Distribution Agreement

Signature:

Yize Zhao   Date

Bayesian Feature Selection Methods for Complex Biomedical Data

By

Yize Zhao

Doctor of Philosophy

Biostatistics

Kang, Jian, Ph.D.
Advisor

Long, Qi, Ph.D.
Advisor

Chang, Howard H., Ph.D.
Committee Member

Hu, Xiaoping P., Ph.D.
Committee Member

Waller, Lance A., Ph.D.
Committee Member

Yu, Tianwei, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Bayesian Feature Selection Methods for Complex Biomedical Data

By

Yize Zhao
M.S., Emory University, 2013
B.S., Zhejiang University, 2010

Co-advisor: Jian Kang, Ph.D. and Qi Long, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2014

Abstract

Motivated by three different biomedical studies, this dissertation investigates novel Bayesian feature selection methods to analyze complex biomedical data.

In the first project, motivated by the colorectal cancer study, we propose a unified Bayesian approach for hierarchical feature selection of structured functional predictors in Generalized Functional Linear Models (GFLMs). Feature selection here is inherently hierarchical, involving selection of functional predictors and selection of regions within them. To achieve hierarchical feature selection, we construct a class of mixture priors for functional coefficients based on Gaussian processes. In addition, we use Ising priors on the model space to incorporate hierarchical structural information. Applying our approach to the motivating study, we find that one functional biomarker and its expression level in the transitional region between the proliferation and differentiation zones are associated with the risk for colorectal cancer.

In the second project, motivated by the Autism Brain Imaging Data Exchange (ABIDE) study, we are interested in identifying important biomarkers for early detection of the ASD under high resolution brain. We propose a novel multiresolution variable selection procedure under a Bayesian probit regression framework and it recursively uses posterior samples for variable selection at a lower resolution to guide variable selection at a higher resolution. The proposed algorithms are computationally feasible for ultra-high dimensional data. In addition, we also incorporate two levels of structural information into variable selection. Applied to the resting state functional magnetic resonance imaging (R-fMRI) data in the ABIDE study, our methods identify imaging biomarkers predictive of the ASD in several brain regions, which are biologically meaningful and interpretable.

Finally, with the goal to select gene and gene subnetworks with periodic behavior in a microarray dataset, we propose a nonparametric Bayesian model incorporating network information. In addition to identifying genes that have a strong association with a clinical outcome, our model can select genes with particular expressional behavior. We show that our proposed model is equivalent to an infinity mixture model for which we develop a posterior computation algorithm. We also propose two fast computing algorithms that approximate the posterior simulation with good gene selection accuracy but low computational cost.

Bayesian Feature Selection Methods for Complex Biomedical Data

By

Yize Zhao
M.S., Emory University, 2013
B.S., Zhejiang University, 2010

Advisor: Jian Kang, Ph.D. and Qi Long, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2014

Acknowledgement

I would like to express my sincere gratitude to my advisors Dr. Jian Kang and Dr. Qi Long for their inspiration and encouragement during the past three years. Dr. Long opened the door for the biostatistical research for me and has helped me find my potential. I will not go this far without his constant support. Dr. Jian Kang not only gave me the best advice with his expertise but also used his generosity and working ethic set me a role model. Without him, my own murky interests on Bayesian statistics would never blossom into a rewarding research agenda.

I would also like to thank my other committee members Dr. Howard Chang, Dr. Xiaoping Hu, Dr. Lance Waller and Dr. Tianwei Yu for their time and effort. Their invaluable suggestions on this dissertation significantly improves the quality of the work.

Lastly, I am deeply indebted to my family: my husband Song, my little girl Kairuo, my parents and parents in-law for their unconditional love and support. No matter how hard time I got through, they were always right behind me and made me fearless. I hope that over the course of my life I am able to repay the love and devotion they have shown me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Recent advances in biomedical technologies enable scientists to produce complex and big data in order to obtain useful information that provides valuable insights into biomedical research. This brings new challenges to develop efficient statistical methods for extracting important features from such data while integrating relevant scientific findings from prior biomedical research. In this dissertation, we propose some new Bayesian methods and develop computational tools to analyze large complex datasets motivated by several biomedical studies.

1.1.1 A Colorectal Adenoma Study

One motivating example is a recent study of protein biomarkers of risk for colorectal cancer (Ahearn et al., 2012). This study is aimed to investigate whether the expression patterns of several markers in normal-appearing colorectal mucosa are associated with the presence of colorectal adenomas, a surrogate for the risk for colorectal cancer. The panel of protein biomarkers under investigation (Table 1.1) represents highlights of features that are related to the known molecular basis in the earliest stages of colorectal carcinogenesis. In addition, colorectal crypts, U-shaped microscopic structures in human colon, are known to be a nice model for the regulation of cell proliferation, differentiation, and apoptosis in a continuously renewing colorectal epithelium. Thus, it is of particular interest to examine whether the profiles of the biomarkers along the length of colon crypts and their features are associated with cancer risk.

In this study, subjects with and without adenomas were considered at higher risk and lower risk for colorectal cancer, respectively. Biopsy samples of normal-appearing colorectal mucosa were collected from participants. Through automated immunohistochemistry (staining) with quantitative image analysis, each crypt was divided into

Biomarker Group	Biomarker Names
Inflammation	COX2
Colon Carcinogenesis Pathway	
APC pathway -	APC, β-catenin, c-myc, cyclin D1, E-cadherin
Mismatch repair pathway -	MSH2, MLH1, TGFβRII, bax
Cell Cycle:	
Proliferation -	mib1
Differentiation -	p21
Apoptosis inhibition & promotion -	bcl2, bak, bax
Crypt Stem Cell Longevity -	Telomerase
Autocrine/Paracrine Growth Factors & Receptors:	TGFα, TGFβ_1, TGFβRII

Table 1.1: Biomarkers of risk in the colorectal adenoma study

two symmetric hemi-crypts, and the entire length of each hemi-crypt was standardized into a fixed number of segments, numbered in an ascending order from the base to the top of a crypt; subsequently, the staining optical density, representing the biomarker expression level, was recorded for each segment and was plotted against the segment location to construct the expression profile along the length of crypts; cf. Figure 1 in Ahearn et al. (2012). The expression profile along the length of crypts forms a natural one-dimensional curve, and is an example of functional data measured over space. We refer to these biomarker profiles/curves as functional biomarkers or functional predictors. The functional biomarkers can be divided into different groups based on biological functions and pathways (Table 1.1), noting that biomarkers (e.g., bax) may belong to multiple groups. In addition, the functional biomarkers were measured at discrete design points subject to measurement error. For each subject, multiple biomarkers were measured; for each biomarker, the expression curve was measured in multiple hemi-crypts from each subject's biopsy tissues.

The goal of the this study is to identify important functional biomarkers and their features that are associated with the risk for colorectal cancer, while also incorporating the structural information. Clearly, feature selection in this case is inherently hierarchical with two levels: first, select important functional biomarkers (i.e., between functional predictors); and second, for each selected biomarker, identify the regions of its profile that are associated with cancer risk (i.e., within functional predictors). Similarly, there are also two levels of structural information that can be incorporated into the feature selection process: the biological structure between biomarkers and the spatial structure within each biomarker curve.

1.1.2 Autism Brain Imaging Data Exchange (ABIDE)

Another interesting example that motivates our methodological development is the Autism Brain Imaging Data Exchange (ABIDE) study (Di Martino et al., 2013). The major goal of the ABIDE study is to explore association of brain activity with the autism spectrum disorder (ASD), a widely recognized disease due to its high prevalence and substantial heterogeneity in children (Rice, 2009). The ABIDE study aggregated 20 resting-state functional magnetic resonance imaging (R-fMRI) data sets from 17 different sites including 539 ASDs and 573 age-matched typical controls. The R-fMRI is a popular non-invasive imaging technique that measures the blood oxygen level to reflect the resting brain activity. For each subject, the R-fMRI signal was recorded for each voxel in the brain over multiple time points (multiple scans). Several standard imaging preprocessing steps (Di Martino et al., 2013) including motion corrections, slice-timing correction, and spatial smoothing have been applied to the R-fMRI data, which were registered into the standard Montreal Neurological Institute (MNI) space consisting of 228,483 voxels. To characterize the localized spontaneous brain activity, we focus on the fractional amplitude of low-frequency fluctuations (fALFF) (Zou et al., 2008) based on the R-fMRI time series at each voxel for each subject. The fALFF is defined as the ratio of the power spectrum of low frequency (0.01-0.08Hz) to the entire frequency range and has been widely used as a voxel-wise measure of the intrinsic functional brain architecture derived from the R-fMRI data (Zuo et al., 2010). In this work, we analyze the voxel-wise fALLF values over 116 regions in the brain involving 185,405 voxels in total, where regions are defined according to the Automated Anatomical Labeling (AAL) system (Hervé et al., 2012). Besides the imaging data and the clinical diagnosis of the ASD, demographical variables were also collected, such as age at scan, sex and intelligence quotient (IQ).

One question of interest in this study is to identify imaging biomarkers, i.e., voxel-wise fALFF values over 116 regions, for detecting the ASD risk. In particular, our goal is to perform two levels of variable selection: at the first level, important regions are selected in relation to the ASD risk; at the second level, a set of important voxels within the selected regions are selected and are referred to as ASD imaging risk factors. Correspondingly, two levels of structural information – functional connectivity among regions and spatial dependence among voxels – can be incorporated to facilitate variable selection and produce biologically more interpretable results. To achieve this goal, we use a Bayesian probit regression model for spatial variable selection, where the binary outcome is the ASD disease status and the predictors include all voxel-level imaging biomarkers from multiple regions. We use Ising prior models to incorporate structural information for the two levels of variable selection. However, it is extremely challenging to perform spatial variable selection in such ultra-high dimensional structured feature space (185,405 voxels within 116 regions) under our modeling framework.

1.1.3 Spellman Yeast Cell Cycle Microarray Data

The third motivating example is the Spellman Yeast Cell Cycle Microarray Data (Spellman et al., 1998). The dataset is intended to detect genes with periodic behavior along the procession of the cell cycle. It has been extensively used in the development of computational methods. The gene network information depicting the biological relationships is summarized from the Database of Interacting Proteins (DIP) (Xenarios et al., 2002). We use the high-confidence connections between yeast proteins from the DIP. Eventually, the network contains 2031 genes, where the mean, median, maximum and minimum edges per gene are 3.948, 2, 57 and 1 respectively.

Different from the previous two motivating examples, there is no outcome variable in the cell-cycle dataset, and we focus on the selection of genes and gene sub-networks with periodic behavior in light of the network. It is known that such genes show different phase shifts along the cell cycle and may not be correlated with each other (Yu, 2010). We perform the Fishers exact G test for periodicity (Wichert et al., 2004) for each gene. In this case, a linear regression or parametric model may not be suitable, and we propose a Bayesian nonparametric mixture model for large scale statistics incorporating network information.

1.2 Literature Review

In this selection, we provide an overview of existing statistical methods on variable selection and discuss their advantages and limitations.

1.2.1 Variable Selection in High-Dimensional Feature Space

We first consider the observed data $(y_i, \boldsymbol{x_i}^T)_{i=1}^n$ with y_i and $\boldsymbol{x_i}$ denoting a clinical outcome and a p-dimensional predictor, respectively. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, then without specifying the distribution of y_i, we construct the following regression model:

$$g\{\mathrm{E}(y_i \mid \boldsymbol{x_i})\} = \boldsymbol{x_i}^T\boldsymbol{\beta}, \quad i = 1, \ldots, n, \tag{1.1}$$

with $g\{\cdot\}$ being a link function and $\mathrm{E}(y \mid \boldsymbol{x})$ denoting the conditional expectation of y given $\boldsymbol{x}$.

Over the past decades, a unified approach to model selection for (1.1) has been suggested–optimization penalized likelihood

$$L(\boldsymbol{\beta}) + P_{\lambda}(\boldsymbol{\beta}), \tag{1.2}$$

with the loss function $L(\cdot)$ and penalized function $P_{\lambda}(\cdot)$. When y_i is from Gaussian distribution with $g\{\cdot\}$ chosen as the identical link, minimization (1.2) based on (1.1) has been well investigated from a frequentist perspective. Starting from the least absolute shrinkage and selection operator (LASSO) using L_1-penalty proposed by Tibshirani (1996), a large number of methods equipped with various penalized functions have been proposed to extend theoretical properties, improve practical performance or accommodate to the emergence of new data structures. These approaches include the nonconcave penalized likelihood variable selection using the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), the least angle regression (LARS) (Efron et al., 2004), the elastic net (Zou and Hastie, 2005), the adaptive LASSO (Zou, 2006), the group LASSO (Yuan and Lin, 2006) and many other extensions (Tibshirani et al., 2005; Li and Li, 2008; Pan et al., 2010; Friedman et al., 2010; Wang et al., 2009; Wu and Wang, 2013). Most of the above LASSO-type optimization problems can be computed through the LARS algorithm by operating the entire solution path.

Compared with the frequentist approach, a huge advantage of Bayesian methods is their ability to quantify uncertainty. Polson and Scott (2010) demonstrated that the estimate based on a LASSO-type optimization method was equivalent to the posterior mode under a Bayesian framework with a global-local (GL) prior

$$\beta_j \;\sim\; \mathrm{N}(0, \psi_j \tau), \quad \psi \sim f, \quad \tau \sim g, \tag{1.3}$$

with properly specified f and g. In (1.3), τ controls the global sparsity (spike) and ψ_j allows deviations (slide). The Bayesian LASSO (Park and Casella, 2008; Hans, 2009) is a canonical application of this type of method by setting an exponential distribution for f to obtain a double-exponential prior.

The GL prior owns computational advantages compared with the other Bayesian variable methods, however, it results in many of the β_js very small but not exact to zero, which limits its application to some extent. The most widely used Bayesian variable selection method for (1.1) in the literature is to set a conditional two components mixture prior for the coefficients with one component concentrated at zero (spike) and the other diffuse (slide). Specifically, we can introduce a latent selection indicator $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ with $\gamma_j \in \{0, 1\}$ indicating the selection status of β_j. Then one can assign a prior for β_j:

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)\mathrm{N}(0, \sigma_{j0}^2) + \gamma_j \mathrm{N}(0, \sigma_{j1}^2), \quad j = 1, \ldots, p, \tag{1.4}$$

where σ_{j0}^2 is a small value for spike and σ_{j1}^2 is a large value for slide. Here σ_{j0}^2 can be viewed as a threshold for selection, and without prior knowledge, one can also simplify (1.4) as

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)\delta(0) + \gamma_j \mathrm{N}(0, \sigma_j^2), \quad j = 1, \ldots, p, \tag{1.5}$$

where $\delta(0)$ represents a point mass at zero. Prior (1.5) indicates the preference to only exclude β_j that exactly equals zero, and any coefficients, no matter how small, can be selected into the model with enough data as long as they are distinct from zero. In practice, it is intractable to calculate the posterior probability of γ_j based on (1.5) and (1.4) exactly. Alternatively, rather than search for the entire model space, George and McCulloch (1993) proposed a stochastic search method named Stochastic

Search Variable Selection (SSVS) to efficiently locate the optimal model in the posterior inference. This technique is further discussed by George and McCulloch (1997); Brown et al. (1998); Chipman et al. (2001). Such two-component variable selection procedure has been extensively adopted by a wide range of applications (Yi et al., 2003; Ishwaran and Rao, 2003; Theo and Mike, 2004; Sha et al., 2006; Li and Zhang, 2010; Stingo et al., 2011; Goldsmith et al., 2012; Huang et al., 2013). Different from the classical slide and spike method, more recently, Johnson et al. (2012) proposed a new model selection procedure by imposing nonlocal prior densities (Johnson and Rossell, 2010) for the coefficients. Such nonlocal prior owns the advantage to have a zero density function in the case of a zero model parameter, and has been shown to have a promising performance in the high-dimensional and ultra high-dimensional cases (Johnson, 2013). However, due the complex prior formulation and potentially intensive posterior computation, both the first and second topics of our work adopt the point mass mixture approach as in (1.5).

Although regression models are widely used for the selection of informative features associated with an outcome variable, in some situations, we are interested in identifying certain type of features, like the third motivating example, to study the periodicity behavior of genes without an outcome variable. To conduct such selection/detection, we assume the following mixture distribution

$$p_0 f_0(x) + p_1 f_1(x) \tag{1.6}$$

of the observed data which forms a classification problem. Bernardo et al. (2003) proposed a two-step procedure to address the problem by reducing the dimension of the data via Principal Component Analysis (PCA) following with a mixture model. By incorporating biological information, Wei and Pan (2012) used a two-component Gaussian mixture model with a Markov random field prior to jointly study multiple gene networks. Tadesse et al. (2005); Hoff et al. (2006) adopted finite mixture of Gaussian distributions with model parameters updated through Markov chain Monte Carlo (MCMC). To further release the model assumption, Kim et al. (2006) conducted clustering via Dirichlet process mixture models which allows the flexility on cluster number and could better fit the data. In the third topic, with the goal to select genes with periodic behaviors, we formulate the problem into a clustering procedure and adopt a Bayesian nonparametric mixture model with network information also incorporated.

1.2.2 Incorporating Biological Information

In biomedical studies, the biological information/knowledge that we learn from previous research findings has always been an important factor that leads the feature selection procedure to a more biologically interpretable direction. For example, in a gene selection problem, incorporating the gene network/pathway information can greatly improve the selection accuracy (Li and Li, 2008; Pan et al., 2010; Stingo et al., 2011) and obtain scientifically meaningful results. In a study on the identification of imaging biomarkers, the relevant brain functional networks are very useful to facilitate the problem solving and provide valuable insights behind the results (Goldsmith et al., 2012; Huang et al., 2013).

From a frequentist perspective, incorporating the biological information into feature selection has been implemented in multiple scenarios by constructing different penalized functions accordingly (Tibshirani et al., 2005; Li and Li, 2008; Pan et al., 2010; Yuan and Lin, 2006; Friedman et al., 2010; Wang et al., 2009; Wu and Wang, 2013). Those approaches have been widely adopted for feature selection or prediction in various applications, particularly in genomic studies, to import gene network/pathway information.

In a Bayesian framework, it is natural to incorporate such biological information through the prior setting for the latent selection indicators or the coefficients with the former approach more often used. Specifically, under model (1.1), with an undirected graph $\mathcal{G}$ representing the biological connectivity information among the p predictors, we can specify an Ising / Binary Markov random field (MRF) prior for $\boldsymbol{\gamma}$ as

$$\boldsymbol{\gamma} \mid \eta,\ \xi \ \propto \ \exp\left(\eta \sum_{j=1}^{p} \gamma_j + \xi \sum_{j=1}^{p} \sum_{j \sim k} I[\gamma_j = \gamma_k]\right). \tag{1.7}$$

Here, "$j \sim k$" represents predictors j and k are biologically connected in $\mathcal{G}$. We refer to η as the sparse parameter which controls the overall sparsity among $\boldsymbol{\gamma}$ and ξ as the smooth parameter which regulates the impact from the biological information. In the presence of high-dimensional data/big data, the posterior inference for Ising parameters η and ξ can lead to intractable computation due to the calculation of the normalized constant. Even through the posterior inference can be proceed under path sampling method (Gelman and Meng, 1998) or certain strategies for smooth parameter by Smith and Fahrmeir (2007), most of the previous work (Li and Zhang, 2010; Stingo et al., 2011; Goldsmith et al., 2012; Huang et al., 2013) still fixed these parameters in the posterior inference based on either a biological priori or the cross-validation approach. In the absence of prior knowledge, particularly, one can also assign a zero external field–remove the sparse parameter η in (1.7) as the previous work Smith and Fahrmeir (2007); Barbu and Zhu (2007); Johnson et al. (2012), which can also ease the computation. When $\xi = 0$, prior (1.7) reduces to independent and identically distributed Bernoulli distributions with no biological information incorporated. In the first topic, we hierarchically incorporate both biological connectivity among the functional predictors and the spatial information within each curve. In the second topic, we consider the functional connectivity among ROIs and the spatial information among the voxels. And in the third topic, the gene network information is directly incorporated. In this dissertation, we adopt different strategies to estimate Ising parameters.

1.3 Outline

The remainder of the dissertation proposal is organized as follows. In Chapter 2, we propose a unified Bayesian framework for hierarchical feature selection of structured functional predictors measure with error. In Chapter 3, we develop a Bayesian feature selection method for ultra high-dimensional data based on a multiresolution approach. In Chapter 4, we investigate gene selection via Bayesian nonparametric method. We finally conclude this dissertation with discussion and future plans in Chapter 5.

Chapter 2

Bayesian Hierarchical Feature Selection of Structured Functional Predictors Measured with Error

2.1 Introduction

Functional data, measured temporally or spatially, have been regularly collected in many biomedical and epidemiological studies. In these studies, it is often of interest to investigate the association between a scalar outcome and functional predictors while also conducting feature selection between and within functional predictors.

2.1.1 Functional Data Analysis

Ramsay and Silverman (2005) provides a nice, thorough review of different types of models for functional data. In particular, when assessing the relationship between multiple functional predictors (say, $\theta_j(\cdot)$, $j = 1, \ldots, m$) and a scalar response (say, y), the generalized functional linear models (GFLMs), an extension of the Generalized Linear Models (GLMs) in the presence of functional predictors, are a natural choice,

$$g\left[E\left\{y|\theta_j(\cdot), j = 1, \ldots, m\right\}\right] = \alpha + \sum_{j=1}^{m} \int \theta_j(t)\beta_j(t)dt$$

where $\beta_j(\cdot)$'s are the functional coefficients and $g(\cdot)$ is a monotonic and smooth link function. GFLMs have been intensively investigated in recent years (James, 2002; Müller and Stadtmüller, 2005; Malloy et al., 2010; Yao et al., 2005; Reiss and Ogden, 2007; Krämer et al., 2008; Reiss and Ogden, 2009). In many studies such as our motivating study, functional predictors are not fully observed, in which case $\int \theta_j(t)\beta_j(t)dt$ is intractable and needs to be approximated in order to fit GFLMs. To this end, one approach is to approximate $\theta_j(\cdot)$ using a truncated series expansion based on a set of basis functions; for example, James (2002) used cubic spline bases and Yao et al. (2005) used a set of parsimonious bases obtained from functional principal component analysis (FPCA) (Yao et al., 2005). Alternatively, one can fit discretized

GFLMs (Reiss and Ogden, 2007, 2009; Malloy et al., 2010). In the presence of measurement error in functional predictors, a popular, desirable approach is to jointly model the scalar outcome and the error-contaminated functional predictors (James, 2002; Crainiceanu et al., 2009); a less desirable approach is to estimate functional predictors first and plug their estimates into GFLMs, which ignores the uncertainty of estimating functional predictors and underestimates the true sampling variance.

Functional data analysis has been extended to multi-level functional data such as the biomarker curve data in the motivating study. For example, Morris and Carroll (2006) adopted a wavelet-based Bayesian approach for modeling multi-level functional data and applied their methods to a colon carcinogenesis study, and Baladandayuthapani et al. (2007) considered a similar setting and proposed a hierarchical Bayesian model to account for spatial correlation among functions measured from the same study unit; however, these authors focused on models with functional biomarkers as outcomes, which is fundamentally different from our setting. Crainiceanu et al. (2009) have extended GFLMs to handle multi-level functional data that were measured temporally; their approach used multi-level functional principal component analysis (MFPCA) (Di et al., 2009) to derive basis functions and also accounted for measurement error in functional predictors.

2.1.2 Feature Selection in GFLMs

There has been limited work on feature selection in GFLMs and most existing methods focus on the case of no measurement error in functional predictors. For feature selection between functional predictors, Zhu et al. (2009) proposed a Bayesian hierarchical model for classification which also accounted for batch effect; their approach for variable selection was to specify a hierarchical mixture prior on functional coefficients $\beta_j(\cdot)$ along the lines of stochastic search variable selection (SSVS) proposed by

George and McCulloch (1997). A more recent work by Lian (2011) adopted a regularization approach to select functional predictors in functional linear models using group smoothly clipped absolute deviation penalty (SCAD), where the coefficients of basis functions for the same functional predictor were grouped together.

To conduct feature selection within functional predictors, i.e., identify regions where $\beta(\cdot) = 0$. James et al. (2009) proposed to enforce sparsity in the derivatives of $\beta(\cdot)$ and Tian and James (2012) proposed a different approach by expanding $\beta(\cdot)$ based on a set of piecewise constants or linear basis functions that are interpretable. Alternatively, regularization methods such as lasso and SCAD have been used in functional linear models to realize selection of basis functions for modeling $\beta(\cdot)$ (Zhao et al., 2012; Lee and Park, 2012), which does not directly lead to feature selection within functional predictors; however, if B-spline basis functions are used, such an approach can induce sparsity in the estimate of $\beta(\cdot)$ and hence feature selection within functional predictors (Zhou et al., 2012).

To the best of our knowledge, no methods have been proposed to conduct hierarchical feature selection between and within functional predictors, not to mention incorporating structural information. In particular, different from Zhu et al. (2009), we investigate feature selection between and within functional predictors that is guided by structural information through a class of Ising priors (Li and Zhang, 2010). While there has been considerable interest in incorporating structural or biological information in feature selection in recent years (Li and Li, 2008; Pan et al., 2009; Li and Zhang, 2010; Stingo et al., 2011), this approach has not been adopted for functional data.

The novel contributions of our work are several-fold. First, it makes the first attempt to propose a unified framework for hierarchical feature selection between and within functional predictors in GFLMs. Second, with a consideration for measurement error, we incorporate a hierarchical Bayesian model for multi-level functional data into the joint-modeling framework. Third, our approach incorporates two levels of structural information into feature selection, leading to more biologically meaningful and interpretable results. Lastly, we assign Gaussian process priors to multi-level functional predictors in the model and adopt a discrete approximation, which is more amenable to our hierarchical feature selection procedure. More importantly, this approach circumvents the issue of varying variable dimensions and simplifies the posterior computation as an alternative to more complicated trans-dimensional MCMC algorithms. In addition, different covariance functions such as the exponential kernel and theM atérn kernel can be specified for Gaussian process priors in light of different degrees of smoothness.

2.2 Model Formulation

2.2.1 Basic Structure

To fix ideas, suppose that the observed data have the same multi-level structure as the motivating data described in Section 1.1.1. For subject i ($i = 1, \ldots, n$), let y_i denote the binary outcome, $\mathbf{s}_i$ denote a set of p scalar predictors including an intercept term, and $\{X_{ijk}(t_l) : j = 1, \ldots, m; k = 1, \ldots, q_{ij}; l = 1, \ldots, L\}$ denote the observed, error-contaminated functional data at design point l (e.g., crypt location) of replicate curve k (e.g., crypt k) for functional predictor j. Without loss of generality, we focus on the case of a balanced design, i.e., the set of design points $\{t_1, t_2, \ldots, t_L\}$ is the same for all functional predictors and for all subjects. It is straightforward to extend our method to data with unbalanced design or data with a different multi-level structure.

For the multi-level functional data, we denote by $\theta_{ijk}(\cdot)$ the true replicate curve k

for predictor j from subject i, by $\theta_{ij}(\cdot)$ the true curve for predictor j from subject i, and by $\theta_j(\cdot)$ the true mean curve for predictor j across all subjects. Let $\mathcal{T}$ denote the compact domain of all functional predictors; without loss of generality, we take $\mathcal{T} = [0, 1]$, $t_1 = 0$ and $t_L = 1$. We model y_i through a GFLM with the probit link by introducing a latent variable z_i as follows,

$$\begin{aligned} y_i &= I[z_i > 0], \\ z_i &= \mathbf{s}_i^T \boldsymbol{\alpha} + \sum_{j=1}^{m} \int_{\mathcal{T}} \beta_j(t)\theta_{ij}(t)dt + \varepsilon_i, \qquad (2.1) \\ X_{ijk}(t_l) &= \theta_{ijk}(t_l) + \epsilon_{ijkl}, \end{aligned}$$

where $I(\mathcal{A}) = 1$ if event $\mathcal{A}$ is true and 0 if otherwise, $\varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0,\ 1)$ with $\mathrm{N}(\mu, \sigma^2)$ denoting a normal distribution with mean μ and standard deviation σ, and $\boldsymbol{\epsilon}_{ijk} = (\epsilon_{ijk1}, \epsilon_{ijk2}, \ldots, \epsilon_{ijkL})^T \sim \mathrm{N}(\mathbf{0}, \Omega)$ with $\Omega = \sigma^2 I_L$ and I_L denoting an identity matrix of $L \times L$. The parameters of interest include the scalar coefficients $\boldsymbol{\alpha}$ and the functional coefficients $\{\beta_1(\cdot), \ldots, \beta_m(\cdot)\}$. Write $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$, $\boldsymbol{X}_{ijk} = (X_{ijk}(t_1), \ldots, X_{ijk}(t_L))^T$, $\boldsymbol{X} = \{\boldsymbol{X}_{ijk}, i = 1, \ldots, n; j = 1, \ldots, m; k = 1, \ldots, q_{ij}\}$.

We adopt a Bayesian hierarchical model with Gaussian process priors for the functional predictors at each level of the hierarchy. Specifically, the proposed priors are:

$$\begin{aligned} \boldsymbol{\alpha} &\sim \mathrm{N}(\mathbf{0}, \sigma_0^2 I_p), && (2.2) \\ \theta_{ijk}(\cdot) &\sim \mathcal{GP}(\theta_{ij}(\cdot), \mathcal{K}_1(\cdot,\cdot)), && (2.3) \\ \theta_{ij}(\cdot) &\sim \mathcal{GP}(\theta_j(\cdot), \mathcal{K}_2(\cdot,\cdot)), && (2.4) \\ \theta_j(\cdot) &\sim \mathcal{GP}(\mathbf{0}, \mathcal{K}_3(\cdot,\cdot)), && (2.5) \end{aligned}$$

where $\mathcal{GP}(\mu(\cdot), \mathcal{K}(\cdot,\cdot))$ denotes a Gaussian process with mean function $\mu(\cdot)$ and covariance function $\mathcal{K}(\cdot,\cdot)$.

2.2.2 Priors for Functional Coefficients: Feature Selection

We investigate three prior models for $\beta_j(\cdot)$ $(j = 1, \ldots, m)$ and their utility in feature selection. We start with a model assuming that $\beta_j(\cdot)$ is a Gaussian process, i.e.,

$$\beta_j(\cdot) \ \sim \ \mathcal{GP}(\mathbf{0}, \mathcal{K}_4(\cdot, \cdot)).$$

Under this model, $\beta_j(\cdot)$ is almost surely non-zero in any specific region of $\mathcal{T}$; thus, this model does not directly allow for feature selection at any level.

To enable feature selection between functional predictors, we assume

$$\beta_j(\cdot) = C_j \tilde{\beta}_j(\cdot), \qquad \tilde{\beta}_j(\cdot) \sim \mathcal{GP}(\mathbf{0},\ \mathcal{K}_4(\cdot, \cdot)), \tag{2.6}$$

where $C_j \in \{0, 1\}$ is a latent selection indicator. If $C_j = 1$, $\beta_j(\cdot) = \tilde{\beta}_j(\cdot)$ is a Gaussian process, indicating that functional predictor j is selected. If $C_j = 0$, $\beta_j(\cdot) \equiv 0$, indicating that predictor j is not related to the outcome. It follows that model (2.6) is equivalent to a mixture of Gaussian process and a point mass concentrated at a function that equals zero everywhere, denoted by $\delta(0)$,

$$\beta_j(\cdot) \mid C_j \ \sim \ (1 - C_j)\delta(0) + C_j \mathcal{GP}(\mathbf{0},\ \mathcal{K}_4(\cdot, \cdot)). \tag{2.7}$$

We refer to the model based on (2.6) as the Selection Between-Predictor Model (SBPM).

To conduct hierarchical feature selection between and within functional predictors, we need to modify prior (2.6) further. To facilitate feature selection within functional predictors, we introduce a set of grid points and, to simplify exposition, we consider the case where the set of grid points for feature selection is the same as the set of design points in the observed data $\{t_1 = 0, t_2, \ldots, t_L = 1\}$. However, our method can be readily extended to conduct feature selection in a set of arbitrarily chosen subintervals by introducing additional grid points and we provide a brief discussion on this extension in Section 2.6. Given the set of grid points $\boldsymbol{t} = \{t_l, l = 1, \ldots, L\}$, the compact domain $\mathcal{T} = [0, 1]$ is divided into $L-1$ subintervals, $\{[t_l, t_{l+1}) : l = 1, 2, \ldots, L-2; [t_{L-1}, t_L]\}$. We then introduce a set of lower level latent indicator variables for each functional coefficient $\beta_j(\cdot)$, namely, $\boldsymbol{\gamma}_j = \{\gamma_{jl} \in \{0, 1\} : l = 1, \ldots, L-1\}$, where $\gamma_{jl} = 0$ indicates that $\beta_j(\cdot) \equiv 0$ in the subinterval $[t_l, t_{l+1})$. To ensure selection consistency at both levels, we impose a constraint on $\boldsymbol{\gamma}_j$ as

$$\max\{\boldsymbol{\gamma}_j\} \;=\; C_j, \quad j = 1, \ldots, m, \tag{2.8}$$

which essentially avoids the situation that a functional predictor is selected but none of its regions is selected.

Given $\boldsymbol{\gamma}_j$, the functional coefficient j is divided into R_j regions and each region includes a set of contiguous subintervals that have the same γ value; we denote by K_j the number of such regions with $\gamma = 1$, i.e., the number of selected regions where $\beta_j(\cdot)$ is allowed to be nonzero. It can be shown that $K_j = \lfloor (R_j + \gamma_{j1})/2 \rfloor$, where $\lfloor x \rfloor$ is the largest integer not greater than x, and

$$R_j = \sum_{l=1}^{L-2} I(\mid \gamma_{j(l+1)} - \gamma_{jl} \mid = 1) + 1, \qquad K_j = \sum_{l=1}^{L-2} I(\gamma_{j(l+1)} - \gamma_{jl} = 1) + r_{j1}.$$

For example, as illustrated in Figure 2.1, given $L = 8$, $\boldsymbol{\gamma}_j = \{0, 0, 1, 1, 0, 1, 1\}$ indicates that $\beta_j(\cdot) = 0$ in $[t_1, t_3) \cup [t_5, t_6)$ and $\beta_j(\cdot) \neq 0$ in $[t_3, t_5) \cup [t_6, t_8]$ with $R_j = 4$ and $K_j = 2$. We define an index set $\tilde{\boldsymbol{b}}_j = \{b_{jr}\}_{r=1}^{2K_j}$ based on $\boldsymbol{\gamma}_j$ as follows,

$$\begin{aligned} b_{j1} &= \min\{l : \gamma_{jl} = 1\}, \\ b_{jr} &= \min_{l > b_{j(r-1)}} \{l : \gamma_{jl} = 1 - \gamma_{jb_{j(r-1)}}\}, \quad r = 2, \ldots, 2K_j - 1, \end{aligned}$$

and $b_{j(2K_j)} = \min_{l > b_{j(2K_j-1)}}\{l : \gamma_{jl} = 1 - \gamma_{jb_{j(2K_j-1)}}\}$ if $R_j - \gamma_{j1}$ is an odd number and $b_{j(2K_j)} = L-1$ if $R_j - \gamma_{j1}$ is an even number. The set, $\tilde{\boldsymbol{b}}_j$, essentially includes the indices of grid points that are at the two ends of each selected region with $\gamma = 1$ for functional predictor j and there is a one-to-one mapping between $\boldsymbol{\gamma}_j$ and $\tilde{\boldsymbol{b}}_j$ $(j = 1, \ldots, m)$. In the aforementioned example (Figure 2.1), with $\boldsymbol{\gamma}_j = \{0, 0, 1, 1, 0, 1, 1\}$, the corresponding index set is $\tilde{\boldsymbol{b}}_j = \{3, 5, 6, 8\}$. Given $\tilde{\boldsymbol{b}}_j$, we can write the set of selected (active) regions for predictor j as $\mathcal{R}_j = \bigcup_{k=1}^{K_j}[t_{b_{j(2k-1)}}, t_{b_{j(2k)}}) = \bigcup_{k=1}^{K_j}\mathcal{R}_{jk}$.

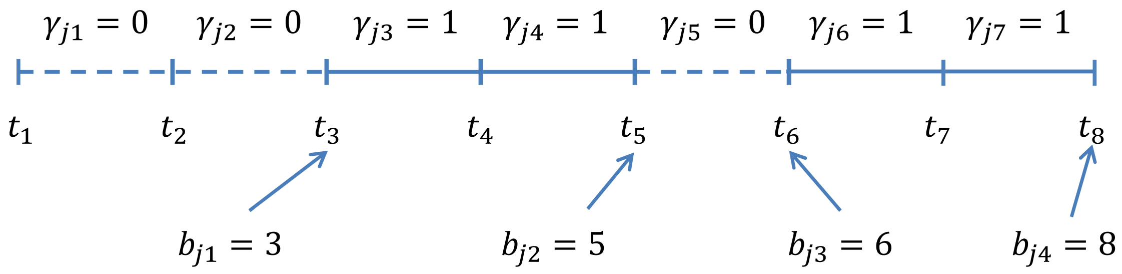


Figure 2.1: Relationship between $\boldsymbol{t}$, $\boldsymbol{\gamma}_j$ and $\tilde{\boldsymbol{b}}_j$. The dashed lines represent the non-selected regions and solid lines represent the selected regions.

To enable selection within functional predictors, we modify prior (2.6) given C_j and $\boldsymbol{\gamma}_j$ as follows,

$$\beta_j(\cdot) = C_j \sum_{k=1}^{K_j} I_{\mathcal{R}_{jk}}(\cdot)\tilde{\beta}_j(\cdot), \qquad \tilde{\beta}_j(\cdot) \sim \mathcal{GP}(\mathbf{0},\ \mathcal{K}_4(\cdot,\cdot)), \tag{2.9}$$

where the indicator function $I_{\mathcal{R}_{jk}}(t) = 1$ if $t \in \mathcal{R}_{jk}$ and 0 if otherwise. Given $C_j = 1$, model (2.9) implies that $\beta_j(\cdot)$ is allowed to be nonzero in $\mathcal{R}_j$ and $\beta_j(\cdot) \equiv 0$ in $\overline{\mathcal{R}_j}$ where $\overline{\mathcal{R}_j}$ denotes the set complement of $\mathcal{R}_j$ in $\mathcal{T}$. Similar to (2.7), integrating out $\tilde{\beta}_j(\cdot)$, prior (2.9) is equivalent to a mixture prior distribution with its mixture components indexed by different values of $\mathcal{R}_j$ (or equivalently, $\boldsymbol{\gamma}_j$); for a given $\mathcal{R}_j$, $\sum_{k=1}^{K_j} I_{\mathcal{R}_{jk}}(\cdot)\tilde{\beta}_j(\cdot)$ defines the form of the distribution for the corresponding mixture component, which is constructed from a Gaussian process. For example, for $\mathcal{R}_j = [0, 1]$, the corresponding mixture component is $\mathcal{GP}(\mathbf{0},\ \mathcal{K}_4(\cdot,\cdot))$; for $\mathcal{R}_j = \emptyset$, the corresponding mixture component is $\delta(0)$. It follows that the number of components is the number of values that $\mathcal{R}_j$ can take. We refer to the model based on (2.9) as the Hierarchical Feature Selection Model (HFSM). Of note, while this prior assumes that $\beta_j(\cdot)$ is continuous within each selected region $\mathcal{R}_{jk}$, it does not impose continuity conditions at the boundary points $\{t_{b_{jr}} : r = 1, \ldots, 2K_j\}$. The SBPM is a special case of the HFSM, since prior (2.9) reduces to (2.6) when $\gamma_{jl} \equiv 1$ for all $l = 1, \ldots, L-1$ or equivalently $\mathcal{R}_j$ can only take the value of $[0, 1]$.

