ON THE BAYESIAN MULTIPLE INDEX MODELS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Zhengkang Liang
August 2022

Examining Committee Members:

Dr. Zhigen Zhao, Advisory Chair, Department of Statistics, Operations, and Data Science
Dr. Yuexiao Dong, Department of Statistics, Operations, and Data Science
Dr. Kenichiro Mcalinn, Department of Statistics, Operations, and Data Science
Dr. Xiaojing Wang, External Reader, Department of Statistics, University of Connecticut

ABSTRACT

In modern statistical applications when the dimension is relatively large, it is a common practice to reduce the dimension using methods such as principal component analysis (PCA), sliced inverse regression and others before applying any statistical models. In this article, we synthetically combine these two steps by considering three Bayesian multi-index models: Bayesian multi-index additive model (BMIAM) for continuous response variable, Bayesian single-index model for binary response variable, and Bayesian multi-index model for categorical response variable. The indexes are parametrized by the hyper-spherical coordinates. The ridge functions are modeled using the Bayesian B-splines, which could be easily extended to other non-parametric methods. We have shown that the posterior consistency holds under certain conditions for the BMIAM. Further, we have developed the Markov chain Monte Carlo (MCMC) algorithm to sample the posterior of the proposed methods. It has been demonstrated through both simulation and real data analysis that the proposed methods provide a reliable estimation of indexes, dimension reduction space and good predictions for the responses.

ACKNOWLEDGEMENTS

This research includes calculations carried out on HPC resources supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189.

This research is based upon work supported in part by the National Science Foundation under grant No. IIS-1633283.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

INTRODUCTION

In modern statistical applications when the number of covariates is large, it is a common practice to reduce the dimension for the purposes of mitigating the effects of colinearity and facilitating model specification, and even visualizing in low dimensions (D. R. Cook, 1998; R. D. Cook, 2007). Among many models which trade off the model flexibility and the reduction of dimensionality, the multi-index model is one of the most popular choices and has been well investigated and widely applied in many areas such as econometrics and biometrics. A general multi-index regression model has the form

$$Y_i = f\left(\boldsymbol{X}_i'\boldsymbol{\beta}_1, \boldsymbol{X}_i'\boldsymbol{\beta}_2, \cdots, \boldsymbol{X}_i'\boldsymbol{\beta}_D\right) + \epsilon_i, \tag{1.0-1}$$

where Y_i is the ith scalar response, $\boldsymbol{X}_i$ is a p-dimensional covariates for $i = 1, \cdots, n$, $\boldsymbol{\beta}_d$, $d = 1, \cdots, D$, denotes a p-dimensional vector of unknown coefficients, and ϵ_i represents the ith random error, which is assumed to be independent of $\boldsymbol{X}_i$'s. Here, the response variable depends on the covariates only through D numbers of indexes $\boldsymbol{\beta}_d$'s and the associated link function $f(\cdot)$.

Let $\mathcal{S}$ be the central space spanned by the vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_D$. In the existing literature on sufficient dimension reduction, many methods, following by the pioneering work of the sliced inverse regression (SIR, Li (1991)), are proposed to estimate the space $\mathcal{S}$ without knowing the link function. In the work of Li (1991), they sliced the data according to the value of the response variable and used the PCA to estimate

the space $\mathcal{S}$. Under the linearity and coverage condition, the SIR has been shown to produce consistent estimator of the true space $\mathcal{S}$ (Hsing & Carroll (1992); Xia et al. (2002); L.-X. Zhu & Ng (1995); L. Zhu et al. (2006); Lin et al. (2018)). After the seminal work of Li (1991), many other methods are proposed to deal with different issues arising from both theoretical and application studies (D. R. Cook (2000); D. R. Cook et al. (2007); Ni et al. (2005); Li (1992); Lin et al. (2018, 2019); Reich et al. (2011)).

In all these work above, their primary goal is to estimate the central space $\mathcal{S}$ without assuming the link function. After that, it serves as an intermediate step towards the final modeling between the response and the covariates. In many practice, the estimation of the central space and the estimation of the link function are split into two separate steps. It is not clear whether a consistent estimator of the central space could lead to a consistent estimator of the link function. In addition, the estimation of the central space does not rely on the link function which could be beneficial to indexes or central space estimation if modeled correctly. In this paper, we synthetically combine these two steps using the Bayesian modeling technique. The proposed approach has two-fold advantages: (i) explicitly modeling the link function could benefit the estimation of the dimension reduction space; and (ii) a reduced dimension could improve the model prediction.

In the current literature, there are several attempts to combine the two steps together in the single-index model, which is a special case of the multi-index model. To name a few, Antoniadis et al. (2004) considered a single-index regression model under the Bayesian framework. Namely, it is assumed that $\mathrm{E}(Y_i \mid X_i) = f(\boldsymbol{X}_i'\boldsymbol{\beta})$ where the unknown function $f(\cdot)$ is approximated by a regularly spaced knots B-spline. Park et al. (2005) proposed a Bayesian single-index model using wavelet to model the link function. Choi et al. (2011) developed a single-index model using Gaussian process regression. More recently, Dhara et al. (2020) modeled a Bayesian

single-index model using Ornstein-Uhlenbeck process prior. Both Choi et al. (2011) and Dhara et al. (2020) have established theoretical properties about the posterior consistency under certain regularity conditions.

There are some further attempts trying to work with multi-index models. Lang & Brezger (2004) considered a Bayesian generlized additive model with predictors: $\mathrm{E}(Y_i \mid X_i) = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon$ without taking the structure of indexes. McGee et al. (2021) proposed a Bayesian multi-index model (BMIM) in which the covariates are divided into different groups without overlapping.

In this article, we consider the following multi-index additive model (MIAM)

$$Y_i = f_1\left(\boldsymbol{X}_i'\boldsymbol{\beta}_1\right) + f_2(\boldsymbol{X}_i'\boldsymbol{\beta}_2) + \ldots + f_D(\boldsymbol{X}_i'\boldsymbol{\beta}_D) + \epsilon_i, \tag{1.0-2}$$

where the indexes satisfy $\|\boldsymbol{\beta}_d\|_2 = 1$ with $\|\cdot\|_2$ indicating the Euclidean norm, and $f_d(\cdot)$ is the corresponding link function for the index $\boldsymbol{\beta}_d$, also called the ridge function in the literature. The aforementioned multi-index additive models are just special cases of our defined multi-index additive model in (1.0-2).

We want to point out that the model in (1.0-2) has been considered in the existing literature, which is also known as the projection pursuit (J. Friedman & Tukey (1974); J. H. Friedman & Stuetzle (1981)). J. H. Friedman et al. (1983) first introduced projection pursuit regression using a B-spline approximation, where the parameter $\boldsymbol{\beta_d}$'s and the ridge function $f_d(\cdot)$'s are estimated using an alternative algorithm. In fact, this alternative algorithm is to estimate the indexes using SIR and then applied the projection pursuit model, which is indeed two separate steps. In contrast, we will build up the projection pursuit under the Bayesian framework and introduce a method to estimate the indexes and the ridge functions simultaneously. Specifically, the indexes $\boldsymbol{\beta}_d$'s are re-parameterized using the spherical coordinates. The ridge functions

$f_d(\cdot)$ of each additive component is approximated by the Bayesian B-splines with regularly spaced knots. In this paper, we have further investigated the asymptotic property on the posterior consistency. To the best of our knowledge, this is so far the first posterior consistency result for multi-index additive model under the Bayesian framework. It has been shown that the proposed method works better in comparison to existing methods through both simulations and real data analysis.

All the multi-index models mentioned above are used to deal with continuous response variables. Generalized linear model (GLM) is one the most commonly used technique in binary and multinomial regression. However, notice that there is no conjugate prior exists for the parameters in the model except for normal regression, inference in Bayesian GLMs is difficult because of the computation challenge in posterior sampling. Albert & Chib (1993) developed an auxiliary variable approach for binary probit regression models such that the conditional distributions of the model parameters is equivalent to those under the Bayesian normal linear regression model with Gaussian noise:

$$\begin{aligned} Y_i &= \mathbb{1}(Z_i > 0) \\ Z_i &= \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i \\ \epsilon_i &\sim \mathbf{N}(0, \sigma^2) \end{aligned} \tag{1.0-3}$$

Here, $\boldsymbol{X}_i$ is a p-dimensional vector of the covariates, $\boldsymbol{\beta}$ represents a p-dimensional vector of the loading of each index. The random error term ϵ_i is assumed to follow a normal distribution with zero mean. O'Brien & Dunson (2004) extended this linear regression model to deal with binary multiple response variable. They also proposed an idea about how their model can potentially handle categorical response variables (more than two categories). However, they did not provide a practical implementation. Sun & Jiang (2020) proposed a semi-parametric logistic regression using

Gaussian process prior. We extend the Bayesian normal linear regression model to multi-index nonlinear regression for binary and categorical response variables. Similar to what we do in the Bayesian multi-index additive model, the indexes $\boldsymbol{\beta}$'s are parameterized using the spherical coordinates. The ridge functions $f(\cdot)$ of each additive component is approximated by the Bayesian B-splines with regularly spaced knots.

This dissertation is organized as follows. We present the Bayesian multi-index additive model for continuous response variable in Chapter 2. The Bayesian single-index model for binary response variable is presented in Chapter 3. The Bayesian multi-index model for categorical response variable is presented in Chapter 4. We show the advantage of our methods comparing with some well-known competing methods through Simulation studies and a real data analysis in Chapters 5 and 6. In Chapter 7, we conclude the article with some discussions. The technical proofs are found in the appendix.

CHAPTER 2

BAYESIAN MULTI-INDEX ADDITIVE MODEL FOR CONTINUOUS RESPONSE VARIABLE

In this Chapter, we have re-parameterized our proposed Bayesian multi-index additive model (BMIAM) to facilitate the computation and prior specification, and then we develop the sampling scheme for the posterior of unknowns in the model and discuss the choice for the number of knots and indexes in the end.

2.1 Model Representation and Parametrization

The proposed Bayesian multi-index additive model is shown as following

$$Y_i = \sum_{d=1}^{D} f_d\left(\boldsymbol{X}_i'\boldsymbol{\beta}_d\right) + \mu + \epsilon_i. \tag{2.1-1}$$

Here, $\boldsymbol{X}_i$ is a p-dimensional vector of the covariates, $\boldsymbol{\beta}_d$ represents a p-dimensional vector of the loading of each index, $f_d(\cdot)$ is the ridge function corresponding to the d-th index. The random error term ϵ_i is assumed to follow a normal distribution with zero mean. The BMIAM includes the Bayesian generalized additive model (Liang & Zeger (1986)) and the Bayesian single-index model (Dhara et al. (2020)) as special cases.

2.1.1 Hyper-Spherical Parametrization of Index Parameters

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_D)$ be a $p \times D$ matrix by combining all the column vectors $\boldsymbol{\beta}_d$ together. Note that the L_2 norm of each column of $\boldsymbol{\beta}$ is one, i.e., $\|\boldsymbol{\beta}_d\|_2 = 1$. We thus can express $\boldsymbol{\beta}_d$ by using the hyper-spherical coordinate. Namely, let $\boldsymbol{\theta}_d = (\theta_{1,d}, \theta_{2,d}, \cdots, \theta_{p-1,d})$ be the angular coordinates corresponding to the d-th column of $\boldsymbol{\beta}$ and then $\boldsymbol{\beta}_d = \boldsymbol{T}(\boldsymbol{\theta}_d)$, where

$$\boldsymbol{T}(\boldsymbol{\theta}_d) = \begin{pmatrix} \cos\theta_{p-1,d}\cos\theta_{p-2,d}\cdots\cos\theta_{2,d}\cos\theta_{1,d} \\ \cos\theta_{p-1,d}\cos\theta_{p-2,d}\cdots\cos\theta_{2,d}\sin\theta_{1,d} \\ \cos\theta_{p-1,d}\cos\theta_{p-2,d}\cdots\sin\theta_{2,d} \\ \cdots \\ \sin\theta_{p-1,d} \end{pmatrix} \tag{2.1-2}$$

2.1.2 B-spline Representation of BMIAM

For the ridge functions $f_d(\cdot)$, we model them using Bayesian B-spline technique. Let b be the degree of the chosen B-spline basis and $b_{h,b}(\cdot)$ be the h-th B-spline basis function of degree b, which is a function of $\boldsymbol{X}_i'\boldsymbol{\beta}_d$. Define $\boldsymbol{B}_{id}$ to be an m-dimension column vector, where $m = \#\{\text{knots}\} + b - 1$ with $\#\{\text{knots}\}$ being the number of the knots for the spline and the h-th element of $\boldsymbol{B}_{id}$ is given by $b_{h,b}(\boldsymbol{X}_i'\boldsymbol{\beta}_d)$. Define $\boldsymbol{B}_d = (\boldsymbol{B}_{1d}, \cdots, \boldsymbol{B}_{nd})'$ to be a $n \times m$ matrix, where the i-th row is $\boldsymbol{B}_{id}'$. Let $\boldsymbol{\eta}_d$ be the B-spline regression parameters corresponding to the d-th ridge function $f_d(\cdot)$, then the B-spline representation of BMIAM can be written as:

$$\mathrm{E}(Y_i \mid X_i) = \mu + \sum_{d=1}^{D} \boldsymbol{B}_{id}'\boldsymbol{\eta}_d. \tag{2.1-3}$$

2.2 Likelihood Function and Prior Distributions

2.2.1 Likelihood Function

Let $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)'$ and $\boldsymbol{Y} = (Y_1, \cdots, Y_n)'$. According to (2.1-1), (2.1-2), and (2.1-3), the likelihood function of the proposed BMIAM is

$$\mathcal{L}(\sigma, \mu, \boldsymbol{\eta}, \boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n} (Y_i - \sum_{d=1}^{D} \boldsymbol{B}'_{id}\boldsymbol{\eta}_d - \mu)^2}{2\sigma^2} \right), \quad (2.2\text{-}4)$$

where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_D)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_D)$ and $\boldsymbol{\Theta} = (\sigma, \mu, \boldsymbol{\eta}, \boldsymbol{\theta})$ are all unknown parameters of the model.

