solène Frismand, and Eric Guedj. Single photon emission computed tomography/positron emission tomography molecular imaging for parkinsonism: A fast-developing field. *Annals of neurology*, 90(5):711–719, 2021.

Michael Kelberman, Shella Keilholz, and David Weinshenker. What's that (blue) spot on my mri? multimodal neuroimaging of the locus coeruleus in neurodegenerative disease. *Frontiers in neuroscience*, 14:583421, 2020.

Xin Niu, Fengqing Zhang, John Kounios, and Hualou Liang. Improved prediction of brain age using multimodal neuroimaging data. *Human brain mapping*, 41(6):1626–1643, 2020.

Vince D Calhoun and Jing Sui. Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(3): 230–244, 2016.

Emine Elif Tulay, Barış Metin, Nevzat Tarhan, and Mehmet Kemal Arıkan. Multimodal neuroimaging: basic concepts and classification of neuropsychiatric diseases. *Clinical EEG and neuroscience*, 50(1):20–33, 2019.

Hongtu Zhu, Tengfei Li, and Bingxin Zhao. Statistical learning methods for neuroimaging data analysis with applications. *arXiv preprint arXiv:2210.09217*, 2022.

Kyungho Won, Moonyoung Kwon, Minkyu Ahn, and Sung Chan Jun. Eeg dataset for rsvp and p300 speller brain-computer interfaces. *Scientific Data*, 9(1):388, 2022.

Jonathan R Wolpaw, Richard S Bedlack, Domenic J Reda, Robert J Ringer, Patricia G Banks, Theresa M Vaughan, Susan M Heckman, Lynn M McCane, Charles S Carmack, Stefan Winden, et al. Independent home use of a brain-computer interface by people with amyotrophic lateral sclerosis. *Neurology*, 91(3): e258–e267, 2018.

Yujun Gao, Xinfu Zhao, JiChao Huang, Sanwang Wang, Xuan Chen, Mingzhe Li, Fengjiao Sun, Gaohua Wang, and Yi Zhong. Abnormal regional homogeneity in right caudate as a potential neuroimaging biomarker for mild cognitive impairment: A resting-state fmri study and support vector machine analysis. *Frontiers in Aging Neuroscience*, 14:979183, 2022.

Monika Sethi, Shalli Rani, Aman Singh, and Juan Luis Vidal Mazón. A cad system for alzheimer's disease classification using neuroimaging mri 2d slices. *Computational and Mathematical Methods in Medicine*, 2022, 2022.

Ben Wu, Subhadip Pal, Jian Kang, and Ying Guo. Distributional independent component analysis for diverse neuroimaging modalities. *Biometrics*, 78(3):1092–1105, 2022a.

P Boonyakitanont, B Gabrielson, I Belyaeva, P Olikkal, J Songsiri, YP Wang, TW Wilson, VD Calhoun, JM Stephen, and T Adalı. An ica-based framework for joint analysis of cognitive scores and meg event-related fields. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3594–3598. IEEE, 2022.

PP Mini, Tessamma Thomas, and R Gopikakumari. Eeg based direct speech bci system using a fusion of smrt and mfcc/lpcc features with ann classifier. *Biomedical Signal Processing and Control*, 68:102625, 2021.

Andrea Cimmino, Angelo Ciaramella, Giovanni Dezio, and Pasquale Junior Salma. Non-linear pca neural network for eeg noise reduction in brain-computer interface. *Progresses in Artificial Intelligence and Neural Systems*, pages 405–413, 2021.

Neng Xu, Xiaorong Gao, Bo Hong, Xiaobo Miao, Shangkai Gao, and Fusheng Yang. Bci competition 2003-data set iib: enhancing p300 wave detection using ica-based subspace projections for bci applications. *IEEE transactions on biomedical engineering*, 51(6):1067–1072, 2004.

Matthias Kaper, Peter Meinicke, Ulf Grossekathoefer, Thomas Lingner, and Helge Ritter. Bci competition 2003-data set iib: support vector machines for the p300 speller paradigm. *IEEE Transactions on biomedical Engineering*, 51(6):1073–1076, 2004.