2.2.3 Hyperpriors: Incorporating Structural Information

Structural information is incorporated through hyperpriors on $\boldsymbol{C} = (C_1, \ldots, C_m)^T$ and $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_j, j = 1, \ldots, m\}$. Suppose that the structure of the functional predictors is represented by an undirected graph $\mathcal{G}$ where an edge indicates that two predictors are connected biologically either in the same pathway or with the same biological function as shown in Table 1.1. Based on $\mathcal{G}$, we define an $m \times m$ connection matrix $\boldsymbol{R} = (r_{st})$ where $r_{st} = 1$ if functional predictors s and t are connected in $\mathcal{G}$ and 0 if otherwise. We introduce an Ising prior for $\boldsymbol{C}$, the indicator for feature selection between functional predictors, as follows,

$$\boldsymbol{C} \mid \eta \ \propto \ \exp\left(\eta \sum_{j=1}^{m} \sum_{k: r_{kj}=1} I[C_j = C_k]\right), \tag{2.10}$$

which naturally incorporates the structural information in $\mathcal{G}$. In the absence of prior information, we here assign a zero external field in (2.10) (Smith and Fahrmeir, 2007; Johnson et al., 2012; Barbu and Zhu, 2007), resulting in the marginal probability of $P(C_j \mid \eta) = 0.5$ for each $j = 1, \ldots, m$. The parameter $\eta > 0$ controls the degree of smoothness of $\boldsymbol{C}$ over $\mathcal{G}$. When $\eta = 0$, the prior (2.10) reduces to independently identically distributed Bernoulli priors, which does not incorporate any structural information between functional predictors. We can assign a prior for η, $\eta \sim \text{Unif}(0, U_\eta)$ with a pre-specified U_η, e.g. $U_\eta = 10$, where $\text{Unif}(a, b)$ denotes a uniform distribution in (a, b). When m is not too large, we can calculate the normalizing constant in the Ising prior, namely, $h(\eta) = \sum_{\boldsymbol{C}\in\Xi} \exp(\eta \sum_{j=1}^{m} \sum_{k:r_{kj}=1} I[C_j = C_k])$ with Ξ being the domain of $\boldsymbol{C}$, which allows for posterior inference for η. This approach is adopted in our numerical studies.

To incorporate the spatial information within each functional predictor, we assume a conditional Ising prior for $\boldsymbol{\gamma}$, the indicator for feature selection within each functional predictor, given $\boldsymbol{C}$, as follows,

$$\boldsymbol{\gamma} \mid \boldsymbol{C} \;\; \propto \;\; \exp\left(\xi \sum_{j=1}^{m} \sum_{l=1}^{L-2} I[\gamma_{jl} = \gamma_{j(l+1)}]\right) \prod_{j=1}^{m} I(\max_{l} \gamma_{jl} = C_j), \qquad (2.11)$$

where ξ controls the degree of smoothness of $\boldsymbol{\gamma}$ over the spatial structure of each functional predictor. Essentially, this prior treats two adjacent subintervals as connected. Similarly, when $\xi = 0$, prior (2.11) reduces to independently identically distributed Bernoulli priors with no spatial information incorporated. Since $\boldsymbol{\gamma}$ is often of high dimension even for moderate m, we choose to pre-specify ξ to avoid intractable computation. We propose to run MCMC for different ξ values and select an optimal value of ξ based on a chosen criterion. In our numerical studies, we use the posterior Bayes factor (Aitkin, 1991) for its ease of computation.

The hyperpriors for the remaining parameters introduced in Sections 2.2.1 and 2.2.2 are as follows. The covariance functions for Gaussian processes (2.3) – (2.5) and (2.9) are assumed to be $\mathcal{K}_p(s, t) = \tau_p^2 \exp\{-\rho(s - t)^2\}, \forall s, t \in \mathcal{T},\ p = 1, 2, 3, 4$. Other covariance functions can also be used depending on the specific property of the data. To set a fairly noninformative prior for $\boldsymbol{\alpha}$, we specify a large value for its variance σ_0^2, e.g., $\sigma_0^2 = 20$. The hyperpriors for other parameters are

$$\sigma^2 \sim \text{IG}(\alpha_1, \zeta_1), \qquad \rho \sim \text{Unif}(0, U_\rho), \qquad \tau_p^2 \sim \text{IG}(\alpha_2, \zeta_2),\ p = 1, 2, 3, 4$$

where $\text{IG}(\alpha, \zeta)$ denotes an inverse gamma distribution with shape α and scale ζ and α_1, α_2, ζ_1, and ζ_2 are pre-specified.

2.2.4 Model Approximation

To conduct posterior inference for the SBPM and the HFSM, we adopt a discretized representation of model (2.1); specifically we approximate $\int \beta_j(t)\theta_{ij}(t)dt$ using the trapezoidal rule, similar in spirit to the discrete model formulation in Malloy et al. (2010). Such a model approximation, though not as widely used as the approach of basis expansion, facilitates one essential idea in our approach – feature selection within each functional predictor. Since the SBPM is a special case of the HFSM, we only discuss the model approximation for the HFSM, which can be readily modified to accommodate the SBPM. Similar to what is encountered in Section 2.2.2, we need to choose a set of grid points for model approximation. Again, we focus on the case where the set of grid points for model approximation is the same as the set of design points, i.e., $\{t_1, t_2, \ldots, t_L\}$, and our numerical studies in Sections 2.4 and 2.5 show that this approach performs well when the number of design points in the observed data is moderate to large, e.g., $L = 20$.

We denote functional predictors $\theta_j(\cdot)$, $\theta_{ij}(\cdot)$, $\theta_{ijk}(\cdot)$ and functional coefficients $\beta_j(\cdot)$ at design points by $\boldsymbol{\theta}_j = (\theta_j(t_1), \ldots, \theta_j(t_L))^T$, $\boldsymbol{\theta}_{ij} = (\theta_{ij}(t_1), \ldots, \theta_{ij}(t_L))^T$, $\boldsymbol{\theta}_{ijk} = (\theta_{ijk}(t_1), \ldots, \theta_{ijk}(t_L))^T$, and $\boldsymbol{\beta}_j = (\beta_j(t_1), \ldots, \beta_j(t_L))^T$, respectively. We denote the index set of selected (active) functional predictors by $\mathcal{S} = \{j : C_j = 1\}$, and the index set of selected (active) grid points for predictor j by $\mathcal{Q}_j = \bigcup_{k=1}^{K_j}(b_{j(2k-1)} : b_{j(2k)})$

where $(a : a^{'}) = \{a, a+1, \ldots, a^{'}\}$ for integers $a < a^{'}$. It follows that $\gamma_{jl} = 1$ if $l \in \mathcal{Q}_j$, otherwise $\gamma_{jl} = 0$. Let $d_j = |\mathcal{Q}_j|$ denote the number of active grid points for predictor j where $|\cdot|$ is the cardinality of a set. Given $\boldsymbol{C}$ and $\boldsymbol{\gamma}$, we denote selected functional coefficients at design points in selected regions by $\boldsymbol{\beta}_j^* = (\beta_j(t_l), l \in \mathcal{Q}_j)^T$ (of dimension $d_j \times 1$), and write $\boldsymbol{\beta}_{\mathcal{S}} = (\boldsymbol{\beta}_j^{*T}, j \in \mathcal{S})^T$ (of dimension $(\sum_{j\in\mathcal{S}} d_j) \times 1$) representing a collection of functional coefficients corresponding to all the selected functional predictor segments. Similarly, we write selected functional predictors at design points in selected regions by $\boldsymbol{\theta}_{ij}^* = (\theta_{ij}(t_l), l \in \mathcal{Q}_j)^T$ (of dimension $d_j \times 1$). We write $\boldsymbol{\theta}_{\mathcal{S}i} = (\boldsymbol{\theta}_{ij}^{*T}, j \in \mathcal{S})^T$ (of dimension $(\sum_{j\in\mathcal{S}} d_j) \times 1$) and $\boldsymbol{\theta}_{\mathcal{S}} = (\boldsymbol{\theta}_{\mathcal{S}1}, \boldsymbol{\theta}_{\mathcal{S}2}, \ldots, \boldsymbol{\theta}_{\mathcal{S}n})$ (of dimension $(\sum_{j\in\mathcal{S}} d_j) \times n$). Under the SBPM, $\boldsymbol{\beta}_j^*$ and $\boldsymbol{\theta}_{ij}^*$ simply become $\boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_{ij}$, respectively. Denote by $\triangle \boldsymbol{t}_{g,f}$ a diagonal matrix with the diagonal elements as $\big(\triangle t_{g(g+1)}, \triangle t_{g(g+2)}, \triangle t_{(g+1)(g+3)}, \ldots, \triangle t_{(f-3)(f-1)}, \triangle t_{(f-2)(f-1)}, \triangle t_{(f-1)f}\big)$, where $\triangle t_{gf} = \frac{t_f - t_g}{2}$. Define $\triangle \boldsymbol{T}_j$ as a block diagonal matrix with K_j blocks and the kth block being $\triangle \boldsymbol{t}_{b_{j(2k-1)}, b_{j(2k)}}$, and $\triangle \boldsymbol{T}$ as a block diagonal matrix with $|\mathcal{S}|$ blocks and the diagonal blocks being $\triangle \boldsymbol{T}_j$ $(j \in \mathcal{S})$. Then, model (2.1) is discretized according to the trapezoidal rule as follows,

$$z_i \;=\; \mathbf{s}_i^T \boldsymbol{\alpha} + \sum_{j \in \mathcal{S}} (\boldsymbol{\theta}_{ij}^*)^T \triangle \boldsymbol{T}_j \boldsymbol{\beta}_j^* + \varepsilon_i,$$

or in a more compact form,

$$\boldsymbol{Z} \;=\; \boldsymbol{S}^T \boldsymbol{\alpha} + (\boldsymbol{\theta}_{\mathcal{S}})^T \triangle \boldsymbol{T} \boldsymbol{\beta}_{\mathcal{S}} + \boldsymbol{\varepsilon}, \tag{2.12}$$

where $\boldsymbol{Z} = (z_1, \ldots, z_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, and $\boldsymbol{S} = (\mathbf{s}_i^T, \ldots, \mathbf{s}_n^T)^T$. In addition, it follows from models (2.3)–(2.5) that the priors for the discretized functional predictors are as follows,

$$\boldsymbol{\theta}_{ijk} \sim \mathrm{N}(\boldsymbol{\theta}_{ij}, \tau_1^2 H(\rho)), \quad \boldsymbol{\theta}_{ij} \sim \mathrm{N}(\boldsymbol{\theta}_j, \tau_2^2 H(\rho)), \quad \boldsymbol{\theta}_j \sim \mathrm{N}(\mathbf{0}, \tau_3^2 H(\rho)),$$

where $H(\rho)$ is the correlation matrix and its (p, q)th element is $\exp\{-\rho(t_p - t_q)^2\}$ $(p, q = 1, \ldots, L)$. From model (2.9), the priors for the discretized functional coefficients $\boldsymbol{\beta}_j^*$ are

$$[\boldsymbol{\beta}_j^* \mid C_j = 1, \boldsymbol{\gamma}_j\ ] \ \sim \ \mathrm{N}(\mathbf{0}, \tau_4^2 H_j^*(\rho)), j = 1, \ldots, m$$

where $H_j^*(\rho)$ is a sub-matrix of $H(\rho)$, namely, $H(\rho)[\mathcal{Q}_j, \mathcal{Q}_j]$, noting that $\mathcal{Q}_j$ is the index set of the active grid points in all selected regions of functional predictor j and is used to index the corresponding rows and columns of $H(\rho)$.

2.3 Posterior Inference

To conduct posterior inference, we adopt the Swendsen-Wang algorithm (Swendsen and Wang, 1987) by introducing two sets of auxiliary variable $\boldsymbol{u} = \{u_{st}; r_{st} = 1\}$ and $\boldsymbol{v} = \big\{v_{jl}, j = 1, \ldots, m; l = 1, \ldots, L - 2\big\}$ for $\boldsymbol{C}$ and $\boldsymbol{\gamma}$. Each u_{st} corresponds to an edge in $\mathcal{G}$ that connects the functional predictor pair (s, t) and each v_{jl} corresponds to an edge in functional predictor j that connects two adjacent subintervals indexed by l and $l + 1$. Specifically, given $\boldsymbol{C}$ and $\boldsymbol{\gamma}$, all the auxiliary variables are assumed to be mutually independent and their conditional distributions are given by

$$\pi(u_{st} \mid \boldsymbol{C}, \eta) \ = \ \exp(-\eta I(C_s = C_t)) \cdot I(0 \leq u_{st} \leq \exp(\eta I(C_s = C_t))), \tag{2.13}$$

$$\pi(v_{jl} \mid \gamma_j) \ = \ \exp(-\xi I(\gamma_{jl} = \gamma_{j(l+1)})) \cdot I(0 \leq v_{jl} \leq \exp(\xi I(\gamma_{jl} = \gamma_{j(l+1)}))), \tag{2.14}$$

noting that ξ is pre-specified. The full list of parameters in our model is $\boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_2$, $\boldsymbol{\Theta}_3$, $\boldsymbol{\beta}$, σ^2, $\boldsymbol{\alpha}$, $\boldsymbol{\tau}^2$, ρ, $\boldsymbol{C}$, $\boldsymbol{\gamma}$, $\boldsymbol{u}$, $\boldsymbol{v}$ and η, where $\boldsymbol{\Theta}_1 = (\boldsymbol{\theta}_j, j = 1, \ldots, m)$, $\boldsymbol{\Theta}_2 = (\boldsymbol{\theta}_{ij}, i = 1, \ldots, n; j = 1, \ldots, m)$, $\boldsymbol{\Theta}_3 = (\boldsymbol{\theta}_{ijk}, i = 1, \ldots, n; j = 1, \ldots, m; k = 1, \ldots, q_{ij})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_j, j = 1, \ldots, m)$, and $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_4^2)$. To speed up the convergence of the posterior simulation, we integrate out parameters $\boldsymbol{\Theta}_3$, $\boldsymbol{\Theta}_1$, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in the model. This leads to the target distribution of our MCMC algorithm:

$$\begin{aligned} & \pi(\boldsymbol{Z}, \boldsymbol{\Theta}_2, \boldsymbol{\tau}^2, \sigma^2, \rho, \boldsymbol{C}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \eta \mid Y, \boldsymbol{X}) \\ \propto \; & \pi(Y \mid \boldsymbol{Z}) \pi(\boldsymbol{Z} \mid \boldsymbol{\Theta}_2, \boldsymbol{C}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2, \rho) \pi(\boldsymbol{X} \mid \boldsymbol{\Theta}_2, \boldsymbol{\tau}^2, \rho, \sigma^2) \pi(\boldsymbol{\Theta}_2, \boldsymbol{\tau}^2, \rho, \sigma^2) \pi(\boldsymbol{C}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \eta) \end{aligned} \tag{2.15}$$

A detailed formulation for (2.15) is provided in Appendix 2.7.1. The posterior inference is similar for the SBPM and the HFSM and the main difference is in the updating scheme for the latent indicators $\boldsymbol{C}$ and $\boldsymbol{\gamma}$, noting that the SBPM does not involve $\boldsymbol{\gamma}$. Thus, we focus Sections 2.3.1 and 2.3.2 on the posterior inference for $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ under each model. The details of the complete MCMC algorithm are provided in Appendix 2.7.2. Given simulated samples from the posterior distribution (2.15), it is straightforward to make posterior inference on other parameters in the model by conditional sampling. For example, recall that each sample of $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ partitions the functional coefficient $\boldsymbol{\beta}$ into two parts $\boldsymbol{\beta}_{\mathcal{S}}$ and $\boldsymbol{\beta}_{-\mathcal{S}} = [(\boldsymbol{\beta}_{-j}^{*T}, j \in \mathcal{S}), (\boldsymbol{\beta}_j^T, j \notin \mathcal{S})]^T$ where $\boldsymbol{\beta}_{-j}^{*T} = (\beta_j(t_l), l \notin \mathcal{Q}_j)^T$. Thus, to obtain a sample from the posterior distribution of $\boldsymbol{\beta}$, we draw $\boldsymbol{\beta}_{\mathcal{S}}$ from $\mathrm{N}(\mu_\beta, \Sigma_\beta)$ and set $\boldsymbol{\beta}_{-\mathcal{S}} = \boldsymbol{0}$, where $\Sigma_\beta = [\triangle \boldsymbol{T} \boldsymbol{\theta}_{\mathcal{S}} \{I_n - \boldsymbol{S}(I_p \sigma_0^{-2} + \boldsymbol{S}^T \boldsymbol{S})^{-1} \boldsymbol{S}^T\} \boldsymbol{\theta}_{\mathcal{S}}^T \triangle \boldsymbol{T} + (\tau_4^2 H^*(\rho))^{-1}]^{-1}$ and $\mu_\beta = \Sigma_\beta \cdot (\triangle \boldsymbol{T} \boldsymbol{\theta}_{\mathcal{S}}) \{I_n - \boldsymbol{S}(I_p \sigma_0^{-2} + \boldsymbol{S}^T \boldsymbol{S})^{-1} \boldsymbol{S}^T\} \cdot \boldsymbol{Z}$. We note that our posterior simulation algorithm is similar in spirit to the SSVS algorithm proposed by George and McCulloch (1997) and the Bayesian variable selection approach by Li and Zhang (2010) and the dimension of the parameter space in our MCMC algorithm does not change.

2.3.1 Posterior inference for $\boldsymbol{C}$ under SBPM

Note that in (2.21) a value $u_{st} > 1$ indicates $C_s = C_t$. Hence, given a sample $\boldsymbol{u}$, we partition the m functional predictors into G classes ($\mathcal{F}_g, g = 1, \ldots, G$) with l_g functional predictors in class g ($g = 1, \ldots, G$), and the predictors in the same class are either added to or dropped from the active set $\mathcal{S}$ together. The full conditional distribution of $\boldsymbol{C}$ follows a discrete distribution with $N = 2^G$ elements which is given by

$$\begin{aligned}&\pi(\boldsymbol{C} = (\boldsymbol{c}_{\mathcal{F}_1}, \ldots, \boldsymbol{c}_{\mathcal{F}_G}) \mid \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{\theta}_{\mathcal{S}}, \boldsymbol{\tau}^2, \rho, \boldsymbol{u}) \\ \propto\ & |\Sigma_t|^{\frac{1}{2}} |\Sigma_\alpha|^{\frac{1}{2}} \exp\left[\frac{\boldsymbol{Z}^T\{\boldsymbol{W} + (\boldsymbol{S} - \boldsymbol{S}\boldsymbol{W})\Sigma_\alpha(\boldsymbol{S} - \boldsymbol{S}\boldsymbol{W})^T\}\boldsymbol{Z}}{2}\right] |\tau_4^2 H^*(\rho)|^{-\frac{1}{2}}\end{aligned} \tag{2.16}$$

where $\boldsymbol{c}_{\mathcal{F}_g} = c_g \mathbf{1}_{l_g}$ with $c_g \in \{0, 1\}$ ($g = 1, \ldots, G$), $\mathbf{1}_d$ is a d-vector with all elements equal to 1, $\Sigma_t = \left\{(\triangle\boldsymbol{T} \cdot \boldsymbol{\theta}_{\mathcal{S}})(\triangle\boldsymbol{T} \cdot \boldsymbol{\theta}_{\mathcal{S}})^T + (\tau_4^2 H^*(\rho))^{-1}\right\}^{-1}$, $\boldsymbol{W} = (\triangle\boldsymbol{T} \cdot \boldsymbol{\theta}_{\mathcal{S}})^T \Sigma_t (\triangle\boldsymbol{T} \cdot \boldsymbol{\theta}_{\mathcal{S}})$, and $\Sigma_\alpha = (\boldsymbol{S}^T\boldsymbol{S} - \boldsymbol{S}^T\boldsymbol{W}\boldsymbol{S} + I_p\sigma_0^{-2})^{-1}$. Here, $H^*(\rho)$ and $\triangle\boldsymbol{T}$ are defined as those in Section 2.2.4 with $\mathcal{Q}_j = (1 : L)$ for $j \in \mathcal{S}$, i.e., without feature selection within functional predictors.

The above algorithm becomes computationally intensive when G becomes large. As an alternative, we update each element in $\boldsymbol{C}_{\mathcal{F}_g} = (C_j, j \in \mathcal{F}_g)$ for $g = 1, \ldots, G$ directly from a Bernoulli distribution:

$$\begin{aligned}&\pi(\boldsymbol{C}_{\mathcal{F}_g} = k\mathbf{1}_{l_g} \mid \boldsymbol{C}_{-\mathcal{F}_g}, \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{\theta}_{\mathcal{S}}, \boldsymbol{\tau}^2, \rho, \boldsymbol{u}) \\ &\propto |\Sigma_t|^{\frac{1}{2}} |\Sigma_\alpha|^{\frac{1}{2}} \exp\left[\frac{\boldsymbol{Z}^T\{\boldsymbol{W} + (\boldsymbol{S} - \boldsymbol{S}\boldsymbol{W})\Sigma_\alpha(\boldsymbol{S} - \boldsymbol{S}\boldsymbol{W})^T\}\boldsymbol{Z}}{2}\right] |\tau_4^2 H^*(\rho)|^{-\frac{1}{2}}\end{aligned} \tag{2.17}$$

where k takes values 0 or 1, and $\boldsymbol{C}_{-\mathcal{F}_g} = (\boldsymbol{C}_{\mathcal{F}_g}, \ldots, \boldsymbol{C}_{\mathcal{F}_{g-1}}, \boldsymbol{C}_{\mathcal{F}_{g+1}}, \ldots, \boldsymbol{C}_{\mathcal{F}_G})$.

2.3.2 Posterior Inference for $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ under HFSM

Under the HFSM, both $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ are included in posterior inference to realize hierarchical feature selection. We consider two algorithms for updating $\boldsymbol{C}$ and $\boldsymbol{\gamma}$. One is a nested Swendsen-Wang algorithm through which we jointly update $(\boldsymbol{C}, \boldsymbol{\gamma})$. In (2.21), a value $v_{jl} > 1$ indicates $\gamma_{jl} = \gamma_{j(l+1)}$, thus, $\boldsymbol{u}$ and $\boldsymbol{v}$ divides all functional predictors into G classes; it also divides the set of $L-1$ subintervals for each functional predictor j into H_j groups with their index sets defined as $\{\psi_{jh}, h = 1, \ldots, H_j\}$ and group h has l_{jh} elements $(h = 1, \ldots, H_j)$ with $\sum_{h=1}^{H_j} l_{jh} = L - 1$. Let $\boldsymbol{\gamma}_{\mathcal{F}_g} = (\boldsymbol{\gamma}_j, j \in \mathcal{F}_g)$. Given the constraint (2.8) on $\boldsymbol{\gamma}$, a natural extension of (2.17) for the posterior inference for $(\boldsymbol{C}_{\mathcal{F}_g}, \boldsymbol{\gamma}_{\mathcal{F}_g})$ is implemented by drawing from a discrete distribution with $N' = 1 + \prod_{j \in \mathcal{F}_g}(2^{H_j} - 1)$ elements. However, this algorithm becomes computationally infeasible even for moderate G and H_j's.

To mitigate this problem, we consider an alternative approach by using a Metropolis-Hastings (MH) algorithm to jointly update $(\boldsymbol{C}, \boldsymbol{\gamma})$. We choose the following proposal distribution:

$$g(\boldsymbol{C}^*, \boldsymbol{\gamma}^* \mid \boldsymbol{C}^{(o)}, \boldsymbol{\gamma}^{(o)}, \boldsymbol{\mathcal{P}}, \boldsymbol{S}, \boldsymbol{R}) \;=\; \pi(\boldsymbol{\gamma}^* \mid \boldsymbol{C}^*, \boldsymbol{\mathcal{P}}, \boldsymbol{S}) \cdot p(\boldsymbol{C}^* \mid \boldsymbol{S}, \boldsymbol{R}), \qquad (2.18)$$

where the superscript "$*$" and "(o)" denote the proposed and the current parameter value respectively, and $\boldsymbol{\mathcal{P}} = \{\boldsymbol{Z}, \boldsymbol{\theta}_{\mathcal{S}}, \boldsymbol{\tau}^2, \boldsymbol{v}, \rho, \eta\}$. The term $p(\boldsymbol{C}^* \mid \boldsymbol{S}, \boldsymbol{R})$ is the marginal posterior probability of $\boldsymbol{C}$ defined under the SBPM and $\pi(\boldsymbol{\gamma}^* \mid \boldsymbol{C}^*, \boldsymbol{\mathcal{P}}, \boldsymbol{S})$ is the conditional posterior probability of $\boldsymbol{\gamma}^*$ defined under the HFSM. Drawing $(\boldsymbol{C}^*, \boldsymbol{\gamma}^*)$ from this proposal entails two steps. We first draw $\boldsymbol{C}^*$ from the posterior sample of $\boldsymbol{C}$ generated using the algorithms in Section 2.3.1 based on (2.16) or (2.17). Subsequently, given $\boldsymbol{C}^*$, we draw $\boldsymbol{\gamma}^*$ from its full conditional (provided in Appendix 2.7.2) with $\boldsymbol{C}$ fixed at $\boldsymbol{C}^*$. One can show that $p(\boldsymbol{C} \mid \boldsymbol{S}, \boldsymbol{R}) > 0$ for any $\boldsymbol{C}$, so the sample space for $\boldsymbol{C}$ under this proposal function is the same as that for $\boldsymbol{C}$ under the HFSM, making (2.18) a valid proposal. In addition, our numerical studies show that this proposal achieves satisfactory performance.

2.4 Simulation Studies

We conduct simulation studies to assess the performance of the proposed approach in terms of hierarchical feature selection for structured functional predictors. Since no existing methods can handle hierarchical feature selection between and within functional predictors, we focus on comparing the HFSM and the SBPM and note that the approach by Zhu et al. (2009) is similar to the SBPM but cannot incorporate the biological information between functional predictors. Each Monte Carlo data set, containing $n = 50$ or $n = 100$ subjects, is generated based on models (2.1) and (2.3)-(2.5) including eight functional predictors and one scalar predictor. The scalar predictor is generated from a uniform distribution on $(-15, 25)$. The intercept and the scalar coefficient are set to $\boldsymbol{\alpha} = (1, 1)$. The underlying true curves for each functional predictor – namely, $\theta_{ijk}(\cdot)$, $\theta_{ij}(\cdot)$ and $\theta_j(\cdot)$ – are generated from models (2.3)-(2.5) with $\tau_1^2 = \tau_2^2 = \tau_3^2 = 0.5$ and $\rho = 36$, where $i = 1, \ldots, n$, $j = 1, \ldots, 8$, and $k = 1, \ldots, 10$. Given $\theta_{ijk}(\cdot)$, the observed functional data $\boldsymbol{X}$ are generated at 20 equally spaced design points between 0 to 1 with independent measurement error added. To examine the effect of measurement error, all components of σ^2 are set to 0.1 or 0.2. The functional predictors are structured through the network that contains edges $\{1 \sim 4,\ 2 \sim 4,\ 3 \sim 4,\ 4 \sim 5,\ 5 \sim 6,\ 5 \sim 7,\ 5 \sim 8\}$ with functional predictors 4 and 5 as the central nodes. Following this structure, we set the true values for $\boldsymbol{C}$ as $(1, 1, 1, 1, 0, 0, 0, 0)$. The binary outcome, $\boldsymbol{Y}$, is generated from model (2.1) where the integrals are calculated using the approach of Gaussian quadrature. The true functional coefficients are set as $\beta_1(t) = 0.5 + 9\sin(4t+1)$, $\beta_2(t) = 1.5 - 8\sin(6t)$, $\beta_3(t) = 1 - 8\sin(ct+1)$ with $c = 19/14 \times (\arcsin(0.125) + 4)$, $\beta_4(t) = 1 + 8\sin(4t - 1.05)$, and $\beta_5(\cdot) = \cdots = \beta_8(\cdot) = 0$. In addition, to evaluate feature selection for both discontinuous ($\beta_1(\cdot)$ and $\beta_2(\cdot)$) and continuous ($\beta_3(\cdot)$ and $\beta_4(\cdot)$) functional coefficients, we set $\beta_1(\cdot) = 0$ in $(0.000, 0.315)$ and $(0.632, 1.000)$, $\beta_2(\cdot) = 0$ in $(0.263, 0.789)$, and $\beta_3(\cdot) = 0$ in $(0.000, 0.737)$, whereas $\beta_4(\cdot)$ remains unchanged.

Given the simulated data $(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X})$, the analyses are conducted using both the HFSM and the SBPM, where we use the set of design points as the set of grid points for discretizing the GFLM and for feature selection. In both models, we use a flat prior for $\boldsymbol{\alpha}$ with $\sigma_0^2 = 20$ and set hyper-parameters $\alpha_1, \alpha_2, \zeta_1$ and ζ_2 to 1 and U_ρ to 1000. We let the smoothing parameter ξ in the Ising prior for $\boldsymbol{\gamma}$ vary in $(0.0, 1.0, 1.5, 2.0, 2.5, 3.0)$ to investigate the effect of different ξ. We use the marginal posterior mode of $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ to conduct feature selection. In each simulation scenario, multiple chains with random initial values are run for 5000 iterations with the first 2000 as burn-in. Our results show that the posterior inference is insensitive to initial values and a proper mixing for each parameter is verified by the trace plots.

In all settings, the posterior samples of $\boldsymbol{C}$ converge to its true value within 30 iterations, indicating a good performance of our method on feature selection between functional predictors. Therefore, we focus on feature selection within predictors and we calculate the sensitivity (Sens) and the specificity (Spec) for feature selection within $\boldsymbol{\gamma}$,

$$\text{Sens} = \frac{\sum_{(j,l)\in\mathcal{S}^0} \mathrm{I}(\hat{pr}(\gamma_{jl} = 1) \geq 0.5)}{|\mathcal{S}^0|}, \qquad \text{Spec} = \frac{\sum_{(j,l)\notin\mathcal{S}^0} \mathrm{I}(\hat{pr}(\gamma_{jl} = 1) < 0.5)}{mL - |\mathcal{S}^0|},$$

where $\mathcal{S}^0 = \{(j,l) : \gamma_{jl} = 1\}$ is the true active set in $\boldsymbol{\gamma}$ and $\hat{pr}(\gamma_{jl} = 1)$ is the marginal posterior probability of $\gamma_{jl} = 1$.

Table 2.1 summarizes the simulation results for different settings. The proposed approach achieves a very good performance under $n = 50$ and a larger sample size ($n = 100$) leads to only modest improvement in performance. When ξ is between 1.0 and 2.0, the proposed approach achieves a satisfactory performance in feature selection with both the sensitivity and the specificity approaching 1. As ξ increases, the sensitivity further approaches or remains at 1, whereas the specificity gradually decreases and approaches that of the model with only feature selection between predictors (i.e., the SBPM). This is expected since larger values of ξ induce stronger effect of the Ising prior on the posterior inference of $\boldsymbol{\gamma}$; eventually, the Ising prior would dominate the likelihood and, in our case, result in the same values in the posterior samples for all γ's that are connected in a network. The impact of σ^2 is moderate in our simulation studies and the results in the cases of $\sigma^2 = 0.2$ and 0.1 are comparable given the same sample size, showcasing the ability of our model to handle measurement error. The criterion of the posterior Bayes factor is presented in Table 2.1 with the model under $\xi = 0$ as the reference; an optimal value of $\xi = 1.5$ is chosen for both σ^2 values under $n = 50$ and $\xi = 1.5$ ($\xi = 1.0$) is chosen in the case of $\sigma^2 = 0.1$ ($\sigma^2 = 0.2$) under $n = 100$. As shown in Table 2.1, $\xi = 1.0$ and $\xi = 1.5$ lead to two of the best results in feature selection, confirming the usefulness of this tuning strategy.

To further evaluate the performance on feature selection within functional predictors, Figure 2.2 presents the marginal posterior probability of $\gamma = 1$ for the selected functional predictors (i.e., $\boldsymbol{\gamma}_1$ through $\boldsymbol{\gamma}_4$) under $n = 50$ in the case of $\sigma^2 = 0.1$ and $\xi = 1.5$. Figure 2.2 shows that our model correctly selects most of the nonzero regions for $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_3$, which have considerably higher posterior probabilities of being selected than the zero regions. The only false positive in the second curve is located closely to the transition point between zero and nonzero regions. For $\beta_4(\cdot)$ which has

		$\sigma^2 = 0.1$			$\sigma^2 = 0.2$		
n	ξ	Sens	Spec	PBF	Sens	Spec	PBF
50	0.0	0.750	0.982	*referent*	0.725	1.000	*referent*
	1.0	0.900	0.991	46.244	0.800	1.000	-47.314
	1.5	0.925	0.991	195.809	0.900	0.991	26.954
	2.0	0.950	0.973	185.245	0.925	0.991	-55.004
	2.5	0.950	0.929	66.696	0.950	0.884	-69.436
	3.0	1.000	0.804	131.257	0.950	0.839	-25.775
	SBPM	1.000	0.679	-2.195	1.000	0.679	-308.282
100	0.0	0.800	0.955	*referent*	0.850	0.964	*referent*
	1.0	0.900	0.982	-114.533	0.950	0.982	338.773
	1.5	1.000	0.982	280.001	1.000	0.982	-70.689
	2.0	1.000	0.964	118.216	1.000	0.938	-14.098
	2.5	1.000	0.848	91.300	1.000	0.884	-62.905
	3.0	1.000	0.759	167.309	1.000	0.804	-40.732
	SBPM	1.000	0.679	-338.464	1.000	0.679	-29.413

Table 2.1: Simulation results under different σ^2 and n. Sens and Spec, sensitivity and specificity for $\boldsymbol{\gamma}$; and PBE, log posterior Bayes factor with the model under $\xi = 0$ as the referent.

no pre-specified zero region, the posterior mode of $\boldsymbol{\gamma}_4$ is exactly equal to the truth; however, the posterior probabilities of being selected are lower in the region (0.579, 0.737) than in the other regions. A closer examination reveals that this region is close to the point where the true functional coefficient $\beta_4(\cdot)$ crosses zero.

Finally, Figure 2.3 presents the posterior means and credible intervals for $\beta_1(\cdot)$ through $\beta_4(\cdot)$ with a comparison between the HFSM under $\xi = 1.5$ (right panel) and the SBPM (left panel) in the case of $\sigma^2 = 0.1$ and $n = 50$, illustrating the differences between the two models. As shown in Figure 2.3, for all four functional predictors, the 95% credible intervals estimated by the HFSM cover the true curves and the posterior means are fairly close to the truth. In the case of the SBPM, however, the posterior inference fails to identify the regions in which each functional coefficient is 0, resulting in over-smoothed posterior means for such functional coefficients.

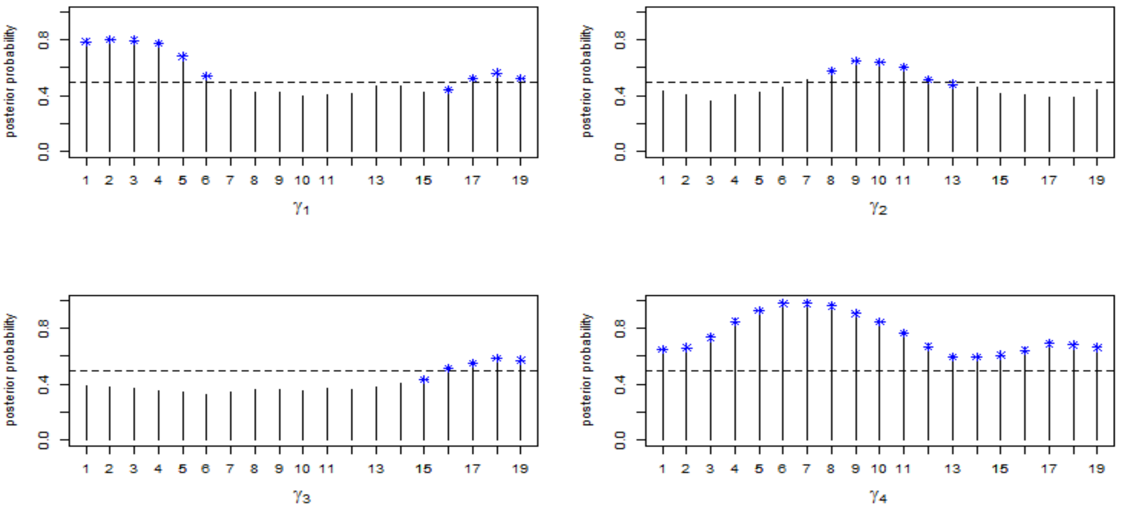


Figure 2.2: Posterior probability of $\gamma = 1$ for the functional coefficients that are selected in the case of $\sigma^2 = 0.1$ and $\xi = 1.5$ (n=50). The symbol $*$ indicates a true value of 1 for the corresponding indicator γ. The dotted horizontal line highlights the cutoff of 0.5.

2.5 Application to the Colorectal Adenoma Data

We apply the proposed approach to the motivating study introduced in Section 1.1.1. In this study, a total of 17 functional biomarkers (Table 1.1) were measured, though not all biomarkers were measured for each participant. Since the data processing is still ongoing, the final results will be reported elsewhere once all data become available. Nevertheless, our current analysis of the data represents the first attempt to identify functional biomarkers and their features that are associated with the risk for colorectal cancer through hierarchical feature selection while incorporating biological information. We conduct an analysis of seven functional biomarkers (namely, APC, MSH2, bax, β-catenin, E-cadherin, MLH1, and TGFβ_1), for which data processing has been completed. In the data set for our analysis, we have complete data for all seven biomarkers in a subset of 44 subjects and the number of hemicrypts scored per biomarker and per subject ranges between 1 to 76. Functional data measured at 24 equally-spaced design points between 0 and 1 are used to model biomarker profiles with 0 representing the base of crypts and 1 representing the top of crypts. The biological information listed in (Table 1.1) and the spatial information are incorporated through Ising priors (2.10) and (2.11), respectively, in the analyses. The outcome variable, y, is binary with 1 for presence of adenoma and 0 for absence of adenoma; our GFLM also includes a scalar predictor, age.

Both the SBPM and the HFSM are applied to the data set, where we again use the set of design points as the set of grid points for discretizing the GFLM and for feature selection. Similar to the simulation studies, we set $\sigma_0^2 = 20$, $(\alpha_1, \alpha_2, \zeta_1, \zeta_2) = (1, 1, 1, 1)$ and $U_\rho = 1000$. We fit the HFSM with different values of the smoothing parameter ξ in the Ising prior (2.11) for $\boldsymbol{\gamma}$, ranging between $\xi = 1.0$ to 3.0 with a step size of 0.1 as well as $\xi = 0$, and we use the posterior Bayes factor to choose an optimal ξ value. For each model, starting with random initial values, we run MCMC chains for 5,000 iterations with a burn-in period of 2,000 iterations. The trace plots for all parameters are checked for convergence, all showing satisfactory mixing.

The hierarchical feature selection is conducted based on the marginal posterior mode of $\boldsymbol{C}$ and $\boldsymbol{\gamma}$. TGFβ_1 is the only biomarker selected by the SBPM and the HFSM with different ξ values and the marginal posterior probability for selecting TGFβ_1 is greater than 0.9 in all models. For the HFSM, $\xi = 2.5$ is chosen as the optimal value based on the criterion of the posterior Bayes factor. Under this model, the region between 0.435 and 0.826 for TGFβ_1 is selected. The right panel in Figure 2.4 presents the posterior mean and 95% credible interval of the functional coefficient $\beta(\cdot)$ for TGFβ_1; it shows that the 95% credible interval of $\beta(\cdot)$ excludes 0 between 0.565 and 0.696, indicating that higher expression level of TGFβ_1 in this region is associated with lower risk for colorectal cancer. In addition, the lower 60% and the upper 40% of colon crypts are known as the proliferation zone and the differentiation zone, respectively (Gerdes et al., 1993); the expression levels of biomarkers that are involved in cell cycle change in the transitional region. Our results indicate that the expression level of TGFβ_1 near this transitional region is likely predictive of the risk for colorectal cancer. To evaluate the goodness of fit of the model, a posterior predictive assessment (Gelman et al., 1996) is conducted using a discrepancy measure, the sum of mean χ^2 discrepancy for y_i and mean χ^2 discrepancy for X_{ijk}. The resulting posterior predictive p-value of 0.542 suggests a good fit to the data.