2.2.2 Prior Distributions

To perform the Bayesian analysis of BMIAM model, we need to assume prior distributions on all these unknown parameters. We start with the prior specification of $\boldsymbol{\eta}_d$'s. A popular choice for $\boldsymbol{\eta}_d$'s is the g-prior proposed by Zellner (1986). Further, smoothness of the B-spline approximation can be achieved by adding penalty on adjacent B-spline coefficients (Lang & Brezger (2004)). Thus, we specify a prior distribution $\pi(\boldsymbol{\eta}_d)$ as follows for $\boldsymbol{\eta}_d$, $d = 1, \cdots, D$:

$$\boldsymbol{\eta}_d \mid \tau_d \sim \mathcal{N}_m \left(\mathbf{0}, \left(\frac{\boldsymbol{M}\boldsymbol{M}' + \boldsymbol{K}}{\tau_d^2} \right)^{-1} \right), \quad (2.2\text{-}5)$$

where $\mathcal{N}_m(\cdot,\cdot)$ indicates a multivariate normal distribution,

$$\boldsymbol{M} = \begin{pmatrix} -1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}_{m\times m}, \boldsymbol{K} = \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & \vdots & 0 \\ 0 & \cdots & \frac{\tau_d^2}{\sigma_\eta^2} & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}_{m\times m},$$

and τ_d is a hyper-parameter in the prior distribution. Note that this prior is equivalent to assume

$$\eta_{d,\ell} - \eta_{d,\ell-1} \sim \mathcal{N}(0,\tau_d^2), \text{ and } \eta_{d,\frac{m+1}{2}} \sim \mathcal{N}(0,\sigma_\eta^2),$$

where $\ell = 2,\cdots,m$. The amount of smoothness is controlled by the variance parameter τ_d^2. We set the variance $\sigma_\eta^2 = 100$ to make the prior less informative.

For τ_d^2 and σ^2, we choose the inverse gamma priors $\pi(\tau_d^2)$ for τ_d^2, $d = 1,\cdots,D$, and $\pi(\sigma^2)$ for σ^2, respectively, i.e.,

$$\tau_d^2 \sim IG\left(a_\tau, b_\tau\right), \ \sigma^2 \sim IG\left(a_\sigma, b_\sigma\right). \tag{2.2-6}$$

where $IG(\cdot,\cdot)$ represents an inverse gamma prior with the first dot as a shape parameter and the second dot as a scale parameter. Follow the recommendation by Lang & Brezger (2004), we choose $a_\tau = a_\sigma = 1$, and $b_\tau = b_\sigma = 0.005$ such that the prior is a nearly diffuse prior and highly dispersed.

For the intercept parameter μ, we assume a non-informative prior $\pi(\mu)$ for μ, i.e.,

$$\mu \propto const. \tag{2.2-7}$$

For the index parameter, we assign priors $\pi(\boldsymbol{\theta})$ for the angular coordinates $\boldsymbol{\theta}$ as:

$$\theta_{1,d} \sim U(\phi_{center} - \frac{\pi}{2}, \phi_{center} + \frac{\pi}{2}), \text{ and } \theta_{j,d} \sim U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right), \tag{2.2-8}$$

where $j = 2, 3, \cdots, p-1$, $d = 1, 2, \cdots, D$ and $U(\cdot, \cdot)$ stands for the uniform distribution. A natural and common choice for ϕ_{center} is to set $\phi_{center} = \frac{\pi}{2}$. Based on our experience, the sequence reaches to a good mixture faster if ϕ_{center} is set to be closer to the true value of $\theta_{1,d}$. Thus, we get a vague estimate of ϕ_{center} by running a pre-burning period and use it as ϕ_{center} in the MCMC.

Denote $\boldsymbol{\Lambda}$ to be a collection of unknown parameters $(\mu, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\tau}, \sigma)$ for our proposed BMIAM, then the joint prior $p(\boldsymbol{\Lambda})$ is equal to

$$p(\boldsymbol{\Lambda}) = p(\mu) \times p(\boldsymbol{\eta}) \times p(\boldsymbol{\theta}) \times p(\boldsymbol{\tau}) \times p(\sigma). \tag{2.2-9}$$

2.3 Sampling from Posterior Distribution

In this section, we will develop the MCMC algorithm. By applying the prior in (2.2-9) and using the likelihood in (2.2-4), we obtain the joint posterior $p(\boldsymbol{\Lambda} \mid \boldsymbol{X}, \boldsymbol{Y})$

given the observed data $\boldsymbol{X}, \boldsymbol{Y}$ as

$$
\begin{aligned}
p(\boldsymbol{\Lambda} \mid \boldsymbol{X}, \boldsymbol{Y}) &\propto \mathcal{L}(\sigma, \mu, \boldsymbol{\eta}, \boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y}) \times p(\boldsymbol{\Lambda}) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(Y_i - \mu - \sum_{d=1}^{D} \boldsymbol{B}'_{id}\boldsymbol{\eta}_d)^2}{2\sigma^2}\right) \\
&\times \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\tau_d^2}} \exp\left(-\frac{\boldsymbol{\eta}'_d(\boldsymbol{M}\boldsymbol{M}' + \boldsymbol{K})\boldsymbol{\eta}_d}{2\tau_d^2}\right) \\
&\times \prod_{d=1}^{D} \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} \tau_d^{2^{-a_\tau-1}} \exp\left(-\frac{b_\tau}{\tau_d^2}\right) \\
&\times \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \sigma^{2^{-a_\sigma-1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \\
&\times \mathbb{1}(0 < \theta_{1,1}, ..., \theta_{1,D} < \pi) \\
&\times \prod_{d=1}^{D} \mathbb{1}(-\frac{\pi}{2} < \theta_{2,d}, ..., \theta_{p-1,d} < \frac{\pi}{2}),
\end{aligned}
\tag{2.3-10}
$$

noticing that $\boldsymbol{B}_{id}$ is a function of $\boldsymbol{X}'_i\boldsymbol{\beta}_d$, where $\boldsymbol{\beta}_d = \boldsymbol{T}(\boldsymbol{\theta}_d)$, and $\mathbb{1}(\cdot)$ is an indicator function.

Based on the joint posterior (2.3-10), we can obtain the conditional distribution of the joint posterior distribution for each parameter and derive the corresponding MCMC sampling procedure. To simplify the notations, we use '$-$' to denote the observed data and all the other parameters in our derivation.

Step 1 : Given $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, μ and the hyperparameters a_σ and b_σ, we have the full conditional distribution of the variance parameter σ^2 is an inverse gamma as

$$
[\sigma^2 \mid -] \sim IG\left(a_\sigma + n/2, b_\sigma + \boldsymbol{z}'\boldsymbol{z}/2\right),
$$

where $\boldsymbol{z} = \boldsymbol{Y} - \mu - \sum_{d=1}^{D} \boldsymbol{B}'_d\boldsymbol{\eta}_d$.

Step 2 : For any $\boldsymbol{\eta}_d$ (i.e., the coefficients corresponding to the B-splines of the d-th

ridge function $f_d(\cdot)$), $d = 1, \cdots, D$, provided that $\boldsymbol{\eta}$, σ^2 and τ_d^2 are known, its full conditional distribution is a multivariate normal distribution, that is,

$$[\boldsymbol{\eta}_d \mid -] \sim \mathcal{N}_m \left(\boldsymbol{P}_d^{-1} \frac{1}{\sigma^2} \boldsymbol{B}_d' \left(\boldsymbol{Y} - \tilde{\boldsymbol{\zeta}}_d \right), \boldsymbol{P}_d^{-1} \right),$$

where $\boldsymbol{P}_d$ and $\tilde{\boldsymbol{\zeta}}_d$ are defined as following,

$$\boldsymbol{P}_d = \frac{1}{\sigma^2} \boldsymbol{B}_d' \boldsymbol{B}_d + \frac{1}{\tau_d^2} \boldsymbol{M}\boldsymbol{M}', \text{ and } \tilde{\boldsymbol{\zeta}}_d = \sum_{i \neq d, i=1}^{D} \boldsymbol{B}_i' \boldsymbol{\eta}_i.$$

Step 3 : For the intercept μ, given η and σ^2, its full conditional distribution is an univariate normal distribution as below,

$$[\mu \mid -] \sim \mathcal{N} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \sum_{d=1}^{D} \boldsymbol{B}_{id}' \boldsymbol{\eta}_d), \ \frac{\sigma^2}{n} \right).$$

Step 4 : The hyperparameters a_τ and b_τ are specified as well as $\boldsymbol{\eta}$ is given, then the full conditional distribution of τ_d is an inverse gamma distribution, i.e.,

$$[\tau_d^2 \mid -] \sim IG \left(a_\tau', b_\tau' \right),$$

where $a_\tau' = a_\tau + \text{rank} \left(\boldsymbol{M}\boldsymbol{M}' \right)/2$ and $b_\tau' = b_\tau + \boldsymbol{\eta}_d' \boldsymbol{M}\boldsymbol{M}' \boldsymbol{\eta}_d/2$.

Step 5 : Finally, we need to derive the sampling scheme for the angular coordinate $\boldsymbol{\theta}_d$. By given μ, σ^2 and $\boldsymbol{\theta}_d$ except $\theta_{j,d}$ are known, the full conditional distribution

of $\theta_{j,d}$ is proportional to

$$\begin{aligned} p(\theta_{j,d} \mid -) &\propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(Y_i - \sum_{d=1}^{D} \boldsymbol{B}'_{id}(\boldsymbol{X}_i\boldsymbol{T}(\boldsymbol{\theta}_d))\boldsymbol{\eta}_d - \mu)^2}{2\sigma^2}\right) \\ &\times \mathbb{1}(\phi_{center} - \frac{\pi}{2} < \theta_{1,1}, ..., \theta_{1,D} < \phi_{center} + \frac{\pi}{2}) \\ &\times \prod_{d=1}^{D} \mathbb{1}(-\frac{\pi}{2} < \theta_{2,d}, ..., \theta_{p-1,d} < \frac{\pi}{2}), \end{aligned}$$

which is not in a closed form. Therefore, we apply the Metropolis-Hastings algorithm to sample $\theta_{j,d}$. Notice that the ranges for those $\theta_{j,d}$'s are $(0, \pi)$ for $j = 1$ and $(-\frac{\pi}{2}, \frac{\pi}{2})$ for $j > 1$. Hence, it is natural to sample $\theta_{j,d}^{(new)}$ from a truncated normal distribution centered at $\theta_{j,d}^{(old)}$, with a proposal variance $\sigma^2_{proposal}$. To be specific, for $\theta_{1,d}$, we sample the $\theta_{1,d}^{(new)}$ from

$$\theta_{1,d}^{(new)} \mid \theta_{1,d}^{(old)} \sim \mathcal{N}_{(0,\pi)}(\theta_{1,d}^{(old)}, \sigma^2_{proposal}),$$

with $\mathcal{N}_{(0,\pi)}(\cdot, \cdot)$ denoting a normal distribution truncated by zero on the left and by π on the right. Then, according to the acceptance ratio

$$\alpha_1 = \min\left\{1, e^{\frac{2(\boldsymbol{Y}-\mu\mathbb{1}_n)'(\sum_{d=1}^{D}(\boldsymbol{B}_d^{(new)} - \boldsymbol{B}_d^{(old)})\boldsymbol{\eta}_d) - \sum_{d=1}^{D}\boldsymbol{\eta}'_d(\boldsymbol{B}_d'^{(new)}\boldsymbol{B}_d^{(new)} - \boldsymbol{B}_d'^{(old)}\boldsymbol{B}_d^{(old)})\boldsymbol{\eta}_d}{2\sigma^2}}\right\},$$

we will accept $\theta_{1,d}^{(new)}$ or we will remain at $\theta_{1,d}^{(old)}$. Notice that $\boldsymbol{B}_d$ is a function of $\boldsymbol{X}\boldsymbol{T}(\boldsymbol{\theta}_d)$, thus $\boldsymbol{B}_d^{(new)} = \boldsymbol{B}_d(\boldsymbol{X}\boldsymbol{T}(\boldsymbol{\theta}_d^{(new)}))$ and $\boldsymbol{B}_d^{(old)} = \boldsymbol{B}_d(\boldsymbol{X}\boldsymbol{T}(\boldsymbol{\theta}_d^{(old)}))$, where $\boldsymbol{\theta}_d^{(new)}$ is the new hyper-spherical coordinate vector drawn from the proposal distribution, and $\boldsymbol{\theta}_d^{(old)}$ is the hyper-spherical coordinate vector from the previous draws. Similarly, for $\theta_{j,d}$ with $j > 1$, we accept the samples $\theta_{j,d}^{(new)}$ from

$$\theta_{j,d}^{(new)} \mid \theta_{j,d}^{(old)} \sim \mathcal{N}_{(-\frac{\pi}{2}, \frac{\pi}{2})}(\theta_{j,d}^{(old)}, \sigma^2_{proposal})$$

with the acceptance ratio

$$\alpha_2 = \min\left\{1, e^{\frac{2(\boldsymbol{Y}-\mu\mathbb{1}_n)'(\sum_{d=1}^{D}(\boldsymbol{B}_d^{(new)}-\boldsymbol{B}_d^{(old)})\boldsymbol{\eta}_d)-\sum_{d=1}^{D}\boldsymbol{\eta}_d'(\boldsymbol{B}_d'^{(new)}\boldsymbol{B}_d^{(new)}-\boldsymbol{B}_d'^{(old)}\boldsymbol{B}_d^{(old)})\boldsymbol{\eta}_d}{2\sigma^2}}\right\}.$$

2.4 Selection of Number of Knots and Number of indexes

So far, we have assumed that the number of knots for the B-spline and the number of index components D in (1.0-2) are known. However, often in the practice, we need to estimate these numbers. In the literature, generalized cross-validation (GCV) is the most commonly used approach to determine the 'number of knots' and the 'number of index components'. However, GCV is computationally extensive in our model settings. Thus, to determine these numbers, we consider a modified Bayesian information criterion (BIC), i.e., :

$$\text{M-BIC} = n\log\left(\frac{1}{n}\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2\right) + qh(n).$$

Here, q is the total number of parameters needed to estimate in the proposed BMIAM model, and n is the sample size. There are different choices of the function $h(\cdot)$ in the literature. For instance, Huang & Yang (2004) suggested to use $h(n) = \log(n)$. In Hannan & Quinn (1979), the authors suggested to apply the Hannan–Quinn information criteria by setting $h(n) = \log(\log n)$. When $h(n) = \log(n)$, it penalizes too much on the number of model parameters. In this paper, we employ $h(n) = \log(\log n)$ in the numerical studies.

2.5 Posterior Consistency of Model Prediction

For Bayesian single-index models using Gaussian process prior, methods proposed by Choi et al. (2011) and Dhara et al. (2020) have been shown to possess posterior consistency of model prediction. For BMIAM, we also investigate the theoretical property of posterior consistency of model prediction. Notice that the $\boldsymbol{Y}$ is usually standardized, then we focus on the regression function $g(\boldsymbol{X}) = \sum_{d=1}^{D} f_d(\boldsymbol{X}\boldsymbol{\beta}_d)$.