Yuhui Du, Zening Fu, Jing Sui, Shuang Gao, Ying Xing, Dongdong Lin, Mustafa Salman, Anees Abrol, Md Abdur Rahaman, Jiayu Chen, et al. Neuromark: An automated and adaptive ica based pipeline to identify reproducible fmri markers of brain disorders. *NeuroImage: Clinical*, 28:102375, 2020.

Hossam H Sultan, Nancy M Salem, and Walid Al-Atabany. Multi-classification of brain tumor images using deep neural network. *IEEE access*, 7:69215–69225, 2019.

Hassan Ali Khan, Wu Jue, Muhammad Mushtaq, and Muhammad Umer Mushtaq. Brain tumor classification in mri image using convolutional neural network. *Math. Biosci. Eng*, 17(5):6203–6216, 2020.

Apiwat Ditthapron, Nannapas Banluesombatkul, Sombat Ketrat, Ekapol Chuangsuwanich, and Theerawit Wilaiprasitporn. Universal joint feature extraction for p300 eeg classification using multi-task autoencoder. *IEEE Access*, 7:68415–68428, 2019.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Ghanahshyam B Kshirsagar and Narendra D Londhe. Weighted ensemble of deep convolution neural networks for single-trial character detection in devanagari-script-based p300 speller. *IEEE Transactions on Cognitive and Developmental Systems*, 12(3):551–560, 2019.

José L Marroquín, Baba C Vemuri, Salvador Botello, E Calderon, and Antonio Fernandez-Bouzas. An accurate and efficient bayesian method for automatic segmentation of brain mri. *IEEE transactions on medical imaging*, 21(8):934–945, 2002.

S Jayachitra and A Prasanth. Multi-feature analysis for automated brain stroke classification using weighted gaussian naïve bayes classifier. *journal of circuits, systems and computers*, 30(10):2150178, 2021.

Seyed Mostafa Kia, Hester Huijsdens, Richard Dinga, Thomas Wolfers, Maarten Mennes, Ole A Andreassen, Lars T Westlye, Christian F Beckmann, and Andre F Marquand. Hierarchical bayesian regression for multi-site normative modeling of neuroimaging data. In *Medical Image Computing and Computer*

Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23, pages 699–709. Springer, 2020.

Quentin Barthélemy, Sylvain Chevallier, Raphaëlle Bertrand-Lalo, and Pierre Clisson. End-to-end p300 bci using bayesian accumulation of riemannian probabilities. *Brain-Computer Interfaces*, 10(1):50–61, 2023.

Tianwen Ma, Yang Li, Jane E Huggins, Ji Zhu, and Jian Kang. Bayesian inferences on neural activity in eeg-based brain-computer interface. *Journal of the American Statistical Association*, 117(539): 1122–1133, 2022.

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. *arXiv preprint arXiv:1206.6380*, 2012.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

Akihiko Nishimura, David B Dunson, and Jianfeng Lu. Discontinuous hamiltonian monte carlo for discrete parameters and discontinuous likelihoods. *Biometrika*, 107(2):365–380, 2020.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

Bertil Matérn. *Spatial variation*, volume 36. Springer Science & Business Media, 2013.

Alex J Smola and Bernhard Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

Joseph Futoma, Sanjay Hariharan, Katherine Heller, Mark Sendak, Nathan Brajer, Meredith Clement, Armando Bedoya, and Cara O'brien. An improved multi-output gaussian process rnn with real-time validation for early sepsis detection. In *Machine Learning for Healthcare Conference*, pages 243–254. PMLR, 2017.

Yikuan Li, Shishir Rao, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Gholamreza Salimi-Khorshidi, Mohammad Mamouei, Thomas Lukasiewicz, and Kazem Rahimi. Deep bayesian gaussian processes for uncertainty estimation in electronic health records. *Scientific reports*, 11(1):1–13, 2021.

Ben Wu, Ying Guo, and Jian Kang. Bayesian spatial blind source separation via the thresholded gaussian process. *Journal of the American Statistical Association*, In Press, 2022b.

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.

Seyed Mostafa Kia and Andre Marquand. Normative modeling of neuroimaging data using scalable multi-task gaussian processes. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, pages 127–135. Springer, 2018.

Seyed Mostafa Kia, Christian F Beckmann, and Andre F Marquand. Scalable multi-task gaussian process tensor regression for normative modeling of structured variation in neuroimaging data. *arXiv preprint arXiv:1808.00036*, 2018.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, pages 63–71. Springer, 2004.