When the HFSM with $\xi = 0$ is used (i.e., feature selection is conducted within functional predictors without incorporating the spatial information), no region within the expression profile of TGFβ_1 is selected with the posterior probabilities of $\boldsymbol{\gamma} = 1$ all less than 0.5; evidently, without incorporating the spatial information, signals selected under the HFSM with $\xi = 2.5$ – the region between 0.435 and 0.826 – are lost. When the SBPM is used (i.e., feature selection is not conducted within functional predictors), the estimated functional coefficient is fairly flat with its 95% credible interval covering 0 throughout (the left panel of Figure 2.4), underestimating the strength of the association between TGFβ_1 and the outcome. In addition, we also fit a GFLM without feature selection at any level and find no statistically significant results, further demonstrating the importance of feature selection between and within functional predictors.

2.6 Discussion

In this article, we propose a unified Bayesian framework for hierarchical feature selection of structured functional predictors in GFLMs, which also accommodates multi-level functional data and measurement error. We present our method in a simplified setup where the set of grid points for feature selection within functional predictors and for model discretization are the same as the set of design points in the observed data. We briefly describe here how to extend our method to conduct posterior inference on an arbitrary set of grid points. Suppose that we consider a set of grid points that include all design points in the observed data and L^a additional points $\{t_l^a, l = 1, \ldots, L^a\}$. A careful examination of our method shows that we can conduct posterior inference using this set of grid points as long as we can generate the posterior samples for $\boldsymbol{\theta}_{ij}^a = (\theta_{ij}(t_l^a), l = 1, \ldots, L^a)$. This can be readily achieved through the posterior predictive distribution for $\boldsymbol{\theta}_{ij}^a$ by recognizing that $(\boldsymbol{\theta}_{ij}^a, \boldsymbol{\theta}_{ij})$ follows a multivariate normal distribution based on model (2.4). Specifically, we draw $\boldsymbol{\theta}_{ij}^a$ from a conditional normal distribution given $\boldsymbol{\theta}_{ij}$ with $\boldsymbol{\theta}_{ij}$ drawn from its posterior distribution.

In our numerical studies, we choose the marginal posterior mode of $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ to conduct hierarchical feature selection in our model. A different threshold can also be used, in particular, if it is motivated by prior knowledge. In addition, the smoothing parameter ξ in the Ising prior (2.11) is fixed in posterior inference to avoid potentially intractable computation. We use the posterior Bayes factor (Aitkin, 1991) to choose an optimal value for ξ for its ease of computation, which is shown to achieve good performance in our simulation studies. In practice, we can also use other criteria such as variations of the Bayes factor as discussed in Kadane and Lazar (2004).

Our model does not impose any continuity condition on functional coefficients at the boundaries of selected and unselected regions, essentially allowing for jumps within a functional coefficient at these boundaries. Such jumps may capture a sudden change of biological events and are potentially biologically meaningful. In cases where continuity holds true, our model is expected to still work given that our model relies on a weaker assumption, as shown in our simulation studies. If such prior knowledge is provided, our model can be extended to impose this condition by modifying the mixture prior (2.9).

2.7 Appendix

2.7.1 Marginalizing the Likelihood with respect to $\boldsymbol{\Theta}_3$, $\boldsymbol{\Theta}_1$, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$

The posterior likelihood of all the parameters $(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\tau}^2, \rho, \boldsymbol{C}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \eta)$ is

$$\begin{aligned}
&\pi(\boldsymbol{Z}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\tau}^2, \rho, \boldsymbol{C}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \eta \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}) \\
&\quad \propto \prod_{i=1}^{n} (I[z_i > 0] I[y_i = 1] + I[z_i \leq 0] I[y_i = 0]) \exp\Big\{ - \frac{(\boldsymbol{Z} - \boldsymbol{S}^T \boldsymbol{\alpha} - \boldsymbol{\theta}_{\mathcal{S}}^T \triangle \boldsymbol{T} \boldsymbol{\beta}_{\mathcal{S}})^{\otimes 2}}{2} \Big\} \\
&\quad \cdot \prod_{i=1}^{n} \prod_{j=1}^{m} \prod_{k=1}^{q_{ij}} \Bigg[\sigma^{-L} \exp\left\{ - \frac{(\boldsymbol{X}_{ijk} - \boldsymbol{\theta}_{ijk})^T (\boldsymbol{X}_{ijk} - \boldsymbol{\theta}_{ijk})}{2\sigma^2} \right\} |\tau_1^2 H(\rho)|^{-\frac{1}{2}} \exp\Big\{ - \frac{1}{2\tau_1^2} (\boldsymbol{\theta}_{ijk} - \boldsymbol{\theta}_{ij})^T \\
&\quad H(\rho)^{-1} (\boldsymbol{\theta}_{ijk} - \boldsymbol{\theta}_{ij}) \Big\} \Bigg] \cdot \prod_{i=1}^{n} \prod_{j=1}^{m} |\tau_2^2 H(\rho)|^{-\frac{1}{2}} \exp\left\{ - \frac{(\boldsymbol{\theta}_{ij} - \boldsymbol{\theta}_j)^T H(\rho)^{-1} (\boldsymbol{\theta}_{ij} - \boldsymbol{\theta}_j)}{2\tau_2^2} \right\} \\
&\quad \prod_{j=1}^{m} |\tau_3^2 H(\rho)|^{-\frac{1}{2}} \exp(- \frac{\boldsymbol{\theta}_j^T H(\rho)^{-1} \boldsymbol{\theta}_j}{2\tau_3^2}) \cdot |2\pi\tau_4^2 H^*(\rho)|^{-\frac{1}{2}} \exp(- \frac{\boldsymbol{\beta}_{\mathcal{S}}^T H^*(\rho)^{-1} \boldsymbol{\beta}_{\mathcal{S}}}{2\tau_4}) \exp(- \frac{\alpha^T \alpha}{2\sigma_0^2}) \\
&\quad (\sigma^2)^{-\alpha_1 - 1} \exp(- \frac{\gamma_1}{\sigma^2}) \prod_{p=1}^{4} (\tau_p^2)^{-\alpha_2 - 1} \exp(- \frac{\gamma_2}{\tau_p^2}) \cdot \prod_{s,t: r_{st} = 1} I(0 \leq u_{st} \leq \exp(\eta I(C_s = C_t))) \\
&\quad \cdot \prod_{j=1}^{m} \prod_{l=1}^{L-2} I(0 \leq v_{jl} \leq \exp(\xi I(\gamma_{jl} = \gamma_{j(l+1)}))) \cdot \prod_{j=1}^{m} I(\max_l \gamma_{jl} = C_j) \cdot \frac{I[0 < \rho < U_\rho]}{U_\rho} \cdot \frac{I[0 < \eta < U_\eta]}{U_\eta}.
\end{aligned}$$

After integrating out $\boldsymbol{\Theta}_3$, $\boldsymbol{\Theta}_1$, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we obtain $\pi(\boldsymbol{Z}, \boldsymbol{\Theta}_2, \boldsymbol{\tau}^2, \sigma^2, \rho, \boldsymbol{C}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \eta \mid$

$\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S})$

$$
\begin{aligned}
&\pi(\boldsymbol{Z}, \boldsymbol{\Theta}_2, \boldsymbol{\tau}^2, \sigma^2, \rho, \boldsymbol{C}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}, \eta \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}) \\
&\quad \propto \prod_{i=1}^{n} (I[z_i > 0] I[y_i = 1] + I[z_i \le 0] I[y_i = 0]) \exp\Big[-\frac{1}{2} \boldsymbol{Z}^T \{ I_n - \boldsymbol{W} - (\boldsymbol{S} - \boldsymbol{W} \cdot \boldsymbol{S}) \Sigma_\alpha (\boldsymbol{S} \\
&\quad -\boldsymbol{W} \cdot \boldsymbol{S})^T \} \boldsymbol{Z} \Big] \cdot \exp\Bigg[-\frac{1}{2} \sum_{j=1}^{m} \Bigg\{ \sum_{i=1}^{n} \frac{\tau_1^2 + q_{ij}\tau_2^2}{(\tau_2\tau_1)^2} \boldsymbol{\theta}_{ij}^T H(\rho)^{-1} \boldsymbol{\theta}_{ij} - \frac{\tau_3^2}{\tau_2^4 + n(\tau_3\tau_2)^2} (\sum_{i=1}^{n} \boldsymbol{\theta}_{ij})^T H(\rho)^{-1} \\
&\quad (\sum_{i=1}^{n} \boldsymbol{\theta}_{ij}) \Bigg\} \Bigg] \cdot \exp\Bigg\{ \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{q_{ij}} \frac{1}{2} \left(\frac{H(\rho)^{-1}}{\tau_1^2} \boldsymbol{\theta}_{ij} + \sigma^{-2} \boldsymbol{X}_{ijk} \right)^T \left(\frac{H(\rho)^{-1}}{\tau_1^2} + \sigma^{-2} I_L \right)^{-1} \Big(\frac{H(\rho)^{-1}}{\tau_1^2} \boldsymbol{\theta}_{ij} \\
&\quad +\sigma^{-2} \boldsymbol{X}_{ijk} \Big) \Bigg\} \prod_{i=1}^{n} \prod_{j=1}^{m} \prod_{k=1}^{q_{ij}} \sigma^{-L} \exp(-\frac{\boldsymbol{X}_{ijk}^T \boldsymbol{X}_{ijk}^T}{2\sigma^2}) \cdot (\sigma^2)^{-\alpha_1 - 1} \exp(-\frac{\gamma_1}{\sigma^2}) \prod_{p=1}^{4} (\tau_p^2)^{-\alpha_2 - 1} \exp(-\frac{\gamma_2}{\tau_p^2}) \\
&\quad \prod_{s,t:r_{st}=1} I(0 \le u_{st} \le \exp(\eta I(C_s = C_t))) \cdot \prod_{j=1}^{m} \prod_{l=1}^{L-2} I(0 \le v_{jl} \le \exp(\xi I(\gamma_{jl} = \gamma_{j(l+1)}))) \\
&\quad \prod_{j=1}^{m} I(\max_l \gamma_{jl} = C_j) \cdot \frac{I[0 < \rho < U_\rho]}{U_\rho} \cdot \frac{I[0 < \eta < U_\eta]}{U_\eta} \frac{|H(\rho)|^{-\frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (q_{ij}+1)}{2}}}{\tau_1^{\sum_{i=1}^{n} \sum_{j=1}^{m} q_{ij} L} \tau_2^{m(n-1)L} (\tau_2^2 + n\tau_3^2)^{\frac{mL}{2}}} \\
&\quad \cdot \left| \frac{H(\rho)^{-1}}{\tau_1^2} + \sigma^{-2} I_L \right|^{-\frac{\sum_{i=1}^{n} \sum_{j=1}^{m} q_{ij}}{2}} \cdot |\Sigma_t|^{\frac{1}{2}} |\Sigma_\alpha|^{\frac{1}{2}} |\tau_4^2 H^*(\rho)|^{-\frac{1}{2}}.
\end{aligned}
$$

2.7.2 MCMC Algorithm

We provide here the details of the Metropolis–Hastings within Gibbs sampling algorithm for posterior inference under the HFSM . The MCMC algorithm under the SBPM is a special case of this algorithm.

Scheme for sampling $\boldsymbol{Z}$: The full conditional for $\boldsymbol{Z}$ is as follows:

$$\pi(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\Theta}_2, \tau_4^2, \rho, \boldsymbol{C}, \boldsymbol{\gamma}) \sim \mathrm{TN}_n(\boldsymbol{0}, \boldsymbol{\mu}^{Z-}, \boldsymbol{\mu}^{Z+}, \Sigma^Z), \tag{2.19}$$

with $\mathrm{TN}_n(\boldsymbol{\mu}, \boldsymbol{\mu}^-, \boldsymbol{\mu}^+, \Sigma)$ denoting a n dimensional truncated normal distribution with mean $\boldsymbol{\mu}$ (of dimension $n \times 1$), element-wise lower and upper bounds $\boldsymbol{\mu}^-$ (of dimension

$n\times1$) and $\boldsymbol{\mu}^+$ (of dimension $n\times1$), and covariance matrix Σ (of dimension $n\times n$). Here, $\boldsymbol{\mu}^{Z-} = (\mu^{z_1-},\dots,\mu^{z_n-})^T = (\frac{I[y_1=1]-1}{I[y_1=1]},\dots,\frac{I[y_n=1]-1}{I[y_n=1]})^T$, $\boldsymbol{\mu}^{Z+} = (\mu^{z_1+},\dots,\mu^{z_n+})^T = (\frac{1-I[y_1=0]}{I[y_1=0]},\dots,\frac{1-I[y_n=0]}{I[y_n=0]})^T$ and $\Sigma^Z = \{I_n - \boldsymbol{W} - (\boldsymbol{S}-\boldsymbol{W}\cdot\boldsymbol{S})\Sigma_\alpha(\boldsymbol{S}-\boldsymbol{W}\cdot\boldsymbol{S})^T\}^{-1}$.

Based on (2.19), we perform an element-wise update for $\boldsymbol{Z}$ from conditional univariate truncated normal distributions through a Gibbs sampler with

$$\pi(z_i \mid \boldsymbol{Z}_{-i}, y_i, \boldsymbol{S}, \boldsymbol{\Theta}_2, \tau_4^2, \rho, \boldsymbol{C}, \boldsymbol{\gamma}) \;\sim\; \mathrm{TN}_1(\mu_{z_i|\boldsymbol{Z}_{-i}}, \mu^{z_i-}, \mu^{z_i+}, \sigma_i^2), \quad i = 1,\dots,n,$$

where $\boldsymbol{Z}_{-i} = (z_1,\dots,z_{i-1},z_{i+1},\dots,z_n)$, $\mu_{z_i|\boldsymbol{Z}_{-i}} = (\Sigma^Z_{i,-i})^T(\Sigma^Z_{-i,-i})^{-1}\boldsymbol{Z}_{-i}$ and $\sigma_i^2 = \Sigma^Z_{i,i} - (\Sigma^Z_{i,-i})^T\Sigma^Z_{-i,-i}\Sigma^Z_{i,-i}$. Here, $\Sigma_{-i,-i}$ (of dimension $(n-1)\times(n-1)$) denotes the sub-matrix of Σ by eliminating the ith row and ith column; $\Sigma_{i,-i}$ (of dimension $(n-1)\times 1$) denotes the sub-vector of the ith column of Σ by eliminating its ith element.

Scheme for sampling $\boldsymbol{\Theta}_2$: The full conditional for $\boldsymbol{\theta}_{ij}$ is as follows:

$$\pi(\boldsymbol{\theta}_{ij} \mid \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{\theta}_{-i-j}, \boldsymbol{\tau}^2, \rho, \boldsymbol{C}, \boldsymbol{\gamma}) \;\propto\; \mathrm{N}(\mu_\theta, \Sigma_\theta)\cdot Q_\theta, \tag{2.20}$$

with $\boldsymbol{\theta}_{-i-j} = (\boldsymbol{\theta}_{ij}, i = 1,\dots,i-1,i+1,\dots,n; j = 1,\dots,j-1,j+1,\dots,m)$,

$$Q_\theta \;=\; \exp\Big[\frac{\boldsymbol{Z}^T\cdot\big\{\boldsymbol{W} + (\boldsymbol{S}-\boldsymbol{W}\cdot\boldsymbol{S})\Sigma_\alpha(\boldsymbol{S}-\boldsymbol{W}\cdot\boldsymbol{S})^T\big\}\cdot\boldsymbol{Z}}{2}\Big]\mid\Sigma_t\mid^{\frac{1}{2}}\mid\Sigma_\alpha\mid^{\frac{1}{2}},$$

$$\begin{aligned}\Sigma_\theta &= \Big[(\frac{q_{ij}}{\tau_1^2}+\frac{1}{\tau_2^2}-\frac{\tau_3^2}{n\tau_3^2\tau_2^2+\tau_2^4})\cdot H^{-1}(\rho)\\ &\qquad -q_{ij}(\tau_1^2H(\rho))^{-1}\{(\sigma^{-2}I_L+(\tau_1^2H(\rho))^{-1})^{-1}\}(\tau_1^2H(\rho))^{-1}\Big]^{-1},\end{aligned}$$

$$\begin{aligned}\mu_\theta &= \Sigma_1\cdot\left[H(\rho)^{-1}\sum_{w\neq i}\frac{\tau_3^2}{n\tau_3^2\tau_2^2+\tau_2^4}\boldsymbol{\theta}_{wj}\right.\\ &\qquad \left.+(\tau_1^2H(\rho))^{-1}\left\{(\sigma^{-2}I_L+(\tau_1^2H(\rho))^{-1})^{-1}\sigma^{-2}\sum_{k=1}^{q_{ij}}\boldsymbol{X}_{ijk}\right\}\right].\end{aligned}$$

Based on (2.20), we update θ_{ij} as follows:

1. Draw $\boldsymbol{\theta}_{ij}^p \sim \mathrm{N}(\mu_\theta,\Sigma_\theta)$.
2. Set $\boldsymbol{\theta}_{ij}=\boldsymbol{\theta}_{ij}^p$ with probability $(1-C_j)+C_j\min\{1,R_{\boldsymbol{\theta}}\}$, where

$$R_{\boldsymbol{\theta}} = \frac{\pi(\boldsymbol{\theta}_{ij}^p \mid \boldsymbol{X},\boldsymbol{S},\boldsymbol{\theta}_{-i-j},\boldsymbol{Z},\boldsymbol{\tau}^2,\rho,\boldsymbol{C},\boldsymbol{\gamma})}{\pi(\boldsymbol{\theta}_{ij} \mid \boldsymbol{X},\boldsymbol{S},\boldsymbol{\theta}_{-i-j},\boldsymbol{Z},\boldsymbol{\tau}^2,\rho,\boldsymbol{C},\boldsymbol{\gamma})}.$$

Scheme for sampling $\boldsymbol{u}$ and $\boldsymbol{v}$: As stated in the paper, the conditional uniform distribution for each u_{st} is

$$\pi(u_{st}\mid \boldsymbol{C},\eta)=\exp(-\eta I(C_s=C_t))\cdot I(0\le u_{st}\le \exp(\eta I(C_s=C_t))),$$

where functional predictor pair (s,t) satisfies $r_{st}=1$, corresponding to each edge in $\mathcal{G}$. Similarly, for each v_{jl} with $j=1,\dots,m; l=1,\dots,L-2$, we have the conditional density function

$$\pi(v_{jl}\mid\boldsymbol{\gamma}_j)=\exp(-\xi I(\gamma_{jl}=\gamma_{j(l+1)}))\cdot I(0\le v_{jl}\le\exp(\xi I(\gamma_{jl}=\gamma_{j(l+1)}))).$$

Scheme for sampling $\boldsymbol{C}$ and $\boldsymbol{\gamma}$:

As stated in the paper, one approach is to jointly update $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ by drawing from the full conditional

$$\begin{aligned}
&\pi(\boldsymbol{C}_{\digamma_g} = k\mathbf{1}_{l_g}, \{\boldsymbol{\gamma}_{j\psi_{jh}} = k_{j\psi_{jh}}\mathbf{1}_{l_{jh}}, j \in \digamma_g, h = 1, \dots, H_j\} \mid \boldsymbol{C}_{-\digamma_g}, \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{\theta}_{\mathcal{S}}, \boldsymbol{\tau}^2, \rho, \boldsymbol{u}, \boldsymbol{v}) \\
&\qquad \propto \prod_{j \in \digamma_g} I(\max_l \gamma_{jl} = C_j)|\Sigma_t|^{\frac{1}{2}}|\Sigma_\alpha|^{\frac{1}{2}} \exp\Big[\frac{1}{2}\boldsymbol{Z}^T\big\{\boldsymbol{W} \\
&\qquad\qquad + (\boldsymbol{S} - \boldsymbol{S}\boldsymbol{W})\Sigma_\alpha(\boldsymbol{S} - \boldsymbol{S}\boldsymbol{W})^T\big\}\boldsymbol{Z}\Big]|\tau_4^2 H^*(\rho)|^{-\frac{1}{2}}, \qquad (2.21)
\end{aligned}$$

where $k_{j\psi_{jh}}$ takes value 0 or 1, and $\boldsymbol{\gamma}_{j\psi_{jh}} = (\gamma_{jl}, l \in \psi_{jh})$. Based on (2.21), one can update each element $(\boldsymbol{C}_{\digamma_g}, \boldsymbol{\gamma}_{\digamma_g})$ individually $(g = 1, \dots, G)$.

The alternative approach is to sample proposal $(\boldsymbol{C}^*, \boldsymbol{\gamma}^*)$ through $g(\boldsymbol{C}^*, \boldsymbol{\gamma}^* \mid \boldsymbol{C}^{(o)}, \boldsymbol{\gamma}^{(o)}, \boldsymbol{\mathcal{P}}, \boldsymbol{S}, \boldsymbol{R})$ with the full conditional $\pi(\boldsymbol{\gamma}^* \mid \boldsymbol{C}^*, \boldsymbol{\mathcal{P}}, \boldsymbol{S})$ depending on (2.21). Given the proposed values $(\boldsymbol{C}^*, \boldsymbol{\gamma}^*)$, the MH acceptance ratio can be calculated as follows:

$$\begin{aligned}
R(\boldsymbol{C}^*, \boldsymbol{\gamma}^* \mid \boldsymbol{\gamma}^{(o)}, \boldsymbol{C}^{(o)}) &= \frac{\pi(\boldsymbol{C}^*, \boldsymbol{\gamma}^* \mid \boldsymbol{\mathcal{P}}, \boldsymbol{S}, \boldsymbol{R})}{\pi(\boldsymbol{C}^{(o)}, \boldsymbol{\gamma}^{(o)} \mid \boldsymbol{\mathcal{P}}, \boldsymbol{S}, \boldsymbol{R})} \cdot \frac{g(\boldsymbol{C}^{(o)}, \boldsymbol{\gamma}^{(o)} \mid \boldsymbol{C}^*, \boldsymbol{\gamma}^*, \boldsymbol{\mathcal{P}}, \boldsymbol{S}, \boldsymbol{R})}{g(\boldsymbol{C}^*, \boldsymbol{\gamma}^* \mid \boldsymbol{C}^{(o)}, \boldsymbol{\gamma}^{(o)}, \boldsymbol{\mathcal{P}}, \boldsymbol{S}, \boldsymbol{R})}, \\
&= \frac{\pi(\boldsymbol{C}^* \mid \boldsymbol{S}, \boldsymbol{R}, \boldsymbol{\mathcal{P}})}{\pi(\boldsymbol{C}^{(o)} \mid \boldsymbol{S}, \boldsymbol{R}, \boldsymbol{\mathcal{P}})} \cdot \frac{p(\boldsymbol{C}^{(o)} \mid \boldsymbol{S}, \boldsymbol{R})}{p(\boldsymbol{C}^* \mid \boldsymbol{S}, \boldsymbol{R})}. \qquad (2.22)
\end{aligned}$$

In (2.22), the second ratio in the right-hand side is calculated directly from the posterior sample under the SBPM; for the first ratio

$$\pi(\boldsymbol{C}^* \mid \boldsymbol{S}, \boldsymbol{R}, \boldsymbol{\mathcal{P}}) = \int \pi(\boldsymbol{C}^*, \boldsymbol{\gamma} \mid \boldsymbol{S}, \boldsymbol{R}, \boldsymbol{\mathcal{P}}) d\boldsymbol{\gamma}. \qquad (2.23)$$

Integral (2.23) is computed using importance sampling with a given instrumental distribution and the sample size n_γ (e.g., $\gamma_{jl} \overset{i.i.d.}{\sim}$ Bernoulli(0.5), $n_\gamma = 10000$).

Scheme for sampling η : The full conditional for η is as follows:

$$\pi(\eta|\boldsymbol{C},\boldsymbol{R}) \;\propto\; \frac{\exp(\eta\sum_j\sum_{k:r_{kj}=1} I[C_j = C_k])}{\int\exp(\eta\sum_j\sum_{k:r_{kj}=1} I[C_j = C_k])d\boldsymbol{C}} I[0 < \eta < U_\eta]. \tag{2.24}$$

Based on (2.24), we update η by sampling a proposal $\eta^p \sim \mathrm{N}(\eta, \sigma^2_\eta)$ and setting $\eta = \eta^p$ with probability $\min\{1, R_\eta\}$, where $R_\eta = \frac{\pi(\eta^P|\boldsymbol{C},\boldsymbol{R})}{\pi(\eta|\boldsymbol{C},\boldsymbol{R})}$.

Scheme for sampling σ^2 : The full conditional for σ^2 is as follows:

$$\begin{aligned}
&\pi(\sigma^2|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\tau_1^2)\\
&\quad\propto \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q_{ij}}\frac{1}{2}\left(\frac{H(\rho)^{-1}}{\tau_1^2}\boldsymbol{\theta}_{ij} + \sigma^{-2}\boldsymbol{X}_{ijk}\right)^T\left(\frac{H(\rho)^{-1}}{\tau_1^2} + \sigma^{-2}I_L\right)^{-1}\Big(\frac{H(\rho)^{-1}}{\tau_1^2}\boldsymbol{\theta}_{ij}\right.\\
&\quad\left. + \sigma^{-2}\boldsymbol{X}_{ijk}\Big)\right\}\sigma^{-\sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij}L - 2\alpha_1 - 2}\exp\left(-\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q_{ij}}\boldsymbol{X}_{ijk}^T\boldsymbol{X}_{ijk}^T + 2\gamma_1}{2\sigma^2}\right)\\
&\quad\cdot\left|\frac{H(\rho)^{-1}}{\tau_1^2} + \sigma^{-2}I_L\right|^{-\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij}}{2}} I[\sigma^2 > 0].
\end{aligned} \tag{2.25}$$

Based on (2.25), we update σ^2 by sampling a proposal $\sigma^{2(p)} \sim \mathrm{N}(\sigma^2, \sigma^2_\sigma)$ and setting $\sigma^2 = \sigma^{2(p)}$ with probability $\min\{1, R_\sigma\}$, where $R_\sigma = \frac{\pi(\sigma^{2(p)}|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\tau_1^2)}{\pi(\sigma^2|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\tau_1^2)}$.

Scheme for sampling τ_1^2 : The full conditional for τ_1^2 is as follows:

$$\begin{aligned}
&\pi(\tau_1^2|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\sigma^2)\\
&\quad\propto \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q_{ij}}\frac{1}{2}\left(\frac{H(\rho)^{-1}}{\tau_1^2}\boldsymbol{\theta}_{ij} + \sigma^{-2}\boldsymbol{X}_{ijk}\right)^T\left(\frac{H(\rho)^{-1}}{\tau_1^2} + \sigma^{-2}I_L\right)^{-1}\Big(\frac{H(\rho)^{-1}}{\tau_1^2}\boldsymbol{\theta}_{ij}\right.\\
&\quad\left. + \sigma^{-2}\boldsymbol{X}_{ijk}\Big) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{q_{ij}}{\tau_1^2}\boldsymbol{\theta}_{ij}^T H(\rho)^{-1}\boldsymbol{\theta}_{ij} - \frac{\gamma_2}{\tau_1^2}\right\}\cdot\tau_1^{-\sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij}\cdot L - 2\alpha_2 - 2}\\
&\quad\cdot\left|\frac{H(\rho)^{-1}}{\tau_1^2} + \sigma^{-2}I_L\right|^{-\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij}}{2}} I[\tau_1^2 > 0].
\end{aligned} \tag{2.26}$$

Based on (2.26), we update τ_1^2 by sampling a proposal ${\tau_1}^{2(p)} \sim \mathrm{N}(\tau_1^2, \sigma_{\tau_1}^2)$ and setting $\tau_1^2 = {\tau_1}^{2(p)}$ with probability $\min\{1, R_{\tau_1}\}$, where $R_{\tau_1} = \frac{\pi({\tau_1}^{2(p)}|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\sigma^2)}{\pi(\tau_1^2|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\sigma^2)}$.

Scheme for sampling τ_2^2 : The full conditional for τ_2^2 is as follows:

$$\begin{aligned}
&\pi(\tau_2^2|\boldsymbol{X}, \boldsymbol{\Theta}_2, \rho, \tau_3^2) \\
&\quad \propto \cdot \exp\Bigg[-\frac{1}{2}\sum_{j=1}^{m}\Bigg\{ \sum_{i=1}^{n} \boldsymbol{\theta}_{ij}^T(\tau_2^2 H(\rho))^{-1}\boldsymbol{\theta}_{ij} - \frac{\tau_3^2}{\tau_2^4 + n\tau_3^2\tau_2^2}\left(\sum_{i=1}^{n}\boldsymbol{\theta}_{ij}\right)^T H(\rho)^{-1} \\
&\left(\sum_{i=1}^{n}\boldsymbol{\theta}_{ij}\right)\Bigg\} - \frac{\gamma_2}{\tau_2^2}\Bigg] \cdot {\tau_2}^{-m(n-1)L-2\alpha_2-2} \cdot (\tau_2^2 + n\tau_3^2)^{-\frac{mL}{2}} I[\tau_2^2 > 0] \qquad (2.27)
\end{aligned}$$

Based on (2.27), we update τ_2^2 by sampling a proposal ${\tau_2}^{2(p)} \sim \mathrm{N}(\tau_2^2, \sigma_{\tau_2}^2)$ and setting $\tau_2^2 = {\tau_2}^{2(p)}$ with probability $\min\{1, R_{\tau_2}\}$, where $R_{\tau_2} = \frac{\pi({\tau_2}^{2(p)}|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\tau_3^2)}{\pi(\tau_2^2|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\tau_3^2)}$.

Scheme for sampling τ_3^2 : The full conditional for τ_3^2 is as follows:

$$\begin{aligned}
&\pi(\tau_3^2|\boldsymbol{X}, \boldsymbol{\Theta}_2, \rho, \tau_2^2) \\
&\quad \propto \cdot \exp\left[\sum_{j=1}^{m}\left\{\frac{\tau_3^2}{2(\tau_2^4 + n\tau_3^2\tau_2^2)}(\sum_{i=1}^{n}\boldsymbol{\theta}_{ij})^T H(\rho)^{-1}(\sum_{i=1}^{n}\boldsymbol{\theta}_{ij})\right\} - \frac{\gamma_2}{\tau_3^2}\right] \\
&\quad \cdot {\tau_3}^{-2\alpha_2-2} \cdot (\tau_2^2 + n\tau_3^2)^{-\frac{mL}{2}} I[\tau_3^2 > 0] \qquad (2.28)
\end{aligned}$$

Based on (2.28), we update τ_3^2 by sampling a proposal ${\tau_3}^{2(p)} \sim \mathrm{N}(\tau_3^2, \sigma_{\tau_3}^2)$ and setting $\tau_3^2 = {\tau_3}^{2(p)}$ with probability $\min\{1, R_{\tau_3}\}$, where $R_{\tau_3} = \frac{\pi({\tau_3}^{2(p)}|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\tau_2^2)}{\pi(\tau_3^2|\boldsymbol{X},\boldsymbol{\Theta}_2,\rho,\tau_2^2)}$.

Scheme for sampling τ_4^2 : The full conditional for τ_4^2 is given by $\pi(\tau_4^2 \mid \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{\Theta}_2, \rho, \boldsymbol{C}, \boldsymbol{\gamma}) \propto Q_\theta \cdot \exp(\frac{\gamma_2}{\tau_4^2}) \cdot (\tau_4^2)^{-\alpha_2-1-\frac{\sum_{j\in\mathcal{S}} d_j}{2}} I[\tau_4^2 > 0]$. We update τ_4^2 by sampling a proposal ${\tau_4}^{2(p)} \sim \mathrm{N}(\tau_4^2, \sigma_{\tau_4}^2)$ and setting $\tau_4^2 = {\tau_4}^{2(p)}$ with probability $\min\{1, R_{{\tau_4}^{2(p)}}\}$, where $R_{{\tau_4}^{2(p)}} = \frac{\pi({\tau_4}^{2(p)}|\boldsymbol{Z},\boldsymbol{S},\boldsymbol{\Theta}_2,\rho,\boldsymbol{C},\boldsymbol{\gamma})}{\pi(\tau_4^2|\boldsymbol{Z},\boldsymbol{S},\boldsymbol{\Theta}_2,\rho,\boldsymbol{C},\boldsymbol{\gamma})}$.

Scheme for sampling ρ : The full conditional for ρ is given by

$$
\begin{aligned}
&\pi(\rho \mid \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{\Theta}_2, \boldsymbol{\tau}^2, \boldsymbol{C}, \boldsymbol{\gamma}) \\
&\propto \quad Q_\theta \cdot \exp\Bigg[-\frac{1}{2}\sum_{j=1}^{m}\Bigg\{ \sum_{i=1}^{n} \frac{\tau_1^2 + q_{ij}\tau_2^2}{(\tau_2\tau_1)^2}\boldsymbol{\theta}_{ij}^T H(\rho)^{-1}\boldsymbol{\theta}_{ij} - \frac{\tau_3^2}{\tau_2^4 + n(\tau_3\tau_2)^2}(\sum_{i=1}^{n}\boldsymbol{\theta}_{ij})^T H(\rho)^{-1} \\
&\qquad (\sum_{i=1}^{n}\boldsymbol{\theta}_{ij})\Bigg\}\Bigg] \cdot \exp\Bigg\{ \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q_{ij}} \frac{1}{2}\left(\frac{H(\rho)^{-1}}{\tau_1^2}\boldsymbol{\theta}_{ij} + \sigma^{-2}\boldsymbol{X}_{ijk}\right)^T \left(\frac{H(\rho)^{-1}}{\tau_1^2} + \sigma^{-2}I_L\right)^{-1} \Bigg(\frac{H(\rho)^{-1}}{\tau_1^2}\boldsymbol{\theta}_{ij} \\
&\qquad +\sigma^{-2}\boldsymbol{X}_{ijk}\Bigg)\Bigg\} I(0 < \rho < U_\rho)|H(\rho)|^{-\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(q_{ij}+1)}{2}} \cdot \left|\frac{H(\rho)^{-1}}{\tau_1^2} + \sigma^{-2}I_L\right|^{-\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}q_{ij}}{2}} |H^*(\rho)|^{-\frac{1}{2}}.
\end{aligned}
$$

We update ρ by sampling a proposal $\rho^p \sim \mathrm{N}(\rho, \sigma_\rho^2)$ and setting $\rho = \rho^p$ with probability $\min\{1, R_{\rho^p}\}$, where $R_{\rho^p} = \frac{\pi(\rho^p | \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{\Theta}_2, \boldsymbol{\tau}^2, \boldsymbol{C}, \boldsymbol{\gamma})}{\pi(\rho | \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{\Theta}_2, \boldsymbol{\tau}^2, \boldsymbol{C}, \boldsymbol{\gamma})}$.

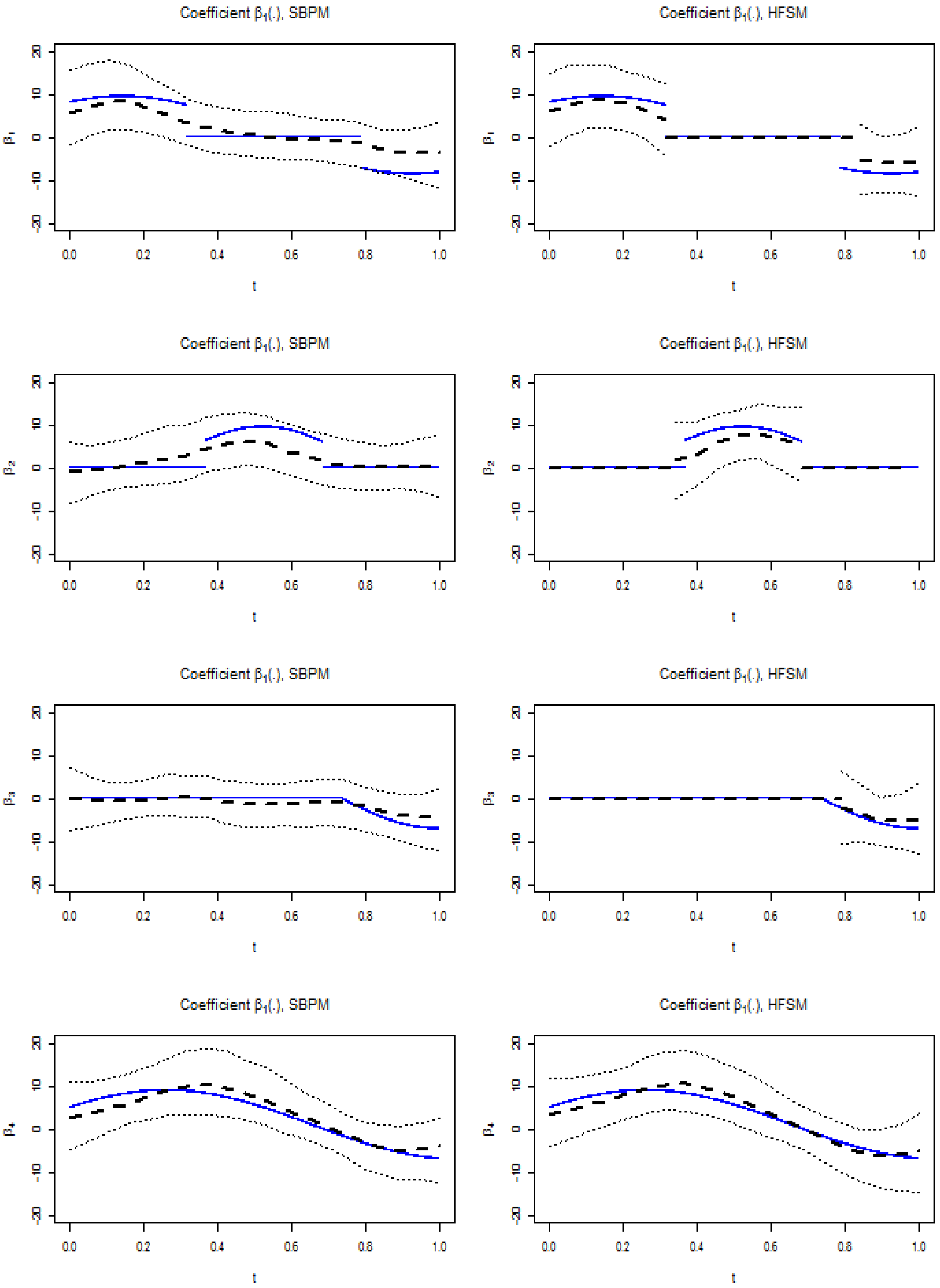


Figure 2.3: Posterior inference of the functional coefficients under the mode of $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ in the case of $\sigma^2 = 0.1$ (n=50). The left panel is for the case under the SBPM and the right panel is for the case under the HFSM ($\xi = 1.5$). The solid blue lines represent the true functional coefficients; the dashed lines represent the posterior means; and the dotted lines represent the corresponding 95% credible intervals.

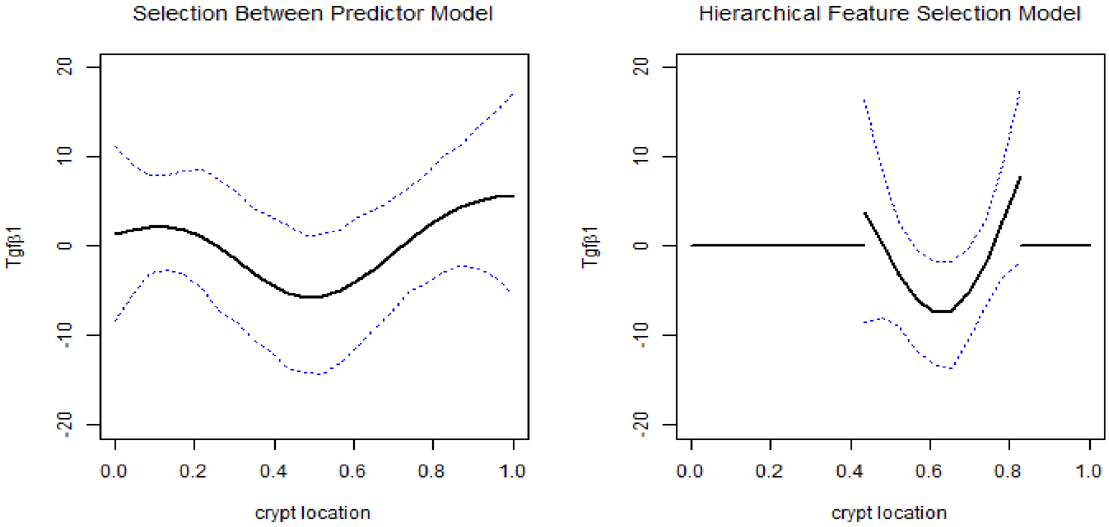


Figure 2.4: Posterior mean (solid line) and 95% credible interval (dotted lines) of the functional biomarker TGFβ_1 under the SBPM (left panel) and the HFSM (right panel).