Denote $(g_0(\cdot), \boldsymbol{\beta}^0)$ be the true value of the parameters that generate the data, where $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^0, \cdots, \boldsymbol{\beta}_D^0)'$ is the vector of true values of index parameters and $g_0(\boldsymbol{X}_i) = \sum_{d=1}^{D} f_d(\boldsymbol{X}_i'\boldsymbol{\beta}_d^0) = \sum_{d=1}^{D} \boldsymbol{B}_d^0\boldsymbol{\eta}_d^0 + \delta$ in the proposed BMIAM model with δ presenting the bias induced from the approximation of B-spline basis expansion and $\boldsymbol{\eta}^0 = (\boldsymbol{\eta}_1^0, \cdots, \boldsymbol{\eta}_D^0)'$ being the true values of B-spline parameters. Notice according to our definition earlier, $\boldsymbol{B}_d$ is a function of $\boldsymbol{X}_i'\boldsymbol{\beta}_d$ and similarly, $\boldsymbol{B}_d^0$ is a function of $\boldsymbol{X}_i'\boldsymbol{\beta}_d^0$. Further, we define $\boldsymbol{B}^0 = (\boldsymbol{B}_1^0, \boldsymbol{B}_2^0, \cdots, \boldsymbol{B}_D^0)$ and denote $g(\boldsymbol{X}_i) = \sum_{d=1}^{D} \boldsymbol{B}_d\boldsymbol{\eta}_d$, which in fact is used in our procedure to estimate $g_0(\boldsymbol{X}_i)$. However, when the B-spline basis expansion is implemented, we are expected that under some mild conditions, the prediction performance of $\sum_{d=1}^{D} \boldsymbol{B}_d\boldsymbol{\eta}_d$ is asymptotically concentrated around the additive mean function under the truth. To prove this, we have to first set up some assumptions as follows.

Assumptions:

A1. $\|\boldsymbol{B}(x)\| = O(m)$, for $x \in (-L, L)$, where $L = \max_i\{\|\boldsymbol{X_i}\|\}$, $i = 1, ..., n$;

A2. $\max_d\{\|\boldsymbol{\eta}_d^0\|_\infty\} \leq \gamma_3 W$ for fixed $\gamma_3 \in (0, 1), d = 1, ..., D$, and W is non-decreasing with n;

A3. All ridge functions are κ-times continuously differentiable;

The assumption **A1** is imposed to restrict the maximum L^1-norm of the row vectors in all B-spline basis design matrices are on the same order of m. Assumption **A2** is imposed such that ridge functions are bounded. In **A3**, κ defines the smoothness for all ridge functions, so the bias induced by a m-dimensional basis expansion satisfies $\delta = O(m^{-\kappa})$. Similar assumption is given in Wei et al. (2020).

Further, notice that the BMIAM model defined in (2.1-1) is not identifiable if there are no restrictions on the ridge functions $f_d(\cdot)$'s. We assume the following conditions to make the BMIAM model is identifiable, which have been used in Yuan (2011).

B1. $f_d(0) = 0$, for $d = 1, \cdots, D$;

B2. There is at most one linear ridge function; furthermore, The projection indexes matrix is of column full rank. If $f_d(\cdot)$ is a linear ridge function, then $\boldsymbol{\beta}_d'\boldsymbol{\beta}_k = 0$ for all $k \neq d$;

B3. There is at most one quadratic ridge function.

We then have the following theorem.

Theorem 1. *Let* $\boldsymbol{g}_0 = [g_0(\boldsymbol{X}_1), g_0(\boldsymbol{X}_2), \cdots, g_0(\boldsymbol{X}_n)]$, *and* $\boldsymbol{g} = [g(\boldsymbol{X}_1), g(\boldsymbol{X}_2), \cdots, g(\boldsymbol{X}_n)]$. *Assume the regression model* $\boldsymbol{Y} = \boldsymbol{g}_0 + \boldsymbol{\epsilon} = \sum_{d=1}^{D} \boldsymbol{B}_d(\boldsymbol{X}\boldsymbol{\beta}_d^0)\boldsymbol{\eta}_d^0 + \boldsymbol{\delta} + \boldsymbol{\epsilon}$. *If conditions* **A1** - **A3**, *and* **B1** - **B3** *hold, and* $\boldsymbol{g}$ *follows prior* (2.2-9), *then*

$$\mathbb{E}_0\mathbb{P}\left(\|\boldsymbol{g} - \boldsymbol{g}_0\|_{2,n}^2 \leq M\epsilon_n^2 \mid \boldsymbol{X}, \boldsymbol{Y}\right) \to 1 \tag{2.5-11}$$

is satisfied with $\epsilon_n^2 = \frac{1}{n}\left[[log(n)]^{D(p-1)+1} + log\left(\frac{(2\pi)^{\frac{m}{2}}\sqrt{m}(b_\tau + \frac{m}{2})^{a_\tau + \frac{m}{2}}}{(\frac{m}{2})!}\right)\right]$, *for a suitably positive number* M. *Here,* $\mathbb{E}_0$ *denotes the expectation under the true data generating mechanism, and* $\|\boldsymbol{g} - \boldsymbol{g}_0\|_{2,n}^2 = \frac{1}{n}\sum_{i=1}^{n}\{g(\boldsymbol{X}_i) - g_0(\boldsymbol{X}_i)\}^2$.

This theorem proves that, when the B-spline basis expansion is implemented, the prediction is asymptotically concentrated around the additive mean function under the truth.

CHAPTER 3

BAYESIAN SINGLE-INDEX MODEL FOR BINARY RESPONSE VARIABLE

3.1 Model Representation and Parametrization

The proposed Bayesian multi-index additive model (BMIAM) is shown as follow:

$$\begin{aligned} Y_i &= \mathbb{1}(Z_i > 0) \\ Z_i &= f(\mathbf{X}_i\boldsymbol{\beta}) + \epsilon_i \\ \epsilon_i &\sim \mathbf{N}(0, \sigma^2) \end{aligned} \tag{3.1-1}$$

Here, $\boldsymbol{X}_i$ is a p-dimensional vector of the covariates, $\boldsymbol{\beta}$ represents a p-dimensional vector of the loading of each index, and $f(\cdot)$ is the link function. The random error term ϵ_i is assumed to follow a normal distribution with zero mean.

3.1.1 Hyper-Spherical Parametrization of Index Parameters

We express $\boldsymbol{\beta}$ by using the hyper-spherical coordinate. Namely, let $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_{p-1})$ be the angular coordinates and then $\boldsymbol{\beta} = \boldsymbol{T}(\boldsymbol{\theta})$, where

$$\boldsymbol{T}(\boldsymbol{\theta}) = \begin{pmatrix} \cos\theta_{p-1}\cos\theta_{p-2}\cdots\cos\theta_2\cos\theta_1 \\ \cos\theta_{p-1}\cos\theta_{p-2}\cdots\cos\theta_2\sin\theta_1 \\ \cos\theta_{p-1}\cos\theta_{p-2}\cdots\sin\theta_2 \\ \cdots \\ \sin\theta_{p-1} \end{pmatrix} \tag{3.1-2}$$

3.1.2 B-spline Representation of Auxiliary Variable

For the link function, we model them using Bayesian the B-spline approach. Let b be the degree of the chosen B-spline basis and $B_{h,b}(\cdot)$ be the h-th B-spline basis function of degree b, which is a function of $\boldsymbol{X}_i'\boldsymbol{\beta}$. Define $\boldsymbol{B}_i$ to be an m-dimension vector, where $m = \#\{\text{knots}\} + b - 1$, $\#\{\text{knots}\}$ is the number of the knots for the spline and the h-th element of $\boldsymbol{B}_i$ is given by $B_{h,b}(\boldsymbol{X}_i'\boldsymbol{\beta})$. Define $\boldsymbol{B}$ to be a $n \times m$ matrix, where the i-th row is $\boldsymbol{B}_i'$. Let $\boldsymbol{\eta}$ be the B-spline regression parameters, then the B-spline representation of Bayesian single-index model for binary response can be written as:

$$E(Z_i|X_i) = \boldsymbol{B}_i'\boldsymbol{\eta}. \tag{3.1-3}$$

3.2 Likelihood Function and Prior Distributions

3.2.1 Likelihood Function

According to (3.1-1), (3.1-2), and (3.1-3), we have likelihood functions:

$$P(\boldsymbol{Y}|\boldsymbol{Z},\sigma,\boldsymbol{X},\boldsymbol{\eta},\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(Y_i \mathbb{1}(Z_i > 0) + (1 - Y_i)\mathbb{1}(Z_i \leq 0)\right), \tag{3.2-4}$$

$$P(\boldsymbol{Z}|\sigma,\boldsymbol{X},\boldsymbol{\eta},\boldsymbol{\theta}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(z_i - \boldsymbol{B}_i'\boldsymbol{\eta})^2}{2\sigma^2}\right), \tag{3.2-5}$$

where $(\sigma, \boldsymbol{\eta}, \boldsymbol{\theta})$ are unknown parameters of the model.

3.2.2 Prior Distributions

Following prior distributions are assumed on those unknown parameters. We start with the prior of $\boldsymbol{\eta}$'s. Similar to BMIAM, we specific a normal prior $\Pi_{\boldsymbol{\eta}}$ for $\boldsymbol{\eta}$:

$$\boldsymbol{\eta}|\tau \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta\right), \tag{3.2-6}$$

where $\boldsymbol{\Sigma}_\eta = \left(\frac{\boldsymbol{MM'}+\boldsymbol{K^*}}{\tau^2}\right)^{-1}$.

$$\boldsymbol{M} = \begin{pmatrix} -1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \ \boldsymbol{K^*} = \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & \vdots & 0 \\ 0 & \cdots & \frac{\tau^2}{k^2} & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The amount of smoothness is controlled by the variance parameter τ^2. We set the variance $k^2 = 100$ to avoid any informative choices.

For τ^2 and σ^2, we choose the Inverse Gamma prior Π_{τ^2} for τ^2, and Π_{σ^2} for σ^2:

$$\tau^2 \sim IG\left(a_\tau, b_\tau\right), \ \sigma^2 \sim IG\left(a_\sigma, b_\sigma\right). \tag{3.2-7}$$

Similar to, we choose $a_\tau = a_\sigma = 1$, and $b_\tau = b_\sigma = 0.005$ such that the prior is a nearly diffuse prior and highly dispersed.

For the index parameter, we assign priors $\Pi_{\boldsymbol{\theta}}$ for the angular coordinates $\boldsymbol{\theta}$ as:

$$\theta_1 \sim U(\phi_{center} - \frac{\pi}{2}, \phi_{center} + \frac{\pi}{2}), \text{ and } \theta_j \sim U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right), \tag{3.2-8}$$

where $j = 2, 3, \cdots, p-1$, and $U(a, b)$ stands for the uniform distribution from a to b. A natural and common choice for ϕ_{center} is to set $\phi_{center} = \frac{\pi}{2}$. Similar to BMIAM, the sequence reaches to a good mixture faster if ϕ_{center} is set to be closer to the true value of θ_1. Thus, we get a vague estimate of ϕ_{center} by running a pre-burning period

and use it as ϕ_{center} in the MCMC.

Therefore, the joint prior $\Pi(\boldsymbol{\Lambda})$ of the parameters $\boldsymbol{\Lambda} = (\boldsymbol{\theta}, \boldsymbol{\eta}, \tau^2, \sigma^2)$ of our BMIAM is

$$\Pi(\boldsymbol{\Lambda}) = \Pi_{\boldsymbol{\eta}} \times \Pi_{\boldsymbol{\theta}} \times \Pi_{\tau^2} \times \Pi_{\sigma^2} \tag{3.2-9}$$

3.3 Sampling from Posterior Distribution

In this section, we will develop the MCMC algorithm. According to the joint prior in (3.2-9) and the likelihood in (3.2-5), the joint posterior $p(\Lambda|\boldsymbol{X}, \boldsymbol{Y})$ given observed data $\boldsymbol{X}, \boldsymbol{Y}$ is

$$p(\Lambda|\boldsymbol{X}, \boldsymbol{Y}) \propto P(\boldsymbol{Y}|\boldsymbol{Z}, \sigma, \boldsymbol{X}, \boldsymbol{\eta}, \boldsymbol{\theta}) \times P(\boldsymbol{Z}|\sigma, \boldsymbol{X}, \boldsymbol{\eta}, \boldsymbol{\theta}) \times \Pi(\boldsymbol{\Lambda}) \tag{3.3-10}$$

Based on the joint posterior (3.3-10), we obtain the conditional distribution of the joint posterior distribution for each parameter.

Step 1 : Given $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, σ^2, and $\boldsymbol{Y}$ we have the full conditional distribution of the auxiliary variable Z_i, for $i = 1, 2, ..., n$, is a truncated normal distribution as

$$Z_i|- \sim \begin{cases} \boldsymbol{\mathcal{N}}_{(0,+\infty)}\left(\boldsymbol{B}_i'\boldsymbol{\eta}, \sigma^2\right), & Y_i = 1 \\ \boldsymbol{\mathcal{N}}_{(-\infty,0)}\left(\boldsymbol{B}_i'\boldsymbol{\eta}, \sigma^2\right), & Y_i = 0 \end{cases} \tag{3.3-11}$$

Step 2 : For the variance parameter σ^2, provided that $\boldsymbol{Z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are known, its full conditional distribution is an inverse gamma distribution, that is,

$$\sigma^2|- \sim IG\left(a_\sigma + n/2, b_\sigma + \boldsymbol{\epsilon}'\boldsymbol{\epsilon}/2\right), \tag{3.3-12}$$

where $\boldsymbol{\epsilon} = \boldsymbol{Z} - \boldsymbol{B}'(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\eta}$.