Sathiya Keerthi and Wei Chu. A matching pursuit approach to sparse gaussian process regression. *Advances in neural information processing systems*, 18, 2005.

Neil Lawrence, Matthias Seeger, and Ralf Herbrich. Fast sparse gaussian process methods: The informative vector machine. *Advances in neural information processing systems*, 15, 2002.

Arman Melkumyan and Fabio Tozeto Ramos. A sparse covariance function for exact gaussian process inference in large datasets. In *Twenty-first international joint conference on artificial intelligence*, 2009.

Kohei Hayashi, Masaaki Imaizumi, and Yuichi Yoshida. On random subsampling of gaussian process regression: A graphon-based analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 2055–2065. PMLR, 2020.

Dongbin Xiu. *Numerical methods for stochastic computations: a spectral method approach*. Princeton university press, 2010.

Robert B Gramacy and Daniel W Apley. Local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578, 2015.

Chiwoo Park and Jianhua Z Huang. Efficient computation of gaussian process regression for large spatial data sets by patching local gaussian processes. *1foldr Import 2019-10-08 Batch 6*, 2016.

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

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *The Artificial Intelligence Review*, 42(2):275, 2014.

Yanshuai Cao and David J Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Jouchi Nakajima and Mike West. Bayesian analysis of latent threshold dynamic models. *Journal of Business & Economic Statistics*, 31(2):151–164, 2013.

Yang Ni, Francesco C Stingo, and Veerabhadran Baladandayuthapani. Bayesian graphical regression. *Journal of the American Statistical Association*, 114(525):184–197, 2019.

Qingpo Cai, Jian Kang, and Tianwei Yu. Bayesian network marker selection via the thresholded graph laplacian gaussian prior. *Bayesian Analysis*, 15(1):79, 2020.

Taeryon Choi. *Posterior consistency in nonparametric regression problems under Gaussian process priors*. PhD thesis, Carnegie Mellon University, 2005.

Subhashis Ghosal and Anindya Roy. Posterior consistency of gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5):2413–2429, 2006.

Surya T Tokdar and Jayanta K Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137(1):34–42, 2007.

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.

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

Ido Tavor, O Parker Jones, Rogier B Mars, SM Smith, TE Behrens, and Saad Jbabdi. Task-free mri predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

O Parker Jones, NL Voets, JE Adcock, R Stacey, and S Jbabdi. Resting connectivity predicts task activation in pre-surgical populations. *NeuroImage: Clinical*, 13:378–385, 2017.

Alexander D Cohen, Ziyi Chen, Oiwi Parker Jones, Chen Niu, and Yang Wang. Regression-based machine-learning approaches to predict task activation using resting-state fMRI. *Human brain mapping*, 41(3): 815–826, 2020.

Chlöé Farrer, Scott H Frey, John D Van Horn, Eugene Tunik, David Turk, Souheil Inati, and Scott T Grafton. The angular gyrus computes action awareness representations. *Cerebral cortex*, 18(2):254–261, 2008.

Mohamed L Seghier. The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist*, 19(1):43–61, 2013.

Daniel J Acheson and Peter Hagoort. Stimulating the brain's language network: syntactic ambiguity resolution after tms to the inferior frontal gyrus and middle temporal gyrus. *Journal of cognitive neuroscience*, 25 (10):1664–1677, 2013.

Michael Koenigs, Aron K Barbey, Bradley R Postle, and Jordan Grafman. Superior parietal cortex is critical for the manipulation of information in working memory. *Journal of Neuroscience*, 29(47):14980–14986, 2009.

Eric D Leshikar, Audrey Duarte, and Christopher Hertzog. Task-selective memory effects for successfully implemented encoding strategies. *PloS one*, 7(5):e38160, 2012.

Dean J Krusienski, Eric W Sellers, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. Toward enhanced p300 speller performance. *Journal of neuroscience methods*, 167(1):15–21, 2008.

Michael T McCann, David E Thompson, Zeeshan H Syed, and Jane E Huggins. Electrode subset selection methods for an eeg-based p300 brain-computer interface. *Disability and Rehabilitation: Assistive Technology*, 10(3):216–220, 2015.

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

Andre G Journel and Charles J Huijbregts. Mining geostatistics. 1976.