Chapter 3

Bayesian Spatial Variable Selection for Ultra-High Dimensional Neuroimaging Data: A Multiresolution Approach

3.1 Introduction

3.1.1 Variable Selection in Ultra-high dimensionality

As discussed in Section 1, variable selection is a widely encountered issue in biomedical studies to facilitate comprehensive statistical learning and biological discovery. Regularization methods have been developed to conduct variable selection and extended to handle high-dimensional feature spaces. Alternatively, Bayesian methods also play a prominent role in solving the variable selection problem.

Although the aforementioned methods have been successful for variable selection in relatively high-dimensional feature space (e.g., the number of predictors is on the order of thousands), these methods become infeasible due to their prohibitive computational costs when faced with a problem such as our motivating study involving hundreds of thousands or even millions of predictors. This has stimulated the development of variable selection techniques for ultra-high dimensional problems. Fan and Lv (2008) proposed the Sure Independence Screening (SIS) approach often used in conjunction with regularization methods. This method does not require intensive computations and has good theoretical properties. Although it is applicable to a probit regression model, the SIS does not explicitly model the dependence among variables and cannot assess the uncertainty of variable selection. In a Bayesian modeling framework, Bottolo and Richardson (2010) developed a powerful sampling scheme to accommodate the high-dimensional multimodal model space based on the evolutionary Monte Carlo. This method has been shown to be able to handle up to 10,000 predictors, but it is still computationally inefficient when applied to our motivating study with almost 200,000 predictors. More recently, by assigning nonlocal priors to model parameters, Johnson and Rossell (2012) proposed a novel Bayesian model selection method that possesses the posterior selection consistency when the number of predictors is smaller than the sample size. Johnson (2013) demonstrated that it can achieve high selection accuracy in ultra high-dimensional problems, comparable to the SIS combined with regularization methods. However, their method is not directly applicable to our problem in that it was developed for a linear regression model without incorporating any structural information in the covariate space. Goldsmith et al. (2012) and Huang et al. (2013) developed a single-site Gibbs sampler for Bayesian spatial variable selection using Ising priors with application to neuroimaging studies. This algorithm is able to fit linear regression models with ultra-high dimensional imaging biomarkers, i.e. "scalar-on-image regression" models, however, the single-site updating scheme leads to a very slow mixing of the Markov chain in the posterior computation for a probit regression model (Lamnisos et al., 2009, 2012). Thus, there are needs for developing more efficient posterior computation algorithms that can be applied to our motivating problem. Particularly, we resort to a multiresolution approach.

3.1.2 Multiresolution Approach

The idea of multiresolution, which facilitates the information transition through a construction of coarse-and-fine-scale model parameters, has been adopted to optimize algorithms successfully in data mining and machine learning. The pioneer work of utilizing the multiresolution idea for Bayesian computation traces back to a multi-grid MCMC method proposed by Liu and Sabatti (2000). This approach was originally adopted by Goodman and Sokal (1989) to solve a problem in statistical physics. Motivated by image denoising problems, Higdon et al. (2002) proposed a coupled MCMC algorithm with the coarsened-scale Markov chains serving as a guide to the original fine-scale chains. The coupled Markov chains can better explore the entire sample space and avoid getting trapped at local maxima of the posterior distribution. Holloman et al. (2006) further proposed a multiresolution genetic algorithm to reduce computational burden, provide more accurate solution of maximization problem, and improve mixing of the MCMC sampling. In a similar fashion, Koutsourelakis (2009) adopted a multiresolution idea to estimate spatially-varying parameters in PDE-based models with the salient features detected by the coarse solvers. From the computational perspective, Giles (2008) showed that the computational complexity for estimating the expected value from a stochastic differential equation could be reduced by a multiresolution Monte Carlo simulation. More recently, Kou et al. (2012) applied a multiresolution method to diffusion process models for discrete data and showed that their approach improves computational efficiency and estimation accuracy. From the perspective of model construction, Fox and Dunson (2012) adopted the multiresolution idea in Gaussian process models to capture both long-range dependencies and abrupt discontinuities.

In this chapter, we develop efficient multiresolution MCMC algorithms for variable selection in the ultra-high dimensional image feature space. In contrast to the coupled Markov chain methods (Higdon et al., 2002; Holloman et al., 2006; Kou et al., 2012) that alternate between different resolutions in posterior simulation, we construct and conduct posterior computations for a sequence of nested auxiliary models for variable selection from the coarsest scale to the finest scale. Our goal is to conduct variable selection at the finest scale – the resolution in the observed data. The MCMC algorithm for the model at each resolution depends on the posterior inclusion probabilities obtained from fitting the auxiliary model at the previous, resolution through the use of a "smart" proposal distribution that allows the algorithm to explore the entire sample space more efficiently. This avoids the complication of alternating between resolutions for a large number of selection indicators in our problem.

3.2 Model Formulation

3.2.1 A Probit Regression Model for Variable Selection

Suppose there are n subjects in the data. For $i = 1, \ldots, n$, let $y_i \in \{0, 1\}$ be the binary outcome representing the disease status of subject i (disease = 1, control = 0). Assume that the whole brain $\mathcal{B}$ consists of R regions and region r contains V_r voxels, for $r = 1, \ldots, R$, with $V = \sum_{r=1}^{R} V_r$ representing the total number of voxels in the brain. Let x_{irv} denote the imaging biomarker at voxel v within region r for subject i and s_{ij} denote clinical variable j for subject i ($j = 1, \ldots, p$). We consider a probit regression model for variable selection

$$y_i = I[z_i \geq 0], \qquad z_i = \alpha_0 + \sum_{j=1}^{p} \alpha_j s_{ij} + \sum_{r=1}^{R} c_r \sum_{v=1}^{V_r} \gamma_{rv} \beta_{rv} x_{irv} + \epsilon_i, \quad \text{and} \quad \epsilon_i \sim \mathrm{N}(0, 1) \tag{3.1}$$

where indicator function $I(\mathcal{A}) = 1$ if event $\mathcal{A}$ occurs and 0 otherwise, α_j and β_{rv} are coefficients of clinical variable s_{ij} and imaging biomarker x_{irv}, respectively, $c_r \in \{0, 1\}$ is the selection indicator for region r, and $\gamma_{rv} \in \{0, 1\}$ is the selection indicator for voxel v in region r. Thus, the imaging biomarker x_{irv} is excluded from the model if at least one of c_r and γ_{rv} is zero.

We further denote by $\mathbf{e}_{mk} = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ an $m \times 1$ vector with the kth element of 1 and all other elements of 0, by $\mathbf{0}_m = (0, \ldots, 0)^\top$ an all-zero vector of dimension $m \times 1$, by $\mathbf{1}_m = \sum_{k=1}^{m} \mathbf{e}_{mk}$ an all-one vector, and by $\mathbf{I}_m = \sum_{k=1}^{m} \mathbf{e}_{mk}\mathbf{e}_{mk}^\top$ an $m \times m$ identity matrix. Define $\mathbf{M}_r = (\mathbf{0}_{\underline{V}_r}^\top, \mathbf{1}_{V_r}^\top, \mathbf{0}_{\overline{V}_r}^\top)^\top$ (of dimension $V \times 1$) and $\mathbf{M} = (\mathbf{M}_1, \cdots, \mathbf{M}_R)$ (of dimension $V \times R$), where $\underline{V}_r = \sum_{r'=1}^{r-1} V_{r'}$ and $\overline{V}_r = \sum_{r'=r+1}^{R} V_{r'}$. It follows that $\mathbf{M}$ represents an index map between voxel and region and model (3.1) can be rewritten in a compact form,

$$\mathbf{y} = I[\mathbf{z} \geq \mathbf{0}_n], \quad \mathbf{z} = \mathbf{S}\boldsymbol{\alpha} + \mathbf{X}\left(\boldsymbol{\lambda} \circ \boldsymbol{\beta}\right) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}_n, \mathbf{I}_n) \tag{3.2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{z} = (z_1, \ldots, z_n)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, $\mathbf{x}_{rv} = (x_{1rv}, \ldots, x_{nrv})^\top$, $\mathbf{X}_r = (\mathbf{x}_{r1}, \ldots, \mathbf{x}_{rV_r})$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_R)$, $\mathbf{s}_j = (s_{1j}, \ldots, s_{nj})^\top$, $\mathbf{S} = (\mathbf{1}_n, \mathbf{s}_1, \ldots, \mathbf{s}_p)$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)^\top$, $\boldsymbol{\beta}_r = (\beta_{r1}, \ldots, \beta_{rV_r})^\top$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_R^\top)^\top$, $\mathbf{c} = (c_1, \ldots, c_R)^\top$, $\boldsymbol{\gamma}_r = (\gamma_{r1}, \ldots, \gamma_{rV_r})^\top$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_R^\top)^\top$, and $\boldsymbol{\lambda} = (\mathbf{Mc}) \circ \boldsymbol{\gamma}$ with "$\circ$" representing the Hadamard product (or entry-wise product) (Styan, 1973). It is worth noting that $\boldsymbol{\lambda}$, the V dimensional binary vector, defines the set of important voxels.

3.2.2 Prior Specifications

We assign the Gaussian priors to the regression coefficients in model (3.2),

$$\boldsymbol{\alpha} \sim \mathrm{N}(\mathbf{0}_{p+1}, \sigma_\alpha^2 \mathbf{I}_{p+1}) \quad \text{and} \quad \boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}_V, \sigma_\beta^2 \mathbf{I}_V), \tag{3.3}$$

where σ_α^2 and σ_β^2 are the prior variances of the regression coefficients. Given a network configuration matrix $\mathbf{W} = \{w_{ij}\}$ for a multivariate binary random variable $\mathbf{d} = (d_1, \ldots, d_m)^\top \in \{0,1\}^m$, we denote by $\mathbf{d} \sim \mathrm{Ising}(a, b, \mathbf{W})$ an Ising distribution with a sparse parameter a and a smooth parameter b and the probability mass function of $\mathbf{d}$ is proportional to $\exp\left(a \sum_{i=1}^m I[d_i = 0] + b \sum_{i=1}^m \sum_{j=1}^m w_{ij} I[d_i = d_j]\right)$. The prior specifications for $\mathbf{c}$ and $\boldsymbol{\gamma}_r$ are

$$\mathbf{c} \sim \mathrm{Ising}(\eta_1, \xi_1, \mathbf{F}) \quad \text{and} \quad \boldsymbol{\gamma}_r \overset{\text{iid}}{\sim} \mathrm{Ising}(\eta_2, \xi_2, \mathbf{L}_r), \text{ for } r = 1, \ldots, R, \tag{3.4}$$

where $\mathbf{F} = \{f_{r'r}\}$ with $f_{r'r} \in \mathbb{R}$ representing the population-level functional connectivity between region r' and region r and $\mathbf{L}_r = \{l_{rv'v}\}$ with $l_{rv'v} \in \{0,1\}$ indicating whether voxels v' and v are neighbors in region r. In our case, $\mathbf{F}$ can be estimated separately from the R-fMRI time series or obtained from existing literature. For the hyper-prior specifications in (3.3) and (3.4), we have

$$\sigma_\beta^2 \sim \mathrm{IG}(a_\beta, b_\beta), \quad \eta_k \sim \mathrm{U}(a_\eta, b_\eta), \text{ and } \xi_k \sim \mathrm{U}(a_\xi, b_\xi), \text{ for } k = 1, 2, \tag{3.5}$$

where $\mathrm{IG}(a, b)$ denotes an inverse gamma distribution with shape a and rate b, and $\mathrm{U}(a, b)$ represents a uniform distribution on region $[a, b]$.

3.2.3 Standard Posterior Computation

In a standard MCMC algorithm for posterior computation of models (3.2)–(3.5), each parameter in $\mathbf{z}, \mathbf{c}, \boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and σ_β^2 can be directly sampled from its full conditional. The sparse and smooth parameters in the Ising priors, η_k and ξ_k for $k = 1, 2$, can be updated using the auxiliary variable method by Møller et al. (2006). The details of the MCMC algorithm are provided in Appendix 3.7.1.

In the case of high or ultra-high dimensional data, we suggest a block update of $\boldsymbol{\beta}$. The full conditional of $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\alpha}, \mathbf{c}, \boldsymbol{\gamma}, \sigma_\beta^2) \propto \prod_{r=1}^{R} \prod_{v=1}^{V_r} \phi\left(\beta_{rv}/\sigma_\beta\right) \exp\left\{-\frac{1}{2} \left\|\mathbf{z} - \mathbf{S}\boldsymbol{\alpha} - \mathbf{X}\left\{\boldsymbol{\lambda} \circ \boldsymbol{\beta}\right\}\right\|^2\right\}, \tag{3.6}$$

where $\phi(\cdot)$ is the standard normal density function and $\|\cdot\|$ denotes the Euclidean vector norm. Given $\boldsymbol{\lambda}$, the block update entails drawing $\boldsymbol{\beta}_1$ (the coefficients corresponding to the selected predictors with $\lambda = 1$) and $\boldsymbol{\beta}_0$ (the coefficients corresponding to the unselected predictors with $\lambda = 0$) separately from

$$\boldsymbol{\beta}_1 \;\sim\; \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}_1}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}\right) \quad \text{and} \quad \boldsymbol{\beta}_0 \sim \mathrm{N}\left(\mathbf{0}_{m_0}, \sigma_\beta^2 \mathbf{I}_{m_0}\right), \tag{3.7}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1} = \left(\sigma_\beta^{-2}\mathbf{I}_{m_1} + \mathbf{X}_{\boldsymbol{\lambda}}^\top \mathbf{X}_{\boldsymbol{\lambda}}\right)^{-1}$, $\boldsymbol{\mu}_{\boldsymbol{\beta}_1} = \boldsymbol{\Sigma}_\beta \mathbf{X}_{\boldsymbol{\lambda}}^\top (\mathbf{z} - \mathbf{S}\boldsymbol{\alpha})$, $m_1 = \|\boldsymbol{\lambda}\|^2$, $m_0 = V - m_1$, and $\mathbf{X}_{\boldsymbol{\lambda}}$ includes the columns of $\mathbf{X}$ corresponding to the important voxels defined by $\boldsymbol{\lambda}$. The computational complexity of computing $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1}$ is $O(nm_1^2)$. Also, we implement the rank-one downrate algorithm of Cholesky decomposition on sampling $\boldsymbol{\beta}_1$ with an order of $O(nm_1^2)$. While m_1 changes from one MCMC iteration to another, the posterior samples of m_1 are likely concentrated on values substantially smaller than V when the true model is sparese, i.e., the number of important voxels is small.

Compared with the single-site Gibbs sampling approach (Goldsmith et al., 2012; Huang et al., 2013), the block update of $\boldsymbol{\beta}$ reduces the computational costs and improves Markov chains mixing and hence is more appealing for high-dimensional problems where the number of predictors is on the order of thousands. However, for ultra-high dimensional problems such as imaging data in a standard brain space with around 200,000 voxels, this algorithm is still very inefficient. To address this challenge, we propose a novel multiresolution posterior computation approach.

3.3 Multiresolution Approach

The basic steps of our multiresolution approach include first carefully constructing a sequence of partitions of brain regions from the pre-defined coarsest scale to the finest scale – the resolution in the observed data – and subsequently defining and fitting a sequence of auxiliary models for variable selection from the coarsest scale to the finest scale. The key idea is that the posterior samples on coarse scale variable selection are used to create a “smart” proposal for the MCMC posterior computation for the model in the next, finer scale resolution, allowing the MCMC algorithm to explore the entire sample space for model selection more efficiently.

3.3.1 Partition and Auxiliary Models

Suppose that we define K resolutions with resolution K being the target resolution in the observed data. At resolution k, the R brain regions are grouped into $G^{(k)}$ mutually exclusive partitions with $1 = G^{(0)} < G^{(1)} < G^{(2)} < \cdots < G^{(K)} = R$ and each partition is a collection of regions based on spatial contiguity or functional connectivity. The partitions at resolution k are nested within the partitions at resolution $k-1$. Let $\mathbf{B}^{(k)} = (b_{rg}^{(k)})$ be an $R \times G^{(k)}$ matrix with $b_{rg}^{(k)} \in \{0, 1\}$ indicating whether region r is located in partition g at resolution k, and $\widetilde{\mathbf{B}}^{(k)} = (\widetilde{b}_{gg'}^{(k)})$ be a $G^{(k)} \times G^{(k-1)}$ matrix with $\widetilde{b}_{gg'}^{(k)} \in \{0, 1\}$ indicating whether partition g at resolution k is located in partition g' at resolution $k-1$. We have $\mathbf{B}^{(k)}\mathbf{1}_{G^{(k)}} = \mathbf{1}_R$ due to mutually exclusive partitions at each resolution and $\mathbf{B}^{(k-1)} = \mathbf{B}^{(k)}\widetilde{\mathbf{B}}^{(k)}$ due to nested partitions across resolutions. In addition, $\mathbf{B}^{(K)} = \mathbf{I}_R$, $\widetilde{\mathbf{B}}^{(1)} = \mathbf{1}_{G^{(1)}}$, $\widetilde{\mathbf{B}}^{(k)}\mathbf{1}_{G^{(k-1)}} = \mathbf{1}_{G^{(k)}}$ and $\{\mathbf{B}^{(k)}\}_{k=1}^{K}$ is uniquely determined by $\{\widetilde{\mathbf{B}}^{(k)}\}_{k=1}^{K}$. Figure 3.1 provides a detailed illustration on the partitions of a two-dimensional rectangle area in one slice of brain at three resolutions. Of note, $\mathbf{B}^{(k)}$ defines the partitions at resolution k.

In a similar fashion, at resolution k, we divide region r with a total of V_r voxels into $H_r^{(k)}$ mutually exclusive subregions with $1 = H_r^{(0)} < H_r^{(1)} < H_r^{(2)} < \ldots < H_r^{(K)} = V_r$ and each subregion is a collection of contiguous voxels. The subregions in resolution k are nested within the subregions in resolution $k-1$. Let $\mathbf{A}_r^{(k)} = (a_{rvh}^{(k)})$ denote a $V_r \times H_r^{(k)}$ matrix with $a_{rvh}^{(k)} \in \{0, 1\}$ indicating whether voxel v is located in subregion h at resolution k and let $\widetilde{\mathbf{A}}_r^{(k)} = (\widetilde{a}_{rhh'}^{(k)})$ denote an $H_r^{(k)} \times H_r^{(k-1)}$ matrix with $\widetilde{a}_{rhh'}^{(k)} \in \{0, 1\}$ indicating whether subregion h at resolution k is located in subregion h' at resolution $k-1$. Similarly, we have $\mathbf{A}_r^{(k)}\mathbf{1}_{H_r^{(k)}} = \mathbf{1}_{V_r}$ due to mutually exclusive subregions at each resolution and $\mathbf{A}_r^{(k-1)} = \mathbf{A}_r^{(k)}\widetilde{\mathbf{A}}_r^{(k)}$ due to subregions nested across resolutions. In addition, $\mathbf{A}_r^{(K)} = \mathbf{I}_{V_r}$, $\widetilde{\mathbf{A}}_r^{(1)} = \mathbf{1}_{H_r^{(1)}}$, $\widetilde{\mathbf{A}}_r^{(k)}\mathbf{1}_{H_r^{(k-1)}} = \mathbf{1}_{H_r^{(k)}}$, and $\{\mathbf{A}_r^{(k)}\}_{k=1}^{K}$ is uniquely determined by $\{\widetilde{\mathbf{A}}_r^{(k)}\}_{k=1}^{K}$. It follows that $\mathbf{A}^{(k)} = \mathrm{diag}\{\mathbf{A}_1^{(k)}, \ldots, \mathbf{A}_R^{(k)}\}$ defines the subregions at resolution k.

Given the partitions defined by $\mathbf{B}^{(k)}$ and the subregions defined $\mathbf{A}^{(k)}$ which could be specified in light of the brain anotomy, we define an auxiliary probit model for variable selection at resolution k, denoted as $\mathcal{M}^{(k)}$, which is given by

$$\mathbf{y} = I[\mathbf{z}^{(k)} \geq \mathbf{0}_n], \quad \mathbf{z}^{(k)} = \mathbf{S}\boldsymbol{\alpha}^{(k)} + \mathbf{X}\left\{\boldsymbol{\lambda}^{(k)} \circ \boldsymbol{\beta}^{(k)}\right\} + \boldsymbol{\epsilon}^{(k)}, \tag{3.8}$$

where $\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\epsilon}^{(k)}$ have the same definitions and dimensions as $\mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ in the target model (3.2). The binary indictor vector $\boldsymbol{\lambda}^{(k)} = (\mathbf{M}\mathbf{B}^{(k)}\mathbf{c}^{(k)}) \circ (\mathbf{A}^{(k)}\boldsymbol{\gamma}^{(k)})$, where $\mathbf{c}^{(k)} = (c_g^{(k)})$ is of dimension $G^{(k)} \times 1$ with $c_g^{(k)}$ denoting the selection indicator for partition g; $\boldsymbol{\gamma}_r^{(k)} = (\gamma_{rh}^{(k)})$ is of dimension $H_r^{(k)} \times 1$ with $\gamma_{rh}^{(k)}$ representing the selection indicator for subregion h; and $\boldsymbol{\gamma}^{(k)} = (\boldsymbol{\gamma}_1^{(k)\top}, \ldots, \boldsymbol{\gamma}_R^{(k)\top})^\top$ is of dimension $H^{(k)} \times 1$ with $H^{(k)} = \sum_{r=1}^{R} H_r^{(k)}$. By this definition, $\mathcal{M}^{(K)}$ is equivalent to model (3.2) including the prior specifications Section 3.2.2. The main difference between $\mathcal{M}^{(k)}$ $(k < K)$ and $\mathcal{M}^{(K)}$ is that variable selection is conducted at the partition level and the subregion level in $\mathcal{M}^{(k)}$ as opposed to the region level and the voxel level in $\mathcal{M}^{(K)}$, reflected by the definitions of the selection indicators in $\mathcal{M}^{(k)}$, i.e. $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$.

The dimensions of $\mathbf{c}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ increase as the resolution k increases and eventually become equal to the dimensions of $\mathbf{c}$ and $\boldsymbol{\gamma}$ in the target model (3.2). In other words, the large number of latent indicators $\mathbf{c}$ and $\boldsymbol{\gamma}$ in the target model are replaced by a smaller number of latent indicators $\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}$ $(k < K)$ in the auxiliary model $\mathcal{M}^{(k)}$ particularly in the initial resolutions. In ultra-high dimensional problems, this dimension reduction can be very significant and is exploited in our proposed sampling schemes in Sections 3.3.2 and 3.3.3 to allow for efficient posterior computations for the sequence of auxiliary models $\mathcal{M}^{(k)}$ $(k < K)$ and the target model $\mathcal{M}^{(K)}$.

To complete the specification for auxiliary models $\mathcal{M}^{(k)}$ $(k < K)$, we assign the same priors to $\boldsymbol{\alpha}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ in (3.8) as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in (3.2) and denote by $\sigma_{\beta}^{2(k)}$ the prior variance for $\boldsymbol{\beta}^{(K-1)}$, and we assign i.i.d Bernoulli priors with a probability 0.5 to $\mathbf{c}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$. Under such prior specifications, it can be shown that the posterior inclusion probability for each voxel or region is always positive in each auxiliary model.

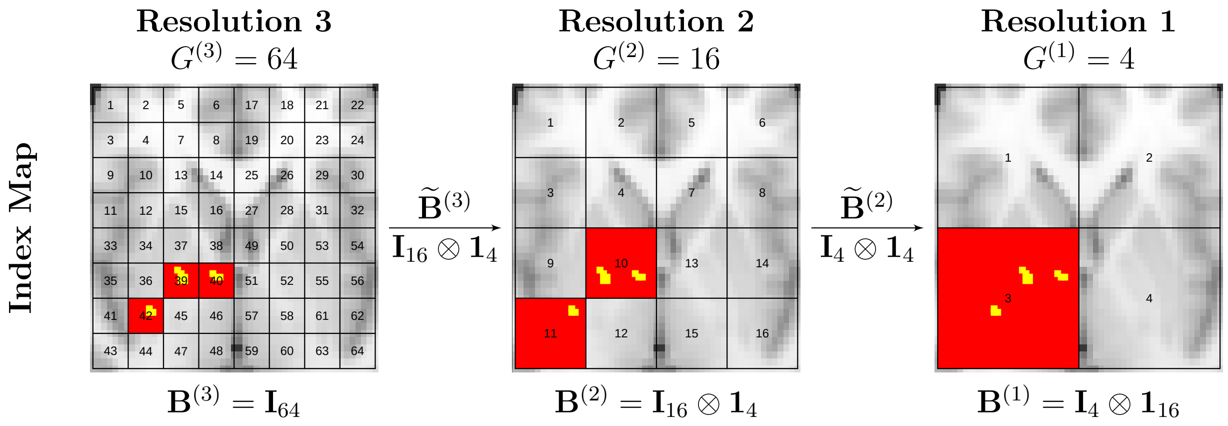


Figure 3.1: An example of multiresolution partitions and variable selection. Suppose a rectangle area in one axial slice cutting through brain that contains 64 regions ($R = 64$) is of interest. We consider three resolutions ($K = 3$). Three images in the right, middle, and left panels are labeled with the partition indices for the nested partitions at resolutions 3, 2 and 1 respectively. At the highest resolution (Resolution 3) there are 64 partitions ($G^{(3)} = 64$) with each partition including only one region and the partition indices are the same as the region indices, thus $\mathbf{B}^{(3)} = \mathbf{I}_{64}$. Resolution 2 has 16 partitions ($G^{(2)} = 16$) where each partition g contains four regions indexed by $4g - 3, 4g - 2, 4g - 1$ and $4g$, for $g = 1, \ldots, 16$, indicating $\widetilde{\mathbf{B}}^{(3)} = \mathbf{B}^{(2)} = \mathbf{I}_{16} \otimes \mathbf{1}_4$, where $\otimes$ is Kronecker product. Resolution 1 has four partitions ($G^{(1)} = 4$) where each partition g' contains four finer-scale partitions at resolution 2 indexed by $4g' - 3, 4g' - 2, 4g' - 1$ and $4g'$, for $g' = 1, \ldots, 4$, resulting in $\widetilde{\mathbf{B}}^{(2)} = \mathbf{I}_4 \otimes \mathbf{1}_4$; thus it contains 16 regions indexed by $16g' - 15, 16g' - 14, \ldots, 16g'$, for $g' = 1, \ldots, 16$, leading to $\mathbf{B}^{(1)} = \mathbf{I}_4 \otimes \mathbf{1}_{16}$. Suppose the true important voxels (yellow) are located in regions 39, 40 and 41. Valid posterior inferences for models at different resolutions produce high posterior inclusion probabilities of imaging biomarkers in the corresponding partitions (red) at all resolutions.

3.3.2 Sequential Resolution Sampling

In the analysis of ultra-high dimensional imaging data, it is reasonable to assume that the signals (i.e., important voxels and regions) are sparse and the vast majority of voxels in the brain are not associated with the outcome. Typically, many of the unimportant voxels/regions, providing little information on prediction of disease risk, are included in the model at each iteration of a standard MCMC algorithm such as the one in Section 3.2.3, resulting in potentially intractable posterior computations. To construct an efficient and computationally feasible MCMC algorithm, one solution is to specify a good proposal distribution in the Metropolis–Hastings (M-H) step for voxel/region selection. Ideally, this proposal distribution should possess two properties:

P1: It assigns large probabilities for excluding unimportant voxels and including important voxels, which substantially reduces the number of selected voxels and simplifies computations in most MCMC iterations.

P2: It still assigns a positive probability for including each voxel in the model, ensuring that the simulated Markov chain is able to explore the entire sample space of the voxel selection.

In other words, we want to construct a "smart" proposal distribution that concentrates on the true model with sparse signals. To this end, we resort to the multiresolution auxiliary models $\mathcal{M}^{(k)}$ defined in Section 3.3.1, based on which we develop a *sequential resolution sampling (SRS)* procedure. Specifically, we conduct the posterior computations for each auxiliary model $\mathcal{M}^{(k)}$ sequentially from resolution 1 to resolution K. At resolution 1, we use the standard MCMC algorithm for posterior simulation on model $\mathcal{M}^{(1)}$. At resolution k, for $k = 2, \ldots, K$, we propose a resolution dependent MCMC algorithm for posterior simulation on model $\mathcal{M}^{(k)}$, referred to as the SRS–MCMC. The essential step is an M-H step for sampling selection indicators $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$, where the "smart" proposal distribution is constructed using the posterior distribution (samples) of $\{\mathbf{c}^{(k-1)}, \boldsymbol{\gamma}^{(k-1)}\}$ in $\mathcal{M}^{(k-1)}$ at resolution $k-1$. Of note, based on the SRS procedure, at resolution K, we can obtain posterior samples on voxel/region selection at the finest scale, i.e. our target resolution.

The SRS procedure is illustrated in Figures 3.1 and 3.2. Figure 3.1 presents an example where the location information of the important voxels is passed along from resolution 1 to resolution 3, becoming more and more precise. Figure 3.2 provides the details on the SRS procedure. Specifically, to construct the "smart" proposal distributions in the M-H step of the SRS–MCMC, we first introduce auxiliary variable selection indicators $\widetilde{\mathbf{c}}^{(k-1)}$ and $\widetilde{\boldsymbol{\gamma}}^{(k-1)}$ in $\mathcal{M}^{(k)}$ at resolution k,

$$\widetilde{c}_{g'}^{(k-1)} = \max\left\{c_g^{(k)} : \widetilde{b}_{gg'}^{(k)} = 1\right\}, \quad \text{and} \quad \widetilde{\gamma}_{rh'}^{(k-1)} = \max\left\{\gamma_{rh}^{(k)} : \widetilde{a}_{rhh'}^{(k)} = 1\right\}, \tag{3.9}$$

for $g' = 1, \ldots, G^{(k-1)}$, $r = 1, \ldots, R$ and $h' = 1, \ldots, H_r^{(k-1)}$. $\{\widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}\}$ are completely determined by $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$ and can be considered as a "coarse-scale summary" of the variable selection indicators in $\mathcal{M}^{(k)}$. In particular, $\{\widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}\}$ in $\mathcal{M}^{(k)}$ are of the same dimension and structure as the variable selection indicators $\{\mathbf{c}^{(k-1)}, \boldsymbol{\gamma}^{(k-1)}\}$ in $\mathcal{M}^{(k-1)}$. The key idea is to use the posterior distribution of $\{\mathbf{c}^{(k-1)}, \boldsymbol{\gamma}^{(k-1)}\}$ in $\mathcal{M}^{(k-1)}$ as the proposal distribution for $\{\widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}\}$ in $\mathcal{M}^{(k)}$, which subsequently is served as a guide to the construction of the proposal distribution for $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$ in $\mathcal{M}^{(k)}$.

The posterior distribution of the parameters and latent quantities in $\mathcal{M}^{(k)}$ is

$$\begin{aligned}
&\pi(\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \sigma_\beta^{2(k)}, \mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)} \mid \mathbf{S}, \mathbf{X}, \mathbf{y}) \\
&\quad = \pi(\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \sigma_\beta^{2(k)}, \mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)} \mid \mathbf{S}, \mathbf{X}, \mathbf{y}) \pi(\widetilde{\mathbf{c}}^{(k-1)} \mid \mathbf{c}^{(k)}) \pi(\widetilde{\boldsymbol{\gamma}}^{(k-1)} \mid \boldsymbol{\gamma}^{(k)}),
\end{aligned} \tag{3.10}$$

where $\pi(\widetilde{\mathbf{c}}^{(k-1)} \mid \mathbf{c}^{(k)})$ and $\pi(\widetilde{\boldsymbol{\gamma}}^{(k-1)} \mid \boldsymbol{\gamma}^{(k)})$ are equal to 1 if (3.9) holds and 0 otherwise. In the SRS–MCMC, the updating scheme for $\{\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \sigma_\beta^{2(k)}\}$ given all other parameters is the same as the Gibbs sampling scheme for $\{\mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\beta^2\}$ for model (3.2); see Appendix 3.7.2 for details. However, the updating scheme for $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}\}$ is more elaborate and is described in detail as follows.

In the M-H step of the SRS-MCMC, we introduce the subscripts "$*$" and "o" to represent the proposed and current values of the corresponding parameters, respectively. Denote by "$\bullet$" all other parameters $\{\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}\}$ and data $\{\mathbf{S}, \mathbf{X}, \mathbf{y}\}$. A proposal distribution for updating $\{\mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \tilde{\mathbf{c}}_o^{(k-1)}, \tilde{\boldsymbol{\gamma}}_o^{(k-1)}\}$ is given by

$$\begin{aligned}
&\mathrm{T}[\{\mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \widetilde{\mathbf{c}}_o^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_o^{(k-1)}\} \to \{\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)}, \widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\} \mid \bullet] \\
&\quad = \mathrm{H}(\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)} \mid \mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \widetilde{\mathbf{c}}_o^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_o^{(k-1)}, \widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}) \mathrm{P}_{k-1}(\widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)} \mid \mathbf{S}, \mathbf{X}, \mathbf{y}) \quad (3.11)
\end{aligned}$$

where $\mathrm{P}_{k-1}(\cdot \mid \cdot)$, specifying the sampling scheme for $\{\widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\}$, is the posterior distribution of the variable selection indicators $\{\mathbf{c}^{(k-1)}, \boldsymbol{\gamma}^{(k-1)}\}$ in the model $\mathcal{M}^{(k-1)}$ at resolution $k-1$ and $\mathrm{H}(\cdot \mid \cdot)$ specifies the sampling scheme for $\{\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)}\}$ given the sampled $\{\widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\}$ from $\mathrm{P}_{k-1}(\cdot \mid \cdot)$ and the current state of the Markov chain, i.e., $\{\mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \tilde{\mathbf{c}}_o^{(k-1)}, \tilde{\boldsymbol{\gamma}}_o^{(k-1)}\}$. The sampling scheme based on decomposition (3.11) is illustrated in Figure 3.2b. Of note, in the SRS procedure $\mathrm{P}_{k-1}(\cdot \mid \cdot)$ is approximated by the posterior samples of $\{\mathbf{c}^{(k-1)}, \boldsymbol{\gamma}^{(k-1)}\}$ in the model $\mathcal{M}^{(k-1)}$ at resolution $k-1$. One choice of $\mathrm{H}(\cdot \mid \cdot)$ that has performed well in our numerical studies is

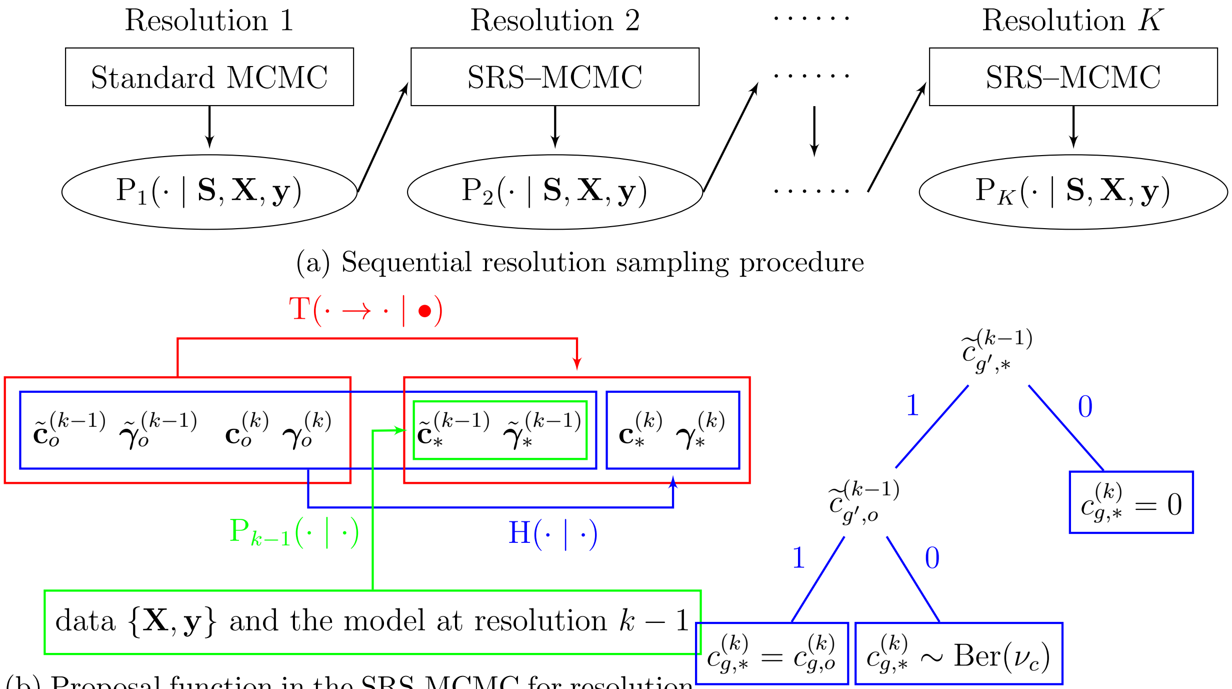
$$\begin{aligned}
&\mathrm{H}(\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)} \mid \mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \tilde{\mathbf{c}}_o^{(k-1)}, \tilde{\boldsymbol{\gamma}}_o^{(k-1)}, \tilde{\mathbf{c}}_*^{(k-1)}, \tilde{\boldsymbol{\gamma}}_*^{(k-1)}) \\
&\quad = \prod_{\widetilde{b}_{gg'}^{(k)}=1} \mathrm{h}(c_{g,*}^{(k)} \mid c_{g,o}^{(k)}, \tilde{c}_{g',*}^{(k-1)}, \tilde{c}_{g',o}^{(k-1)}, \nu_c) \prod_{r=1}^{R} \prod_{\widetilde{a}_{rhh'}^{(k)}=1} \mathrm{h}(\gamma_{rh,*}^{(k)} \mid \gamma_{rh,o}^{(k)}, \gamma_{rh',*}^{(k-1)}, \gamma_{rh',o}^{(k-1)}, \nu_\gamma) \quad (3.12)
\end{aligned}$$

where $\mathrm{h}(\cdot \mid \cdot)$ is a probability mass function for binary random variable defined as

$$\mathrm{h}(x \mid y, a, b, \nu) = (1-a)\delta_0(x) + a[(1-b)\nu^x(1-\nu)^{1-x} + b\delta_y(x)], \text{ for } x \in \{0,1\}$$

with $\nu \in (0,1)$ and $a, b \in \{0,1\}$. The indicator $\delta_y(x) = 1$ if $x = y$ and $\delta_y(x) = 0$ otherwise. Figure 3.2c presents a binary tree to illustrate the sampling scheme for $c_{g,*}^{(k)}$ based on the $\mathrm{h}(\cdot \mid \cdot)$ function and the sampling scheme for $\gamma_{rh,*}^{(k)}$ according to $\mathrm{h}(\cdot \mid \cdot)$

is along the same lines.