Step 3 : For the B-spline regression coefficients $\boldsymbol{\eta}$, given $\boldsymbol{Z}$, $\boldsymbol{\theta}$, and σ^2, its full conditional distribution is an multivariate normal distribution as below:

$$\boldsymbol{\eta}|- \sim \mathbf{N}\left(\widetilde{\boldsymbol{\mu}}_\eta, \widetilde{\boldsymbol{\Sigma}}_\eta\right), \tag{3.3-13}$$

where

$$\widetilde{\boldsymbol{\Sigma}}_\eta = \left(\boldsymbol{\Sigma}_\eta^{-1} + \sigma^{-2}\sum_{i=1}^{n} \boldsymbol{B}(\mathbf{X}_i\boldsymbol{\beta})\boldsymbol{B}'(\mathbf{X}_i\boldsymbol{\beta})\right)^{-1}, \tag{3.3-14}$$

$$\widetilde{\boldsymbol{\mu}}_\eta = \widetilde{\boldsymbol{\Sigma}}_\eta\left(\sigma^{-2}\sum_{i=1}^{n} Z_i\boldsymbol{B}'(\mathbf{X}_i\boldsymbol{\beta})\right) \tag{3.3-15}$$

Step 4 : For τ^2, given the hyperparameters a'_τ, b'_τ, and the B-spline regression coefficients $\boldsymbol{\eta}$, τ^2 is drawn from an inversed gamma distribution,

$$\tau^2 \sim IG\left(a'_\tau, b'_\tau\right), \tag{3.3-16}$$

where

$$a'_\tau = a_\tau + \operatorname{rank}\left(\boldsymbol{M}\boldsymbol{M}'\right)/2, \tag{3.3-17}$$

$$b'_\tau = b_\tau + \boldsymbol{\eta}'\boldsymbol{M}\boldsymbol{M}'\boldsymbol{\eta}/2. \tag{3.3-18}$$

Step 5 : Finally, we need to derive the sampling scheme for the angular coordinate $\boldsymbol{\theta}$. By given $\boldsymbol{Z}$, σ^2, $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ except θ_j are known, the full conditional distribution of θ_j is proportional to

$$\begin{aligned} p(\theta_j \mid -) \;&\propto\; \exp\left(-\frac{\sum_{i=1}^{n}(Z_i - \boldsymbol{B}'_i(\boldsymbol{X}_i\boldsymbol{T}(\boldsymbol{\theta}))\boldsymbol{\eta})^2}{2\sigma^2}\right) \\ &\times\; \mathbb{1}(\phi_{center} - \frac{\pi}{2} < \theta_1 < \phi_{center} + \frac{\pi}{2}) \\ &\times\; \mathbb{1}(-\frac{\pi}{2} < \theta_2, ..., \theta_{p-1} < \frac{\pi}{2}), \end{aligned}$$

which is not in a closed form. Therefore, we apply the Metropolis-Hastings algorithm to sample θ_j. Notice that the ranges for those θ_j's are $(0, \pi)$ for $j = 1$ and $(-\frac{\pi}{2}, \frac{\pi}{2})$ for $j > 1$. Hence, it is natural to sample $\theta_j^{(new)}$ from a truncated normal distribution centered at $\theta_j^{(old)}$, with a proposal variance $\sigma^2_{proposal}$. To be specific, for θ_1, we sample the $\theta_1^{(new)}$ from

$$\theta_1^{(new)} \mid \theta_1^{(old)} \sim \mathcal{N}_{(0,\pi)}(\theta_1^{(old)}, \sigma^2_{proposal}),$$

with $\mathcal{N}_{(0,\pi)}$ denoting a normal distribution truncated by 0 on the left and by π on the right. Then, according to the acceptance ratio α_1, we will accept $\theta_1^{(new)}$ or we will remain at $\theta_1^{(old)}$. Similarly, for θ_j with $j > 1$, we accept the samples $\theta_j^{(new)}$ from

$$\theta_j^{(new)} \mid \theta_j^{(old)} \sim \mathcal{N}_{(-\frac{\pi}{2}, \frac{\pi}{2})}(\theta_j^{(old)}, \sigma^2_{proposal})$$

with the acceptance ratio α_2.

$$\alpha_1 = \min \left\{ 1, \frac{\exp\left(\frac{\sum_{i=1}^n (Z_i - \boldsymbol{B}_i'(\boldsymbol{X}_i'\boldsymbol{T}(\boldsymbol{\theta}^{(old)}))\boldsymbol{\eta})^2}{2\sigma^2} \right)}{\exp\left(\frac{\sum_{i=1}^n (Z_i - \boldsymbol{B}_i'(\boldsymbol{X}_i'\boldsymbol{T}(\boldsymbol{\theta}^{(new)}))\boldsymbol{\eta})^2}{2\sigma^2} \right)} \right\},$$

$$\alpha_2 = \min \left\{ 1, \frac{\exp\left(\frac{\sum_{i=1}^n (Z_i - \boldsymbol{B}_i'(\boldsymbol{X}_i'\boldsymbol{T}(\boldsymbol{\theta}^{(old)}))\boldsymbol{\eta})^2}{2\sigma^2} \right)}{\exp\left(\frac{\sum_{i=1}^n (Z_i - \boldsymbol{B}_i'(\boldsymbol{X}_i'\boldsymbol{T}(\boldsymbol{\theta}^{(new)}))\boldsymbol{\eta})^2}{2\sigma^2} \right)} \right\},$$

where $\boldsymbol{\theta}^{(new)}$ is the new hyper-spherical coordinate vector drawn from the proposal distribution.

3.4 Selection of Number of Knots

To determine the 'number of knots', we consider the same modified BIC in BMIAM:

$$\text{M-BIC} = n \log \left(\frac{1}{n} \| \boldsymbol{Y} - \hat{\boldsymbol{Y}} \|^2 \right) + q f(n).$$

Here, q is the total number of parameters estimated in the model, and n is the sample size. In the proposed method, we used the Hannan-Quinn information criteria in the numerical studies. The number of knots with the minimum M-BIC will be selected.

CHAPTER 4

BAYESIAN MULTI-INDEX MODEL FOR CATEGORICAL RESPONSE VARIABLE

4.1 Model Representation and Parametrization

The proposed Bayesian multi-index model for categorical response variable is shown as follows:

$$
Y_i = \begin{cases} 1, & \mathbb{1}(Z_{i,1} > 0) \\ 2, & \mathbb{1}(Z_{i,1} \leq 0, Z_{i,2} > 0) \\ 3, & \mathbb{1}(Z_{i,1} \leq 0, Z_{i,2} \leq 0, Z_{i,3} > 0) \\ \vdots & \\ D-1, & \mathbb{1}(Z_{i,1} \leq 0, Z_{i,2} \leq 0, ..., Z_{i,D-2} \leq 0, Z_{i,D-1} > 0) \\ D, & \mathbb{1}(Z_{i,1} \leq 0, Z_{i,2} \leq 0, ..., Z_{i,D-1} \leq 0) \end{cases} \tag{4.1-1}
$$

$$
Z_{i,d} = f_d(\mathbf{X}_i \boldsymbol{\beta}_d) + \epsilon_{i,d}
$$

$$
\epsilon_{i,d} \sim \mathbf{N}(0, \sigma_d^2)
$$

Here, Y_i is a categorical response variable with range $Y_i = \{1, 2, ..., D\}$, $\boldsymbol{X}_i$ is a p-dimensional vector of the covariates, $\boldsymbol{\beta}_d$ represents a p-dimensional index vector corresponding to the d-th link function, and $f_d(\cdot)$ is the link function corresponding to the d-th index, $d = 1, 2, ..., D-1$. The random error term $\epsilon_{i,d}$ is assumed to follow a normal distribution with zero mean. For all the $d's$ in this chapter, $d = 1, 2, ..., D-1$.

4.1.1 Hyper-Spherical Parametrization of Index Parameters

We express $\boldsymbol{\beta}_d$ by using the hyper-spherical coordinate. Namely, let $\boldsymbol{\theta}_d = (\theta_{1,d}, \theta_{2,d}, \cdots, \theta_{p-1,d})$ be the hyper-spherical coordinates corresponding to the d-th index $\boldsymbol{\beta}_d$. Let $\boldsymbol{\beta}_d = \boldsymbol{T}(\boldsymbol{\theta}_d)$, where

$$\boldsymbol{T}(\boldsymbol{\theta}_d) = \begin{pmatrix} \cos\theta_{p-1,d}\cos\theta_{p-2,d}\cdots\cos\theta_{2,d}\cos\theta_{1,d} \\ \cos\theta_{p-1,d}\cos\theta_{p-2,d}\cdots\cos\theta_{2,d}\sin\theta_{1,d} \\ \cos\theta_{p-1,d}\cos\theta_{p-2,d}\cdots\sin\theta_{2,d} \\ \cdots \\ \sin\theta_{p-1,d} \end{pmatrix} \tag{4.1-2}$$

4.1.2 B-spline Representation of Auxiliary Variables

For the link functions, we model them using Bayesian the B-spline approach. Let b be the degree of the chosen B-spline basis and $b_{h,b}(\cdot)$ be the h-th B-spline basis function of degree b, which is a function of $\boldsymbol{X}_i'\boldsymbol{\beta}_d$. Define $\boldsymbol{B}_{id}$ to be an m-dimension column vector, where $m = \#\{\text{knots}\} + b - 1$ with $\#\{\text{knots}\}$ being the number of the knots for the spline and the h-th element of $\boldsymbol{B}_{id}$ is given by $b_{h,b}(\boldsymbol{X}_i'\boldsymbol{\beta}_d)$. Define $\boldsymbol{B}_d = (\boldsymbol{B}_{1d}, \cdots, \boldsymbol{B}_{nd})'$ to be a $n \times m$ matrix, where the i-th row is $\boldsymbol{B}_{id}'$. Let $\boldsymbol{\eta}_d$ be the B-spline regression parameters corresponding to the d-th index, then the B-spline representation of the auxiliary variable in Bayesian multi-index model for categorical response can be written as:

$$E(Z_{i,d}|X_i) = \boldsymbol{B}_{id}'\boldsymbol{\eta}_d. \tag{4.1-3}$$

4.2 Likelihood Function and Prior Distributions

4.2.1 Likelihood Function

Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n)'$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_{D-1})$, $\boldsymbol{\sigma} = (\sigma_1^2, \sigma_2^2, ..., \sigma_{D-1}^2)$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_{D-1})$. Let $\boldsymbol{Z}_d^* = (Z_{1,d}^*, Z_{2,d}^*, ..., Z_{n_d,d}^*)$ be the auxiliary variable vector obtained from $\boldsymbol{Z}_d$ by taking out all the $Z_{i,d}$ whose corresponding Y_i is less than d. n_d is the number of elements in vector $\boldsymbol{Z}_d^*$. Let $Y_{i,d}^*$ be the corresponding response variable of Z_{id}^*.

According to (4.1-1), (4.1-2), and (4.1-3), we have likelihood functions

$$\begin{aligned} P(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{\eta}, \boldsymbol{\theta}) = & \prod_{i:1<Y_i<D} \mathbb{1}(z_{i,Y_i} > 0)\mathbb{1}(z_{i,1}, ..., z_{i,Y_i-1} \leq 0) \\ & \times \prod_{i:Y_i=1} \mathbb{1}(z_{i,1} > 0) \\ & \times \prod_{i:Y_i=D} \mathbb{1}(z_{i,1}, ..., z_{i,D-1} \leq 0) \end{aligned} \tag{4.2-4}$$

$$\begin{aligned} P(\boldsymbol{Z}|\boldsymbol{\sigma}, \boldsymbol{X}, \boldsymbol{\eta}, \boldsymbol{\theta}) &= \prod_{d=1}^{D-1} P(\boldsymbol{Z}_d^*|\boldsymbol{\sigma}_d, \boldsymbol{X}, \boldsymbol{\eta}_d, \boldsymbol{\theta}_d) \\ &= \prod_{d=1}^{D-1} (2\pi\sigma_d^2)^{-\frac{n_d}{2}} \exp\left(-\frac{\sum_{i=1}^{n_d}(z_{id}^* - \boldsymbol{B}_{id}^{*\prime}\boldsymbol{\eta}_d)^2}{2\sigma_d^2}\right), \end{aligned} \tag{4.2-5}$$

where $(\boldsymbol{\sigma}, \boldsymbol{\eta}, \boldsymbol{\theta})$ are unknown parameters of the model.

4.2.2 Prior Distributions

Following prior distributions are assumed on those unknown parameters. We start with the prior of $\boldsymbol{\eta}_d$'s.

Similar to BMIAM, we specific a normal prior $\Pi_{\boldsymbol{\eta}_d}$ for $\boldsymbol{\eta}_d$:

$$\boldsymbol{\eta}_d|\tau_d \sim \boldsymbol{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\eta_d}\right), \tag{4.2-6}$$

where $\boldsymbol{\Sigma}_{\eta_d} = \left(\frac{\boldsymbol{MM'}+\boldsymbol{K^*}}{\tau_d^2}\right)^{-1}$.

$$\boldsymbol{M} = \begin{pmatrix} -1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \ \boldsymbol{K^*} = \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & \vdots & 0 \\ 0 & \cdots & \frac{\tau_d^2}{k^2} & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The amount of smoothness is controlled by the variance parameter τ^2. We set the variance $k^2 = 100$ to avoid any informative choices.

For τ_d^2 and σ_d^2, we choose the Inverse Gamma prior $\Pi_{\tau_d^2}$ for τ_d^2, and $\Pi_{\sigma_d^2}$ for σ_d^2:

$$\tau_d^2 \sim IG\left(a_\tau, b_\tau\right), \ \sigma_d^2 \sim IG\left(a_\sigma, b_\sigma\right). \tag{4.2-7}$$

Similar to, we choose $a_\tau = a_\sigma = 1$, and $b_\tau = b_\sigma = 0.005$ such that the prior is a nearly diffuse prior and highly dispersed.

For the index parameter, we assign priors $\Pi_{\boldsymbol{\theta}}$ for the angular coordinates $\boldsymbol{\theta}$ as:

$$\theta_{d,1} \sim U(\phi_{center} - \frac{\pi}{2}, \phi_{center} + \frac{\pi}{2}), \text{ and } \theta_{d,j} \sim U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right), \tag{4.2-8}$$

where $j = 2, 3, \cdots, p-1$, $d = 1, 2, \cdots, D-1$ and $U(a, b)$ stands for the uniform distribution from a to b. A natural and common choice for ϕ_{center} is to set $\phi_{center} = \frac{\pi}{2}$. Based on our experience, the sequence reach to a good mixture faster if ϕ_{center} is set to be closer to the true value of $\theta_{d,1}$. Thus, we get a vague estimate of ϕ_{center} by running a pre-burning period and use it as ϕ_{center} in the MCMC.

Therefore, the joint prior $\Pi(\boldsymbol{\Lambda})$ of the parameters $\boldsymbol{\Lambda} = (\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\sigma})$ of our BMIAM is

$$\Pi(\boldsymbol{\Lambda}) = \Pi_{\boldsymbol{\eta}} \times \Pi_{\boldsymbol{\theta}} \times \Pi_{\boldsymbol{\tau}} \times \Pi_{\boldsymbol{\sigma}} \tag{4.2-9}$$

4.3 Sampling from Posterior Distribution

In this section, we will develop the MCMC algorithm. According to the joint prior in (4.2-9) and the likelihood in (4.2-5), the joint posterior $p(\Lambda|\boldsymbol{X}, \boldsymbol{Y})$ given observed data $\boldsymbol{X}, \boldsymbol{Y}$ is

$$p(\Lambda|\boldsymbol{X}, \boldsymbol{Y}) \propto P(\boldsymbol{Y}|\boldsymbol{Z}, \sigma, \boldsymbol{X}, \boldsymbol{\eta}, \boldsymbol{\theta}) \times P(\boldsymbol{Z}|\sigma, \boldsymbol{X}, \boldsymbol{\eta}, \boldsymbol{\theta}) \times \Pi(\boldsymbol{\Lambda}) \tag{4.3-10}$$

Based on the joint posterior (4.3-10), we obtain the conditional distribution of the joint posterior distribution for each parameter. For all the $d's$ in this section, $d = 1, 2, ..., D-1$; For all the $i's$, $i = 1, 2, ..., n_d$.