Bradley Jones and Rachel T Johnson. Design and analysis for the gaussian process model. *Quality and Reliability Engineering International*, 25(5):515–524, 2009.

Chih-Li Sung, Ying Hung, William Rittase, Cheng Zhu, and CF Jeff Wu. A generalized gaussian process model for computer experiments with binary time series. *Journal of the American Statistical Association*, 115(530):945–956, 2020.

Jay M Ver Hoef and Ronald Paul Barry. Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, 69(2):275–294, 1998.

Dave Higdon. Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, pages 37–56. Springer, 2002.

Mauricio A Álvarez and Neil D Lawrence. Computationally efficient convolved multiple output gaussian processes. *Journal of Machine Learning Research*, 12(May):1459–1500, 2011.

Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Thomas E Fricker, Jeremy E Oakley, and Nathan M Urban. Multivariate gaussian process emulators with nonseparable covariance structures. *Technometrics*, 55(1):47–56, 2013.

Douglas Nychka, Soutir Bandyopadhyay, Dorit Hammerling, Finn Lindgren, and Stephan Sain. A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015.

Andreas C Damianou, Michalis K Titsias, and Neil D Lawrence. Variational inference for latent variables and uncertain inputs in gaussian processes. *The Journal of Machine Learning Research*, 17(1):1425–1486, 2016.

Robert B Gramacy. lagp: large-scale spatial modeling via local approximate gaussian processes in r. *Journal of Statistical Software*, 72(1):1–46, 2016.

Robert B Gramacy and Benjamin Haaland. Speeding up neighborhood search in local gaussian process prediction. *Technometrics*, 58(3):294–303, 2016.

David R Burt, Carl E Rasmussen, and Mark Van Der Wilk. Rates of convergence for sparse variational gaussian process regression. *arXiv preprint arXiv:1903.03571*, 2019.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18:1257, 2006.

Chih-Li Sung, Robert B Gramacy, and Benjamin Haaland. Potentially predictive variance reducing subsample locations in local gaussian process regression. *arXiv preprint arXiv:1604.04980*, 2016.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α -variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020.

Jing Zhao and Shiliang Sun. Variational dependent multi-output gaussian process dynamical systems. *The Journal of Machine Learning Research*, 17(1):4134–4169, 2016.

Trung V Nguyen, Edwin V Bonilla, et al. Collaborative multi-output gaussian processes. In *UAI*, pages 643–652, 2014.

Pablo Moreno-Muñoz, Antonio Artés, and Mauricio Álvarez. Heterogeneous multi-output gaussian process prediction. In *Advances in neural information processing systems*, pages 6711–6720, 2018.

Ke Wang, Geoff Pleiss, Jacob Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson. Exact gaussian processes on a million data points. In *Advances in Neural Information Processing Systems*, pages 14622–14632, 2019.

Duc-Trung Nguyen, Maurizio Filippone, and Pietro Michiardi. Exact gaussian process regression with distributed computations. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1286–1295, 2019.

Boya Zhang, D Austin Cole, and Robert B Gramacy. Distance-distributed design for gaussian process surrogates. *Technometrics*, 63(1):40–52, 2021.

Andrew Gordon Wilson. *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes*. PhD thesis, Citeseer, 2014.

Ning Zhang and Daniel W Apley. Brownian integrated covariance functions for gaussian process modeling: Sigmoidal versus localized basis functions. *Journal of the American Statistical Association*, 111(515): 1182–1195, 2016.

Gabriel Parra and Felipe Tobar. Spectral mixture kernels for multi-output gaussian processes. In *Advances in Neural Information Processing Systems*, pages 6681–6690, 2017.

Kai Chen, Twan van Laarhoven, Perry Groot, Jinsong Chen, and Elena Marchiori. Multioutput convolution spectral mixture for gaussian processes. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

Kyle R Ulrich, David E Carlson, Kafui Dzirasa, and Lawrence Carin. Gp kernels for cross-spectrum analysis. In *Advances in neural information processing systems*, pages 1999–2007, 2015.

Min Li, Min-Qian Liu, Xiao-Lei Wang, and Yong-Dao Zhou. Prediction for computer experiments with both quantitative and qualitative factors. *Statistics & Probability Letters*, 165:108858, 2020a.

Qiang Zhou, Peter ZG Qian, and Shiyu Zhou. A simple approach to emulation for computer models with qualitative and quantitative factors. *Technometrics*, 53(3):266–273, 2011.