Resolution 1
Resolution 2
Resolution K
Standard MCMC
SRS–MCMC
SRS–MCMC
$\mathrm{P}_1(\cdot \mid \mathbf{S}, \mathbf{X}, \mathbf{y})$
$\mathrm{P}_2(\cdot \mid \mathbf{S}, \mathbf{X}, \mathbf{y})$
$\mathrm{P}_K(\cdot \mid \mathbf{S}, \mathbf{X}, \mathbf{y})$
$\mathrm{T}(\cdot \to \cdot \mid \bullet)$
$\widetilde{\mathbf{c}}_o^{(k-1)}$ $\widetilde{\boldsymbol{\gamma}}_o^{(k-1)}$ $\mathbf{c}_o^{(k)}$ $\boldsymbol{\gamma}_o^{(k)}$
$\widetilde{\mathbf{c}}_*^{(k-1)}$ $\widetilde{\boldsymbol{\gamma}}_*^{(k-1)}$
$\mathbf{c}_*^{(k)}$ $\boldsymbol{\gamma}_*^{(k)}$
$\mathrm{P}_{k-1}(\cdot \mid \cdot)$
$\mathrm{H}(\cdot \mid \cdot)$
data $\{\mathbf{X}, \mathbf{y}\}$ and the model at resolution $k-1$
$\widetilde{c}_{g',*}^{(k-1)}$
1
0
$\widetilde{c}_{g',o}^{(k-1)}$
$c_{g,*}^{(k)} = 0$
1
0
$c_{g,*}^{(k)} = c_{g,o}^{(k)}$
$c_{g,*}^{(k)} \sim \mathrm{Ber}(\nu_c)$

(a) Sequential resolution sampling procedure

(b) Proposal function in the SRS-MCMC for resolution k

(c) Sampling $c_{g,*}^{(k)}$ via $\mathrm{h}(\cdot \mid \cdot)$

Figure 3.2: Illustration of sequential resolution sampling. (a) Initially, we utilize the standard MCMC algorithm to produce the posterior distribution of the selection indicators in $\mathcal{M}^{(1)}$ at resolution 1, i.e. $\mathrm{P}_1(\cdot \mid \mathbf{S}, \mathbf{X}, \mathbf{y})$, which is then used to guide the construction of the proposal function in the SRS-MCMC algorithm to produce $\mathrm{P}_2(\cdot \mid \mathbf{S}, \mathbf{X}, \mathbf{y})$ for $\mathcal{M}^{(2)}$ at resolution 2. This procedure is performed sequentially until resolution K to generate the posterior distribution $\mathrm{P}_K(\cdot \mid \mathbf{S}, \mathbf{X}, \mathbf{y})$ for our target model $\mathcal{M}^{(K)}$. (b) Decomposition of the proposal function $\mathrm{T}(\cdot \to \cdot \mid \bullet)$ (red) includes two steps for drawing a proposed sample. Step 1 (green): draw $\{\widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\}$ from the posterior distribution $\mathrm{P}_{k-1}(\cdot \mid \cdot)$ under the model $\mathcal{M}^{(k-1)}$ at resolution $k-1$. Step 2 (blue): sample $\{\mathbf{c}_*^{(k-1)}, \boldsymbol{\gamma}_*^{(k-1)}\}$ given $\{\widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\}$ in step 1 and the current state of the Markov chain using $\mathrm{H}(\cdot \mid \cdot)$. (c) A binary tree represents the sampling scheme for $c_{g,*}^{(k)}$ based on the probability mass function $\mathrm{h}(\cdot \mid \cdot)$. It is determined by $\widetilde{c}_{g',*}^{(k)}$ and $\widetilde{c}_{g',o}^{(k)}$ for g' satisfying $\widetilde{b}_{gg'}^{(k)} = 1$, and $c_{g,o}^{(k)}$.

In addition to the above M-H step, we suggest a moving step with full conditional updates for each element of $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$ given $\{\widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}\}$ and all other parameters using Gibbs sampling. In this step, we only need to update the selection for the fine-scale partitions/subregions that are nested within the selected coarse-scale partitions/subregions. Thus, this step does not require extensive computations and it improves the mixing of the entire Markov chain. To recapitulate, the updating scheme in SRS-MCMC is as follows.

<u>Updating Scheme for $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$ in SRS-MCMC</u>

<u>M-H Step</u>: Set $\{\mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \widetilde{\mathbf{c}}_o^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_o^{(k-1)}\} = \{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}\}$

– Draw $(\widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}) \sim \mathrm{P}_{k-1}(\cdot \mid \mathbf{S}, \mathbf{X}, \mathbf{y})$;

– Draw $(\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)}) \sim \mathrm{H}(\cdot \mid \mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \widetilde{\mathbf{c}}_o^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_o^{(k-1)}, \widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)})$;

– Draw $r \sim \mathrm{U}[0,1]$. Set $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}\} = \{\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)}, \widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\}$ if $r < R$, where

$$R = \frac{\pi(\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)}, \widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)} \mid \bullet)}{\pi(\mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \widetilde{\mathbf{c}}_o^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_o^{(k-1)} \mid \bullet)} \frac{\mathrm{T}[\{\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)}, \widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\} \rightarrow \{\mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \widetilde{\mathbf{c}}_o^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_o^{(k-1)}\} \mid \bullet]}{\mathrm{T}[\{\mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \widetilde{\mathbf{c}}_o^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_o^{(k-1)}\} \rightarrow \{\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)}, \widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\} \mid \bullet]}.$$

<u>Moving Step</u>: Full conditional updates of $\{c_g^{(k)}, \gamma_{rh}^{(k)}\}$ via Gibbs sampling.

– For g' with $\widetilde{c}_{g'}^{(k-1)} = 1$ and g with $\widetilde{b}_{gg'}^{(k)} = 1$,

– if $\mathbf{c}_{[-g]}^{(k)} \neq \mathbf{0}_{G^{(k)}-1}$ then draw $c_g^{(k)} \sim \pi(\cdot \mid \mathbf{c}_{[-g]}^{(k)}, \boldsymbol{\gamma}^{(k)}, \bullet)$, else set $c_g^{(k)} = 1$;

– For r with $b_{rg}^{(k)} = 1$, h' with $\widetilde{\gamma}_{rh'}^{(k-1)} = 1$ and h with $\widetilde{a}_{rhh'}^{(k)} = 1$,

if $\boldsymbol{\gamma}_{r[-h]}^{(k)} \neq \mathbf{0}_{H_r^{(k)}-1}$, then draw $\gamma_{rh}^{(k)} \sim \pi(\cdot \mid \boldsymbol{\gamma}_{r[-h]}^{(k)}, \boldsymbol{\gamma}_{[-r]}^{(k)}, \mathbf{c}^{(k)}, \bullet)$, else set $\gamma_{rh}^{(k)} = 1$.

where $\mathbf{c}_{[-g]}^{(k)} = (c_1^{(k)}, \ldots, c_{g-1}^{(k)}, c_{g+1}^{(k)}, \ldots, c_{G^{(k)}}^{(k)})^\top$, $\boldsymbol{\gamma}_{r[-h]}^{(k)} = (\gamma_{r,1}^{(k)}, \ldots, \gamma_{r,h-1}^{(k)}, \gamma_{r,h+1}^{(k)}, \ldots \gamma_{r,H^{(k)}}^{(k)})^\top$, and $\boldsymbol{\gamma}_{[-r]}^{(k)} = (\boldsymbol{\gamma}_1^{(k)\top}, \ldots, \boldsymbol{\gamma}_{r-1}^{(k)\top}, \boldsymbol{\gamma}_{r+1}^{(k)\top}, \ldots, \boldsymbol{\gamma}_R^{(k)\top})^\top$.

3.3.3 Fast Sequential Resolution Sampling

The SRS procedure in Section 3.3.2 provides a general framework for posterior computations for variable selection in a ultra-high dimensional feature space. The choice of auxiliary models over resolutions and the corresponding MCMC algorithms can be flexible as long as they reduce the total computational cost and improve the mixing of the Markov chains. As an example, we consider two modifications that can potentially further improve computation efficiency: 1) Gaussian quadrature approximation in the auxiliary models $\mathcal{M}^{(k)}$ ($k < K$) that further reduces the number of parameters in the models and 2) a joint updating scheme for the regression coefficients and the selection indicators in $\mathcal{M}^{(k)}$. Combining both, we develop a fast sequential resolution sampling (fastSRS) algorithm.

Gaussian Quadrature Approximation in Auxiliary Models The element-wise representation of auxiliary model (3.8) at resolution k is given by

$$z_i^{(k)} = \alpha_0^{(k)} + \sum_{j=1}^{p} \alpha_j^{(k)} s_{ij} + \sum_{g=1}^{G^{(k)}} \left[c_g^{(k)} \sum_{r=1}^{R} \left\{ b_{rg}^{(k)} \sum_{h=1}^{H_r^{(k)}} \left(\gamma_{rh}^{(k)} \sum_{v=1}^{V_r} a_{rvh}^{(k)} \beta_{rv}^{(k)} x_{irv} \right) \right\} \right] + \epsilon_i^{(k)} \tag{3.13}$$

To introduce an approximation to summation $\sum_{v=1}^{V_r} a_{rvh}^{(k)} \beta_{rv}^{(k)} x_{irv}$, we first define two integrable functions $\beta_r^{(k)}(\cdot)$ and $\mathrm{x}_{ir}(\cdot)$ defined on $\mathcal{B}$ with constraints that $\beta_r^{(k)}(\mathbf{s}_v) = \beta_{rv}^{(k)}$ and $\mathrm{x}_{ir}^{(k)}(\mathbf{s}_v) = x_{irv}$, where the coordinate $\mathbf{s}_v \in \mathcal{B}$ represents the location of voxel v. Denote by $\mathcal{S}_{rh}^{(k)}$ the compact domain of subregion h in region r. Let $\delta\mathbf{s}$ represent the volume of a voxel in the brain. Based on the definition of the Riemann integral, we have

$$\int_{\mathcal{S}_{rh}^{(k)}} \beta_r^{(k)}(\mathbf{s}) \mathrm{x}_{ir}(\mathbf{s}) d\mathbf{s} \approx \delta\mathbf{s} \sum_{v=1}^{V_r} a_{rvh}^{(k)} \beta_{rv}^{(k)} x_{irv}. \tag{3.14}$$

When $\delta\mathbf{s}$ is small, this approximation is accurate. If both $\beta_r^{(k)}(\cdot)$ and $\mathrm{x}_{ir}(\cdot)$ are smooth over $\mathcal{B}$, we can further approximate the integral using Gaussian quadrature on a set of sparse grids given by

$$\int_{\mathcal{S}_{rh}^{(k)}} \beta_r^{(k)}(\mathbf{s}) \mathrm{x}_{ir}(\mathbf{s}) d\mathbf{s} \approx \sum_{v \in \mathcal{Q}_{rh}^{(k)}} w_{rvh}^{(k)} \beta_{rv}^{(k)} x_{irv} = \delta\mathbf{s} \sum_{v=1}^{V_r} q_{rvh}^{(k)} \beta_{rv}^{(k)} x_{irv}, \tag{3.15}$$

where $w^{(k)}_{rvh}$ is the weight and $\mathcal{Q}^{(k)}_{rh}$ is a set of voxel indices of the sparse grid points on $\mathcal{S}^{(k)}_{rh}$ based on the Smolyak's construction rule (Gerstner and Griebel, 1998). The term $q^{(k)}_{rvh} = w^{(k)}_{rvh}/\delta\mathbf{s}$ if $v \in \mathcal{Q}^{(k)}_{rh}$, $q^{(k)}_{rvh} = 0$, otherwise. Combining (3.14) and (3.15), we can replace $a^{(k)}_{rvh}$ by $q^{(k)}_{rvh}$ in (3.13) to construct auxiliary models $\mathcal{M}^{(k)}$ based on Gaussian quadrature approximation.

Of note, we only need to conduct such approximation at lower resolutions to reduce computation when each subregion contains a large number of voxels. With $\sum_{r=1}^{R}\sum_{h=1}^{H_r^{(k)}} \mathcal{Q}^{(k)}_{rh}$ approaching V as the resolution increases, the saving in computational costs vanishes and it is recommended to use the original model (3.8) for higher resolutions. Since the auxiliary models are only used to guide the construction of the proposal distributions and our target model $\mathcal{M}^{(K)}$ remains unchanged, such an approximation is still valid.

Joint Updating Scheme We introduce an auxiliary variable defined as

$$\widetilde{\beta}^{(k)}_{rv} = \gamma^{(k)}_{rh}\beta^{(k)}_{rv} \text{ for } r = 1,\dots,R, \text{ and } v,h \text{ with } a^{(k)}_{rvh} = 1. \tag{3.16}$$

Define $\widetilde{\boldsymbol{\beta}}^{(k)}_{rh} = (\widetilde{\beta}^{(k)}_{rv}, a^{(k)}_{rvh} = 1)^\top$, $\widetilde{\boldsymbol{\beta}}^{(k)}_{r} = (\widetilde{\boldsymbol{\beta}}^{(k)\top}_{r1}, \dots, \widetilde{\boldsymbol{\beta}}^{(k)\top}_{rH_r^{(k)}})^\top$ and $\widetilde{\boldsymbol{\beta}}^{(k)} = (\widetilde{\boldsymbol{\beta}}^{(k)\top}_{1}, \dots, \widetilde{\boldsymbol{\beta}}^{(k)\top}_{R})^\top$. It follows that $\widetilde{\boldsymbol{\beta}}^{(k)}$ is completely determined by $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ and the joint posterior distribution of all parameters is given by

$$\begin{aligned}&\pi(\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \sigma^{2(k)}_{\beta}, \mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}, \widetilde{\boldsymbol{\beta}}^{(k)} \mid \mathbf{S}, \mathbf{X}, \mathbf{y}) \\ &\quad = \pi(\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \sigma^{2(k)}_{\beta}, \mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)} \mid \mathbf{S}, \mathbf{X}, \mathbf{y}) \pi(\widetilde{\mathbf{c}}^{(k-1)} \mid \mathbf{c}^{(k)}) \pi(\widetilde{\boldsymbol{\beta}}^{(k)} \mid \boldsymbol{\gamma}^{(k)}, \boldsymbol{\beta}^{(k)}) \pi(\widetilde{\boldsymbol{\gamma}}^{(k-1)} \mid \boldsymbol{\gamma}^{(k)}),\end{aligned} \tag{3.17}$$

where $\pi(\widetilde{\boldsymbol{\beta}}^{(k)} \mid \boldsymbol{\gamma}^{(k)}, \boldsymbol{\beta}^{(k)}) = 1$ if (3.16) holds and $\pi(\widetilde{\boldsymbol{\beta}}^{(k)} \mid \boldsymbol{\gamma}^{(k)}, \boldsymbol{\beta}^{(k)}) = 0$ otherwise. Furthermore, for $r = 1,\dots,R$ and $h = 1,\dots,H^{(k)}_r$, $\pi(\boldsymbol{\beta}^{(k)}_{rh} \mid \gamma^{(k)}_{rh} = 1, \mathbf{S}, \mathbf{X}, \mathbf{y}) = \pi(\widetilde{\boldsymbol{\beta}}^{(k)}_{rh} \mid \gamma^{(k)}_{rh} = 1, \mathbf{S}, \mathbf{X}, \mathbf{y})$ and $\pi(\boldsymbol{\beta}^{(k)}_{rh} \mid \gamma^{(k)}_{rh} = 0, \mathbf{S}, \mathbf{X}, \mathbf{y}) = \pi(\boldsymbol{\beta}^{(k)}_{rh})$, implying that the marginal posterior distribution of $\boldsymbol{\beta}^{(k)}$ is determined by the marginal posterior distribution of $\{\widetilde{\boldsymbol{\beta}}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$ and its prior. Thus, we integrate out $\boldsymbol{\beta}^{(k)}$ in (3.17) and focus on $\pi(\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \sigma_\beta^{2(k)}, \mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}, \widetilde{\boldsymbol{\beta}}^{(k)} \mid \mathbf{S}, \mathbf{X}, \mathbf{y})$, leading to the target distribution of the fastSRS-MCMC algorithm. Compared to the SRS-MCMC algorithm, the updating scheme for $\{\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \sigma_\beta^{2(k)}\}$ is the same but the sampling scheme for $\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}$ and $\widetilde{\boldsymbol{\beta}}^{(k)}$ needs to be modified. For an M-H step, we choose the following proposal distribution

$$\begin{aligned}
&\widetilde{\mathrm{T}}[\{\widetilde{\boldsymbol{\beta}}_o^{(k)}, \mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \widetilde{\mathbf{c}}_o^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_o^{(k-1)}\} \to \{\widetilde{\boldsymbol{\beta}}_*^{(k)}, \mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)}, \widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\} \mid \bullet] \\
&\quad = \mathrm{T}[\{\mathbf{c}_o^{(k)}, \boldsymbol{\gamma}_o^{(k)}, \widetilde{\mathbf{c}}_o^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_o^{(k-1)}\} \to \{\mathbf{c}_*^{(k)}, \boldsymbol{\gamma}_*^{(k)}, \widetilde{\mathbf{c}}_*^{(k-1)}, \widetilde{\boldsymbol{\gamma}}_*^{(k-1)}\} \mid \bullet]\ \widetilde{\mathrm{H}}(\widetilde{\boldsymbol{\beta}}_*^{(k)} \mid \widetilde{\boldsymbol{\beta}}_o^{(k)}, \boldsymbol{\gamma}_*^{(k)}, \boldsymbol{\gamma}_o^{(k)}), \qquad (3.18)
\end{aligned}$$

where $\mathrm{T}[\cdot \to \cdot \mid \bullet]$ is the proposal distribution in SRS-MCMC in Section 3.3.2. The function $\widetilde{\mathrm{H}}(\cdot \mid \cdot)$ is decomposed as

$$\widetilde{\mathrm{H}}(\widetilde{\boldsymbol{\beta}}_*^{(k)} \mid \widetilde{\boldsymbol{\beta}}_o^{(k)}, \boldsymbol{\gamma}_*^{(k)}, \boldsymbol{\gamma}_o^{(k)}) = \prod_{r=1}^{R} \prod_{h=1}^{H_r^{(k)}} \widetilde{\mathrm{h}}(\widetilde{\boldsymbol{\beta}}_{rh*}^{(k)} \mid \widetilde{\boldsymbol{\beta}}_{rh,o}^{(k)}, \gamma_{rh,*}^{(k)}, \gamma_{rh,o}^{(k)}, \sigma_\beta^{2(k)}), \qquad (3.19)$$

Here, $\widetilde{\mathrm{h}}(\cdot \mid \cdot)$ is a probability density function for a d-dimensional random vector

$$\widetilde{\mathrm{h}}(\mathbf{u} \mid \mathbf{v}, a, b, \sigma^2) \quad = \quad (1-a)\delta_{\mathbf{0}}(\mathbf{u}) + a[(1-b)\phi(\mathbf{u}; \mathbf{0}, \sigma^2\mathbf{I}) + b\delta_{\mathbf{v}}(\mathbf{u})],$$

where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ $(d > 1)$, $a, b \in \{0, 1\}$, $\sigma^2 > 0$, and $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a normal density function with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Figure 3.3 illustrates the sampling scheme for $\widetilde{\boldsymbol{\beta}}_{rh,*}^{(k)}$ based on $\widetilde{\mathrm{h}}(\cdot \mid \cdot)$, which depends on $\boldsymbol{\beta}_{rh,o}^{(k)}$, $\gamma_{rh,*}^{(k)}$, $\gamma_{rh,o}^{(k)}$ and $\sigma_\beta^{2(k)}$.

Again, in addition to the M-H step, we suggest a moving step to improve the mixing by updating $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\boldsymbol{\beta}}^{(k)}\}$ given $\widetilde{\mathbf{c}}^{(k-1)}$ and $\widetilde{\boldsymbol{\gamma}}^{(k-1)}$. The moving step for $\mathbf{c}^{(k)}$ is the same as the SRS-MCMC in Section 3.3.2. For $\{\widetilde{\boldsymbol{\beta}}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$, we consider a block updating scheme. For h' with $\widetilde{\gamma}_{rh'}^{(k-1)} = 1$, denote by $\widetilde{\boldsymbol{\beta}}_{rh'}^{(k)} = (\widetilde{\boldsymbol{\beta}}_{rh}^{(k)\top} : \widetilde{a}_{rhh'}^{(k)} = 1)^\top$

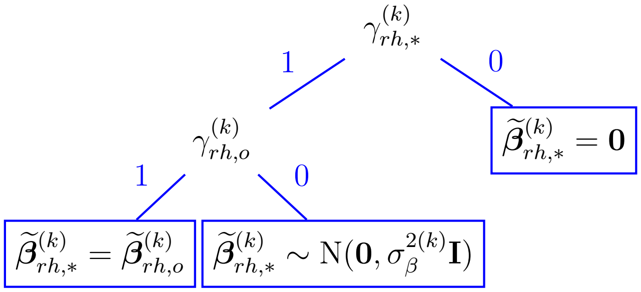

$\gamma^{(k)}_{rh,*}$
1
0
$\gamma^{(k)}_{rh,o}$
$\widetilde{\boldsymbol{\beta}}^{(k)}_{rh,*} = \mathbf{0}$
1
0
$\widetilde{\boldsymbol{\beta}}^{(k)}_{rh,*} = \widetilde{\boldsymbol{\beta}}^{(k)}_{rh,o}$
$\widetilde{\boldsymbol{\beta}}^{(k)}_{rh,*} \sim \mathrm{N}(\mathbf{0}, \sigma^{2(k)}_{\beta}\mathbf{I})$

Figure 3.3: A binary tree to illustrate the sampling scheme of $\widetilde{\boldsymbol{\beta}}^{(k)}_{rh,*}$ via $\widetilde{\mathrm{h}}(\cdot \mid \cdot)$.

and $\boldsymbol{\gamma}^{(k)}_{rh'} = (\gamma^{(k)\top}_{rh} : \widetilde{a}^{(k)}_{rhh'} = 1)^\top$ the collection of the regression coefficients and the collection of the selection indicators in $\mathcal{M}^{(k)}$ for subregion h' at resolution $k-1$, respectively. Similarly, define $\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'1} = (\widetilde{\boldsymbol{\beta}}^{(k)\top}_{rh} : \gamma^{(k)}_{rh} = 1, \widetilde{a}^{(k)}_{rhh'} = 1)^\top$ and $\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'0} = (\widetilde{\boldsymbol{\beta}}^{(k)\top}_{rh} : \gamma^{(k)}_{rh} = 0, \widetilde{a}^{(k)}_{rhh'} = 1)^\top$. The updating scheme for $\{\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'}, \boldsymbol{\gamma}^{(k)}_{rh'}\}$ is based on the following decomposition of the joint full conditional distributions:

$$\pi(\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'}, \boldsymbol{\gamma}^{(k)}_{rh'} \mid \bullet) = \pi(\gamma^{(k)}_{rh'} \mid \bullet)\pi(\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'1} \mid \boldsymbol{\gamma}^{(k)}_{rh'}, \bullet)\pi(\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'0} \mid \boldsymbol{\gamma}^{(k)}_{rh'}, \bullet), \tag{3.20}$$

The details of (3.20) are provided in Appendix 3.7.3.

The updating scheme for the fastSRS-MCMC is summarized as follows.

Updating Scheme for $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\boldsymbol{\beta}}^{(k)}\}$ in fastSRS-MCMC

M-H Step: Set $\{\mathbf{c}^{(k)}_o, \boldsymbol{\gamma}^{(k)}_o, \widetilde{\mathbf{c}}^{(k-1)}_o, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_o, \widetilde{\boldsymbol{\beta}}^{(k)}_o\} = \{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}, \widetilde{\boldsymbol{\beta}}^{(k)}\}$

– Draw $(\mathbf{c}^{(k)}_*, \boldsymbol{\gamma}^{(k)}_*, \widetilde{\mathbf{c}}^{(k-1)}_*, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_*) \sim \mathrm{T}[(\mathbf{c}^{(k)}_o, \boldsymbol{\gamma}^{(k)}_o, \widetilde{\mathbf{c}}^{(k-1)}_o, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_o) \to \cdot \mid \bullet]$;

– Draw $\widetilde{\boldsymbol{\beta}}^{(k)}_* \sim \widetilde{\mathrm{H}}(\cdot \mid \widetilde{\boldsymbol{\beta}}^{(k)}_o, \boldsymbol{\gamma}^{(k)}_*, \boldsymbol{\gamma}^{(k)}_o)$;

– Draw $r \sim \mathrm{U}[0,1]$. Set $\{\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\mathbf{c}}^{(k-1)}, \widetilde{\boldsymbol{\gamma}}^{(k-1)}, \widetilde{\boldsymbol{\beta}}^{(k-1)}\} = \{\mathbf{c}^{(k)}_*, \boldsymbol{\gamma}^{(k)}_*, \widetilde{\mathbf{c}}^{(k-1)}_*, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_*, \widetilde{\boldsymbol{\beta}}^{(k-1)}_*\}$

if $r < R$, where

$$R = \frac{\pi(\mathbf{c}^{(k)}_*, \boldsymbol{\gamma}^{(k)}_*, \widetilde{\mathbf{c}}^{(k-1)}_*, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_*, \widetilde{\boldsymbol{\beta}}^{(k)}_* \mid \bullet)}{\pi(\mathbf{c}^{(k)}_o, \boldsymbol{\gamma}^{(k)}_o, \widetilde{\mathbf{c}}^{(k-1)}_o, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_o, \widetilde{\boldsymbol{\beta}}^{(k)}_o \mid \bullet)} \times \frac{\widetilde{\mathrm{T}}[\{\mathbf{c}^{(k)}_*, \boldsymbol{\gamma}^{(k)}_*, \widetilde{\mathbf{c}}^{(k-1)}_*, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_*, \widetilde{\boldsymbol{\beta}}^{(k)}_*\} \to \{\mathbf{c}^{(k)}_o, \boldsymbol{\gamma}^{(k)}_o, \widetilde{\mathbf{c}}^{(k-1)}_o, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_o, \widetilde{\boldsymbol{\beta}}^{(k)}_o\} \mid \bullet]}{\widetilde{\mathrm{T}}[\{\mathbf{c}^{(k)}_o, \boldsymbol{\gamma}^{(k)}_o, \widetilde{\mathbf{c}}^{(k-1)}_o, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_o, \widetilde{\boldsymbol{\beta}}^{(k)}_o\} \to \{\mathbf{c}^{(k)}_*, \boldsymbol{\gamma}^{(k)}_*, \widetilde{\mathbf{c}}^{(k-1)}_*, \widetilde{\boldsymbol{\gamma}}^{(k-1)}_*, \widetilde{\boldsymbol{\beta}}^{(k)}_*\} \mid \bullet]}.$$

<u>**Moving Step**</u>: Full conditional updates of $\{c_g^{(k)}, \boldsymbol{\gamma}_{rh'}^{(k)}, \widetilde{\boldsymbol{\beta}}_{rh'}^{(k)}\}$ via Gibbs sampling.

- For g' with $\widetilde{c}_{g'}^{(k-1)} = 1$ and g with $\widetilde{b}_{gg'}^{(k)} = 1$,
    - if $\mathbf{c}_{[-g]}^{(k)} \neq \mathbf{0}_{G^{(k)}-1}$ then draw $c_g^{(k)} \sim \pi(\cdot \mid \mathbf{c}_{[-g]}^{(k)}, \boldsymbol{\gamma}^{(k)}, \bullet)$, else set $c_g^{(k)} = 1$;
    - For r with $b_{rg}^{(k)} = 1$, h' with $\widetilde{\gamma}_{rh'}^{(k-1)} = 1$,

        Draw $\{\widetilde{\boldsymbol{\beta}}_{rh'}^{(k)}, \boldsymbol{\gamma}_{rh'}^{(k)}\}$ based on (3.20).

3.4 Application

We analyze the motivating ABIDE study introduced in Section 1.1.2 using the SRS procedure. Our goal is to identify important voxel-wise image biomarkers that are predictive of the ASD risk. After removing missing observations, our analysis includes 831 subjects aggregated from 14 different sites. For each subject, the voxel-wise fALFF values are computed for each of 185,405 voxels over 116 regions in the brain. In addition, three clinical variables, age at scan, sex and IQ, are included in the analysis. Since we observe no substantial difference in the fALFF values and the number of ASDs/TDs between different study sites, site is not included in our analysis, consistent with the previous analysis of the data (Di Martino et al., 2013).

A region-wise functional connectivity network is constructed based on the correlations between the regional R-fMRI time series that are summarized from voxel-wise R-fMRI time series using a single value decomposition approach (Bowman et al., 2012). The neighborhood of each voxel is defined as the set of adjacent voxels from six different directions (top, bottom, front, back, left, and right); voxels are connected to their neighbors in the spatial dependence network. These two levels of structural information are incorporated using the Ising priors for selection indicators $\mathbf{c}$ and $\boldsymbol{\gamma}$. For other prior specifications, we set $\sigma_\alpha^2 = 20$ leading to a fairly flat prior on $\boldsymbol{\alpha}$ and set $a_\beta = 5$ and $b_\beta = 10$ leading to a less-informative prior on σ_β^2; we specify the range of the uniform distribution priors for the sparse parameters in the two levels of Ising model as $[a_\eta, b_\eta] = [-5, 5]$ and for the smooth parameters as $[a_\xi, b_\xi] = [0, 5]$.

In light of the brain anatomy, brain partitions and corresponding subregions at eleven resolutions are used to construct auxiliary multiresolution models, $\{\mathcal{M}^{(k)} : k = 1, \ldots, 11\}$. We utilize the fastSRS-MCMC in Section 3.3.3 to conduct the posterior inference. For each resolution, we run five MCMC chains with random initial values for 2,000 iterations with 1,000 burn-in. The MCMC convergence is assessed by Gelman and Rubin's method (Gelman and Rubin, 1992). For all resolutions, the potential scale reduction factors (PSRF) for the log-likelihood over 1,000 iterations after burn-in are less than 1.1, suggesting convergence.

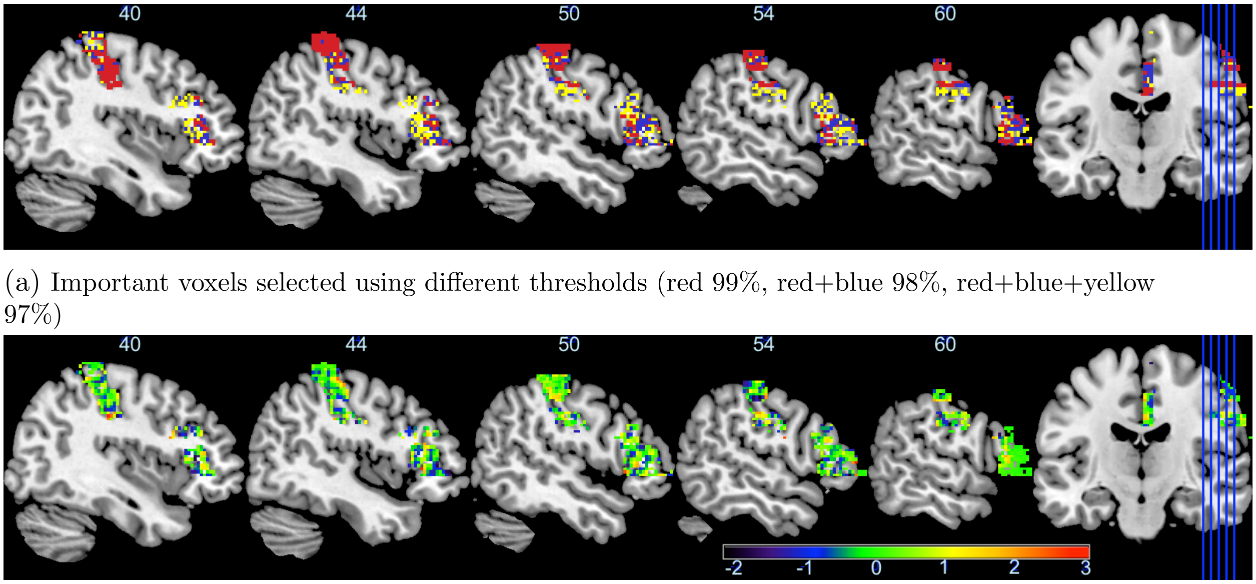
Similar to other works in imaging data analysis, we assume that the true signals are sparse. After obtaining the posterior samples at the highest resolution, the threshold for variable selection is set to 99%, 98% or 97% quantiles of the posterior inclusion probabilities (ranging from 0.000 to 0.380) for all voxels, equivalent to selecting top 1%, 2% and 3% voxels. The selection results based on different thresholds are summarized in Table 4.1. For all thresholds, the selected voxels are mainly located in two regions: the right postcentral gyrus (PoCG-R) and the right inferior frontal gyrus (triangular part) (IFGtriang-R). With a threshold of 97%, our approach selects 2381 and 1424 voxels in the PoCG-R and the IFGtriang-R, respectively. Most of them are spatially clustered and contiguous within a region, as shown in Figure 3.4a. The PoCG is known as the center of the brain for sending and receiving the message and its volume has been shown significantly larger in autism patients compared with controls (Rojas et al., 2006). The IFGtriang is well known for its dominant roles in the cognitive control of language and memory (Foundas et al., 1996; Badre and Wagner, 2007). More recently, several recent task-related fMRI studies (Just and Pelphrey, 2013) showed that autism patients exhibited reduced brain activities in the IFGtriang-R. Our results further suggest that the resting state brain activities (reflected by the fALFF) in the PoCG-R and the IFGTriang-R along with other four regions are highly predictive of the ASD risk. Figure 3.4b presents the posterior means of the regression coefficients for the selected voxels, interestingly, showing both large positive values (red voxels) and large negative values (blue voxels) in the selected regions, especially the IFGtriang-R. Di Martino et al. (2013) reported a negative association of the fALFF in a similar region (the right middle frontal gyrus) with the ASD, suggesting that there may be an anti-correlation brain network located in this region that is predictive of the ASD risk. The posterior mean with 95% credible interval of regression coefficients for age, sex and IQ are respectively -0.132 $(-0.580, 0.352)$, -0.956 $(-2.048, -0.004)$ and -1.504 $(-1.848, -1.133)$, indicating that age is not significantly associated with the ASD, while patients with low IQ and males have a relatively high ASD risk. These findings have the potential to help neuroscientists and epidemiologists better understand the autism etiology.

To evaluate the goodness of fit of our model, we perform a posterior predictive assessment (Gelman et al., 1996) based on the χ^2 discrepancy and obtain a posterior predictive p-value of 0.850, indicating a good fit. To assess the performance on the ASD risk prediction, we use a ten-fold cross-validation approach based on the important sampling method (Vehtari and Lampinen, 2002). Table 4.1 shows that both sensitivity and specificity are greater than 0.9 for all three thresholds, indicating a strong predictive power of our method.

As a comparison, we also analyze the ABIDE data using an alternative approach, namely, SIS+LASSO, implemented by R packages `SIS` and `gl1ce`. This approach first identifies a set of potentially important voxels via the SIS method (Fan and Song, 2010) for a probit regression model and then applies Lasso (Tibshirani, 1996) to the same model using only the voxels selected in the first step. This approach se-

Threshold	Selected AAL Regions	N_{voxel}	P-Sens	P-Spec
99%	IFGtriang-R, PoCG-R, DCG-R,	1,779	0.938	0.918
98%	IFGtriang-R, PoCG-R, DCG-R,	3,494	0.927	0.921
97%	IFGtriang-R, PoCG-R, DCG-R, SFGmed-R, SMA-R, HES-R	5,160	0.901	0.932

Table 3.1: Selection results and prediction accuracy for the ASD risk. The six selected AAL regions are the right postcentral gyrus (PoCG-R), the right inferior frontal gyrus triangular part (IFGtriang-R), the right median cingulate and paracingulate gyri (DCG-R), the right superior frontal gyrus (SFGmed-R), the supplementary motor area (SMA-R) and the right heschl gyrus (HES-R). N_{voxel} is the total number of selected voxels. P-Sens and P-Spec represent sensitivity and specificity in prediction of the ASD risk via a ten-fold cross validation



(a) Important voxels selected using different thresholds (red 99%, red+blue 98%, red+blue+yellow 97%)

(b) Posterior mean of regression coefficients for important voxels selected using a threshold of 97%

Figure 3.4: Five real brain Sagittal (right) slices ($X = 40, 44, 50, 54, 60$ mm) cutting through two regions: IFGtriang-R and PoCG-R.

lects only 99 important voxels, most of which are not located in the regions identified by our method. More notably, when evaluated via a ten-fold cross-validation, the SIS+LASSO approach achieves a considerably lower sensitivity (0.705) and specificity (0.701) in prediction compared to our method, suggesting the superiority of our method in the prediction of the ASD risk.

3.5 Simulation Studies

We conduct three simulations to illustrate the variable selection performance of the proposed methods. In the first simulation, we consider a relatively low dimensional data set (1600 voxels) to compare the standard method and proposed approach (SRS and fastSRS) in terms of the selection accuracy, the computational time and the effective sample size (ESS). The standard method is implemented by the standard posterior computation (SPC) discussed Section 3.2.3. In the second one, in light of the original ABIDE data set, we generate 50 Monte Carlo (MC) data sets with number of signals smaller than sample size to compare the proposed method to the widely used frequentist variable selection method SIS+LASSO (as illustrated in Section 3.4). In the third one, we directly mimic the ABIDE study to illustrate the variable selection performance of the proposed method under the ultra high-dimensional case with number of signals larger than the sample size. All the hyper-prior settings directly follow those in Section 3.4. Similarly, all the MCMC simulations are performed under multiple chains with random initials. The convergence is confirmed by the GR method, where the PSRF is close to 1 for each of the simulations. All algorithms are implemented in `Matlab`. All the simulations are run on a PC with 3.4 GHz CPU, 8GB Memory and Windows System.

3.5.1 Simulation 1

We focus on a 40 × 40 two-dimensional square with 1,600 voxels (Figure 4.6). It consists of four regions (regions 1 – 4) where each of them contains 400 voxels, i.e. $R = 4$, $V_r = 400$, for $r = 1, \ldots, 4$. We set $n = 100$ and jointly simulate imaging biomarkers $\{x_{irv}\}_{r=1}^{R}{}_{v=1}^{V_r}$ from a zero mean Gaussian process with an exponential kernel (variance 0.5, correlation 36). We further set 35 and 50 voxels in regions 1 and 4 to general the true signals (red voxels in Figure 4.6) with the coefficients drawn from Gaussian processes with mean 5 and -6 (variance 0.2, correlation parameter 50). For each of three algorithms, we run 3,000 iterations with 1,000 burn-in.

The variable selection sensitivity and specificity under different thresholds, the area under the curve (AUC), the effective computing time, the resolution related time and the ESS per minute for each algorithm are presented in Table 4.2. Based on the results, with a similar performance of feature selection accuracy, the proposed algorithms (SRS and fastSRS) require a substantial lower computational cost compared to the standard method (SPC); and such a difference is expected to become more remarkable with the number of variables increased. In addition, as shown in Table 4.2, the fastSRS algorithm achieves an around 54 times and 680 times larger ESS per minute than the algorithms SRS and SPC respectively. This is consistent with our expectation that the fastSRS substantially improves the mixing of the Markov chains compared with the other two methods.

The comparable feature selection performance of the three algorithms indicates that our multiresolution approach is a useful tool to improve computational efficiency and speed up the MCMC convergence. When the data dimension is very high, the standard MCMC algorithm suffers a heavy or even intractable computation. In contrast, both the SRS and the fastSRS are still computationally feasible and have a good performance on feature selection, while the fastSRS provides a more appealing

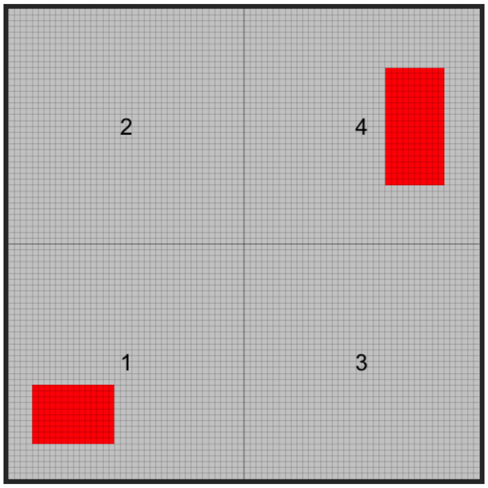


Figure 3.5: Simulation 1 design: two-dimensional square with four regions labeled with texts. Important voxels (red) are located in regions 1 and 4.

ESS. Thus, in the following simulation study, similar to Section **??**, we only conduct posterior inference using the fastSRS.