Step 1 : Given $\boldsymbol{\theta}_d$, $\boldsymbol{\eta}_d$, σ_d^2, and $\boldsymbol{Y}_d^*$ we have the full conditional distribution of the degenerated auxiliary variable Z_{id}^*. It is a truncated normal distribution as

$$Z_{id}^*|- \sim \begin{cases} \boldsymbol{\mathcal{N}}_{(0,+\infty)}\left(\boldsymbol{B}_{id}^{*'}\boldsymbol{\eta}_d, \sigma_d^2\right), & Y_{i,d}^* = d \\ \boldsymbol{\mathcal{N}}_{(-\infty,0)}\left(\boldsymbol{B}_{id}^{*'}\boldsymbol{\eta}_d, \sigma_d^2\right), & otherwise \end{cases} \tag{4.3-11}$$

Step 2 : For the variance parameter σ_d^2, provided that $\boldsymbol{Z}_d^*$, $\boldsymbol{\theta}_d$, and $\boldsymbol{\eta}_d$ are known,

its full conditional distribution is an inverse gamma distribution, that is,

$$\sigma_d^2|- \sim IG\left(a_\sigma + n_d/2, b_\sigma + \boldsymbol{\epsilon}_d'\boldsymbol{\epsilon}_d/2\right), \tag{4.3-12}$$

where $\boldsymbol{\epsilon}_d = \boldsymbol{Z}_d^* - \boldsymbol{B}_d^*(\boldsymbol{X}_d^*\boldsymbol{\beta}_d)\boldsymbol{\eta}_d$.

Step 3 : For the B-spline regression coefficients $\boldsymbol{\eta}_d$, given $\boldsymbol{Z}_d^*$, $\boldsymbol{\theta}_d$, and σ_d^2, its full conditional distribution is an multivariate normal distribution as below,

$$\boldsymbol{\eta}_d|- \sim \mathbf{N}\left(\widetilde{\boldsymbol{\mu}}_{\eta_d}, \widetilde{\boldsymbol{\Sigma}}_{\eta_d}\right), \tag{4.3-13}$$

where

$$\widetilde{\boldsymbol{\Sigma}}_{\eta_d} = \left(\boldsymbol{\Sigma}_{\eta_d}^{-1} + {\sigma_d}^{-2}\boldsymbol{B}_d^{*\prime}\boldsymbol{B}_d^*\right)^{-1}, \tag{4.3-14}$$

$$\widetilde{\boldsymbol{\mu}}_\eta = \widetilde{\boldsymbol{\Sigma}}_\eta\left({\sigma_d}^{-2}\boldsymbol{B}_d^{*\prime}\boldsymbol{Z}_d^*\right) \tag{4.3-15}$$

Step 4 : For τ_d^2, given the hyperparameters a_τ', b_τ', and the B-spline regression coefficients $\boldsymbol{\eta}_d$, τ_d^2 is drawn from an inversed gamma distribution,

$$\tau_d^2 \sim IG\left(a_\tau', b_\tau'\right), \tag{4.3-16}$$

where

$$a_\tau' = a_\tau + \operatorname{rank}\left(\boldsymbol{M}\boldsymbol{M}'\right)/2, \tag{4.3-17}$$

$$b_\tau' = b_\tau + \boldsymbol{\eta}_d'\boldsymbol{M}\boldsymbol{M}'\boldsymbol{\eta}_d/2. \tag{4.3-18}$$

Step 5 : Finally, we need to derive the sampling scheme for the angular coordinate $\boldsymbol{\theta}_d$. Given that $\boldsymbol{Z}_d^*$, σ_d^2, $\boldsymbol{\eta}_d$ and $\boldsymbol{\theta}_d$ except $\theta_{j,d}$ are known, the full conditional

distribution of $\theta_{j,d}$ is proportional to

$$\begin{aligned} p(\theta_j \mid -) &\propto \exp\left(-\frac{\sum_{i=1}^{n_d}(Z^*_{i,d} - \boldsymbol{B}'_{id}(\boldsymbol{X}^*_{id}\boldsymbol{T}(\boldsymbol{\theta}_d))\boldsymbol{\eta}_d)^2}{2\sigma_d^2}\right) \\ &\times \mathbb{1}(\phi_{center} - \frac{\pi}{2} < \theta_{1,d} < \phi_{center} + \frac{\pi}{2}) \\ &\times \mathbb{1}(-\frac{\pi}{2} < \theta_{2,d}, ..., \theta_{p-1,d} < \frac{\pi}{2}), \end{aligned}$$

which is not in a closed form. Therefore, we apply the Metropolis-Hastings algorithm to sample θ_j. Notice that the ranges for those $\theta_{j,d}$'s are $(0, \pi)$ for $j = 1$ and $(-\frac{\pi}{2}, \frac{\pi}{2})$ for $j > 1$. Hence, it is natural to sample $\theta^{(new)}_{j,d}$ from a truncated normal distribution centered at $\theta^{(old)}_{j,d}$, with a proposal variance $\sigma^2_{proposal}$. To be specific, for $\theta_{1,d}$, we sample the $\theta^{(new)}_{1,d}$ from

$$\theta^{(new)}_{1,d} \mid \theta^{(old)}_{1,d} \sim \mathcal{N}_{(0,\pi)}(\theta^{(old)}_{1,d}, \sigma^2_{proposal}),$$

with $\mathcal{N}_{(0,\pi)}$ denoting a normal distribution truncated by 0 on the left and by π on the right. Then, according to the acceptance ratio α_1, we will accept $\theta^{(new)}_{1,d}$ or we will remain at $\theta^{(old)}_{1,d}$. Similarly, for $\theta_{j,d}$ with $j > 1$, we accept the samples $\theta^{(new)}_{j,d}$ from

$$\theta^{(new)}_{j,d} \mid \theta^{(old)}_{j,d} \sim \mathcal{N}_{(-\frac{\pi}{2},\frac{\pi}{2})}(\theta^{(old)}_{j,d}, \sigma^2_{proposal})$$

with the acceptance ratio α_2.

$$\alpha_1 = \min\left\{1, \frac{\exp\left(\frac{\sum_{i=1}^{n_d}(Z^*_{i,d} - \boldsymbol{B}^{*\prime}_{id}(\boldsymbol{X}^{*\prime}_{id}\boldsymbol{T}(\boldsymbol{\theta}^{(old)}_d))\boldsymbol{\eta}_d)^2}{2\sigma_d^2}\right)}{\exp\left(\frac{\sum_{i=1}^{n_d}(Z^*_{i,d} - \boldsymbol{B}^{*\prime}_{id}(\boldsymbol{X}^{*\prime}_{id}\boldsymbol{T}(\boldsymbol{\theta}^{(new)}_d))\boldsymbol{\eta}_d)^2}{2\sigma_d^2}\right)}\right\},$$

$$\alpha_2 = \min\left\{1, \frac{\exp\left(\frac{\sum_{i=1}^{n_d}(Z^*_{i,d} - \boldsymbol{B}^{*\prime}_{id}(\boldsymbol{X}^{*\prime}_{id}\boldsymbol{T}(\boldsymbol{\theta}_d^{(old)}))\boldsymbol{\eta}_d)^2}{2\sigma_d^2}\right)}{\exp\left(\frac{\sum_{i=1}^{n_d}(Z^*_{i,d} - \boldsymbol{B}^{*\prime}_{id}(\boldsymbol{X}^{*\prime}_{id}\boldsymbol{T}(\boldsymbol{\theta}_d^{(new)}))\boldsymbol{\eta}_d)^2}{2\sigma_d^2}\right)}\right\},$$

where $\boldsymbol{\theta}_d^{(new)}$ is the new hyper-spherical coordinate vector drawn from the proposal distribution.

4.4 Categorical Variable Mapping Determination

According to model 4.1-1, whether $Y_i = d$ or not is determined by the values of $(Z_{i,1}, ..., Z_{i,d})$ for $d < D$, and is determined by the values of $(Z_{i,1}, ..., Z_{i,D-1})$ for $d = D$. Therefore, an important issue is to determine how to code the categorical variable. Specifically, how to map each categorical variables to $\{1, 2, ..., D\}$. The posteriors of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ given the data are different for different coding of categorical variables.

To get the posteriors of $\boldsymbol{\beta}_1$ and $\boldsymbol{\eta}_1$, there are D possible choices to code which category to be 1. Then, having determined which category to be 1, to get the posteriors of $\boldsymbol{\beta}_2$ and $\boldsymbol{\eta}_2$, there are $D-1$ possible choices to code which category to be 2. Then, to get the posteriors of $\boldsymbol{\beta}_d$ and $\boldsymbol{\eta}_d$, there are $D-d+1$ possible choices of coding. The procedures to determine the mapping are listed below.

For each way of coding, calculate the within sample classification error according to the following steps:

Step 1 : Let $\boldsymbol{Y}_0^{(-d)}$ be a vector obtained from taking out all the Y_{id} whose value is less than d.

Step 2 : Define $\boldsymbol{Y}^{(-d)}$ be the degenerated outcome vector obtained from $\boldsymbol{Y}_0^{(-d)}$ by letting $\boldsymbol{Y}_i^{(-d)} = 1$ if $\boldsymbol{Y}_{0_i}^{(-d)} = d$, and $\boldsymbol{Y}_i^{(-d)} = 0$ otherwise. Then define $\boldsymbol{X}^{(-d)}$ be the corresponding design matrix to the degenerated outcome vector $\boldsymbol{Y}^{(-d)}$.

Step 3 : Fit the Bayesian single-index model for binary response (almost the same as Bayesian multi-index model for categorical response with D=2). Then calculate the within sample classification error and set *layer-d training error* to be this value.

Among $D-d+1$ types of coding, if coding category-t to d leads to the minimum *layer-d training error*, then we map category-t to d.

4.5 Selection of Number of Knots

To determine the 'number of knots', we consider the same modified BIC in BMIAM:

$$\text{M-BIC} = n \log \left(\frac{1}{n} \sum_{i=1}^{n} \|Y_i - \hat{Y}_i\|^* \right) + kf(n),$$

$$\|Y_i - \hat{Y}_i\|^* = 1, if\ Y_i - \hat{Y}_i \neq 0,$$

$$\|Y_i - \hat{Y}_i\|^* = 0, if\ Y_i - \hat{Y}_i = 0.$$

Here, q is the total number of parameters estimated in the model, and n is the sample size. In the proposed method, we used the Hannan-Quinn information criteria in the numerical studies. The number of knots with the minimum M-BIC will be selected.

CHAPTER 5

SIMULATION STUDIES

5.1 Simulation for Bayesian Multi-Index Additive Model for Continuous Response Variable

One advantage of our proposed method is to provide an estimator of the central space for the additive index model and an estimator of the ridge functions simultaneously. In this section, we numerically compare the performance of the proposed method with some commonly used dimension-reduction methods, such as minimum average variance estimation (MAVE, Xia et al. (2002)), Bayesian dimension reduction (BDR, Reich et al. (2011)), projection pursuit regression (PPR, J. H. Friedman & Stuetzle (1981)), and automatic smoothing spline projection pursuit (PPR-ASS, Roosen & Hastie (1994)). We also compare our approach with some machine learning algorithms such as random forest (RF, Breiman (2001)), boosting (J. H. Friedman (2001)), and Bayesian additive regression trees (BART, Chipman et al. (2010)).

The proposed method provides an estimator of the central space for the additive index model and an estimator of the ridge functions simultaneously. Therefore, performance of both central space estimation and prediction accuracy are investigated.

We consider the following seven settings when simulating the data:

1. $Y = 0.8(\boldsymbol{X}'\boldsymbol{\beta}_1)^2 + 2\sqrt{|\boldsymbol{X}'\boldsymbol{\beta}_2/4|} + \epsilon$

   $p = 6$: $\boldsymbol{\beta}_1 = (1,1,1,0,0,0), \quad \boldsymbol{\beta}_2 = (1,0,0,0,1,3)$

   $p = 10$: $\boldsymbol{\beta}_1 = (1,1,1,0,0,0,0,0,0,0), \quad \boldsymbol{\beta}_2 = (1,0,0,0,1,3,0,0,0,0)$

2. $Y = 1.5\exp(\boldsymbol{X}'\boldsymbol{\beta}_1/10) + 8sin(\boldsymbol{X}'\boldsymbol{\beta}_2/10) + \epsilon$

   $p = 6$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0)$, $\boldsymbol{\beta}_2 = (2, 3, 5, -8, 0, 0)$

   $p = 10$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_2 = (2, 3, 5, -8, 0, 0, 0, 0, 0, 0)$

3. $Y = 1.5\exp(\boldsymbol{X}'\boldsymbol{\beta}_1/10) + 8log(|\boldsymbol{X}'\boldsymbol{\beta}_2/10| + 1) + \epsilon$

   $p = 6$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0)$, $\boldsymbol{\beta}_2 = (8, 0, -2, 0, 0, 5)$

   $p = 10$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_2 = (8, 0, -2, 0, 0, 5, 0, 0, 0, 0)$

4. $Y = 1.5\exp(\boldsymbol{X}'\boldsymbol{\beta}_1/10) + 8sin(\boldsymbol{X}'\boldsymbol{\beta}_2/10) + 2(\boldsymbol{X}'\boldsymbol{\beta}_3/10)^2 + \epsilon$

   $p = 6$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0)$, $\boldsymbol{\beta}_2 = (2, 3, 5, -8, 0, 0)$, $\boldsymbol{\beta}_3 = (8, 0, -2, 0, 0, 5)$

   $p = 10$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_2 = (2, 3, 5, -8, 0, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_3 = (8, 0, -2, 0, 0, 5, 0, 0, 0, 0)$

5. $Y = 1.5\exp(\boldsymbol{X}'\boldsymbol{\beta}_1/10) + (\boldsymbol{X}'\boldsymbol{\beta}_2/10)^3 + 2(\boldsymbol{X}'\boldsymbol{\beta}_3/10)^2 + \epsilon$

   $p = 6$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0)$, $\boldsymbol{\beta}_2 = (2, 3, 5, -8, 0, 0)$, $\boldsymbol{\beta}_3 = (8, 0, -2, 0, 0, 5)$

   $p = 10$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_2 = (2, 3, 5, -8, 0, 0, 0, 0, 0, 0)$, $\boldsymbol{\beta}_3 = (8, 0, -2, 0, 0, 5, 0, 0, 0, 0)$

6. $Y = \boldsymbol{X}'\boldsymbol{\beta}_1(1 + \boldsymbol{X}'\boldsymbol{\beta}_2) + \epsilon$

   $p = 6$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0)$, $\boldsymbol{\beta}_2 = (2, 3, 5, -8, 0, 0)$

7. $Y = \frac{\boldsymbol{X}'\boldsymbol{\beta}_1}{(0.5+(1.5+\boldsymbol{X}'\boldsymbol{\beta}_2)^2)} + \epsilon$

   $p = 6$: $\boldsymbol{\beta}_1 = (5, 0, 0, 5, -3, 0)$, $\boldsymbol{\beta}_2 = (8, 0, -2, 0, 0, 5)$

For Design 1 to Design 5, they are all multi-index additive model and we generate data with $\epsilon \sim \mathcal{N}(0, 1)$. For Design 6 to Design 7, they are multi-index model without additive structure. In these two designs, we assume $\epsilon \sim \mathcal{N}(0, 0.5)$. In all 7 designs, the covariates $\boldsymbol{X}$ are generated according to three different ways: (i) Independent Discrete Uniform, $X_{ij} \overset{i.i.d.}{\sim} DiscreteUnif(-3, 3)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, p$;

(ii) Independent Normal, $X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 4)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, p$; and (iii) Dependent Normal $\boldsymbol{X}_i \overset{i.i.d.}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ with mean of 0 and variance-covariance matrix $\boldsymbol{\Sigma}$ corresponding to the $AR(\rho)$ process with the parameter $\rho = 0.5$.