Peter Z G Qian, Huaiqing Wu, and CF Jeff Wu. Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics*, 50(3):383–396, 2008.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Simon Mak, Chih-Li Sung, Xingjian Wang, Shiang-Ting Yeh, Yu-Hung Chang, V Roshan Joseph, Vigor Yang, and CF Jeff Wu. An efficient surrogate model for emulation and physics extraction of large eddy simulations. *Journal of the American Statistical Association*, 113(524):1443–1456, 2018.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Kai Chen, Perry Groot, Jinsong Chen, and Elena Marchiori. Spectral mixture kernels with time and phase delay dependencies. *arXiv preprint arXiv:1808.00560*, 2018.

Caroline Uhler. Gaussian graphical models: An algebraic and geometric perspective. *arXiv preprint arXiv:1707.04345*, 2017.

Mauricio Alvarez and Neil D Lawrence. Sparse convolved gaussian processes for multi-output regression. In *Advances in neural information processing systems*, pages 57–64, 2009.

Steffen Fieuws and Geert Verbeke. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431, 2006.

Yongxiang Li, Qiang Zhou, Xiaohu Huang, and Li Zeng. Pairwise estimation of multivariate gaussian process models with replicated observations: Application to multivariate profile monitoring. *Technometrics*, 60 (1):70–78, 2018.

Raed Kontar, Garvesh Raskutti, and Shiyu Zhou. Minimizing negative transfer of knowledge in multivariate gaussian processes: A scalable and regularized approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Marc Deisenroth and Jun Wei Ng. Distributed gaussian processes. In *International Conference on Machine Learning*, pages 1481–1490. PMLR, 2015.

David Moore and Stuart J Russell. Gaussian process random fields. In *Advances in Neural Information Processing Systems*, pages 3357–3365, 2015.

Volker Tresp. A bayesian committee machine. *Neural computation*, 12(11):2719–2741, 2000.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Wenjia Wang, Rui Tuo, and CF Jeff Wu. On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530):920–930, 2020.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Yongxiang Li and Qiang Zhou. Pairwise meta-modeling of multivariate output computer models using nonseparable covariance function. *Technometrics*, 58(4):483–494, 2016.

Juan Carlos Reboredo, Miguel A Rivera-Castro, and Gilney F Zebende. Oil and us dollar exchange rate dependence: A detrended cross-correlation approach. *Energy Economics*, 42:132–139, 2014.

Kâmil Uludağ and Alard Roebroeck. General overview on the merits of multimodal neuroimaging data fusion. *Neuroimage*, 102:3–10, 2014.

Sylvia Richardson, George C Tseng, and Wei Sun. Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application*, 3:181–209, 2016.

Dajiang Zhu, Tuo Zhang, Xi Jiang, Xintao Hu, Hanbo Chen, Ning Yang, Jinglei Lv, Junwei Han, Lei Guo, and Tianming Liu. Fusing DTI and fMRI data: a survey of methods and applications. *NeuroImage*, 102: 184–191, 2014a.

Carlo Cavaliere, Sivayini Kandeepan, Marco Aiello, Demetrius Ribeiro de Paula, Rocco Marchitelli, Salvatore Fiorenza, Mario Orsini, Luigi Trojano, Orsola Masotta, Keith St Lawrence, et al. Multimodal neuroimaging approach to variability of functional connectivity in disorders of consciousness: a pet/mri pilot study. *Frontiers in Neurology*, 9:861, 2018.

Lexin Li, Jian Kang, Samuel N Lockhart, Jenna Adams, and William J Jagust. Spatially adaptive varying correlation analysis for multimodal neuroimaging data. *IEEE transactions on medical imaging*, 38(1): 113–123, 2019.

Anita Harrewijn, Rany Abend, Julia Linke, Melissa A Brotman, Nathan A Fox, Ellen Leibenluft, Anderson M Winkler, and Daniel S Pine. Combining fmri during resting state and an attention bias task in children. *Neuroimage*, 205:116301, 2020.

Keith J Worsley, Jonathan E Taylor, Francesco Tomaiuolo, and Jason Lerch. Unified univariate and multivariate random field theory. *NeuroImage*, 23:S189–S195, 2004.

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, 2014b.