	Threshold	SPC	SRS	fastSRS
Sensitivity/Specificity	95%	0.659/0.984	0.625/0.982	0.671/0.985
	90%	0.847/0.941	0.753/0.936	0.847/0.941
	85%	0.941/0.889	0.953/0.895	0.965/0.895
	80%	0.977/0.838	1.000/0.846	1.000/0.844
AUC		0.973	0.970	0.979
Effective Time (mins)[1]		13.100	1.600	2.600
Resolution related Time (mins)[2]		0.000	5.033	7.233
ESS/min		0.669	8.381	454.004

[1] The computational time for resolution K (last resolution) for the SRS and the fastSRS.
[2] The computational time for resolutions $1, \ldots, K-1$ for the SRS and the fastSRS.

Table 3.2: Variable selection performance by three different algorithms in Simulation 1

3.5.2 Simulation 2

We consider an ultra-high dimensional case in simulation 2 and compare the proposed fastSRS method with SIS+LASSO approach in terms of variable selection accuracy. Specifically, we generated 50 MC data sets with the imaging biomarkers obtained by permuting the original ABIDE data ($\mathbf{X}$) over regions, i.e. $X^*_{irv} = X_{\zeta_{ir}rv}$ for

$r = 1, \ldots, R$ and $v = 1, \ldots, V_r$ with X^*_{irv} the observed data form a particular data set and $(\zeta_{1r}, \ldots, \zeta_{nr})$ is one permutation of $(1, \ldots, n)$. In such way, for each data set, we could keep the correlations of fALFF values between voxels within a particular region. We set the true signals located in two regions (IFGtriang-R and PoCG-R) as detected in Section 3.4 with 371 and 241 important voxels that are spatially contiguous. The voxel-wise regression coefficients are drawn from $\mathrm{N}(7, 0.1)$ and $\mathrm{N}(5, 0.1)$ for the important voxels in the IFGtriang-R and the PoCG-R respectively, and are set to be zero for all other voxels. The fastSRS-MCMC is run 2,000 iterations with 1,000 burn-in. Similarly, SIS and LASSO are implemented by R packages `SIS` and `gl1ce`.

Method	Threshold	TP (sd)	TN (sd)	Sensitivity	Specificity
fastSRS	99%	586 (19.535)	183,528 (19.421)	0.958	0.993
	98%	612 (0.000)	181,701 (4.945)	1.000	0.983
	97%	612 (0.000)	180,002 (212.430)	1.000	0.974
SIS+LASSO		65 (5.857)	184,777 (4.840)	0.106	1.000

Table 3.3: Variable selection performance over 50 MC data sets by the fastSRS algorithm and SIS+LASSO approach in Simulation 2

The variable selection performance under different methods are summarized in Table 3.4 based on true positive (TP), true negative (TN), sensitivity and specificity. Here, without pre-specifying the number of selected variables for SIS, the *SIS* function in the `SIS` package results in an extremely small number of selections (around 30), which is far from the truth. To improve the performance of the SIS+LASSO, we specify the number of selected variable in SIS to be 700 for all the 5 data sets, which is larger than the true number of signals and conduct LASSO based on these 700 screening variables. Based on Table 3.4, our methods show a substantial better performance than SIS+LASSO approach in terms of variable selection accuracy.

Threshold	Sensitivity	Specificity
99%	0.719	0.998
98%	0.962	0.990
97%	1.000	0.982

Table 3.4: Variable selection accuracy for different thresholds using fastSRS-MCMC in simulation 3

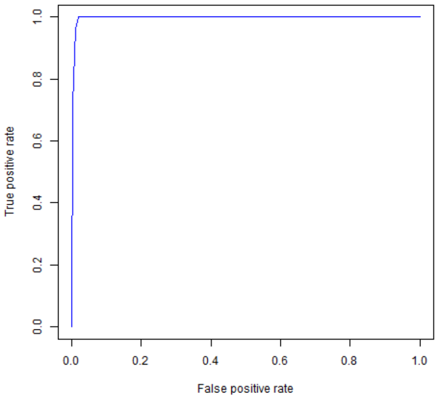


Figure 3.6: Receiver operating characteristic (ROC) curve for the variable selection using fastSRS-MCMC in simulation 3

3.5.3 Simulation 3

In this part of simulation, we directly adopt the voxel-wise fALFF values over the whole brain (116 regions and 185,405 voxels) from the ABIDE data for all 831 subjects containing the region-wise functional connectivity and voxel-wise spatial correlation information. To create sparsity as the case in practice, similar to simulation 2, the true signals are set to be located in the two regions (IFGtriang-R and PoCG-R) as detected in Section 3.4, respectively containing 852 and 1,090 important voxels which are spatially contiguous. We still let the voxel-wise regression coefficients drawn from $N(7, 0.1)$ and $N(5, 0.1)$ for the two pieces of signals, and are set to be zero for the noninformative part. The fastSRS-MCMC is run 2,000 iterations with 1,000 burn-in. Of note, in this set of simulation, the number of signals are larger than the sample size as compared to simulation 2.

The variable selection accuracy under three thresholds are summarized in Table 3.4. The results suggest that our method achieves an extremely high variable selection accuracy: both sensitivity and specificity are close to one for thresholds of 97% and 98%. Also, we obtain a very good receiver operating characteristic (ROC) curve by varying the threshold between 1% and 99% as shown in Figure 3.6. Such satisfactory performance not only shows the feasibility of the proposed method under an ultra high-dimensional case, but also verifies the selection results in data application to certain extent with a consideration of the mimic settings in the current simulation. The computational time for the whole posterior simulation is 2.77 hours, which is extremely remarkable under such an ultra high-dimensionality.

3.6 Discussion

In this chapter, we present a novel Bayesian multiresolution approach for variable selection in a ultra-high dimensional feature space. Our approach is computationally feasible and efficient and it can incorporate both spatial information and functional connectivity information into feature selection, leading to biologically more interpretable results and improved performance. As shown in our numerical studies, it works especially well when the true important voxels are sparse and spatially clustered.

The current multiresolution approach is developed based on the commonly used latent indicator approach in a Bayesian modeling framework. The bottleneck of its posterior computation comes from the inefficiency in sampling multi-level latent selection indicators. One direction of extending our work that will further reduce computational time is to develop parallel computing algorithms for jointly updating high dimensional latent indicators and implement them using the popular General-Purpose computation on Graphics Process Unit (GPGPU) technique (Suchard et al., 2010). Also, the Bayesian shrinkage approach as a different strategy for variable selection has also attracted much attention recently (Park and Casella, 2008; Hans, 2009; Li and Lin, 2010; Hans, 2011; Bhattacharya et al., 2012). This method is closely related to penalized likelihood approaches and it imposes a "weak" sparsity prior assumption that ensures a high probability on the model parameters being close to zero rather than a positive probability of exactly being zero. It avoids introducing latent indictors in the model and the aforementioned complication in posterior computations. Thus, another potentially interesting extension of our work is to develop a multiresolution variable selection procedure using Bayesian shrinkage methods.

3.7 Appendix

3.7.1 Standard Posterior Computation Algorithm

We provide the details of the standard posterior computation algorithm in Section 3.2.3 which is implemented via a Gibbs sampler. The joint posterior distribution of all the parameters given the data is

$$\begin{aligned}&\pi(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{c}, \boldsymbol{\gamma}, \sigma^2_\beta, \eta_1, \eta_2, \xi_1, \xi_2 \mid \mathbf{S}, \mathbf{X}, \mathbf{y}) \\ &\propto \ \pi(\mathbf{y} \mid \mathbf{z})\pi(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{c}, \boldsymbol{\gamma}, \mathbf{S}, \mathbf{X})\pi(\boldsymbol{\beta} \mid \sigma^2_\beta)\pi(\boldsymbol{\alpha})\pi(\mathbf{c} \mid \eta_1, \xi_1)\left[\prod_{r=1}^{R} \pi(\boldsymbol{\gamma}_r \mid \eta_2, \xi_2)\right]\pi(\sigma^2_\beta)\pi(\boldsymbol{\eta})\pi(\boldsymbol{\xi})\end{aligned} \tag{3.21}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2)$ and $\boldsymbol{\xi} = (\xi_1, \xi_2)$.

In the Gibbs sampler, the sampling schemes are as follows.

Sampling scheme for z: for $i = 1, \ldots, n$, draw

$$[z_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{c}, \boldsymbol{\gamma}, y_i, \mathbf{S}, \mathbf{X}\ ] \sim y_i \mathrm{N}_{[0,+\infty)}(\widetilde{\mu}_i, 1) + (1 - y_i)\mathrm{N}_{(-\infty,0)}(\widetilde{\mu}_i, 1), \tag{3.22}$$

where $\mathrm{N}_{\mathcal{A}}(\mu, \Sigma)$ denote a normal distribution with mean μ and covariance Σ truncated on region $\mathcal{A}$, and $\widetilde{\mu}_i = \mathbf{S}_i\boldsymbol{\alpha} - \mathbf{X}_i \{\boldsymbol{\lambda} \circ \boldsymbol{\beta}\}$, where $\mathbf{S}_i, \mathbf{X}_i$ are row i for $\mathbf{S}, \mathbf{X}$.

Sampling scheme for $\boldsymbol{\alpha}$: draw

$$[\boldsymbol{\alpha} \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{c}, \boldsymbol{\gamma}, \mathbf{S}, \mathbf{X}] \sim \mathrm{N}(\widetilde{\boldsymbol{\mu}}_\alpha, \widetilde{\boldsymbol{\Sigma}}_\alpha), \tag{3.23}$$

where $\widetilde{\boldsymbol{\Sigma}}_\alpha = (\mathbf{S}'\mathbf{S} + \sigma_\alpha^{-2}\mathbf{I}_p)^{-1}$ and $\widetilde{\boldsymbol{\mu}}_\alpha = \widetilde{\boldsymbol{\Sigma}}_\alpha \left(\mathbf{z} - \mathbf{X}\left\{\boldsymbol{\lambda} \circ \boldsymbol{\beta}\right\}\right)\mathbf{S}$.

Sampling scheme for σ_β^2: draw

$$[\sigma_\beta^2 \mid \boldsymbol{\beta}] \sim \mathrm{IG}(a_\beta + V/2, b_\beta + (1/2)\sum_{r=1}^{R}\sum_{v=1}^{V_r}\beta_{rv}^2). \tag{3.24}$$

Sampling scheme for $\mathbf{c}$: for $r = 1, \ldots, R$, the full conditional of c_r is given by

$$\begin{aligned} &\pi(c_r \mid \boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\alpha}, \mathbf{c}_{-r}, \boldsymbol{\gamma}, \mathbf{S}, \mathbf{X}) \\ &\quad \propto \; \exp\left(\eta_1 \sum_{r=1}^{R} c_r + \xi_1 \sum_{r'=1}^{R} f_{r'r} I[c_{r'} = c_r]\right) \prod_{i=1}^{n} \phi\left(z_i - \mathbf{S}_i\boldsymbol{\alpha} - \mathbf{X}_i\left\{\boldsymbol{\lambda} \circ \boldsymbol{\beta}\right\}\right), \end{aligned} \tag{3.25}$$

where $\mathbf{c}_{-r} = (c_{r'}, r' \neq r)$.

Sampling scheme for $\boldsymbol{\gamma}$: for $r = 1, \ldots, R$ and $v = 1, \ldots, V_r$, the full conditional of γ_{rv} is given by

$$\begin{aligned} &\pi(\gamma_{rv} \mid \boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\alpha}, \mathbf{c}, \boldsymbol{\gamma}_{-rv}, \mathbf{S}, \mathbf{X}) \\ &\quad \propto \; \exp\left(\eta_2 \sum_{v=1}^{V_r} \gamma_{rv} + \xi_2 \sum_{v'=1}^{V_r} l_{rv'v} I[\gamma_{rv'} = \gamma_{rv}]\right) \prod_{i=1}^{n} \phi\left(z_i - \mathbf{S}_i\boldsymbol{\alpha} - \mathbf{X}_i\left\{\boldsymbol{\lambda} \circ \boldsymbol{\beta}\right\}\right), \end{aligned} \tag{3.26}$$

where $\boldsymbol{\gamma}_{-rv} = (\gamma_{st}, s \neq r \text{ or } t \neq v)$.

Sampling scheme for $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$: the parameters in Ising priors are updated using the auxiliary variable method by Møller et al. (2006).

Sampling scheme for $\boldsymbol{\beta}$: based on the full conditional (3.6), update $\boldsymbol{\beta}$ via a block update by (3.7).

3.7.2 SRS-MCMC Algorithm

The updating scheme for $\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}$, $\sigma_{\beta}^{2(k)}$, $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ follows the standard posterior computation algorithm in Appendix 3.7.1.

Sampling scheme for $\mathbf{c}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$:

- when $k = 1$, for $g = 1, \ldots, G^{(k)}; r = 1, \ldots, R; h = 1, \ldots, H_r^{(k)}$, the full conditionals of $c_g^{(k)}$ and $\gamma_{rh}^{(k)}$ are

$$\pi(c_g^{(k)} \mid \boldsymbol{\beta}^{(k)}, \mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \mathbf{c}_{-g}^{(k)}, \boldsymbol{\gamma}^{(k)}, \mathbf{S}, \mathbf{X}) \;\; \propto \;\; \prod_{i=1}^{n} \phi\left(z_i^{(k)} - \mathbf{S}_i \boldsymbol{\alpha}^{(k)} - \mathbf{X}_i \left\{\boldsymbol{\lambda}^{(k)} \circ \boldsymbol{\beta}^{(k)}\right\}\right); \tag{3.27}$$

$$\pi(\gamma_{rh}^{(k)} \mid \boldsymbol{\beta}^{(k)}, \mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \mathbf{c}^{(k)}, \boldsymbol{\gamma}_{r[-h]}^{(k)}, \boldsymbol{\gamma}_{[-r]}^{(k)}, \mathbf{S}, \mathbf{X}) \;\; \propto \;\; \prod_{i=1}^{n} \phi\left(z_i^{(k)} - \mathbf{S}_i \boldsymbol{\alpha}^{(k)} - \mathbf{X}_i \left\{\boldsymbol{\lambda}^{(k)} \circ \boldsymbol{\beta}^{(k)}\right\}\right); \tag{3.28}$$

- when $1 < k < K$, the sampling scheme is stated in Section 3.3.2 with the full conditional updates of $c_g^{(k)}$ and $\gamma_{rh}^{(k)}$ in the moving step following (3.27) and (3.28).
- when $k = K$, the sampling scheme is stated in Section 3.3.2 with the full conditional updates of $c_g^{(k)}$ and $\gamma_{rh}^{(k)}$ in the moving step following (3.25) and (3.26).

3.7.3 fastSRS-MCMC Algorithm

The updating scheme for $\mathbf{z}^{(k)}, \boldsymbol{\alpha}^{(k)}$, $\sigma_\beta^{2(k)}$, $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ follows the standard posterior computation algorithm in Appendix 3.7.1.

Sampling scheme for $\mathbf{c}^{(k)}, \boldsymbol{\gamma}^{(k)}, \widetilde{\boldsymbol{\beta}}^{(k)}$:

- when $k = 1$, the full conditional of $\mathbf{c}^{(k)}$ follows (3.27). For the full conditional of $(\boldsymbol{\gamma}^{(k)}, \widetilde{\boldsymbol{\beta}}^{(k)})$ in (3.20), $\pi(\gamma^{(k)}_{rh'} \mid \bullet) = \int \pi(\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'}, \boldsymbol{\gamma}^{(k)}_{rh'} \mid \bullet) \mathrm{d}\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'} \propto \phi(\mathbf{0}; \boldsymbol{\mu}^{(k)}_{rh'}, \boldsymbol{\Sigma}^{(k)}_{rh'})$, $\pi(\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'1} \mid \boldsymbol{\gamma}^{(k)}_{rh'}, \bullet) = \phi(\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'1}; \boldsymbol{\mu}^{(k)}_{rh'}, \boldsymbol{\Sigma}^{(k)}_{rh'})$ and $\pi(\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'0} \mid \boldsymbol{\gamma}^{(k)}_{rh'}, \bullet) = \delta_{\mathbf{0}}(\widetilde{\boldsymbol{\beta}}^{(k)}_{rh'0})$ with

$$\boldsymbol{\Sigma}^{(k)}_{rh'} = \left(\sigma_\beta^{-2(k)}\mathbf{I} + \mathbf{X}^\top_{rh'}\mathbf{X}_{rh'}\right)^{-1}; \qquad \boldsymbol{\mu}^{(k)}_{rh'} = \boldsymbol{\Sigma}^{(k)}_{rh'}\mathbf{X}^\top_{rh'}\left(\mathbf{z}^{(k)} - \mathbf{S}\boldsymbol{\alpha}^{(k)}\right), \quad (3.29)$$

  where $\mathbf{X}_{rh'} = (\mathbf{x}_{rv}, a^{(k)}_{rvh} = 1; \gamma^{(k)}_{rh} = 1; \widetilde{a}^{(k)}_{rhh'} = 1)$.

- when $1 < k < K$, the sampling scheme is stated in Section 3.3.3.

- when $k = K$, the sampling scheme is stated in Section 3.3.3 with the full conditional updates in the moving step following of $c^{(k)}_g$ following (3.25), and $(\boldsymbol{\gamma}^{(k)}, \widetilde{\boldsymbol{\beta}}^{(k)})$ following (3.20) with an alternative $\pi(\gamma^{(k)}_{rh'} \mid \bullet) = \pi(\gamma_{rv} \mid \bullet) \propto$ $\exp\left(\eta_2 \sum_{v=1}^{V_r} \gamma_{rv} + \xi_2 \sum_{v'=1}^{V_r} l_{rv'v} I[\gamma_{rv'} = \gamma_{rv}]\right)$
  $\phi(\mathbf{0}; \boldsymbol{\mu}^{(k)}_{rh'}, \boldsymbol{\Sigma}^{(k)}_{rh'})$.

Chapter 4

A Bayesian nonparametric mixture model for selecting genes and gene sub-networks

4.1 Introduction

In high-throughput data analysis, selecting informative features from tens of thousands of measured features is a difficult problem. Incorporating pathway or network information into the analysis has been a promising approach, and the network could contain information such as protein interaction, transcriptional regulation, enzymatic reaction, and signal transduction etc (Cerami et al., 2011).

With the primary goal to identify either the important pathways or the genes that are strongly associated with clinical outcomes of interest, some methods are developed using the available network topology with incorporating the gene-pathway relationships or gene network information into a parametric/regression model. For example, there are a series of works (Wei and Li, 2007, 2008; Wei and Pan, 2010) that model gene-network using a Discrete- or Gaussian-Markov random field (DMRF or GMRF). Li and Li (2008) and Pan et al. (2010) used the gene network to build penalties in a regression model for gene pathway selection. Ma et al. (2010) incorporated the gene co-expression network in identification of caner prognosis markers using a survival model. Li and Zhang (2010) and Stingo et al. (2011) developed Bayesian linear regression models using MRF priors or Ising priors that capture the dependent structure of transcription factors or the gene network/pathway. Recently, Jacob et al. (2012) proposed a powerful graph-structured two-sample test to detect differentially expressed genes.

Although regression models are widely used for the selection of the gene subnetwork that is associated with an outcome variable, in some situations, the question of interest is to study the expressional behavior of genes, e.g. periodicity, without an outcome variable. In other situations, the experimental design is more complex than simple case-control. For example, some gene expression studies involve longitudinal/-

functional measurements for which the parametric models (Leng and Müller, 2006; Zhou et al., 2010; Breeze et al., 2011) or the multivariate testing procedure (Jacob et al., 2012) may not be applicable without a major modification. A straightforward approach to this problem is to perform large-scale simultaneous hypothesis testing on gene behavior. A set of genes can be selected based on the testing statistics or p-values, where a correct choice of a null distribution for those correlated testing statistics (Efron, 2004, 2010) should be used. However, this approach ignores the gene network information that is useful to identify the subnetwork of genes with the particular expressional behavior. Due to the diverse behavior of neighboring genes on the network, it is generally believed that genes in close proximity on a network are likely to have joint effects on biological/medical outcomes or have similar expressional behavior. This motivates the needs of analyzing the large scale testing statistics or statistical estimates incorporating the network information. Another motivation is that a linear regression or parametric model of gene expression levels might not be suitable in some cases. For example, we may be interested in finding subnetworks of genes that have nonlinear relations with an outcome without specifying a parametric form. To address these problems, a simple framework can be adopted. First, a certain statistic is computed for each feature without considering the network structure. The statistic can come from a test of nonlinear association, a test of periodic behavior, or a certain regression model. After obtaining the feature-level statistics, a mixture model that takes into account the network structure can be used to select interesting features/subnetworks.

To mitigate problems of the current methods, we propose a Bayesian nonparametric mixture model for large scale statistics incorporating network information. Specifically, the gene specific statistics are assumed to fall into two classes: "unselected" and "selected", corresponding to whether the statistics are generated from a null distribution, with prior probabilities p_0 and $p_1 = 1 - p_0$. A statistic has density either $f_0(r)$ or $f_1(r)$ depending on its class, where $f_0(r)$ represents "unselected" density and $f_1(r)$ represents "selected" density. Thus, without knowing the classes, the statistics follow a mixture distribution:

$$p_0 f_0(r) + p_1 f_1(r). \tag{4.1}$$

As suggested by Efron (2010), it is reasonable to assume statistics are normally distributed. This justifies the use of a Dirichlet process mixture (DPM) of normal distributions to estimate both $f_0(x)$ and $f_1(x)$. Note that different from Wei and Pan (2012), our model does not assume that f_0 and f_1 directly take the form of a normal density function. The DPM model has been discussed extensively and widely used in Bayesian statistics (Antoniak, 1974; Escobar, 1994; Escobar and West, 1995; Müller and Quintana, 2004; Dunson, 2010), due to the availability of efficient computational techniques (Neal, 2000; Ishwaran and James, 2001; Wang and Dunson, 2011) and the nonparametric nature with good performance on density estimation. The DPM has been extended to make inference for differential gene expression (Do et al., 2005) and estimate positive false discovery rates (Tang et al., 2007) but without incorporating the network information. In our model, we assign an Ising prior (Li and Zhang, 2010) to class labels of all genes according to the dependent structure of the network. As discussed previously, the class label only takes two values: "selected" and "unselected", while a DPM model is equivalent to an infinity mixture model (Neal, 2000; Ishwaran and James, 2001, 2002), based on which we develop a posterior computation algorithm. Our method selects genes and gene subnetworks automatically during the model fitting. To reduce the computational cost, we propose two fast computation algorithms that approximate the posterior distribution either using finite mixture models or guided by a standard DPM model fitting, for which we develop a hierarchical ordered distribution clustering (HODC) algorithm. It essentially performs clustering on ordered density functions. The fast computation algorithms can be tailored from any routine algorithms for the standard DPM model and combined with the HODC algorithm. Also, we suggest two approaches to choosing the hyper-parameters in the model.

4.2 The Model

Let n be the total number of genes in our analysis. For $i = 1, \ldots, n$, let r_i denote a statistic for gene i. It represents either a functional behavior or the association with a clinical outcome. For the association analysis, it is common to have an outcome Y and a gene expression profile X_i for each gene, i. As an alternative to a regression model, we can produce statistics for each gene. i.e. $r_i = s(X_i, Y)$, where $s(\cdot, \cdot)$ can be a covariance function or other dependence test statistics. For a large scale testing problem, we usually obtain p-values, $p_1, \ldots, p_n$, which can be transformed to normally distributed statistics, i.e. $r_i = -\Phi^{-1}(p_i)$, where $\Phi(r)$ denotes the cumulative distribution function for the standard normal distribution. This transformation is a monotone transformation and it ensures the "selected" genes have a larger value of r_i. Let z_i be the class label for gene selection, where $z_i = 1$ if gene i is selected, $z_i = 0$ is unselected. For $i, j = 1, \ldots, n$, let c_{ij} denote the gene network configuration, where $c_{ij} = 1$ if gene i and j are connected, $c_{ij} = 0$ otherwise. Write $\mathbf{r} = (r_1, \ldots, r_n)'$, $\mathbf{z} = (z_1, \ldots, z_n)'$ and $\mathbf{C} = (c_{ij})$. In our model, $\mathbf{r}$ and $\mathbf{C}$ are observed data, $\mathbf{z}$ is a latent vector of our primary interest.

4.2.1 A network based DPM model for gene selection

As suggested by Efron (2010), we assume r_i's are normally distributed. Let $\mathrm{N}(\mu, \sigma^2)$ denote a normal distribution with mean μ and standard deviation σ. Let $\mathcal{DP}(G, \alpha)$ represent a Dirichlet process with base measure G and scalar precision α. Given the class label $\mathbf{z}$, we consider the following DPM model: for $i = 1, 2, \ldots, n$ and $k = 0, 1$,

$$\begin{aligned} [r_i \mid \mu_i, \sigma_i^2\ ] &\sim \mathrm{N}(\mu_i, \sigma_i^2), \qquad (4.2) \\ [(\mu_i, \sigma_i^2) \mid z_i = k, G_k\ ] &\sim G_k, \\ G_k &\sim \mathcal{DP}[G_{0k}, \tau_k], \end{aligned}$$

where μ_i and σ_i^2 are latent mean and variance parameters for each r_i. The random measure G_k and the base measure G_{0k} are both defined on $(-\infty, +\infty) \times (0, +\infty)$. We specify $G_{0k} = \mathrm{N}(\gamma_k, \xi_k^2) \times \mathrm{IG}(\alpha_k, \beta_k)$, where $\mathrm{IG}(\alpha, \beta)$ denotes an inverse gamma distribution with shape α and scale β. Note that given latent parameters μ_i, σ_i^2, the statistic r_i is conditionally independent of z_i. By integrating out (μ_i, σ_i^2), we build the conditional density of r_i given $z_i = k$ in (4.1), i.e.

$$f_k(r) = \int \pi(r \mid \boldsymbol{\theta}) dG_k(\boldsymbol{\theta}), \quad \pi(r \mid \boldsymbol{\theta}) = \frac{1}{\sigma}\phi\left(\frac{r-\mu}{\sigma}\right), \qquad (4.3)$$

where $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $\phi(r)$ is the standard Gausian density function. This provides a Bayesian nonparametric construction of $f_k(r)$.

To incorporate the network structure, we assign a weighted Ising prior to $\mathbf{z}$:

$$\pi(\mathbf{z} \mid \boldsymbol{\pi}, \boldsymbol{\varrho}, \boldsymbol{\omega}, \mathbf{C}) \propto \exp\left[\sum_{i=1}^{n}\left(\widetilde{\omega}_i \log(\pi_{z_i}) + \varrho_{z_i}\sum_{j\neq i}\omega_j c_{ij} I[z_i = z_j]\right)\right], \qquad (4.4)$$

where $\boldsymbol{\pi} = (\pi_0, \pi_1)$ with $0 < \pi_1 = 1 - \pi_0 < 1$, $\boldsymbol{\varrho} = (\varrho_0, \varrho_1)$ with $\varrho_k > 0$ for $k = 0, 1$,

$\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)'$ with $\omega_i > 0$ for $i = 1, \ldots, n$, and $\widetilde{\omega}_i = \sum_{j=1}^n c_{ij}\omega_j / \sum_{j=1}^n c_{ij}$. The indicator function $I[\mathcal{A}] = 1$ if event $\mathcal{A}$ is true, $I[\mathcal{A}] = 0$, otherwise. The parameter $\boldsymbol{\pi}$ controls the sparsity of $\mathbf{z}$, and the parameter $\boldsymbol{\varrho}$ characterizes the smoothness of $\mathbf{z}$ over the network. For each gene i, a weight ω_i is introduced to control the information inflow to gene i from other connected genes, which can adjust the prior distribution of z_i based on biologically meaningful knowledge, if any. The term $\widetilde{\omega}_i$ is introduced to balance the contribution from $\boldsymbol{\pi}$ and $\boldsymbol{\varrho}$ to the prior probability of $\mathbf{z}$. When $\boldsymbol{\varrho} = (0, 0)$ and $\boldsymbol{\omega} = (1, \ldots, 1)'$, the latent class labels z_i's are independent identically distributed as Bernoulli with parameter π_1.

4.2.2 Model Representations

As discussed by Neal (2000), the DPM models can also be obtained by taking the limit as the number of components goes to infinity. With a similar fashion, we construct an equivalent model representation of (4.2) for efficient posterior computations. Let Discrete$(\mathbf{a}, \mathbf{b})$ denote a discrete distribution taking values in $\mathbf{a} = (a_1, \ldots, a_L)'$ with probability $\mathbf{b} = (b_1, \ldots, b_L)'$, i.e. if $\xi \sim$ Discrete$(\mathbf{a}, \mathbf{b})$, then $\Pr(\xi = a_l) = b_l$, for $l = 1, \ldots, L$. Let Dirichlet$(\boldsymbol{\alpha})$ denote a Dirichlet distribution with parameter $\boldsymbol{\alpha}$. Let L_k, for $k = 0, 1$, represent the number of components for density $f_k(r)$. We define the index sets $\mathbf{a}_0 = (-L_0 + 1, -L_0 + 2, \ldots, 0)$ and $\mathbf{a}_1 = (1, 2, \ldots, L_1)$. Let $\mathbf{q}_0 = (q_{-L_0+1}, q_{-L_0+2}, \ldots, q_0)$ and $\mathbf{q}_1 = (q_1, \ldots, q_{L_1})$ with $\sum_{g \in \mathbf{a}_k} q_g = 1$. Let $\mathbf{1}_n = (\underbrace{1, \ldots, 1}_{n})$. Then model (4.2) is equivalent to the following model, as $L_0 \to \infty$ and $L_1 \to \infty$,

$$\begin{aligned} [r_i \mid g_i, \widetilde{\boldsymbol{\theta}}\,] &\overset{i.i.d.}{\sim} \mathrm{N}(\widetilde{\mu}_{g_i}, \widetilde{\sigma}^2_{g_i}), \\ [g_i \mid z_i = k, \mathbf{q}_k\,] &\overset{i.i.d.}{\sim} \mathrm{Discrete}(\mathbf{a}_k, \mathbf{q}_k), \\ \widetilde{\boldsymbol{\theta}}_g &\sim G_{0k}, \quad \text{for } g \in \mathbf{a}_k, \\ \mathbf{q}_k &\sim \mathrm{Dirichlet}(\tau_k \mathbf{1}_{L_k}/L_k), \end{aligned} \tag{4.5}$$

where $\widetilde{\boldsymbol{\theta}} = \{\widetilde{\boldsymbol{\theta}}_g\}_{g \in \mathbf{a}_0 \cup \mathbf{a}_1}$ and $\widetilde{\boldsymbol{\theta}}_g = (\widetilde{\mu}_g, \widetilde{\sigma}_g^2)$. The index g_i indicates the latent class associated with each data point r_i. Write $\mathbf{g} = (g_1, \dots, g_n)$ and $\mathbf{z} = (z_1, \dots, z_n)$. For each class, g, the parameter $\widetilde{\boldsymbol{\theta}}_c$ determines the distribution of r_i from that class. The conditional distributions of g_i and $\widetilde{\boldsymbol{\theta}}_{g_i}$ given $z_i = 0$ and $z_i = 1$ are different. Based on model (4.5), the conditional density of $f_k(r)$ in (4.3) becomes

$$f_k(r) = \sum_{g \in \mathbf{a}_k} \frac{q_g}{\widetilde{\sigma}_g} \phi \left(\frac{r - \widetilde{\mu}_g}{\widetilde{\sigma}_g} \right). \tag{4.6}$$

This further implies that given L_0 and L_1, the marginal distribution of r_i also has a form of finite mixture normals. i.e.,

$$\pi(r) = \sum_{k=0}^{1} p_k f_k(r) = \sum_{g=-L_0+1}^{L_1} \frac{\widetilde{q}_g}{\widetilde{\sigma}_g} \phi \left(\frac{r - \widetilde{\mu}_g}{\widetilde{\sigma}_g} \right), \tag{4.7}$$

where $\widetilde{q}_g = p_0 q_g$ if $g \leq 0$, $\widetilde{q}_g = p_1 q_g$, otherwise.

Model (4.5) is not identifiable for z_i in the sense that if we switch the gene selection class label "0" and "1", the marginal distribution of r_i (4.7) is unchanged. Without loss of generality, we assume that the "selected gene" should be more likely to have large statistics compared to the "unselected genes". Thus, we impose an order restriction on the parameter $\widetilde{\boldsymbol{\theta}}$, for $g = -L_0 + 1, \dots, L_1$,

$$\widetilde{\mu}_g < \widetilde{\mu}_{g+1}. \tag{4.8}$$

This also sorts out the non-identifiability of parameter $\widetilde{\boldsymbol{\theta}}$. In many cases, the functional behaviors of some genes are strongly evident from prior biological knowledge. Whether or not those genes are selected is not necessarily determined by other genes in the network. Those genes are likely to be the hubs of the networks, thus the determination of the status of these genes might help select genes in their neighborhood.

This suggests that it is reasonable to pre-select a small amount of genes that can be surely elictied by biologists from their experience and knowledge. We refer to them as "surely selected" (SS) genes. These genes are usually associated with very large statistics. We evaluate the performance via the simulation studies in Section 4.4.2.

4.2.3 Posterior Computation

In model (4.5), given L_0 and L_1, we have the full conditional distribution of $g_i = g$ and $z_i = k$ given $\mathbf{g}_{-i} = (g_1, \ldots, g_{i-1}, g_{i+1}, \ldots g_n)$, $\mathbf{z}_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$ and data $\mathbf{r}$:

$$\begin{aligned} &\pi(g_i = g, z_i = k \mid \mathbf{g}_{-i}, \mathbf{z}_{-i}, \mathbf{r}, \widetilde{\boldsymbol{\theta}}) \\ &\quad \propto \frac{1}{\widetilde{\sigma}_g}\phi\left(\frac{r_i - \widetilde{\mu}_g}{\widetilde{\sigma}_g}\right)\frac{n_{-ig} + \tau_k/L_k}{\tau_k + m_k - 1}\exp\left(\widetilde{\omega}_i \log(\pi_k) + \varrho_k \sum_{j \neq i} \omega_j c_{ij} I[z_j = k]\right), \end{aligned} \tag{4.9}$$

where $m_k = \sum_{i=1}^{n} I[z_i = k]$ is the number of genes in class k and $n_{-ig} = \sum_{j \neq i} I[g_j = g]$ represents the number of g_j for $j \neq i$ that are equal to g.

As $L_0 \to \infty$ and $L_1 \to \infty$, if $(g, k) = (g_j, z_j)$ for some $j \neq i$, then

$$\begin{aligned} &\pi(g_i = g, z_i = k \mid \mathbf{g}_{-i}, \mathbf{z}_{-i}, \mathbf{r}, \widetilde{\boldsymbol{\theta}}) \\ &\quad \propto \frac{n_{-ig}}{\tau_k + m_k - 1}\exp\left(\widetilde{\omega}_i \log(\pi_k) + \varrho_k \sum_{j \neq i} \omega_j c_{ij} I[z_j = k]\right)\frac{1}{\widetilde{\sigma}_g}\phi\left(\frac{r_i - \widetilde{\mu}_g}{\widetilde{\sigma}_g}\right), \end{aligned} \tag{4.10}$$

and

$$\begin{aligned} &\pi(g_i \neq g_j, z_i \neq z_j, \text{for all } j \neq i \mid \mathbf{g}_{-i}, \mathbf{z}_{-i}, \mathbf{r}, \widetilde{\boldsymbol{\theta}}) \\ &\quad \propto \frac{\tau_k}{\tau_k + m_k - 1}\exp\left(\widetilde{\omega}_i \log(\pi_k) + \varrho_k \sum_{j \neq i} \omega_j c_{ij} I[z_j = k]\right) \\ &\quad \times \frac{\Gamma(\alpha_k + 1/2)\beta_k^{\alpha_k}}{\sqrt{2\pi}\Gamma(\alpha_k)\xi_k} \int \phi\left(\frac{\mu - \gamma_k}{\xi_k}\right)\left(\beta_k + \frac{1}{2}(r_i - \mu)^2\right)^{-(\alpha_k + 1/2)} d\mu, \end{aligned} \tag{4.11}$$

where the integral can be efficiently computed by the Gaussian quadrature method in practice.

The full conditionals of $\widetilde{\mu}_g$ and $\widetilde{\sigma}_g^2$, for $g \in \{g_1, \dots, g_n\}$ are given by,

$$[\widetilde{\mu}_g \mid \widetilde{\sigma}_g^2, \mathbf{r}\ ] \sim \mathrm{N}\left(\frac{\widetilde{\sigma}_g^2\gamma_k + \xi_k^2 \sum_{i:g_i=g} r_i}{\widetilde{\sigma}_g^2 + \xi_k^2 n_g}, \frac{\widetilde{\sigma}_g^2\xi_k^2}{\widetilde{\sigma}_g^2 + \xi_k^2 n_g}\right), \tag{4.12}$$

$$[\widetilde{\sigma}_g^2 \mid \widetilde{\mu}_g, \mathbf{r}\ ] \sim \mathrm{IG}\left(\alpha_k + \frac{n_g}{2}, \beta_k + \frac{1}{2}\sum_{i:g_i=g}(r_i - \widetilde{\mu}_g)^2\right), \tag{4.13}$$

where $k = I[g > 0]$ and $n_g = \sum_{i=1}^n I[g_i = g]$. We summarize this algorithm in Appendix 4.6.2 and refer to it as NET-DPM-1. It is computationally intensive when n is very large. To mitigate this problem, we propose two fast algorithms to fit finite mixture models (FMM) with appropriate choices of the number of components.

4.2.4 Fast Computation Algorithms

FMM Approximation

When L_1 and L_0 fit the data well, we can accurately approximate the infinite mixture model (4.2) by the FMM (4.5). Given a fixed L_0 and L_1, it is straightforward to perform posterior computation for model (4.5) based on (4.9). We refer to this algorithm as NET-DPM-2 (see Appendix 4.6.2 for details). This algorithm does not change the dimension of $\widetilde{\boldsymbol{\theta}}$ over iterations. In this sense, it simplifies the computation. Also, in order to keep computation efficient, we search for smaller values of L_0 and L_1 which fit the data well. This can be achieved under the guidance of a DPM density fitting for which we introduce an algorithm in the next section.

Hierarchical Ordered Density Clustering

Without using the network information, a DPM model fitting on data $\mathbf{r}$ provides an approximation to the marginal density (4.7). It generates posterior samples for mixture densities, where the mean number of components should be close to $L_0 + L_1$. Let us focus on one sample. Suppose $L_0 + L_1$ is equal to the number of components in this sample. To further obtain an estimate of L_0 and L_1 for this sample, we need to partition the $L_0 + L_1$ components into two classes. Thus, we propose an algorithm to cluster a set of ordered densities. We call it hierarchical ordered density clustering (HODC). Here, the density order is determined by the mean location of that density. For example, a set of Gaussian density functions are sorted according to their mean parameters. Similar to the classical hierarchical clustering analysis, we define a distance metric of density functions:

$$d(f, f') = \int_{-\infty}^{+\infty} [f(x) - f'(x)]^2 dx, \tag{4.14}$$

where f and f' are two univariate density functions. Let $\mathcal{P} = \{(\widehat{\mu}_g, \widehat{\sigma}_g^2, \widehat{p}_g)\}_{g=1}^{L_0+L_1}$ denote parameters for $L_0 + L_1$ Gaussian densities, where $\widehat{\mu}_g < \widehat{\mu}_{g+1}, g = 1, 2, \ldots, L_0 + L_1 - 1$. This is the input data to the HODC algorithm totally consisting of $L_0 + L_1 - 2$ steps. At the m step, there are $L_0 + L_1 - m$ clusters of densities and let $\mathbf{s}_l^{(m)}$, for $l = 1, \ldots, L_0 + L_1 - m$, denote the density indices in cluster l. For simple, we define

$$\widetilde{\phi}(r; \mathbf{s}, \mathcal{P}) = \sum_{g \in \mathbf{s}} \frac{\widehat{p}_g}{\widehat{\sigma}_g} \phi\left(\frac{r - \widehat{\mu}_g}{\widehat{\sigma}_g}\right) \Bigg/ \sum_{g \in \mathbf{s}} \widehat{p}_g, \tag{4.15}$$

which represents a mixture of Gaussian densities, where the components indexed by $\mathbf{s}$ are a subset of $\{\phi[(r - \hat{\mu}_g)/\hat{\sigma}_g]/\widehat{\sigma}_g\}_{g=1}^{L_0+L_1}$.