The sample size n is chosen as 500 when $p = 6$, and 1000 when $p = 10$. The simulation is replicated 100 times to compute the average matrix distance. In each replication, 20000 MCMC samples are drawn for each chain, and the first 10000 are discarded as burn-in. Convergence is monitored using Gelman-Rubin diagnostic (Gelman & Rubin (1992)). Specifically, parallel chains from different initial values are used to calculate the ratio of between-chain variance estimate and within-chain variance estimate. The MCMC is stopped when the ratio is less than or equal to 1.06. The number of knots and the number of directions are selected by the M-BIC as described in Section 2.4.

5.1.1 Simulation for Comparisons of Central Space Estimation

When comparing the accuracy of the estimation of the central space, we compute the matrix distance between the projections of the true and the estimated dimension reduction space. To be specific, let $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ be the true central space and and its estimator, let $\boldsymbol{P} = \boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\beta})^{-1}\boldsymbol{\beta}$ and $\hat{\boldsymbol{P}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}}$ be the corresponding projection matrix. Define the distance as $trace\{(\boldsymbol{P} - \hat{\boldsymbol{P}})(\boldsymbol{P} - \hat{\boldsymbol{P}})\}$.

The average matrix distance between the estimated space and true space are reported in Tables 5.1 to 5.3 when the design matrix is generated according to the aforementioned three different settings and the dimension p is 6. Results when $p = 10$ is shown as Tables 5.4 to 5.6.

Table 5.1: Mean and standard error (in parentheses) of the matrix distance. The design matrix is generated from the independent normal distribution, the dimension $p = 6$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
1	0.0094 (0.0059)	1.0709 (0.1590)	0.0313 (0.1951)	0.0104 (0.0095)	0.0120 (0.0049)
2	0.0019 (0.0012)	0.0433 (0.0322)	0.0019 (0.0012)	0.0089 (0.0143)	0.0006 (0.0007)
3	0.0021 (0.0011)	0.0477 (0.0258)	0.0048 (0.0040)	0.0117 (0.0099)	0.0016 (0.0011)
4	0.0024 (0.0012)	0.0532 (0.0577)	0.0129 (0.0150)	0.0042 (0.0043)	0.0061 (0.0050)
5	0.0032 (0.0020)	0.1862 (0.1658)	0.0190 (0.0210)	0.0314 (0.0223)	0.0066 (0.0052)

Table 5.2: Mean and standard error (in parentheses) of the matrix distance. The design matrix is generated from the discrete uniform distribution, the dimension $p = 6$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
1	0.0079 (0.0056)	0.5394 (0.2423)	1.1152 (0.9847)	1.1353 (0.9094)	0.0155 (0.0036)
2	0.0009 (0.0006)	0.0031 (0.0040)	0.0027 (0.0015)	0.0014 (0.0011)	0.0004 (0.0002)
3	0.0017 (0.0011)	0.0061 (0.0039)	0.0031 (0.0023)	0.0024 (0.0016)	0.0009 (0.0005)
4	0.0015 (0.0011)	0.0397 (0.0351)	0.0129 (0.0080)	0.0058 (0.0051)	0.0058 (0.0052)
5	0.0018 (0.0025)	0.1759 (0.1706)	0.0106 (0.0064)	0.0068 (0.0056)	0.0029 (0.0029)

Table 5.3: Mean and standard error (in parentheses) of the matrix distance. The design matrix is generated from the dependent normal distribution, the dimension $p = 6$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
1	0.0131 (0.0094)	0.6815 (0.1306)	0.9329 (0.8694)	0.8410 (0.8444)	0.0141 (0.0129)
2	0.0021 (0.0014)	0.0089 (0.0056)	0.0021 (0.0016)	0.0022 (0.0019)	0.0007 (0.0006)
3	0.0034 (0.0020)	0.0112 (0.0072)	0.0054 (0.0064)	0.0070 (0.0057)	0.0019 (0.0013)
4	0.0034 (0.0023)	0.0426 (0.0457)	0.0124 (0.0108)	0.0040 (0.0035)	0.0064 (0.0041)
5	0.0029 (0.0018)	0.1992 (0.1692)	0.0132 (0.0137)	0.0164 (0.0107)	0.0041 (0.0038)

Table 5.4: Mean and standard error (in parentheses) of the matrix distance. The design matrix is generated from the independent normal distribution, the dimension $p = 10$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
1	0.0060 (0.0030)	0.7379 (0.1776)	0.0085 (0.0044)	0.0055 (0.0028)	0.0085 (0.0049)
2	0.0010 (0.0006)	0.0056 (0.0028)	0.0011 (0.0004)	0.0059 (0.0035)	0.0009 (0.0008)
3	0.0009 (0.0009)	0.0105 (0.0095)	0.0026 (0.0016)	0.0131 (0.0127)	0.0012 (0.0008)
4	0.0017 (0.0011)	0.0534 (0.0778)	0.0069 (0.0043)	0.0019 (0.0015)	0.0024 (0.0026)
5	0.0014 (0.0008)	0.1761 (0.1470)	0.0120 (0.0098)	0.0372 (0.0312)	0.0020 (0.0017)

Table 5.5: Mean and standard error (in parentheses) of the matrix distance. The design matrix is generated from the discrete uniform distribution, the dimension $p = 10$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
1	0.0046 (0.0025)	0.7265 (0.2633)	0.4429 (0.8205)	0.0045 (0.0024)	0.0049 (0.0024)
2	0.0006 (0.0003)	0.0063 (0.0035)	0.0016 (0.0006)	0.0013 (0.0008)	0.0003 (0.0001)
3	0.0013 (0.0007)	0.0055 (0.0032)	0.0024 (0.0011)	0.0021 (0.0013)	0.0008 (0.0003)
4	0.0011 (0.0005)	0.0475 (0.0307)	0.0083 (0.0033)	0.0063 (0.0065)	0.0063 (0.0065)
5	0.0011 (0.0006)	0.1750 (0.1520)	0.0068 (0.0026)	0.0066 (0.0058)	0.0026 (0.0024)

Table 5.6: Mean and standard error (in parentheses) of the matrix distance. The design matrix is generated from the dependent normal distribution, the dimension $p = 10$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
1	0.0074 (0.0047)	0.7866 (0.1475)	0.4401 (0.7699)	0.0241 (0.0196)	0.0075 (0.0048)
2	0.0012 (0.0007)	0.0043 (0.0023)	0.0014 (0.0006)	0.0128 (0.0099)	0.0006 (0.0002)
3	0.0019 (0.0009)	0.0080 (0.0048)	0.0033 (0.0016)	0.0203 (0.0187)	0.0013 (0.0007)
4	0.0020 (0.0010)	0.0457 (0.0496)	0.0069 (0.0046)	0.0066 (0.0065)	0.0066 (0.0065)
5	0.0019 (0.0011)	0.1837 (0.1522)	0.0080 (0.0051)	0.0451 (0.0322)	0.0032 (0.0032)

As seen from the tables, the proposed method works well for all the settings. The average matrix distance of the proposed method is substantially smaller than all its competitors under most cases. For Settings 3, and 4, 5 when they are three indexes, the improvement over the second best methods is at least 30%. These simulation results provide us with some numerical evidence supporting the statement that "explicitly modeling the ridge functions could improve the estimation of the central space".

According to the results from Table 5.7 to Table 5.12, the proposed method works well when the true multi-index model no longer has a additive structure. The top two performance methods in each design are marked in bold font. The average matrix distance of the proposed method in Design 6 and 7 is smaller than all its competitors under most cases. In the meanwhile, BMIAM shows substantially better performance in MSE compared to the other competing methods.

Table 5.7: Mean and standard error (in parentheses) of the matrix distance with non-additive multi-index model. The design matrix is generated from the independent normal distribution, the dimension $p = 6$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
6	0.0046 (0.0024)	0.0130 (0.0115)	0.0044 (0.0028)	0.0033 (0.0023)	0.0029 (0.0021)
7	0.0109 (0.0090)	0.0161 (0.0110)	0.0163 (0.0087)	0.0249 (0.0384)	0.0468 (0.0445)

Table 5.8: Mean and standard error (in parentheses) of the matrix distance with non-additive multi-index model. The design matrix is generated from the discrete uniform distribution, the dimension $p = 6$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
6	0.0042 (0.0026)	0.0183 (0.0158)	0.0048 (0.0025)	0.0043 (0.0026)	0.0034 (0.0020)
7	0.0082 (0.0044)	0.0167 (0.0112)	0.0163 (0.0102)	0.0379 (0.0517)	0.0351 (0.0454)

Table 5.9: Mean and standard error (in parentheses) of the matrix distance with non-additive multi-index model. The design matrix is generated from the correlated normal distribution, the dimension $p = 6$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
6	0.0055 (0.0030)	0.0127 (0.0086)	0.0059 (0.0036)	0.0044 (0.0023)	0.0042 (0.0029)
7	0.0166 (0.0112)	0.0269 (0.0178)	0.0293 (0.0214)	0.0296 (0.0235)	0.0254 (0.0363)

Table 5.10: Mean and standard error (in parentheses) of the matrix distance with non-additive multi-index model. The design matrix is generated from the independent normal distribution, the dimension $p = 10$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
6	0.0034 (0.0014)	0.0089 (0.0074)	0.0038 (0.0016)	0.0028 (0.0011)	0.0098 (0.0040)
7	0.0092 (0.0039)	0.0127 (0.0088)	0.0153 (0.0067)	0.0117 (0.0084)	0.0145 (0.0179)

Table 5.11: Mean and standard error (in parentheses) of the matrix distance with non-additive multi-index model. The design matrix is generated from the discrete uniform distribution, the dimension $p = 10$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
6	0.0029 (0.0012)	0.0215 (0.0165)	0.0048 (0.0025)	0.0029 (0.0015)	0.0026 (0.0011)
7	0.0081 (0.0039)	0.0108 (0.0054)	0.0140 (0.0052)	0.0108 (0.0116)	0.0157 (0.0201)

Table 5.12: Mean and standard error (in parentheses) of the matrix distance with non-additive multi-index model. The design matrix is generated from the correlated normal distribution, the dimension $p = 10$.

Design	BMIAM	BDR	MAVE	PPR	PPR-ASS
6	0.0045 (0.0020)	0.0085 (0.0068)	0.0063 (0.0023)	0.0042 (0.0017)	0.0038 (0.0015)
7	0.0139 (0.0066)	0.0147 (0.0089)	0.0238 (0.0109)	0.0185 (0.0103)	0.0203 (0.0219)

5.1.2 Simulation for Comparisons of Prediction Performance

The prediction performance is benchmarked by mean squared prediction error. We compare the proposed method with its competitor: MAVE, Random Forrest, Boosting, BART, PPR and PPR-ASS under the same simulation settings.

The results are reported in Tables 5.14-5.16 for the case when $p = 6$ and Tables 5.17-5.19 for the case when $p = 10$.

The proposed methods performs very well. It has the smallest mean squared prediction errors under all the settings. We would like to point out that the prediction error of the MAVE could be very large for the cases when it provided a reasonable estimator of the central space. Taking Setting 3 as an example. When the design matrix is generated from the independent normal distribution, the average matrix distance of MAVE is 0.0048. However, the mean squared prediction error based on MAVE is 9.09. The random forest, boosting and BART fail to incorporate the indexes structure and work poorly under the current settings. The Projection Pursuit Regression and PPR-ASS work reasonably well for Settings 1 and 2. However, they work poorly for the other settings. It is clearly seen that the proposed method could fully use both the indexes and the model structure and provide a reliable estimator of the central space and a good prediction.

According to the results from Table 5.20 to Table 5.25, the proposed method works well when the true multi-index model no longer has a additive structure. The average matrix distance of the proposed method in Design 6 and 7 is smaller than all its competitors under most cases. In the meanwhile, BMIAM shows substantially better performance in MSE compared to the other competing methods.

We also perform simulations to study the performance of the M-BIC criterion to select the number of knots and directions. The M-BIC approach is compared

with the cross-validation (5-fold, 100 time random split) method, and the results are reported in Tables 5.13. In this study, the covariates are generated from i.i.d. Normal, $X_{ij} \overset{i.i.d.}{\sim} N(0, 4)$ when $p = 6$. The sample size n is chosen as 500 and the simulation is replicated 100 times. The performance of M-BIC selection procedure is slightly worse than that of the cross-validation approach. However, the computation cost has been substantially reduced.

Table 5.13: Comparisons of knots selection procedures in terms of mean and standard error (in parentheses) of the mean squared prediction error and projection matrix distance.