Jialiang Li, Chao Huang, Zhub Hongtu, and for the Alzheimer's Disease Neuroimaging Initiative. A functional varying-coefficient single-index model for functional response data. *Journal of the American Statistical Association*, 112(519):1169–1181, 2017.

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*, pages 1–15, 2020b.

Anirban Bhattacharya and David B Dunson. Sparse bayesian infinite factor models. *Biometrika*, pages 291–306, 2011.

James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

Dangna Li and Wing H Wong. Mini-batch tempered mcmc. *arXiv preprint arXiv:1707.09705*, 2017.

Tung-Yu Wu, YX Rachel Wang, and Wing H Wong. Mini-batch metropolis–hastings with reversible sgld proposal. *Journal of the American Statistical Association*, 117(537):386–394, 2022c.

Stephen M Smith, Christian F Beckmann, Jesper Andersson, Edward J Auerbach, Janine Bijsterbosch, Gwenaëlle Douaud, Eugene Duff, David A Feinberg, Ludovica Griffanti, Michael P Harms, et al. Resting-state fMRI in the human connectome project. *Neuroimage*, 80:144–168, 2013.

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.

John Mazziotta, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, et al. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322, 2001.

Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6): 510–523, 1988.

Gert Pfurtscheller, Gernot R Müller-Putz, Reinhold Scherer, and Christa Neuper. Rehabilitation with brain-computer interface systems. *Computer*, 41(10):58–65, 2008.

Jan Van Erp, Fabien Lotte, and Michael Tangermann. Brain-computer interfaces: beyond medical applications. *Computer*, 45(4):26–34, 2012.

No-Sang Kwak, Klaus-Robert Müller, and Seong-Whan Lee. A lower limb exoskeleton control system based on steady state visual evoked potentials. *Journal of neural engineering*, 12(5):056009, 2015.

David E Thompson, Kirsten L Gruis, and Jane E Huggins. A plug-and-play brain-computer interface to operate commercial assistive technology. *Disability and Rehabilitation: Assistive Technology*, 9(2): 144–150, 2014.

Jobin T Philip and S Thomas George. Visual p300 mind-speller brain-computer interfaces: a walk through the recent developments with special focus on classification algorithms. *Clinical EEG and neuroscience*, 51(1):19–33, 2020.

Yu Zhang, Guoxu Zhou, Jing Jin, Qibin Zhao, Xingyu Wang, and Andrzej Cichocki. Sparse bayesian classification of eeg for brain–computer interface. *IEEE transactions on neural networks and learning systems*, 27(11):2256–2267, 2015.

Giulio Tononi and Gerald M Edelman. Consciousness and complexity. *science*, 282(5395):1846–1851, 1998.

Karl J Friston, Christian Buechel, Gereon R Fink, Jond Morris, Edmund Rolls, and Raymond J Dolan. Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage*, 6(3):218–229, 1997.

Aya Kabbara, Mohamad Khalil, Wassim El-Falou, Hassan Eid, and Mahmoud Hassan. Functional brain connectivity as a new feature for p300 speller. *PLoS One*, 11(1):e0146282, 2016.

Ning Hao, Yang Feng, and Hao Helen Zhang. Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 113(522):615–625, 2018.

Cheng Wang, Binyan Jiang, and Liping Zhu. Penalized interaction estimation for ultrahigh dimensional quadratic regression. *Statistica Sinica*, 31(3):1549–1570, 2021.

Jim Griffin and Phil Brown. Hierarchical shrinkage priors for regression models. 2017.

Ran Shi and Jian Kang. Thresholded multiscale gaussian processes with application to bayesian feature selection for massive neuroimaging data. *arXiv preprint arXiv:1504.06074*, 2015.

Francisco Ruiz and Michalis Titsias. A contrastive divergence for combining variational inference and mcmc. In *International Conference on Machine Learning*, pages 5537–5545. PMLR, 2019.

Raza Habib and David Barber. Auxiliary variational mcmc. In *7th International Conference on Learning Representations, ICLR 2019*, pages 1–13. International Conference on Learning Representations, 2019.

Aziz Koçanaoğullari, Fernando Quivira, and Deniz Erdoğmuş. Incorporating temporal dependency on erp based bci. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 752–756. IEEE, 2018.

R.J. Adler and J.E. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer New York, 2009.

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.

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.