HODC:

Input: Parameters for a mixture of Gaussian densities, i.e, $\mathcal{P}$.

Initialization: Set $m = 0$ and $\mathbf{s}_l^{(0)} = \{l\}$, for $l = 1, 2, \ldots, L_0 + L_1$;

Repeat the following steps until $m = L_0 + L_1 - 2$:

Step 1: Find

$$l^{(m)} = \arg\min_{l} \; d\left(\widetilde{\phi}(\cdot; \mathbf{s}_l^{(m)}, \mathcal{P}), \widetilde{\phi}(\cdot; \mathbf{s}_{l+1}^{(m)}, \mathcal{P})\right).$$

Step 2: For $l = 1, 2, \ldots, L_0 + L_1 - m - 1$, set

$$\mathbf{s}_l^{(m+1)} = \begin{cases} \mathbf{s}_l^{(m)} & \text{If } l < l^{(m)}, \\ \mathbf{s}_l^{(m)} \cup \mathbf{s}_{l+1}^{(m)} & \text{If } l = l^{(m)}, \\ \mathbf{s}_{l+1}^{(m)} & \text{If } l > l^{(m)}. \end{cases}$$

Step 3: Set $m = m + 1$.

Output: $\{\mathbf{s}_l^{(m)}\}_{l=1}^{L_0+L_1-m}$ for $m = 1, 2, \ldots, L_0 + L_1 - 2$.

Fig. 4.1 illustrates the HODC algorithm. The algorithm stops when $m = L_0 + L_1 - 2$, where the ordered density components are partitionned into two classes indexed by $\mathbf{s}_1^{(m)}$ and $\mathbf{s}_2^{(m)}$. This suggests that the number of indices in $\mathbf{s}_{k+1}^{(m)}$, denoted by $|\mathbf{s}_{k+1}^{(m)}|$, is an estimate for L_k in model (4.5). By running the HODC, we can obtain one L_k estimate for each posterior sample generated from a DPM fitting. We take the average of L_k estimates over all the posterior samples as the input of NET-DPM-2. The HODC also provides an approximation to $f_k(r)$ in (4.6), i.e. $\widetilde{\phi}(r; \mathbf{s}_{k+1}^{(m)}, \mathcal{P})$. This implies that we can further simplify the computation with the algorithm in the following section.

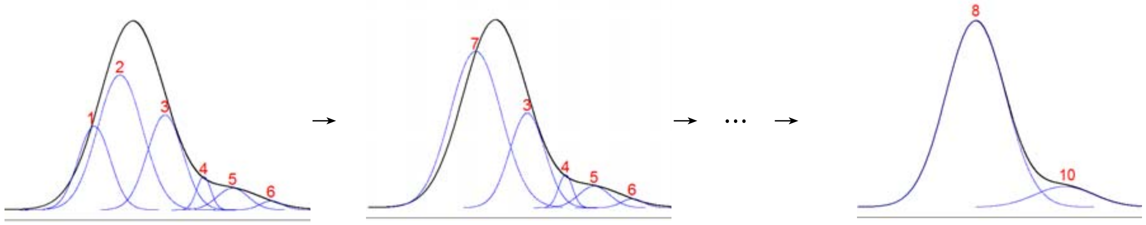


Figure 4.1: An illustration of the HODC algorithm for six density components: the HODC starts with clustering densities 1 and 2 as a mixture density labeled as 7, since the "distance" between 1 and 2 is shorter than all other adjacent density pairs. Then the HODC computes the "distance" between densities 3 and 7, densities 3 and 4, . . . , to proceed the clustering. Following this procedure, the HODC ends up with clustering densities 1,2,3 as a mixture density (labeled as 8) and 4,5,6 as another mixture density (labeled as 10).

FMM guided by a DPM model fitting

From a DPM model fitting, we obtain V posterior samples of the parameters for the marginal density of $\mathbf{r}$. We denote them as $\mathcal{P}_v = \{(\widehat{\mu}_{vg}, \widehat{\sigma}^2_{vg}, \widehat{p}_{vg})\}_{g=1}^{L_{v0}+L_{v1}}$, for $v = 1, 2, \ldots, V$. For each $\mathcal{P}_v$, the HODC algorithm partitions $L_{v0} + L_{v1}$ components into two classes, where the class-specific components are indexed by $\mathbf{a}_{v,0}$ and $\mathbf{a}_{v,1}$. This leads to V approximations of $f_k(r)$, i.e. $\widetilde{\phi}(r; \mathbf{a}_{v,k}, \mathcal{P}_v)$. Given $f_k(r)$, our proposed gene selection model reduces to

$$[r_i \mid z_i = k\ ] \overset{i.i.d.}{\sim} f_k(r), \tag{4.16}$$

for $i = 1, 2, \ldots, n$ and $k = 0, 1$, and $\mathbf{z}$ follows (4.4). To make inference on the posterior distribution of $\mathbf{z}$ by combining all V approximations of $f_k(r)$, we consider

$$\pi(\mathbf{z} \mid \mathbf{r}) \approx \frac{1}{V} \sum_{v=1}^{V} \pi\left(\mathbf{z} \mid \mathbf{r}, \widetilde{\boldsymbol{\phi}}_v\right), \tag{4.17}$$

where $\widetilde{\boldsymbol{\phi}}_v = \{\widetilde{\phi}(r; \mathbf{a}_{v,0}, \mathcal{P}_v), \widetilde{\phi}(r; \mathbf{a}_{v,1}, \mathcal{P}_v)\}$. For each v, the full conditional of z_i is given by

$$\begin{aligned} &\pi(z_i = k \mid \mathbf{z}_{-i}, \mathbf{r}, \widetilde{\boldsymbol{\phi}}_v) \\ &\quad \propto \widetilde{\phi}(r_i; \mathbf{a}_{v,k}, \mathcal{P}_v) \exp\left(\widetilde{\omega}_i \log(\pi_k) + \varrho_k \sum_{j \neq i} \omega_j c_{ij} I[z_j = k]\right). \end{aligned} \tag{4.18}$$

We refer to this algorithm as NET-DPM-3 (see Appendix 4.6.2 for details). It is extremely fast with a moderate V. Since the marginal density is estimated without using the network information, it might introduce bias on the distribution of z_i and underestimate the variability of z_i. From our experience, those issues do not affect the selection accuracy much. Some examples are provided in Section 4.4.

4.2.5 The choice of hyper-parameters

To proceed NET-DPMs, we need to specify the hyper-parameters $\boldsymbol{\pi}$, $\boldsymbol{\varrho}$ and $\boldsymbol{\omega}$ in (4.4). We assume that $\boldsymbol{\omega}$ is pre-specified according to biological information. In this paper, we choose equal weight, i.e. $\boldsymbol{\omega} = \mathbf{1}_n$ without incoorporating any biological prior knowledge. We suggest two approaches to choosing $\boldsymbol{\pi}$ and $\boldsymbol{\varrho}$: 1) we assign hyper-priors on $\boldsymbol{\pi}$ and $\boldsymbol{\varrho}$ and make posterior inference; 2) for a set of possible choices of $\boldsymbol{\pi}$ and $\boldsymbol{\varrho}$, we employ the Bayesian model averaging. The details are provided in Appendix 4.6.3.

4.3 Application

To demonstrate the behavior of our method, we apply the proposed method to the analysis of the Spellman yeast cell cycle microarray dataset (Spellman et al., 1998) as introduced in Section 1.1.3. There is no outcome variable in the cell-cycle dataset. In this demonstration we focus on the selection of genes with periodic behavior in light of the network. It is known that such genes show different phase shifts along the cell cycle and may not be correlated with each other (Yu, 2010). We first perform the Fisher's exact G test for periodicity (Wichert et al., 2004) for each gene. We then transform the p-values to normal quantiles, $r_i = \Phi^{-1}(p_i)$ for gene i. We apply the fully Bayesian inference (NET-DPM-1), one fast computation approach (NET-DPM-3) and the standard DPM model fitting (STD-DPM) to this dataset. For the NET-DPM-1, set $\tau_0 = 10$, $\tau_1 = 2$; following the results by STD-DPM, set $\gamma_k = \overline{\mu}_k, \xi_k^2 = \overline{\sigma}_k^2, \beta_k = 10, \alpha_k = \beta_k/\xi_k^2 + 1$ with $k = 0, 1$, where $\{\overline{\mu}_k\}$ and $\{\overline{\sigma}_k^2\}$ are preliminary estimations by the STD-DPM. We also conduct a sensitivity analysis for the hyper-parameters specification (Appendix 4.6.4) to verify the robustness of the proposed methods. For both methods, the choices of π_0 and $\boldsymbol{\varrho}$ for the model averaging algorithm are $(0.75, 0.8, 0.85, 0.9)$ and $(0.5, 1, 5, 10, 15) \times (0.5, 1, 5, 10, 15)$ with restriction $\varrho_0 < \varrho_1$. We run all the algorithms 5,000 iterations with 2,000 burn-in. In this article, as discussed in Barbieri and Berger (2004), a cutoff 0.5 for the marginal posterior probability of z_i is taken to determine whether gene i is selected or not. The standard DPM fitting is obtained by an R package: `DPpackage` and all the proposed algorithms are implemented in R.

Table 4.1 presents the gene selection results based on three methods in a two-by-two table format. The number of the "selected" genes by the NET-DPM-1, the NET-DPM-3 and the STD-DPM are 201, 216 and 114, respectively. The summation of the diagonal elements of the table comparing the NET-DPM-3 and the NET-DPM-1 is larger than that for NET-DPM-3 and the STD-DPM. This indicates a stronger agreement between the two algorithms for NET-DPM.

We focus our discussion on the NET-DPM-3 results. After removing all unselected genes, as well as selected genes not connected to any other selected genes, 163 of the

		NET-DPM-1		STD-DPM	
		Selected	Unselected	Selected	Unselected
NET-DPM-3	Selected	170	46	100	116
	Unselected	31	1784	14	1801

Table 4.1: The genes selection results by the three methods for the cell cycle dataset

216 genes fall into 11 subnetworks. Of the 11 subnetworks, 10 are very small, each containing 5 or less genes. The remaining subnetwork contains 135 genes. Considering the purpose of the study is to find genes with periodic behavior, and most such genes are functionally related and regulated by the cell cycle process, this result is expected. We present the subnetwork in Fig. 4.2. Sixty-one of the 135 genes belong to the mitotic cell cycle process based on gene ontology (Ashburner et al., 2000). The yeast mitotic cell cycle can be roughly divided into the M phase and the interphase, which contains S and G phases (Ashburner et al., 2000). We do not further divide the interphase because the number of genes annotated to its descendant nodes are small. Among the 135 genes, 45 are annotated to the M phase, and 21 are annotated to the interphase. By coloring the M phase genes in red, the interphase genes in blue, and the genes annotated to both phases in green, we see that the majority of the selected M phase genes are clustered on the subnetwork, while the selected interphase genes are somewhat scattered with 7 falling into a small but tight cluster.

We show part of the subnetwork detected by the NET-DPM-3 with the coresponding one under the STD-DPM in Fig. 4.3, where the genes that are linked by a dashed line are connected to other genes that are not shown in the figure. In this subnetwork, the gene selection results by the NET-DPM-1 agrees with the NET-DPM-3 except only one gene “YML064” for which the NET-DPM-1 does not select it with probability 0.478 while the NET-DPM-3 selects it with probability 0.687. This implies that both methods provide large uncertainty on this gene. Comparing the top-panel (our method, NET-DPM-3) and bottom-panel (STD-DPM), we observe a number of

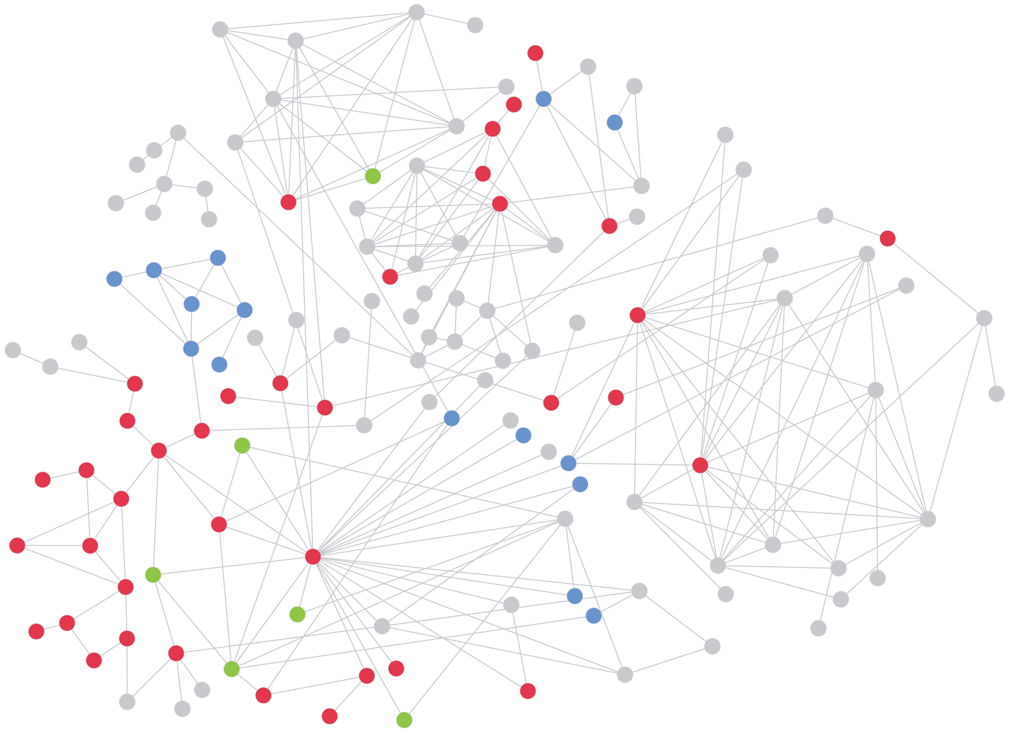


Figure 4.2: A subnetwork composed of genes with periodic behavior. The subnetwork consists of 135 genes. Red nodes: genes functionally involved in the M-phase of cell cycle; blue node: genes functionally involved in the interphase of cell cycle; green nodes: genes functionally involved in both M and interphase of cell cycle.

genes selected by NET-DPM but not by STD-DPM, and almost all such genes are cell cycle-related (denoted by a star by the ORF name). Examples include YAL041W (CLS4), which is required for the establishment and maintenance of polarity and critical in bud formation (Chenevert et al., 1994; Cherry et al., 2012). The gene only shows moderate periodic behavior, as denoted by the color of the node. However due to its links to other genes that have strong periodic behavior, it is selected by our method as an interesting gene. Another example is YFL008W (SMC1). It is a subunit of the cohesion complex, which is essential in sister chromatid cohesion of mitosis and meiosis. The complex is also involved in double-strand DNA break repair (Strunnikov and Jessberger, 1999; Cherry et al., 2012). Similar to CLS4, the periodic behavior of SMC1 is not strong enough. It is only selected when the information is borrowed from linked genes that are functionally related and show strong periodic behavior. A number of other cell cycle-related genes in Fig. 4.3 are in similar situation, e.g. YBR106W, YDR052C, YJL157C, YGL003C, and YMR076C. These examples clearly show the benefit of utilizing the biological information stored in the network structure.

To assess the functional relevance of the selected genes globally, we resort to mapping the genes onto gene ontology biological processes (Ashburner et al., 2000). We limit our search to the GO Slim terms using the mapper of the Saccharomyces Genome Database (Cherry et al., 2012). The full result is listed in the supplementary file. Clearly the over-represented GO Slim terms are centered around cell-cycle. Here we discuss some GO terms that are non-redundant. Among the 216 selected genes, 70 (32.4%, compared to 4.5% among all genes) belong to the process response to DNA damage stimulus (GO:0006974). The term shares a large portion of its genes with DNA recombination (GO:0006310) and DNA replication (GO:0006260) processes, which are integral to the cell cycle. Sixty-seven of the selected genes

(31.0%, compared to 4.7% among all genes) belong to the process mitotic cell cycle (GO:0000278). Twenty-six of the 67 genes are shared with response to DNA damage stimulus (GO:0006974). Forty-one of the selected genes (19.0%, compared to 3.0% among all genes) belong to the process regulation of cell cycle (GO:0051726), among which 29 also belong to mitotic cell cycle (GO:0000278). Thirty-one of the selected genes (14.4%, compared to 2.6% among all genes) belong to the process meiotic cell cycle (GO:0051321), among which 12 are shared with mitotic cell cycle (GO:0000278). Other major enriched terms include chromatin organization (12.5%, compared to 3.5% overall), cytoskeleton organization (12.5%, compared to 3.4% overall), regulation of organelle organization (9.7%, compared to 2.4% overall), and cytokinesis (7.9%, compared to 1.7% overall). These terms clearly show strong relations with the yeast cell cycle.

4.4 Simulation Studies

In this section, we illustrate the performance of our methods (NET-DPMs) using simulation studies with various network structures and data settings compared with other methods. In Simulation 1, we study the similarity between the fully computational algorithm NET-DPM-1 and two fast computation approaches NET-DPM-x, $x = 2, 3$ in terms of gene selection accuracy and uncertainty estimations. Each of the three algorithms can be used along with one of the two methods for choosing hyper-parameters: the posterior inference and model averaging. In Simulation 2, we focus on the gene network selection under a particular network structure and two types of simulated data to demonstrate the flexibility of the proposed methods. In both simulations, we compare the NET-DPMs with a STD-DPM combined with the HODC algorithm without using any network information.

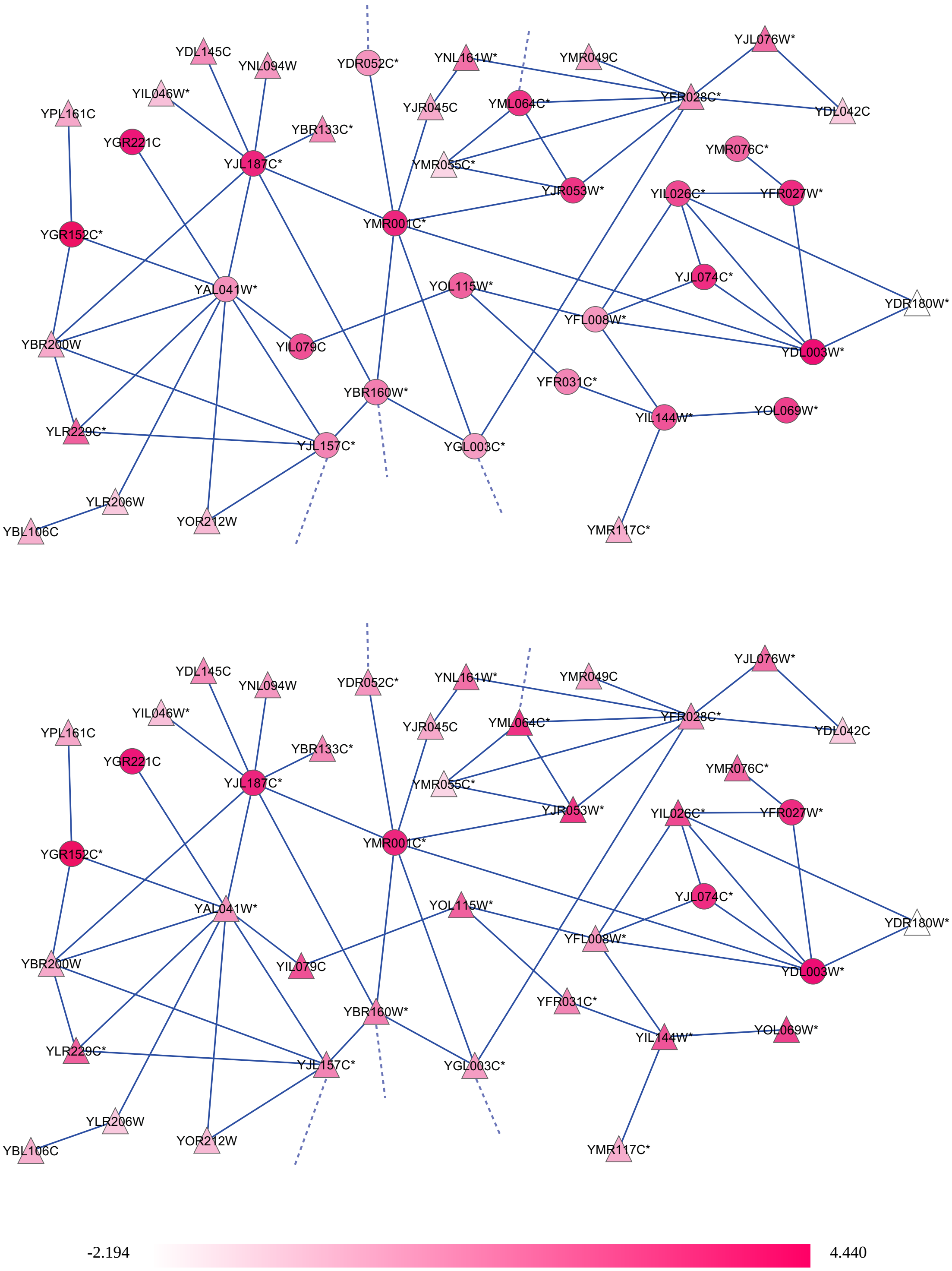


Figure 4.3: A portion of the subnetwork shown in Fig. 4.2, together with the immediate neighbors of the selected genes. Upper panel: NET-DPM-3 results; lower panel: STD-DPM results. The node labels indicate the gene name; circles and triangles represent selected?and unselected?genes; colors denote the value of the normal quantiles; a star in superscript represents the genes functionally annotated to the cell-cycle process. Dash lines denotes connections to genes not shown in the figure.

4.4.1 Simulation 1

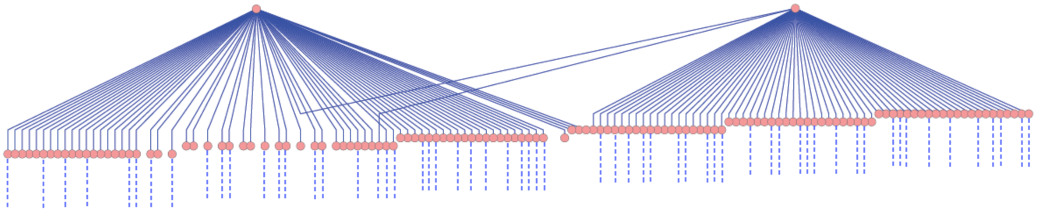


Figure 4.4: Partial network structure with the dash lines representing connections to other nodes not shown in the figure.

In this simulation, we investigate the performance of the proposed algorithms using a simulated dataset that mimic the real data in Section 4.3. We generate a scale-free network with 1,000 genes based on the rich-get-rich algorithm (Barabási and Albert, 1999), i.e. $n = 1,000$. Two hub genes with 64 and 69 connections to other genes are in this network; the mean and median edges per gene are 1.998 and 1. Partial network structure with the two hub genes included is shown in Fig. 4.4. From the network structure, we generate $\mathbf{z}$ from the Ising model (4.4) with the sparsity parameter $\pi_0 = 0.8$, smoothness parameters $\boldsymbol{\varrho} = (\varrho_0, \varrho_1) = (5, 10)$. For $i = 1, \ldots, n$, in light of the results in Section 4.3, we simulate data r_i given z_i from the empirical distributions (Fig. 4.5) of the test statistics for "selected" and "unselected" genes in the Spellman yeast cell cycle microarray data. As shown in Section 4.3, the NET-DPM-3 (Scenario 1) and the STD-DPM (Scenario 2) provide different gene selections results. We set both secenarios as the truth to simulate data.

We apply the NET-DPM-x, for $x = 1, 2, 3$ and the STD-DPM to the simulated dataset. To choose the sparsity and smoothness parameters, the NET-DPM-1 and the NET-DPM-3 are both combined with model averaging, where the possible choices of π_0 and $\boldsymbol{\varrho}$ are $(0.75, 0.8, 0.85, 0.9)$ and $(1, 5, 10, 20, 50) \times (1, 5, 10, 20, 50)$, while the NET-DPM-2 is combined with the posterior inference on $(\pi_0, \boldsymbol{\varrho})$. As for other hyper-parameters, we specify $\tau_k, \xi_k, \gamma_k, \beta_k, \alpha_k; k = 0, 1$ the same way as in the data

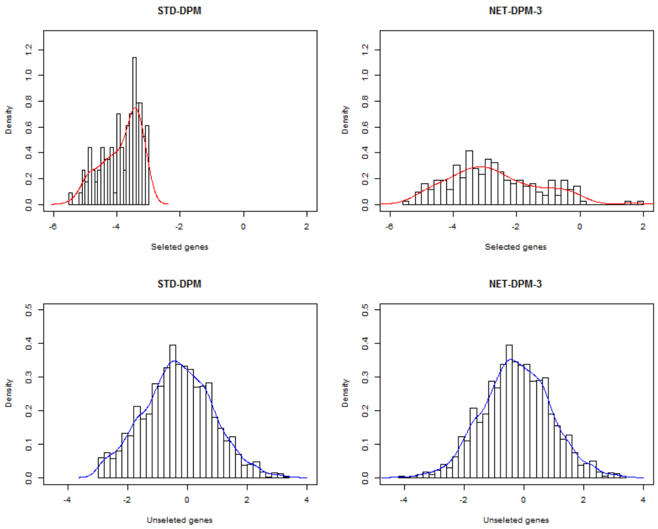


Figure 4.5: Empirical distributions of "selected" genes (upper panel) and "unselected" genes (lower panel) in the Spellman yeast cell cycle data estimated by the NET-DPM-3 (right panel) and the STD-DPM (left panel).

application for the NET-DPM-x, for $x = 1, 2$. With random starting values, each algorithm is run 10 times under 5,000 iterations with 2,000 burn-in. The selection performance for each method based on the average of the 10 runs are presented in Table 4.2. We also compare the posterior probability estimates of $\mathbf{z}$ between different algorithms under Scenario 1 in Fig. 4.6.

	STD-DPM	NET-DPM-1	NET-DPM-2	NET-DPM-3
	Scenario 1			
True Positive Rate	0.893	0.973	0.920	0.920
False Positive Rate	0.292	0.001	0.000	0.006
False Discovery Rate	0.801	0.014	0.000	0.080
	Scenario 2			
True Positive Rate	1.000	1.000	1.000	1.000
False Positive Rate	0.232	0.000	0.000	0.007
False Discovery Rate	0.741	0.000	0.000	0.085
Typical computation time (hrs)	0.100	8.500	2.800	0.150

Table 4.2: Gene selection accuracy in Simulation 1

From Table 4.2, it is clear that the NET-DPMs achieve a better selection perfor-

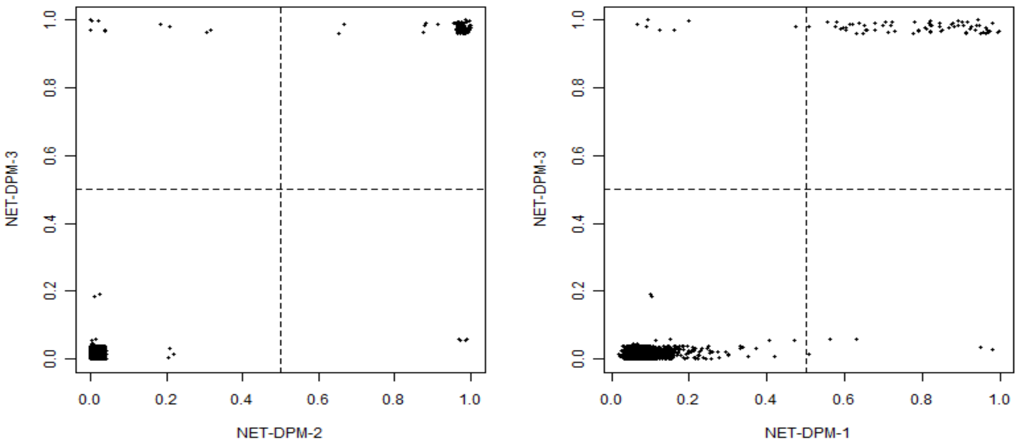


Figure 4.6: Marginal posterior probabilities of the class labels of all 1000 genes by the different methods: NET-DPM-3 vs. NET-DPM-2 (left panel) and NET-DPM-2 vs. NET-DPM-1 (right panel). The probability values are jittered by tiny random noises for better presenting.

mance than the STD-DPM method under both scenarios. The STD-DPM without using the gene network information provides an extreme high false discovery rate in each scenario. This implies that it is critical to incoorporate the gene network information to control FDR. Table 4.2 also suggests the NET-DPM-2 and the NET-DPM-3 approximate the NET-DPM-1 very well in terms of the gene selection accuracy with a substantial lower computational cost (3.4 GHz CPU, 8GB Memory, Windows System). In addition, a comparison between the NET-DPM-2 and the NET-DPM-3 shows that the Bayesian model averaging over hyper-parameters $(\pi_0, \boldsymbol{\varrho})$ provides an efficient alternative to the standard Bayesian posterior inference procedure. For the posterior probability estimates, the NET-DPM-2 and the NET-DPM-3 achieve a good agreement as shown in the left panel of Fig. 4.6. However, in the right panel of Fig. 4.6, compared with the NET-DPM-1, the NET-DPM-3 tends to provide larger probability estimates for the "selected" genes, but smaller probability estimates for "unselected" genes. This implies the fast computation approaches underestimate the uncertainty of gene selection.

4.4.2 Simulation 2

In this simulation, we demonstrate the flexibility of the proposed methods and their ability to identify subnetworks of interest. We consider a 94 gene network which consists of a 11-gene subnetwork by design and a 83-gene scale-free network simulated from the rich-get-rich algorithm. The mean and median edges per node for the whole network are 2.02 and 1. Fig. 4.7 shows the designed 11-gene subnetwork, where genes 5, 6 and 11 are connected with three other genes from the 83 gene scale-free network. Rather than simulating from priors, we directly specify the class label $\mathbf{z}$ as $z_i = 1$ for $i \in \{1, 2, 3, 4, 5, 8, 9, 10\}$, $z_i = 0$, otherwise. In Fig. 4.7, the blue nodes represent the "selected" genes and red nodes are "unselected" genes. In addition, all other genes in the scale-free network (not shown in the figure) are "unselected". The gene subnetwork of interest includes genes 1, 2, 3, 4 and 5, which are encircled by a rectangle frame in Fig. 4.7. The null distribution for "unselected" r_i is specified as a standard normal distribution: $[r_i \mid z_i = 0\ ] \sim \mathrm{N}(0, 1)$. For the distribution of "selected" genes, we consider two settings:

$$\begin{aligned} \text{Gaussian data:} \qquad & [r_i \mid z_i = 1\ ] \sim 0.4 \times \mathrm{N}(3, 1) + 0.6 \times \mathrm{N}(2, 0.5), \\ \text{Non-Gausian data:} \qquad & [r_i \mid z_i = 1\ ] \sim 0.4 \times \mathrm{G}(5, 2) + 0.6 \times \mathrm{G}(6, 3), \end{aligned}$$

where $\mathrm{G}(a, b)$ denotes a gamma distribution with shape a and rate b. According to the above procedure, we simulate 100 datasets for each type of data. We apply the NET-DPM-3 and the STD-DPM to each dataset. We utilize the model averaging for choosing hyper-parameters and a set of possible choices are given by $\{1, 2, 5, 10, 15\}$ for both ϱ_0 and ϱ_1, and $\{0.8,\ 0.85,\ 0.9,\ 0.95\}$ for π_0. We run 10,000 iterations with 2,000 burn-in on each dataset for both methods. In each simulated dataset, we predetermine one gene as a "sure selected" gene. It has the largest number of connections

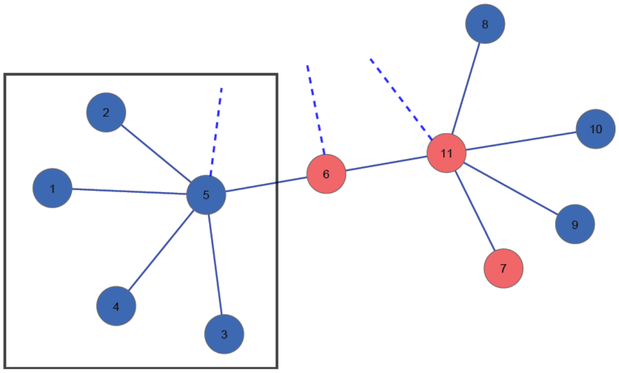


Figure 4.7: Partial simulated gene network structure: the blue nodes represent "selected" genes and red nodes represent "unselected" genes. Dash lines denotes connections to genes not shown in the figure. A subnetwork of interest includes nodes 1,2,3,4 and 5 which are encircled by a rectangle frame.

with the "selected" genes estimated by the STD-DPM model.

Table 4.3 summarizes the selection accuracy of the gene subnetwork based on the 100 simulated datasets for each type of data. It is clear that the NET-DPM-3 provides much higher accuracy of the subnetwork selection than the STD-DPM. The NET-DPM-3 achieves a more than 60% accuracy rate in correctly identifying the subnetwork with an additional low false positive and false negative occurrences regardless of the type of data. This verifies the overall better performance of NET-DPM-3 than the STD-DPM in terms of identifying the gene subnetwork, and the robustness of the proposed methods on different types of data.

Method	TPR	FPR	FDR	TPR	FPR	FDR
	Gaussian data			Non-Gaussian data		
NET-DPM-3	63%	11%	15%	60%	5%	8%
STD-DPM	15%	33%	69%	17%	26%	60%

[1] For gene subnetwork selection, the TPR is defined as the percentage of exactly selecting the correct network. The FPR is the percentage of selecting a larger network containing the correct network and at least one more other gene that has connection to the network. The FDR is the proportion of falsely selecting a larger network among all the network discoveries (selecting a correct or larger network).

Table 4.3: The selection accuracy of gene subnetwork[1] by TPR (true positive rate), FPR (false positive rate) and FDR (false discovery rate) in Simulation 2

4.5 Discussion

In the chapter, we propose a Bayesian nonparametric mixture model for gene/gene subnetwork selection. Our model extends the standard DPM model incorporating the gene network information to significantly improve the accuracy of the gene selections and reduce the false discovery rate. We demonstrate that the proposed method has the ability to identify the subnetworks of genes and individual genes with a particular expressional behavior. We also show that it is able to select genes which are strongly associated with clinical variables. We develop a posterior computation algorithm along with two fast approximation approaches. The posterior inference can produce more accurate uncertainty estimates of gene selection, while the fast computing algorithms can achieve a similar gene selection accuracy. Due to the nonparametric nature, our method has the flexibility to fit various data types and has robustness to model assumptions.

When we observe gene expression data along with measurements of a clinical outcome, we need to create statistics to perform the selection of genes that are strongly associated with the clinical outcome. The choice of the statistics is crucial to the performance of our methods. To model the relationship between the clinical outcome and gene expression data, many literatures suggest a linear regression model (Li and Li, 2008; Pan et al., 2010; Li and Zhang, 2010; Stingo et al., 2011), from which we produce testing statistics or coefficient estimates as the candidates. For instance, the most straightforward approach is to fit simple linear regression on each gene and use the t statistics as the input data to our methods. However, there is no scientific evidence that the relationship between gene expression profiles and the clinical outcome should follow a linear regression model. Without making this assumption, we may test the independence between each gene expression profile and the clinical outcome via a nonparametric model suggested by Einmahl and Van Keilegom (2008) and use our model to fit the testing statistics. Other potential choices of statistics for the non-linear problems include mutual information statistics (Peng et al., 2005) and maximal information coefficient (MIC) statistics (Reshef et al., 2011).

4.6 Appendix

4.6.1 Derivations

Derivations of equation (4.9)

$$\begin{aligned}
&\pi(g_i = g, z_i = k \mid \mathbf{g}_{-i}, \mathbf{z}_{-i}, \mathbf{r}, \widetilde{\boldsymbol{\theta}}) \\
&= \frac{\pi(g_i = g, z_i = k, \mathbf{g}_{-i}, \mathbf{z}_{-i}, \mathbf{r}, \widetilde{\boldsymbol{\theta}})}{\pi(\mathbf{g}_{-i}, \mathbf{z}_{-i}, \mathbf{r}, \widetilde{\boldsymbol{\theta}})} = \frac{\pi(\mathbf{r} \mid \widetilde{\boldsymbol{\theta}}, \mathbf{g}, \mathbf{z})\pi(\mathbf{g}, \mathbf{z})}{\pi(\mathbf{r} \mid \widetilde{\boldsymbol{\theta}}, \mathbf{g}_{-i}, \mathbf{z}_{-i})\pi(\mathbf{g}_{-i}, \mathbf{z}_{-i})} \\
&= \frac{\pi(\mathbf{r} \mid \widetilde{\boldsymbol{\theta}}, \mathbf{g}, \mathbf{z})}{\pi(\mathbf{r} \mid \widetilde{\boldsymbol{\theta}}, \mathbf{g}_{-i}, \mathbf{z}_{-i})} \pi(\mathbf{g}, \mathbf{z} \mid \mathbf{g}_{-i}, \mathbf{z}_{-i}) \propto \pi(r_i \mid \widetilde{\boldsymbol{\theta}}, g_i) \frac{\pi(\mathbf{g} \mid \mathbf{z})}{\pi(\mathbf{g}_{-i} \mid \mathbf{z}_{-i})} \pi(z_i = k \mid \mathbf{z}_{-i}),
\end{aligned}$$

where

$$\begin{aligned}
\frac{\pi(\mathbf{g} \mid \mathbf{z})}{\pi(\mathbf{g}_{-i} \mid \mathbf{z}_{-i})} &= \frac{\int \pi(\mathbf{g} \mid \mathbf{z}, \mathbf{q})\pi(\mathbf{q}_0)\pi(\mathbf{q}_1) d\mathbf{q}}{\int \pi(\mathbf{g}_{-i} \mid \mathbf{z}_{-i}, \mathbf{q})\pi(\mathbf{q}_0)\pi(\mathbf{q}_1) d\mathbf{q}} \\
&= \frac{\int q_{g_1} \cdots q_{g_{i-1}} q_g q_{g_{i+1} \cdots q_{g_n}} \prod_{k'=0}^{1} \Gamma(\tau_{k'})\Gamma(\tau_{k'}/L_{k'})^{-L_{k'}} \prod_{l \in \mathbf{a}_{k'}} q_l^{\tau_{k'}/L_{k'}-1} d\mathbf{q}}{\int q_{g_1} \cdots q_{g_{i-1}} q_{g_{i+1}} \cdots q_{g_n} \prod_{k'=0}^{1} \Gamma(\tau_{k'})\Gamma(\tau_{k'}/L_{k'})^{-L_{k'}} \prod_{l \in \mathbf{a}_{k'}} q_l^{\tau_{k'}/L_{k'}-1} d\mathbf{q}} \\
&= \frac{\Gamma(\tau_k + m_k - 1)}{\Gamma(\tau_k + m_k)} \cdot \frac{\Gamma(\tau_k/L_k + n_{ig} + 1)}{\Gamma(\tau_k/L_k + n_{ig})} = \frac{n_{ig} + \tau_k/L_k}{\tau_k + m_k - 1},
\end{aligned}$$

with $m_k = \sum_{i=1}^{n} I[z_i = k]$ and $n_{ig} = \sum_{i' \neq i} I[g_{i'} = g]$.