	Projection Matrix Distance		MSE	
Design	Cross-Validation	M-BIC	Cross-Validation	M-BIC
1	0.0094 (0.0059)	0.0097 (0.0060)	0.0859 (0.0133)	0.0912 (0.0152)
2	0.0019 (0.0012)	0.0020 (0.0013)	0.0711 (0.0173)	0.0723 (0.0176)
3	0.0021 (0.0011)	0.0024 (0.0013)	0.1291 (0.0197)	0.1458 (0.0221)
4	0.0024 (0.0012)	0.0026 (0.0013)	0.0932 (0.0161)	0.1126 (0.0193)
5	0.0032 (0.0020)	0.0033 (0.0021)	0.0840 (0.0203)	0.0897 (0.0168)
Average	0.0038	0.0040	0.0923	0.1023

Table 5.14: Mean and standard error (in parentheses) of the mean squared prediction error. The design matrix is generated from the independent normal distribution, the dimension $p = 6$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
1	0.08 (0.01)	0.73 (0.27)	10.92 (2.99)	0.99 (0.06)	0.94 (0.07)	0.66 (0.39)	0.10 (0.08)
2	0.07 (0.02)	9.09 (3.81)	24.96 (48.88)	1.00 (0.05)	1.86 (0.81)	4.02 (2.69)	0.14 (0.07)
3	0.13 (0.02)	3.16 (4.05)	13.89 (28.08)	1.00 (0.06)	19.14 (1.14)	21.46 (2.39)	18.99 (1.14)
4	0.09 (0.02)	16.38 (9.53)	13.07 (16.21)	1.00 (0.06)	49.64 (7.47)	57.97 (13.70)	46.98 (8.13)
5	0.08 (0.02)	32.5 (11.29)	44.16 (18.73)	1.00 (0.06)	48.14 (7.00)	69.22 (17.11)	46.01 (10.30)

Table 5.15: Mean and standard error (in parentheses) of the mean squared prediction error for independent discrete predictors. The design matrix is generated from the discrete uniform distribution, the dimension $p = 6$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
1	0.12 (0.02)	1.02 (0.38)	2.88 (0.30)	114.1 (9.27)	0.47 (0.04)	0.68 (0.29)	0.20 (0.06)
2	0.06 (0.02)	5.02 (1.57)	3.90 (0.36)	48.79 (6.26)	1.18 (0.13)	0.61 (0.22)	0.08 (0.02)
3	0.16 (0.03)	1.85 (0.23)	2.11 (0.30)	35.88 (6.49)	20.41 (1.29)	20.60 (1.28)	20.35 (1.29)
4	0.09 (0.02)	25.79 (3.67)	22.77 (2.22)	95.60 (9.31)	34.28 (2.92)	38.83 (7.25)	36.85 (6.58)
5	0.10 (0.02)	34.16 (4.48)	36.05 (3.11)	216 (24.93)	34.74 (2.85)	38.32 (4.60)	30.60 (4.19)

Table 5.16: Mean and standard error (in parentheses) of the mean squared prediction error. The design matrix is generated from the dependent normal distribution, the dimension $p = 6$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
1	0.08 (0.01)	2.57 (0.93)	16.94 (4.61)	0.99 (0.06)	1.05 (0.11)	1.65 (1.04)	0.30 (0.09)
2	0.07 (0.01)	3.96 (1.41)	11.46 (16.23)	0.99 (0.06)	1.52 (0.59)	1.05 (0.54)	0.09 (0.02)
3	0.11 (0.02)	2.39 (0.86)	9.89 (16.41)	0.99 (0.06)	17.63 (1.06)	20.28 (5.49)	17.32 (1.41)
4	0.09 (0.02)	10.92 (2.11)	14.04 (18.40)	1.00 (0.06)	47.27 (7.20)	55.10 (13.80)	46.94 (8.05)
5	0.08 (0.02)	18.56 (6.12)	40.97 (17.36)	1.01 (0.07)	46.74 (6.81)	58.53 (13.07)	45.26 (6.44)

Table 5.17: Mean and standard error (in parentheses) of the mean squared prediction error. The design matrix is generated from the independent normal distribution, the dimension $p = 10$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
1	0.05 (0.01)	0.86 (0.28)	7.65 (4.64)	1.00 (0.05)	0.98 (0.06)	0.36 (0.36)	0.77 (0.46)
2	0.04 (0.01)	12.09 (9.48)	27.97 (52.25)	1.00 (0.04)	2.22 (0.64)	5.48 (4.48)	0.11 (0.05)
3	0.09 (0.02)	4.64 (5.81)	28.21 (50.84)	0.99 (0.05)	19.06 (0.74)	22.92 (3.81)	19.04 (0.79)
4	0.08 (0.01)	17.42 (5.95)	34.69 (48.94)	1.00 (0.04)	49.65 (5.53)	55.93 (10.71)	47.18 (6.80)
5	0.07 (0.01)	37.83 (11.65)	72.93 (47.64)	1.00 (0.04)	51.53 (5.77)	69.46 (15.50)	47.33 (5.78)

Table 5.18: Mean and standard error (in parentheses) of the mean squared prediction error. The design matrix is generated from the discrete uniform distribution, the dimension $p = 10$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
1	0.07 (0.01)	0.85 (0.33)	1.58 (0.13)	115.49 (6.67)	0.40 (0.03)	0.51 (0.29)	0.45 (0.34)
2	0.04 (0.01)	5.70 (1.11)	2.97 (0.21)	48.78 (4.62)	1.03 (0.07)	0.53 (0.13)	0.05 (0.01)
3	0.09 (0.01)	1.89 (0.22)	1.44 (0.17)	35.49 (4.70)	20.42 (0.90)	20.54 (0.89)	20.31 (0.90)
4	0.06 (0.01)	8.58 (2.14)	6.15 (0.39)	96.01 (6.14)	34.39 (2.01)	35.98 (3.53)	35.11 (3.32)
5	0.08 (0.01)	16.49 (2.44)	14.61 (1.08)	216.81 (17.96)	34.85 (2.06)	37.13 (3.55)	34.59 (3.38)

Table 5.19: Mean and standard error (in parentheses) of the mean squared prediction error. The design matrix is generated from the dependent normal distribution, the dimension $p = 10$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
1	0.06 (0.01)	2.01 (0.82)	10.30 (1.85)	0.99 (0.05)	1.11 (0.10)	1.28 (0.62)	0.31 (0.29)
2	0.04 (0.01)	5.43 (2.11)	12.19 (9.97)	1.00 (0.05)	1.70 (0.44)	3.26 (2.20)	0.09 (0.09)
3	0.09 (0.01)	3.68 (0.35)	10.66 (6.29)	1.00 (0.04)	18.44 (0.76)	20.53 (4.34)	18.18 (1.36)
4	0.07 (0.01)	12.53 (2.66)	15.86 (6.90)	1.01 (0.05)	45.90 (4.66)	48.06 (5.47)	45.50 (6.60)
5	0.08 (0.01)	25.99 (6.10)	43.43 (14.64)	1.00 (0.04)	47.00 (4.51)	58.86 (9.75)	44.91 (6.45)

Table 5.20: Mean and standard error (in parentheses) of the mean squared prediction error with non-additive multi-index model. The design matrix is generated from the independent normal distribution, the dimension $p = 6$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
6	0.03 (0.02)	0.10 (0.02)	0.78 (0.11)	0.50 (0.03)	0.35 (0.03)	0.06 (0.02)	0.03 (0.01)
7	0.05 (0.04)	0.30 (0.06)	0.31 (0.03)	0.50 (0.03)	0.36 (0.05)	0.19 (0.19)	0.20 (0.21)

Table 5.21: Mean and standard error (in parentheses) of the mean squared prediction error with non-additive multi-index model. The design matrix is generated from the discrete uniform distribution, the dimension $p = 6$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
6	0.03 (0.02)	0.08 (0.02)	0.43 (0.04)	5.34 (0.52)	0.20 (0.02)	0.04 (0.01)	0.03 (0.01)
7	0.05 (0.01)	0.25 (0.04)	0.27 (0.02)	0.91 (0.10)	0.33 (0.03)	0.20 (0.17)	0.19 (0.18)

Table 5.22: Mean and standard error (in parentheses) of the mean squared prediction error with non-additive multi-index model. The design matrix is generated from the correlated normal distribution, the dimension $p = 6$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
6	0.03 (0.01)	0.05 (0.02)	0.52 (0.06)	0.50 (0.03)	0.32 (0.03)	0.04 (0.01)	0.03 (0.01)
7	0.05 (0.01)	0.26 (0.06)	0.29 (0.02)	0.49 (0.03)	0.29 (0.03)	0.15 (0.10)	0.11 (0.09)

Table 5.23: Mean and standard error (in parentheses) of the mean squared prediction error with non-additive multi-index model. The design matrix is generated from the independent normal distribution, the dimension $p = 10$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
6	0.02 (0.01)	0.09 (0.02)	0.65 (0.07)	0.49 (0.02)	0.33 (0.02)	0.05 (0.01)	0.02 (0.01)
7	0.03 (0.01)	0.29 (0.05)	0.29 (0.02)	0.50 (0.10)	0.31 (0.03)	0.10 (0.09)	0.11 (0.11)

Table 5.24: Mean and standard error (in parentheses) of the mean squared prediction error with non-additive multi-index model. The design matrix is generated from the discrete uniform distribution, the dimension $p = 10$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
6	0.02 (0.01)	0.06 (0.01)	0.37 (0.02)	5.31 (0.35)	0.16 (0.01)	0.02 (0.01)	0.02 (0.01)
7	0.03 (0.01)	0.23 (0.04)	0.26 (0.01)	0.89 (0.06)	0.26 (0.02)	0.08 (0.04)	0.11 (0.11)

Table 5.25: Mean and standard error (in parentheses) of the mean squared prediction error with non-additive multi-index model. The design matrix is generated from the correlated normal distribution, the dimension $p = 10$.

Design	BMIAM	MAVE	Random Forest	Boosting	BART	PPR	PPR-ASS
6	0.02 (0.01)	0.05 (0.02)	0.45 (0.03)	0.50 (0.02)	0.29 (0.02)	0.03 (0.01)	0.02 (0.01)
7	0.03 (0.01)	0.24 (0.05)	0.28 (0.02)	0.50 (0.02)	0.26 (0.02)	0.09 (0.06)	0.09 (0.07)

5.2 Simulation for Bayesian Single-Index Model for Binary Response Variable

In this section, we numerically compare the performance of the Bayesian Single-Index Model for Binary Response Variable with some commonly used binary classification methods, such as logistic regression, k-nearest neighbors algorithm (k-NN), support vector machine (SVM), and linear discriminant analysis (LDA). The performance of both index estimation and prediction accuracy are investigated.

We consider the following three settings when simulating the data:

1. $Z_i = log(|0.6X_{i1} + 0.9X_{i2} + 1.5X_{i3} - 2.4X_{i4}| + 1) - 1.5 + \epsilon_i$,
   $\epsilon_i \sim N(0, 0.1^2)$,
   $\boldsymbol{\beta} = (2, 3, 5, -8, 0, 0)/\sqrt{2^2 + 3^2 + 5^2 + (-8)^2}$.

2. $Z_i = 3sin(0.3X_{i1} + 0.5X_{i2} + 0.8X_{i4} - X_{i5}) - 0.6 + \epsilon_i$,
   $\epsilon_i \sim N(0, 0.1^2)$,
   $\boldsymbol{\beta} = (3, 5, 0, 8, -10, 0)/\sqrt{3^2 + 5^2 + 8^2 + (-10)^2}$.

3. $Z_i = 4(0.2X_{i1} + 0.3X_{i1} + 0.5X_{i4} - 0.7X_{i5})^3 + 8(0.2X_{i1} + 0.3X_{i1} + 0.5X_{i4} - 0.7X_{i5})^2 - 2 + \epsilon_i$,
   $\epsilon_i \sim N(0, 0.1^2)$,
   $\boldsymbol{\beta} = (2, 3, 0, 5, -7, 0)/\sqrt{2^2 + 3^2 + 5^2 + (-7)^2}$.

The covariates are generated from uniform distribution, $X_{ij} \overset{i.i.d.}{\sim} Uni(-3, 3)$. The sample size n is chosen as 500. The training set consists of 475 subjects and the testing set consists of 25 subjects. All results are reported based on 100 independent

repetitions. In each replication, 20000 MCMC samples are drawn for each chain, and the first 10000 are discarded as burn-in.

Convergence is monitored using Gelman-Rubin diagnostic (Gelman & Rubin (1992)). The MCMC is stopped when the ratio is less than or equal to 1.06. The number of knots are selected by the M-BIC as described from Section 3.4.

5.2.1 Simulation Studies on Index Estimation

For index estimation, we report the average bias and the corresponding standard deviation of the estimation for each index parameter for all simulation designs according to the 100 time replication. Table 5.26 shows the simulation results on index estimation. As can be seen from the table, the index estimation is quite accurate. The magnitude of all bias are all less than 0.01.

Table 5.26: Biases and standard error of the estimators from the proposed method for binary response designs.

		β_1	β_2	β_3	β_4	β_5	β_6
Design1	Bias	-0.0066	-0.0091	-0.0034	-0.0060	-0.0034	0.0008
	S.E.	0.0169	0.0174	0.0179	0.0128	0.0169	0.0180
Design2	Bias	-0.0016	-0.0018	0.0006	0.0009	-0.0003	0.0003
	S.E.	0.0083	0.0119	0.0092	0.0111	0.0114	0.0083
Design3	Bias	-0.0033	-0.0027	0.0001	0.0001	-0.0017	0.0008
	S.E.	0.0099	0.0106	0.0079	0.0112	0.0095	0.0092

5.2.2 Simulation Studies on out of Sample Classification (Prediction)

For prediction performance, we report the average cross-validate classification error and its corresponding standard deviation according to the 100 time replication for all simulation designs. Table 5.27 shows the simulation results on cross-validated classification error. As can be seen from the table, the out of sample classification errors are significantly smaller than that of other competing methods.

Table 5.27: Comparison of classification error of different methods for binary response designs.

		Binary BSIM	Logistic	SVM	SVM-nonlinear	k-NN	LDA
Design1	Classification error	0.042	0.466	0.489	0.077	0.267	0.470
	S.E.	0.035	0.094	0.089	0.052	0.084	0.101
Design2	Classification error	0.018	0.408	0.396	0.143	0.247	0.403
	S.E.	0.025	0.093	0.101	0.078	0.086	0.095
Design3	Classification error	0.013	0.372	0.336	0.320	0.264	0.374
	S.E.	0.022	0.090	0.089	0.087	0.084	0.091

5.3 Simulation for Bayesian Single-Index Model for Categorical Response Variable

In this section, we numerically compare the performance of the Bayesian single-index model for categorical response variable with some commonly used classification methods, such as k-nearest neighbors algorithm (k-NN), support vector machine (SVM), and linear discriminant analysis (LDA). The performance of both index estimation and prediction accuracy are investigated.

We consider the following three settings when simulating the data:

1. $Z_{i1} = (0.3X_{i3} + 0.5X_{i5} - 0.8X_{i6})^2 - 2.8 + \epsilon_i,$

   $Z_{i2} = exp(0.2X_{i1} + 0.3X_{i2} + 0.5X_{i3} - 0.8X_{i4}) - 1 + \epsilon_i,$

   $\epsilon_i \sim N(0, 0.1^2),$

   $\boldsymbol{\beta}_1 = (0, 0, 3, 0, 5, -8)/\sqrt{3^2 + 5^2 + (-8)^2}.$

   $\boldsymbol{\beta}_2 = (2, 3, 5, -8, 0, 0)/\sqrt{2^2 + 3^2 + 5^2 + (-8)^2}.$

2. $Z_{i1} = 3sin(0.3X_{i1} + 0.5X_{i2} + 0.8X_{i4} - X_{i5}) - 0.6 + \epsilon_i,$

   $Z_{i2} = 0.5X_{i3} + 0.2X_{i5} - 0.9X_{i6} + 0.4 + \epsilon_i,$

   $\epsilon_i \sim N(0, 0.1^2),$

$$\boldsymbol{\beta}_1 = (3, 5, 0, 8, -10, 0)/\sqrt{3^2 + 5^2 + 8^2 + (-10)^2}.$$
$$\boldsymbol{\beta}_2 = (3, 5, 0, 8, -10, 0)/\sqrt{5^2 + 2^2 + (-9)^2}.$$

The covariates are generated from uniform distribution, $X_{ij} \overset{i.i.d.}{\sim} Uni(-3, 3)$. The sample size n is chosen as 1000. The training set consists of 950 subjects and the testing set consists of 50 subjects. All results are reported based on 100 independent repetitions. In each replication, 20000 MCMC samples are drawn for each chain, and the first 10000 are discarded as burn-in.