Derivations of equation (4.11)

$$
\begin{aligned}
&\pi(g_i \neq g_j, z_i \neq z_j, \text{for all } j \neq i \mid \mathbf{g}_{-i}, \mathbf{z}_{-i}, \mathbf{r}, \widetilde{\boldsymbol{\theta}}) \\
\propto \quad & \frac{\tau_k}{\tau_k + m_k - 1} \exp\left(\widetilde{\omega}_i \log(\pi_k) + \varrho_k \sum_{j \neq i} \omega_j c_{ij} I[z_j = k] \right) \\
& \times \int \phi\left(\frac{r_i - \mu}{\sigma}\right) \frac{1}{\xi_k} \phi\left(\frac{\mu - \gamma_k}{\xi_k}\right) \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)\sigma^{2(\alpha_k+3/2)}} \exp\left(-\frac{\beta_k}{\sigma^2}\right) d\mu d\sigma^2 \\
\propto \quad & \frac{\tau_k}{\tau_k + m_k - 1} \exp\left(\widetilde{\omega}_i \log(\pi_k) + \varrho_k \sum_{j \neq i} \omega_j c_{ij} I[z_j = k] \right) \\
& \times \frac{\beta_k^{\alpha_k}}{\sqrt{2\pi}\Gamma(\alpha_k)\xi_k} \int \phi\left(\frac{\mu - \gamma_k}{\xi_k}\right) \frac{1}{\sigma^{2(\alpha_k+3/2)}} \exp\left(-\frac{\beta_k + \frac{1}{2}(r_i - \mu)^2}{\sigma^2}\right) d\mu d\sigma^2 \\
\propto \quad & \frac{\tau_k}{\tau_k + m_k - 1} \exp\left(\widetilde{\omega}_i \log(\pi_k) + \varrho_k \sum_{j \neq i} \omega_j c_{ij} I[z_j = k] \right) \\
& \times \frac{\Gamma(\alpha_k + 1/2)\beta_k^{\alpha_k}}{\sqrt{2\pi}\Gamma(\alpha_k)\xi_k} \int \phi\left(\frac{\mu - \gamma_k}{\xi_k}\right) \left(\beta_k + \frac{1}{2}(r_i - \mu)^2\right)^{-(\alpha_k+1/2)} d\mu.
\end{aligned}
$$

4.6.2 Algorithms

Let Ξ denote a set of the distinct values in $\{g_1, \ldots, g_n\}$. Let $d_i \in \{1, \ldots, n\}$ denote the rank of r_i. Given the pre-specified hyper-parameters $(\boldsymbol{\pi}, \boldsymbol{\varrho})$, we have the following algorithms to simulate the posterior distribution of $\mathbf{g}$ and $\widetilde{\boldsymbol{\theta}}$:

NET-DPM-1

Input: Data $\mathbf{r}$:

Initialization: Set $g_{d_i} = 0$ for $d_i < \pi_0 n$, and $g_{d_i} = 1$, otherwise. Then $L_k = 1$ and $\Xi = \{0, 1\}$. Set $\widetilde{\mu}_g = \sum_{i:g_i=g} r_i/n_g$ and $\widetilde{\sigma}_g^2 = \sum_{i:g_i=g}(r_i - \widetilde{\mu}_g)^2/(n_g - 1)$.

Repeat the following steps until the distributions of L_0 and L_1 get stable:

```
Step 1: For $i = 1, \ldots, n$:
    Step 1.1: If $n_{-ig_i} = 0$, then update $\Xi$ to $\Xi/\{g_i\}$.
    Step 1.2: Update $(g_i, z_i)$ according to (4.10) and (4.11).
    Step 1.3: If $(g_i, z_i)$ are sampled from (4.11), then
        · set $g_i = (-1)^{1-z_i}(L_{z_i} + z_i)$ and $L_{z_i} = L_{z_i} + 1$;
        · draw $\widetilde{\sigma}^2_{g_i}$ from $\mathrm{IG}(\alpha_k, \beta_k)$;
        · given $\widetilde{\sigma}^2_{g_i}$, sample $\widetilde{\mu}_{g_i}$ from (4.12);
        · update $\Xi$ to $\Xi \cup \{g_i\}$.
Step 2: For $g \in \Xi$, update $\widetilde{\mu}_g$ and $\widetilde{\sigma}^2_g$ from (4.12) and (4.13), respectively.
Step 3: Sort $\widetilde{\boldsymbol{\theta}}_g$ by $\widetilde{\mu}_g$ such that $\widetilde{\mu}_g \leq \widetilde{\mu}_{g'}$ if $g < g'$, for $g, g' \in \Xi$; update $(g_i, z_i)$ accordingly.
```

Output: Simulated samples from the posterior distribution of $\mathbf{z}$.

NET-DPM-2

Input: Data $\mathbf{r}$ and L_k for $k = 0, 1$.

Initialization: Draw $g_i \sim \text{Discrete}[\mathbf{a}_0 \cup \mathbf{a}_1, \mathbf{1}/(L_0 + L_1)]$. Set $\widetilde{\mu}_g = \sum_{i:g_i=g} r_i/n_g$ and $\widetilde{\sigma}^2_g = \sum_{i:g_i=g}(r_i - \widetilde{\mu}_g)^2/(n_g - 1)$. Sort $\widetilde{\boldsymbol{\theta}}_g$ by $\widetilde{\mu}_g$ such that $\widetilde{\mu}_g < \widetilde{\mu}_{g+1}$.

Repeat the following steps until the distributions of $\mathbf{g}$ and $\widetilde{\boldsymbol{\theta}}$ are stable:

```
Step 1: For $i = 1, \ldots, n$, update $g_i$ and $z_i$ according to (4.9).
Step 2: For $g = -L_0 + 1, -L_0, \ldots, L_1$, update $\widetilde{\mu}_g$ and $\widetilde{\sigma}^2_g$ from (4.12) and (4.13), respectively.
Step 3: Sort $\widetilde{\boldsymbol{\theta}}_g$ by $\widetilde{\mu}_g$ such that $\widetilde{\mu}_g \leq \widetilde{\mu}_{g+1}$, for $g = -L_0 + 1, \ldots, L_1 - 1$. Update $(g_i, z_i)$ accordingly.
```

Output: Samples from the posterior distributions of $\mathbf{g}$ and $\widetilde{\boldsymbol{\theta}}$.

NET-DPM-3

Input: Data $\mathbf{r}$, posterior samples of parameters $\{\mathcal{P}_v\}_{v=1}^V$ produced by a DPM fitting and index sets $\{\mathbf{a}_{v,0}, \mathbf{a}_{v,1}\}_{v=1}^V$ generated by the HODC algorithm.

Initialization: for $v = 1, \ldots, V$ and $i = 1, \ldots, n$, draw $z_{vi} \sim \text{Discrete}(\{0, 1\}, \{0.5, 0.5\})$. Write $\mathbf{z}_{(i)} = (z_{1i}, z_{2i}, \ldots, z_{Vi})'$. Sample $z_i \sim \text{Discrete}(\mathbf{z}_{(i)}, \mathbf{1}_V/V)$.

Repeat the following steps until the distribution of $\mathbf{z}$ is stable: for $i = 1, 2, \ldots, n$,

Step 1: For $v = 1, \ldots, V$, update z_{vi} according to (4.18);

Step 2: Sample $z_i \sim \text{Discrete}(\mathbf{z}_{(i)}, \mathbf{1}_V/V)$.

Output: Simulated samples from the posterior distributions of $\mathbf{z}$.

4.6.3 Hyper-parameters

Posterior Inference

We set hyper-priors for π_0 and $\boldsymbol{\varrho}$ in (4.4) as:

$$\pi_0 \propto \text{U}(0,\ 1),\ \ \varrho_k \propto \text{U}(0,\ \infty),\ \text{ for } k = 0, 1,$$

where $\text{U}(a,\ b)$ represents a uniform distribution on $(a,\ b)$. Note that the priors for ϱ_0 and ϱ_1 are improper, however, the corresponding posteriors are proper. A Metropolis-Hastings algorithm can be used to simulate the hyper-parameters $(\boldsymbol{\pi}, \boldsymbol{\varrho})$. We define an unnormalized density of $\mathbf{z}$:

$$h(\mathbf{z} \mid \pi_0, \boldsymbol{\varrho}) = \exp\left[\sum_{i=1}^{n}\left(\widetilde{\omega}_i \log(\pi_{z_i}) + \varrho_{z_i}\sum_{j\neq i}\omega_j c_{ij} I[z_i = z_j]\right)\right], \qquad (4.19)$$

where $\pi_1 = 1-\pi_0$. Then we have the density (4.4), $\pi(\mathbf{z} \mid \boldsymbol{\pi}, \boldsymbol{\varrho}) = h(\mathbf{z} \mid \pi_0, \boldsymbol{\varrho})/Q(\pi_0, \boldsymbol{\varrho})$, where the normalizing constant $Q(\pi_0, \boldsymbol{\varrho}) = \int h(\mathbf{z} \mid \pi_0, \boldsymbol{\varrho}) d\mathbf{z}$. To efficiently compute Q, we rewrite

$$Q(\pi_0, \boldsymbol{\varrho}) = E_{\mathbf{z}}\left[\exp\left(\sum_{i=1}^{n}\varrho_{z_i}\sum_{j\neq i}\omega_j c_{ij} I[z_i = z_j]\right)\right], \qquad (4.20)$$

where $z_i \overset{i.i.d.}{\sim}$ Discrete$(\{0,1\},\{\pi_0,\pi_1\})$ for $i = 1,\ldots,n$. To simulate π_0 and $\boldsymbol{\varrho}$, we consider the following algorithm:

Algorithm for posterior inference on $(\pi_0, \boldsymbol{\varrho})$:

Input: The number of auxiliary variables V and the proposal variance δ:

Initialization: Draw $\widetilde{z}_{vi} \sim$ Discrete$(\{0,1\},\{\pi_0,\pi_1\})$, for $i = 1,2,\ldots,n$ and $v = 1,2,\ldots,V$. Set $\pi_0 = 0.8$ and $\boldsymbol{\varrho} = (0.5, 0.5)'$; Compute $\widehat{Q}(\pi_0, \boldsymbol{\varrho})$ according to

$$\widehat{Q}(\pi_0, \boldsymbol{\varrho}) = \frac{1}{V}\sum_{v=1}^{V}\exp\left(\sum_{i=1}^{n}\varrho_{\widetilde{z}_{vi}}\sum_{j\neq i}\omega_j c_{ij} I[\widetilde{z}_{vi} = \widetilde{z}_{vj}]\right). \qquad (4.21)$$

The following steps can be embedded in each iteration of Algorithm 1–3. Suppose $\mathbf{z}$ is one simulated sample generated from those algorithm.

Step 1: Draw $(\widetilde{\pi_0}, \widetilde{\boldsymbol{\varrho}})' \sim N\left((\pi_0, \boldsymbol{\varrho})', \delta\boldsymbol{I}_3\right)$.

Step 2: If $\widetilde{\pi_0} \in (0,1)$ and $\varrho_0, \varrho_1 > 0$, then compute $\widehat{Q}(\widetilde{\pi_0}, \widetilde{\boldsymbol{\varrho}})$ according to

(4.21) and the acceptance ratio

$$R = \frac{h(\mathbf{z} \mid \widetilde{\pi}_0, \widetilde{\boldsymbol{\varrho}})\widehat{Q}(\pi_0, \boldsymbol{\varrho})}{h(\mathbf{z} \mid \pi, \boldsymbol{\varrho})\widehat{Q}(\widetilde{\pi}_0, \widetilde{\boldsymbol{\varrho}})},$$

and set $(\pi_0, \boldsymbol{\varrho}) = (\widetilde{\pi}_0, \widetilde{\boldsymbol{\varrho}})$ and $\widehat{Q}(\pi_0, \boldsymbol{\varrho}) = \widehat{Q}(\widetilde{\pi}_0, \widetilde{\boldsymbol{\varrho}})$ with probability $\min\{1, R\}$.

Output: Simulated samples for $(\pi_0, \boldsymbol{\varrho})$ along with $\mathbf{z}$.

Note that the primary goal of this algorithm is making posterior inference on $\mathbf{z}$ by taking into account for the uncertainty from the choices of hyper-parameters $(\pi_0, \boldsymbol{\varrho})$.

Bayesian model averaging

In some cases, a set of possible values of $\boldsymbol{\pi}$ and $\boldsymbol{\varrho}$ that can be elicited from biological knowledge. We denote them as $\{\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m\}_{m=1}^{M}$. If we consider each $(\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m)$ as a model choice, then the marginal posterior distribution of $\mathbf{z}$ can be approximated by Bayesian model averaging. i.e.

$$\pi(\mathbf{z} \mid \mathbf{r}) \approx \sum_{m=1}^{M} \pi(\mathbf{z} \mid \boldsymbol{\pi}_m, \boldsymbol{\varrho}_m, \mathbf{r})\pi(\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m \mid \mathbf{r}), \tag{4.22}$$

where $\pi(\mathbf{z} \mid \boldsymbol{\pi}_m, \boldsymbol{\varrho}_m, \mathbf{r})$ refers to the posterior distribution of $\mathbf{z}$ given the hyper-parameters is chosen as $(\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m)$. This can be simulated or approximated by one of Algorithm 1–3. The term $\pi(\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m \mid \mathbf{r})$ is a weight for each choice. It can be represented as

$$\pi(\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m \mid \mathbf{r}) \propto E_{\mathbf{z}|\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m}[\pi(\mathbf{r} \mid \mathbf{z})] := w_m, \tag{4.23}$$

where $E_{\mathbf{z}|\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m}[\cdot]$ is with respect to the prior density of $\mathbf{z}$, i.e., (4.4), with parameter $(\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m)$ and the density $\pi(\mathbf{r} \mid \mathbf{z})$ is the likelihood function of $\mathbf{z}$. This suggests that we might estimate the weight for each $(\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m)$ via Monte Carlo methods. Specifically, we draw $\mathbf{z}_{vm} \sim \pi(\mathbf{z} \mid \boldsymbol{\pi}_m, \boldsymbol{\varrho}_m)$, for $v = 1, \ldots, V$ and $m = 1, \ldots, M$, then the weight estimate $\widehat{w}_m$ and the normalized weight estimate $\widetilde{w}_m$ are respectively given by

$$\widehat{w}_m = \frac{1}{V} \sum_{v=1}^{V} \pi(\mathbf{r} \mid \mathbf{z}_{vm}), \qquad \widetilde{w}_m = \frac{\sum_{v=1}^{V} \pi(\mathbf{r} \mid \mathbf{z}_{vm})}{\sum_{m'=1}^{M} \sum_{v=1}^{V} \pi(\mathbf{r} \mid \mathbf{z}_{vm'})}, \tag{4.24}$$

where $\sum_{m=1}^{M} \widetilde{w}_m = 1$. Write $\widetilde{\mathbf{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_M)$. We consider the following algorithm to simulate $\mathbf{z}$ from (4.22):

Algorithm for model averaging over $(\boldsymbol{\pi}, \boldsymbol{\varrho})$:

```
Input: The possible choices $\{\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m\}_{m=1}^{M}$ and the model averaging weight $\widetilde{\mathbf{w}}$;

Initialization: Draw $z_{mi} \sim \text{Discrete}(\{0,1\}, \{\pi_0, \pi_1\})$, for $i = 1, 2, \ldots, n$ and $m = 1, 2, \ldots, M$. Write $\mathbf{z}_m = (z_{m1}, \ldots, z_{mn})$ for $m = 1, \ldots, M$ and $\mathbf{z}_{(i)} = (z_{1i}, \ldots, z_{Mi})$ for $i = 1, \ldots, n$.

Repeat the following steps until the distribution of $\mathbf{z}$ is stable:

For $i = 1, \ldots, n$,

    Step 1: For $m = 1, \ldots, M$, update $z_{mi}$ according to the corresponding steps in one of Algorithm 1–3 given the hyper-parameter $(\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m)$.

    Step 2: Draw $z_i \sim \text{Discrete}(\mathbf{z}_{(i)}, \widetilde{\mathbf{w}})$.

Output: Simulated samples from the posterior distribution of $\mathbf{z}$.
```

Fig. 2 provides an illustration of this algorithm. One could put further restrictions on the choice of $(\boldsymbol{\pi}, \boldsymbol{\varrho})$ based on background information. For example, we could assume $\varrho_1 > \varrho_0$ to indicate a stronger impact of the "selected" genes than that of the "unselected" ones.

$$\begin{array}{lllllll}
 & \mathbf{z}_{(1)} & \mathbf{z}_{(2)} & \dots & \mathbf{z}_{(n)} & \widetilde{\mathbf{w}} & (\boldsymbol{\pi}, \boldsymbol{\varrho}) \\
\mathbf{z}_1 & z_{11} & z_{12} & \dots & z_{1n} & \widetilde{w}_1 & (\boldsymbol{\pi}_1, \boldsymbol{\varrho}_1) \\
\mathbf{z}_2 & z_{21} & z_{22} & \dots & z_{2n} & \widetilde{w}_2 & (\boldsymbol{\pi}_2, \boldsymbol{\varrho}_2) \\
\vdots & \vdots & \vdots & & \vdots & & \\
\mathbf{z}_M & z_{M1} & z_{M2} & \dots & z_{Mn} & \widetilde{w}_M & (\boldsymbol{\pi}_M, \boldsymbol{\varrho}_M) \\
 & \downarrow & \downarrow & & \downarrow & & \\
\mathbf{z} \leftarrow & z_1 & z_2 & \dots & z_n & &
\end{array}$$

Figure 4.8: An illustration of the algorithm for Bayesian model averaging over $(\boldsymbol{\pi}, \boldsymbol{\varrho})$. In each iteration, $\mathbf{z}_m = (z_{m1}, \dots, z_{mn})$ is simulated from the posterior of $\mathbf{z}$ given hyper-parameters $(\boldsymbol{\pi}_m, \boldsymbol{\varrho}_m)$ for $m = 1, \dots, M$, Resample z_i from $\mathbf{z}_{(i)} = (z_{1i}, \dots, z_{Mi})$ with probability $\widetilde{\mathbf{w}}$, for $i = 1, \dots, n$.

4.6.4 Sensitivity Analysis

We conduct sensitivity analysis for effect of prespecified hyper-parameters, i.e. β_k, τ_k with $k = 0, 1$ on the application dataset. With the presented results in Section 4.3 under $\beta_k = 10, k = 0, 1; \tau_0 = 10, \tau_1 = 2$ as the gold standard, we consider additional 9 realizations of the hyper-parameters with other settings staying the same. The gene selection results for both the whole datasets and the subnetwork of interest illustrated in Figure 4.2 are summarized in Table 4.4 by sensitivity (Sens) and specificity (Spec) against gold standard results.

Based on Table 4.4, the gene selection results for both the whole application dataset and the subnetwork of interest are stable among different combinations. The balance accuracy (arithmetic mean of sensitivity and specificity) for each scenario is always larger than 0.8. Based on this sensitivity analysis, we conclude the robustness of the proposed methods to the hyper-parameters and validate the results in the data application.

	Dataset		Subnetwork	
$(\beta_k, \tau_0, \tau_1)$	Sens	Spec	Sens	Spec
(10, 10, 5)	0.861	0.980	0.857	1.000
(10, 15, 5)	0.826	0.983	0.762	1.000
(10, 15, 5)	0.886	0.979	0.857	1.000
(10, 20, 2)	0.876	0.995	0.762	1.000
(100, 10, 2)	0.761	0.990	0.714	1.000
(100, 10, 5)	0.677	0.990	0.667	1.000
(100, 15, 2)	0.736	0.992	0.762	1.000
(100, 15, 5)	0.910	0.955	1.000	0.737
(100, 20, 2)	0.756	0.987	0.762	1.000

Table 4.4: Sensitivity analysis for the hyper-parameters specification for the application dataset and subnetwork

Chapter 5

Summary and future research

Recent advance in technology brings large numbers of complex biomedical datasets with complicated data structures and comprehensive biological information. This brings new challenges to conduct effective and practical statistical analyses. Among all, the dissertation focuses on the variable selection problem, one of the most important issues to realize statistical learning and biological investigation, for the complex biomedical data. Under Bayesian frameworks, we develop novel statistical methods for the analysis of three different types of data–functional data, neuroimaging data and high-throughput genomics data.

First, motivated by a colorectal adenoma study, with the goal to select informative profiles of functional biomarkers that are highly associated with the clinical outcome, we propose a unified Bayesian approach to conduct feature select under GFLMs. We novelly bring up the hierarchical nature of the feature selection in functional data and a class of mixture priors for the functional biomarkers to perform selection both between and within the functional curves. Accordingly, two levels of biological information are incorporated into the selection procedure to facilitate more biologically meaningful results. Due to some limitations of the current work, future research could be focused on discussing the choosing of grid points of the feature selection and the continuity assumption of functional coefficients at the boundaries points.

Second, to conduct spatial variable selection in the ultra high-dimensional neuroimaging data, we propose a novel multiresolution variable selection procedure under Bayesian probit regression models to make the posterior simulation considerably more efficient. Our approach is to construct auxiliary models at different resolutions and sequentially let the coarse-scale selection serve as a guild to fine-scale selection by constructing a proposal function, which reduces the computation by facilitating the posterior inference concentrates on the true signals more efficiently. As a future work, we plan to replace the point mass mixture priors by the continuous shrinkage priors to further improve the computation efficiency and build up a multiresolution shrinkage effect.

Lastly, to select genes and gene subnetworks with periodic behavior in a microarray dataset, we novelly extend the DPM model by incorporating gene network structure via Ising priors to estimate the selected and unselected densities of gene features. Two fast computational algorithms for the posterior simulation are also developed to dramatically release the computational burden of the standard MCMC algorithm. As a future extension, we plan to extend the current model from one dimensional to multiple dimensions. In addition, the current NET-DPM model focuses on the effect from genes that are directly connected. More biologically meaningful results could be obtained by incorporating the network distance into the priors of the class label, and we will also investigate more on this part.

Bibliography

Ahearn, T. U., Shaukat, A., Flanders, W. D., Seabrook, M. E., and Bostick, R. M. (2012), "Markers of the APC/β-Catenin signaling pathway as potential treatable, preneoplastic biomarkers of risk for colorectal neoplasms," *Cancer Epidemiology Biomarkers & Prevention*, 21, 969–979.

Aitkin, M. (1991), "Posterior Bayes factors," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53, 111–142.

Antoniak, C. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The annals of statistics*, 1152–1174.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000), "Gene Ontology: tool for the unification of biology," *Nature genetics*, 25, 25.

Badre, D. and Wagner, A. D. (2007), "Left ventrolateral prefrontal cortex and the cognitive control of memory," *Neuropsychologia*, 45, 2883–2901.

Baladandayuthapani, V., Mallick, B., Young Hong, M., Lupton, J., Turner, N., and Carroll, R. (2007), "Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis," *Biometrics*, 64, 64–73.

Barabási, A. and Albert, R. (1999), "Emergence of scaling in random networks," *science*, 286, 509–512.

Barbieri, M. M. and Berger, J. O. (2004), “Optimal predictive model selection,” *The Annals of Statistics*, 32, 870–897.

Barbu, A. and Zhu, S.-C. (2007), “Generalizing Swendsen–Wang for Image Analysis,” *Journal of Computational and Graphical Statistics*, 16, 877–900.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003), “Bayesian clustering with variable and transformation selections,” in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, Oxford University Press, USA, p. 249.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2012), “Bayesian shrinkage,” *arXiv preprint arXiv:1212.6088*.

Bottolo, L. and Richardson, S. (2010), “Evolutionary stochastic search for Bayesian model exploration,” *Bayesian Analysis*, 5, 583–618.

Bowman, F. D., Zhang, L., Derado, G., and Chen, S. (2012), “Determining functional connectivity using fMRI data with diffusion-based anatomical weighting,” *NeuroImage*, 62, 1769–1779.

Breeze, E., Harrison, E., McHattie, S., Hughes, L., Hickman, R., Hill, C., Kiddle, S., Kim, Y., Penfold, C., Jenkins, D., et al. (2011), “High-resolution temporal profiling of transcripts during Arabidopsis leaf senescence reveals a distinct chronology of processes and regulation,” *The Plant Cell Online*, 23, 873–894.

Brown, P. J., Vannucci, M., and Fearn, T. (1998), “Multivariate Bayesian variable selection and prediction,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 627–641.

Cerami, E., Gross, B., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G., and Sander, C. (2011), "Pathway Commons, a web resource for biological pathway data," *Nucleic acids research*, 39, D685–D690.

Chenevert, J., Valtz, N., and Herskowitz, I. (1994), "Identification of genes required for normal pheromone-induced cell polarization in Saccharomyces cerevisiae," *Genetics*, 136, 1287–1297.

Cherry, J., Hong, E., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E., Christie, K., Costanzo, M., Dwight, S., Engel, S., et al. (2012), "Saccharomyces Genome Database: the genomics resource of budding yeast," *Nucleic Acids Research*, 40, D700–D705.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., and Stine, R. A. (2001), "The practical implementation of Bayesian model selection," *Lecture Notes-Monograph Series*, 65–134.

Crainiceanu, C., Staicu, A., and Di, C. (2009), "Generalized Multilevel Functional Regression," *Journal of the American Statistical Association*, 104, 1550–1561.

Di, C., Crainiceanu, C., Caffo, B., and Punjabi, N. (2009), "Multilevel functional principal component analysis," *Annals of Applied Statistics*, 3, 458–488.

Di Martino, A., Yan, C., Li, Q., Denio, E., Castellanos, F., Alaerts, K., Anderson, J., Assaf, M., Bookheimer, S., Dapretto, M., et al. (2013), "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry.*

Do, K., Müller, P., and Tang, F. (2005), "A Bayesian mixture model for differential gene expression," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 627–644.

Dunson, D. (2010), “Nonparametric Bayes applications to biostatistics,” in *Bayesian Nonparametrics*, eds. Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., Cambridge University Press, chap. 7, pp. 223–273.

Efron, B. (2004), “Large-scale simultaneous hypothesis testing,” *Journal of the American Statistical Association*, 99, 96–104.

— (2010), “Correlated z-values and the accuracy of large-scale statistical estimates,” *Journal of the American Statistical Association*, 105, 1042–1055.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least angle regression,” *The Annals of statistics*, 32, 407–499.

Einmahl, J. and Van Keilegom, I. (2008), “Tests for independence in nonparametric regression,” *Statistica Sinica*, 18, 601.

Escobar, M. (1994), “Estimating normal means with a Dirichlet process prior,” *Journal of the American Statistical Association*, 268–277.

Escobar, M. and West, M. (1995), “Bayesian density estimation and inference using mixtures,” *Journal of the american statistical association*, 90, 577–588.

Fan, J. and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Lv, J. (2008), “Sure independence screening for ultrahigh dimensional feature space,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Fan, J. and Song, R. (2010), “Sure independence screening in generalized linear models with NP-dimensionality,” *The Annals of Statistics*, 38, 3567–3604.

Foundas, A. L., Leonard, C. M., Gilmore, R. L., Fennell, E. B., and Heilman, K. M. (1996), "Pars triangularis asymmetry and language dominance," *Proceedings of the National Academy of Sciences*, 93, 719–722.

Fox, E. B. and Dunson, D. B. (2012), "Multiresolution gaussian processes," *arXiv preprint arXiv:1209.0833*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*.

Gelman, A. and Meng, X.-L. (1998), "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling," *Statistical Science*, 163–185.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior predictive assessment of model fitness via realized discrepancies," *Statistica Sinica*, 6, 733–759.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical science*, 457–472.

George, E. and McCulloch, R. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

— (1997), "Approaches for Bayesian variable selection," *Statistica Sinica*, 7, 339–374.

Gerdes, H., Gillin, J. S., Zimbalist, E., Urmacher, C., Lipkin, M., and Winawer, S. J. (1993), "Expansion of the epithelial cell proliferative compartment and frequency of adenomatous polyps in the colon correlate with the strength of family history of colorectal cancer," *Cancer research*, 53, 279–282.

Gerstner, T. and Griebel, M. (1998), "Numerical integration using sparse grids," *Numerical algorithms*, 18, 209–232.

Giles, M. B. (2008), "Multilevel monte carlo path simulation," *Operations Research*, 56, 607–617.

Goldsmith, J., Huang, L., and Crainiceanu, C. M. (2012), "Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection," *Journal of Computational and Graphical Statistics*.

Goodman, J. and Sokal, A. D. (1989), "Multigrid monte carlo method. conceptual foundations," *Physical Review D*, 40, 2035.

Hans, C. (2009), "Bayesian lasso regression," *Biometrika*, 96, 835–845.

— (2011), "Elastic net regression modeling with the orthant normal prior," *Journal of the American Statistical Association*, 106, 1383–1393.

Hervé, P.-Y., Razafimandimby, A., Vigneau, M., Mazoyer, B., and Tzourio-Mazoyer, N. (2012), "Disentangling the brain networks supporting affective speech comprehension," *NeuroImage*, 61, 1255–1267.

Higdon, D., Lee, H., and Bi, Z. (2002), "A Bayesian approach to characterizing uncertainty in inverse problems using coarse and fine-scale information," *Signal Processing, IEEE Transactions on*, 50, 389–399.

Hoff, P. D. et al. (2006), "Model-based subspace clustering," *Bayesian Analysis*, 1, 321–344.

Holloman, C. H., Lee, H. K., and Higdon, D. M. (2006), "Multiresolution genetic algorithms and Markov chain Monte Carlo," *Journal of Computational and Graphical Statistics*, 15.

Huang, L., Goldsmith, J., Reiss, P. T., Reich, D. S., and Crainiceanu, C. M. (2013), "Bayesian Scalar-on-Image Regression with Application to Association Between Intracranial DTI and Cognitive Outcomes," *NeuroImage*.

Ishwaran, H. and James, L. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

— (2002), "Approximate Dirichlet Process computing in finite normal mixtures," *Journal of Computational and Graphical Statistics*, 11, 508–532.

Ishwaran, H. and Rao, J. S. (2003), "Detecting differentially expressed genes in microarrays using Bayesian model selection," *Journal of the American Statistical Association*, 98.

Jacob, L., Neuvial, P., and Dudoit, S. (2012), "More power via graph-structured tests for differential expression of gene networks," *The Annals of Applied Statistics*, 6, 561–600.

James, G. (2002), "Generalized linear models with functional predictors," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 411–432.

James, G., Wang, J., and Zhu, J. (2009), "Functional linear regression thats interpretable," *The Annals of Statistics*, 37, 2083–2108.

Johnson, T. D., Liu, Z., Bartsch, A. J., and Nichols, T. E. (2012), "A Bayesian nonparametric Potts model with application to pre-surgical FMRI data," *Statistical Methods in Medical Research*.

Johnson, V. E. (2013), "On Numerical Aspects of Bayesian Model Selection in High and Ultrahigh-dimensional Settings," *Bayesian Analysis*, 7, 1–18.

Johnson, V. E. and Rossell, D. (2010), "On the use of non-local prior densities in Bayesian hypothesis tests," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 143–170.

— (2012), "Bayesian Model Selection in High-Dimensional Settings," *Journal of the American Statistical Association*, 107, 649–660.

Just, M. A. and Pelphrey, K. A. (2013), *Development and Brain Systems in Autism*, Psychology Press.

Kadane, J. and Lazar, N. (2004), "Methods and criteria for model selection," *Journal of the American Statistical Association*, 99, 279–290.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006), "Variable selection in clustering via Dirichlet process mixture models," *Biometrika*, 93, 877–893.

Kou, S., Olding, B. P., Lysy, M., and Liu, J. S. (2012), "A Multiresolution Method for Parameter Estimation of Diffusion Processes," *Journal of the American Statistical Association*, 107, 1558–1574.

Koutsourelakis, P.-S. (2009), "A multi-resolution, non-parametric, Bayesian framework for identification of spatially-varying model parameters," *Journal of computational physics*, 228, 6184–6211.

Krämer, N., Boulesteix, A., and Tutz, G. (2008), "Penalized Partial Least Squares with applications to B-spline transformations and functional data," *Chemometrics and Intelligent Laboratory Systems*, 94, 60–69.

Lamnisos, D., Griffin, J. E., and Steel, M. F. (2009), "Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations," *Journal of Computational and Graphical Statistics*, 18, 592–612.

— (2012), "Adaptive Monte Carlo for Bayesian variable selection in regression models," *Journal of Computational and Graphical Statistics.*

Lee, E. and Park, B. (2012), "Sparse estimation in functional linear regression," *Journal of Multivariate Analysis*, 105, 1–17.

Leng, X. and Müller, H. (2006), “Classification using functional data analysis for temporal gene expression data,” *Bioinformatics*, 22, 68–76.

Li, C. and Li, H. (2008), “Network-constrained regularization and variable selection for analysis of genomic data,” *Bioinformatics*, 24, 1175–1182.

Li, F. and Zhang, N. (2010), “Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics,” *Journal of the American Statistical Association*, 105, 1202–1214.

Li, Q. and Lin, N. (2010), “The Bayesian elastic net,” *Bayesian Analysis*, 5, 151–170.

Lian, H. (2011), “Shrinkage estimation and selection for multiple functional regression,” *arXiv preprint arXiv:1108.3904*.

Liu, J. S. and Sabatti, C. (2000), “Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation,” *Biometrika*, 87, 353–369.

Ma, S., Shi, M., Li, Y., Yi, D., and Shia, B. (2010), “Incorporating gene co-expression network in identification of cancer prognosis markers,” *BMC bioinformatics*, 11, 271.

Malloy, E., Morris, J., Adar, S., Suh, H., Gold, D., and Coull, B. (2010), “Wavelet-based functional linear mixed models: an application to measurement error–corrected distributed lag models,” *Biostatistics*, 11, 432–452.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006), “An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants,” *Biometrika*, 93, 451–458.

Morris, J. and Carroll, R. (2006), “Wavelet-based functional mixed models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 179–199.

Müller, H. and Stadtmüller, U. (2005), "Generalized functional linear models," *The Annals of Statistics*, 33, 774–805.

Müller, P. and Quintana, F. (2004), "Nonparametric Bayesian data analysis," *Statistical science*, 95–110.

Neal, R. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, 249–265.

Pan, W., Xie, B., and Shen, X. (2009), "Incorporating predictor network in penalized regression with application to microarray data," *Biometrics*, 66, 474–484.

— (2010), "Incorporating predictor network in penalized regression with application to microarray data," *Biometrics*, 66, 474–484.

Park, T. and Casella, G. (2008), "The bayesian lasso," *Journal of the American Statistical Association*, 103, 681–686.

Peng, H., Long, F., and Ding, C. (2005), "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27, 1226–1238.

Polson, N. G. and Scott, J. G. (2010), "Shrink globally, act locally: sparse Bayesian regularization and prediction," *Bayesian Statistics*, 9, 501–538.

Ramsay, J. and Silverman, B. (2005), *Functional Data Analysis, 2nd Edition.*, New York: Springer.

Reiss, P. and Ogden, R. (2007), "Functional principal component regression and functional partial least squares," *Journal of the American Statistical Association*, 102, 984–996.

— (2009), “Functional generalized linear models with images as predictors,” *Biometrics*, 66, 61–69.

Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M., and Sabeti, P. (2011), “Detecting novel associations in large data sets,” *science*, 334, 1518–1524.

Rice, C. (2009), “Prevalence of Autism Spectrum Disorders: Autism and Developmental Disabilities Monitoring Network, United States, 2006. Morbidity and Mortality Weekly Report. Surveillance Summaries. Volume 58, Number SS-10.” *Centers for Disease Control and Prevention*.

Rojas, D., Peterson, E., Winterrowd, E., Reite, M., Rogers, S., and Tregellas, J. (2006), “Regional gray matter volumetric changes in autism associated with social and repetitive behavior symptoms,” *BMC psychiatry*, 6, 56.

Sha, N., Tadesse, M. G., and Vannucci, M. (2006), “Bayesian variable selection for the analysis of microarray data with censored outcomes,” *Bioinformatics*, 22, 2262–2268.

Smith, M. and Fahrmeir, L. (2007), “Spatial Bayesian variable selection with application to functional magnetic resonance imaging,” *Journal of the American Statistical Association*, 102, 417–431.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998), “Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization,” *Molecular biology of the cell*, 9, 3273–3297.

Stingo, F., Chen, Y., Tadesse, M., and Vannucci, M. (2011), “Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes," *The Annals of Applied Statistics*, 5, 1978–2002.

Strunnikov, A. and Jessberger, R. (1999), "Structural maintenance of chromosomes (SMC) proteins," *European Journal of Biochemistry*, 263, 6–13.

Styan, G. P. (1973), "Hadamard products and multivariate statistical analysis," *Linear Algebra and Its Applications*, 6, 217–240.

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., Cron, A., and West, M. (2010), "Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures," *Journal of Computational and Graphical Statistics*, 19.

Swendsen, R. and Wang, J. (1987), "Nonuniversal critical dynamics in Monte Carlo simulations," *Physical Review Letters*, 58, 86.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian variable selection in clustering high-dimensional data," *Journal of the American Statistical Association*, 100, 602–617.

Tang, Y., Ghosal, S., and Roy, A. (2007), "Nonparametric Bayesian estimation of positive false discovery rates," *Biometrics*, 63, 1126–1134.

Theo, H. and Mike, E. (2004), "Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data," *Genet. Sel. Evol*, 36, 261–279.

Tian, T. and James, G. (2012), "Interpretable dimension reduction for classifying functional data," *Computational Statistics & Data Analysis.*

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 91–108.

Vehtari, A. and Lampinen, J. (2002), "Bayesian model assessment and comparison using cross-validation predictive densities," *Neural Computation*, 14, 2439–2468.

Wang, L. and Dunson, D. (2011), "Fast Bayesian inference in Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 20, 196–216.

Wang, S., Nan, B., Zhu, N., and Zhu, J. (2009), "Hierarchically penalized Cox regression with grouped variables," *Biometrika*, 96, 307–322.

Wei, P. and Pan, W. (2010), "Network-based genomic discovery: application and comparison of Markov random-field models," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59, 105–125.

— (2012), "Bayesian joint modeling of multiple gene networks and diverse genomic data to identify target genes of a transcription factor," *The annals of applied statistics*, 6, 334.

Wei, Z. and Li, H. (2007), "A Markov random field model for network-based analysis of genomic data," *Bioinformatics*, 23, 1537–1544.

— (2008), "A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data," *The Annals of Applied Statistics*, 2, 408–429.

Wichert, S., Fokianos, K., and Strimmer, K. (2004), "Identifying periodically expressed transcripts in microarray time series data," *Bioinformatics*, 20, 5–20.

Wu, T. T. and Wang, S. (2013), "Doubly regularized Cox regression for high-dimensional survival data with group structures," *STATISTICS AND ITS INTERFACE*, 6, 175–186.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D. (2002), "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic acids research*, 30, 303–305.

Yao, F., Müller, H., and Wang, J. (2005), "Functional linear regression analysis for longitudinal data," *The Annals of Statistics*, 33, 2873–2903.

Yi, N., George, V., and Allison, D. B. (2003), "Stochastic search variable selection for identifying multiple quantitative trait loci," *Genetics*, 164, 1129–1138.

Yu, T. (2010), "An exploratory data analysis method to reveal modular latent structures in high-throughput data," *BMC bioinformatics*, 11, 440.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Zhao, Y., Ogden, R., and Reiss, P. (2012), "Wavelet-based LASSO in functional linear regression," *Journal of Computational and Graphical Statistics*, 21, 600–617.

Zhou, B., Xu, W., Herndon, D., Tompkins, R., Davis, R., Xiao, W., Wong, W., Toner, M., Warren, H., Schoenfeld, D., et al. (2010), "Analysis of factorial time-course microarrays with application to a clinical study of burn injury," *Proceedings of the National Academy of Sciences*, 107, 9923.

Zhou, J., Wang, N., and Wang, N. (2012), "Functional linear model with zero-value coefficient function at sub-regions," *Statistica Sinica*, In press.

Zhu, H., Vannucci, M., and Cox, D. (2009), "A Bayesian hierarchical model for classification with selection of functional predictors," *Biometrics*, 66, 463–473.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

Zou, Q.-H., Zhu, C.-Z., Yang, Y., Zuo, X.-N., Long, X.-Y., Cao, Q.-J., Wang, Y.-F., and Zang, Y.-F. (2008), "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF," *Journal of neuroscience methods*, 172, 137–141.

Zuo, X.-N., Di Martino, A., Kelly, C., Shehzad, Z. E., Gee, D. G., Klein, D. F., Castellanos, F. X., Biswal, B. B., and Milham, M. P. (2010), "The oscillating brain: complex and reliable," *Neuroimage*, 49, 1432–1445.