Convergence is monitored using Gelman-Rubin diagnostic (Gelman & Rubin (1992)). The MCMC is stopped when the ratio is less than or equal to 1.06. The number of knots are selected by the M-BIC as described from Section 4.5.

5.3.1 Simulation Studies on Index Estimation

For index estimation, we report the average bias and the corresponding standard deviation of the estimation for each index parameter for all simulation designs according to the 100 time replication. Table 5.28 shows the simulation results on index estimation. As can be seen from the table, the index estimation is quite accurate. The magnitude of all bias are all less than 0.02, and most of them are less than 0.01.

Table 5.28: Biases and standard error of the estimators from the proposed method for categorical response designs.

		β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
Design1	Bias	-0.0027	0.0019	-0.0019	-0.0049	-0.0086	-0.0058
	S.E.	0.0187	0.0051	0.0055	0.0054	0.0055	0.0040
		β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
	Bias	-0.0086	0.0002	-0.0182	-0.0027	0.0403	-0.0071
	S.E.	0.0391	0.0273	0.0338	0.0179	0.0727	0.0776
		β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
Design2	Bias	0.0021	0.0058	-0.0157	0.0001	0.0039	0.0021
	S.E.	0.0039	0.0035	0.0042	0.0046	0.0042	0.0053
		β_{21}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
	Bias	-0.0045	0.0191	-0.0121	-0.0126	-0.0114	-0.0073
	S.E.	0.0341	0.0091	0.0182	0.0142	0.0188	0.0112

5.3.2 Simulation Studies on out of Sample Classification (Prediction)

For prediction performance, we report the average cross-validate classification error and its corresponding standard deviation according to the 100 time replication for all simulation designs. Table 5.29 shows the simulation results on cross-validated classification error. As can be seen from the table, the out of sample classification errors are significantly smaller than that of other competing methods.

Table 5.29: Comparison of classification error of different methods for categorical response designs.

		Binary BSIM	SVM	SVM-nonlinear	k-NN	LDA
Design1	Classification error	0.040	0.432	0.074	0.230	0.471
	S.E.	0.001	0.004	0.002	0.003	0.005
Design2	Classification error	0.027	0.526	0.146	0.258	0.470
	S.E.	0.001	0.006	0.002	0.003	0.005

CHAPTER 6

REAL DATA EXAMPLES

6.1 Real Data Examples for Bayesian Multi-Index Additive Model for Continuous Response Variable

6.1.1 QSAR Fish Toxicity Data

We apply BMIAM and competing methods MAVE, PPR, PPR-ASS, random forest and boosting here for a real data application using the 'QSAR Fish Toxicity Data' from the UCI machine-learning repository (`https://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity`), which was originally used for predicting acute toxicity towards the fathead minnow (Pimephales promelas) by Cassotti et al. (2015). This data contains 6 continuous independent variables which are molecular descriptors including: MLOGP (molecular properties), CIC0 (information indices), GATS1i (2D autocorrelations), NdssC (atom-type counts), NdsCH (atom-type counts), SM1_Dz(Z) (2D matrix-based descriptors). The response variable is LC50, which is the amount of a toxic substance that causes death in 50% of test fish over a test duration of 96 hours and thus, LC50 is used to evaluate the acute aquatic toxicity. The sample size of this data is $n = 908$.

We randomly split this data into two parts: the training set which consists of 726 subjects and the testing set which consists of 182 subjects. We replicate these steps 100 times ($T = 100$). Let MSE(t) be the mean squared error for the testing set of tth replication based on the model trained from the tth training set, $t = 1, \cdots, T$.

Then, the mean squared cross-validation error is calculated using $\sum_{t=1}^{T} \text{MSE}(t)/T$. We have used the training set to fit the model using our proposed method and listed competitive models. After that, we apply all these methods again to the testing set and then calculate the mean squared cross-validation error for each method in the comparison. We repeat these steps above 100 times to calculate the average mean squared cross-validation error for each method. The results are shown in Table 6.2.

According to 6.2, BMIAM with two directions has the smallest mean squared prediction error among all the competitors. Notice that is no significant difference between BMIAM and machine learning methods, however, BMIAM can provide model interpretation based on index and ridge function estimation. In Table 6.1, we list the loading corresponding to these two indexes and the corresponding 95% credible intervals. We highlight those loadings whose corresponding CI exclude zero. We have further plotted the associated ridge functions in Figure 6.1-6.2.

According to the Table 6.1, NdssC has no significant impact on the response LC50 according to the 95% Bayesian CI. LC50 depends on GATS1i and CIC0 only through the first and the second ridge function respectively.

Table 6.1: Estimated loadings of indexes for the QSAR Fish Toxicity Data.

Indexes	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
CIC0	0.045427	**0.569027**
	(-0.11632, 0.250986)	(0.33369, 0.68888)
SM1_Dz	**0.449078**	**0.468788**
	(0.236985, 0.616016)	(0.212283, 0.590525)
GATS1i	**-0.54576**	0.217133
	(-0.68786, -0.37283)	(-0.00252, 0.416358)
NdsCH	**0.580109**	**-0.15611**
	(0.364892, 0.716213)	(-0.25956, -0.04623)
NdssC	0.090902	-0.04453
	(-0.08159, 0.236061)	(-0.18684, 0.140871)
MLOGP	**0.391939**	**0.618828**
	(0.179499, 0.565665)	(0.423897, 0.815516)

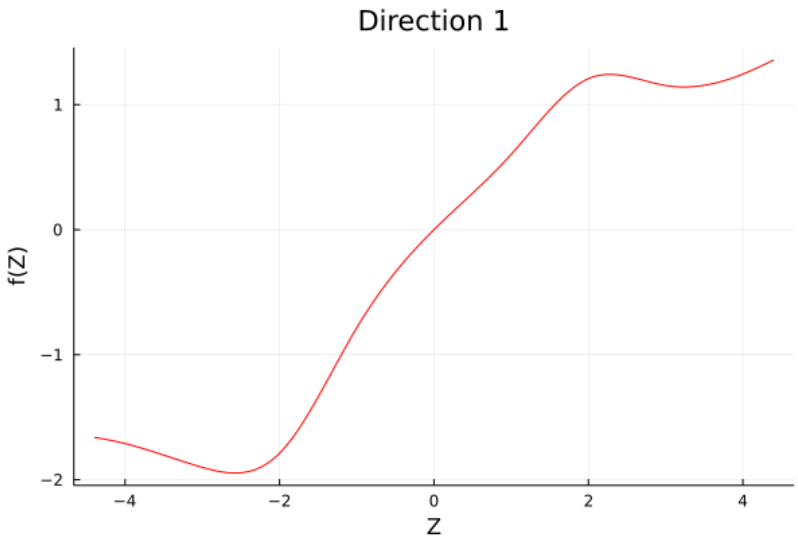

Figure 6.1: Estimated first ridge function for QSAR Fish Toxicity Data by 2-dir BMIAM.

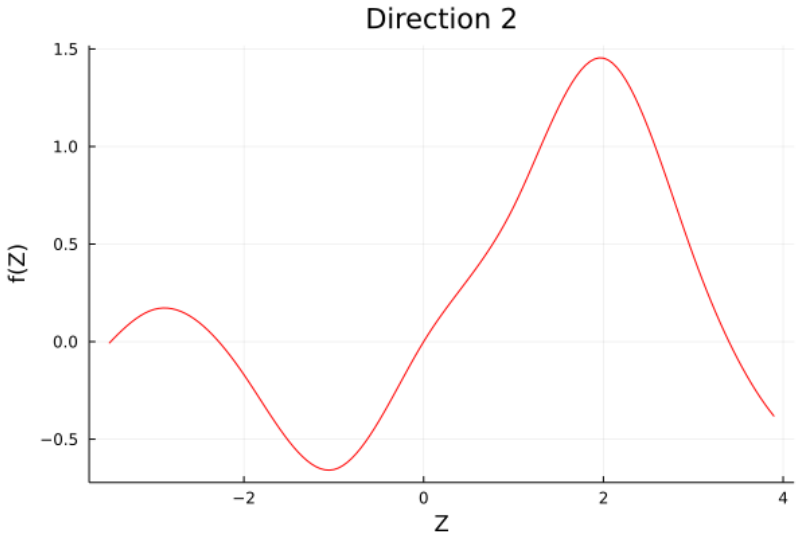

Figure 6.2: Estimated second ridge functions for QSAR Fish Toxicity Data by 2-dir BMIAM.

Table 6.2: Comparison of mean squared cross-validation error of different methods for the QSAR Fish Toxicity Data.

1-dir BMIAM	2-dir BMIAM	MAVE	PPR	PPR-ASS	Random Forest	Boosting	BART
0.91	0.81	0.93	0.91	0.91	0.82	0.84	0.83

6.1.2 Yacht Hydrodynamics Data

The 'Yacht Hydrodynamics Data Set' is available in the UCI machine-learning repository at *http://archive.ics.uci.edu/ml/datasets/yacht+hydrodynamics*. This data set contains 6 continuous independent variables which evaluate the physical characteristics of a yacht. The response variable is the residuary resistance per unit weight of displacement which evaluates the unit residuary resistance of a yacht. The sample size of this data set is $n = 308$. Prediction of residuary resistance of sailing yachts at the initial design stage is of a great value for evaluating the ship's performance and for estimating the required propulsive power. This data set was originally used for optimizing the design of cruising yacht by Gerritsma et al. (1981). We therefore apply BMIAM to this data set to examine how the predictors affect the unit residuary resistance of a yacht.

We compare the performance across different methods in terms of prediction error by cross-validation. We randomly split the data set into two parts: the training set which consists of 250 subjects and the testing set which consists of 58 subjects. We used the training set to fit the model and then calculate the mean squared prediction error of the various methods for the testing set. We replicate these steps 100 times to calculate the average mean squared cross-validation error. The results are shown in Table 6.3.

It is seen that the proposed method with three indexes has the second smallest mean squared prediction error among all the competitors. In Table 6.4, we list the loading corresponding to these three indexes and the corresponding 95% credible

intervals. The associated ridge functions are shown in Figure 6.3.

According to the estimated ridge functions shown in Figure 6.3, the first and the third ridge functions cast more influence on the response, since they are more 'sloppy' in the non-linear part compared with the associated ridge functions for the second direction. From Table 6.4, in these two corresponding indexes, 'Froude Number' is a dominant factor in both directions. In the second direction, 'Froude Number' is still the largest contributor. Therefore, 'Froude Number' is the most influential factor among the 6 individual predictors. By combining all three indexes and ridge functions together (ignoring the statistically insignificant terms in each index), we can conclude that when 'Prismatic Coefficient' decreases, the 'Resistance' increases; when 'Length-Beam Ratio' or 'Froude Number' increase, the 'Resistance' increases. However, the effect of the other 3 predictors are not that simple. As can been seen from Figure 6.4 and Figure 6.5, these two plots illustrate the marginal effect of increasing 'Longitudinal Position' on 'Resistance' by different 'Froude Number' conditioned at $(PC, L/D, B/D, L/B) = (0, 0, 0, 0)$ after standardizing the covariates. For example, the marginal effect of LP given $(PC, L/D, B/D, L/B, Fr) = (0, 0, 0, 0, 0.125)$ is $f(LP, PC = 0, L/D = 0, B/D = 0, L/B = 0, Fr = 0.125)$, where $f(\cdot) = \sum_{d=1}^{D} f_d(\cdot)$. When the 'Froude Number' is relatively large, the marginal effect of 'Longitudinal Position' is not always positive. In addition, all three ridge functions exhibit a 'plateau effect', i.e. the changes reflected on the response is not significant when the $Z = \boldsymbol{X}_i \boldsymbol{\eta}_d$ is less than or greater than a cutoff value in each direction. Realizing that magnitude of 'Froude Number' effect is much heavier than the other variables, we can conclude that our response variable is more sensitive to predictors for large 'Froude Number' compared with the scenario when 'Froude Number' is small.

Table 6.3: Comparison of mean squared cross-validation error of different methods for the Yacht Hydrodynamics Data.

1-dir BMIAM	3-dir BMIAM	MAVE	PPR	PPR-ASS	Random Forest	Boosting	BART
1.18	0.69	10.61	0.74	0.73	1.02	2.03	0.46

Table 6.4: Estimated loadings of indexes for Yacht Hydrodynamics Data.

indexes	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_3$
Longitudinal position (LP)	-0.02387	**-0.52283**	**-0.07288**
	(-0.0393,0.0072)	(-0.6526,-0.3692)	(-0.1144,-0.0385)
Prismatic coefficient (PC)	**0.15869**	**-0.09468**	-0.04753
	(0.0706,0.2171)	(-0.1319,-0.0108)	(-0.1696,0.0064)
Length/Displacement (L/D)	0.099272	**-0.32854**	**-0.14598**
	(-0.0171,0.2571)	(-0.4526,-0.0382)	(-0.2829,-0.0538)
Beam/Draught (B/D)	0.02707	**0.47876**	**0.09802**
	(-0.0892,0.0959)	(0.2986,0.5605)	(0.0010,0.1693)
Length/Beam (L/B)	-0.06093	0.24269	**0.12621**
	(-0.2169,0.0411)	(-0.0375,0.3859)	(0.0452,0.2483)
Froude number (Fr)	**-0.97977**	**0.56713**	**-0.97241**
	(-0.9951,-0.9136)	(0.3794,0.7618)	(-0.9944,-0.8961)

6.2 A Real Data Example for Bayesian Multi-Index Model for Categorical Response Variable

The 'Wholesale Customers Data Set' is available in the UCI machine-learning repository at *https://archive.ics.uci.edu/ml/datasets/Wholesale+customers*. This data set contains 6 continuous independent variables which record the annual spending of a wholesale distributor on different kinds of products in Portugal. The categorical response variable is the region which the wholesale distributor belongs to in Portugal. The sample size of this data set is $n = 440$. The number of categories is 3.

We compare the performance across different methods in terms of prediction error by cross-validation. We randomly split the data set into two parts: the training set with size of 420 subjects and the testing set with size of 20 subjects. We use the training set to fit the model and then calculate the classification error of the various

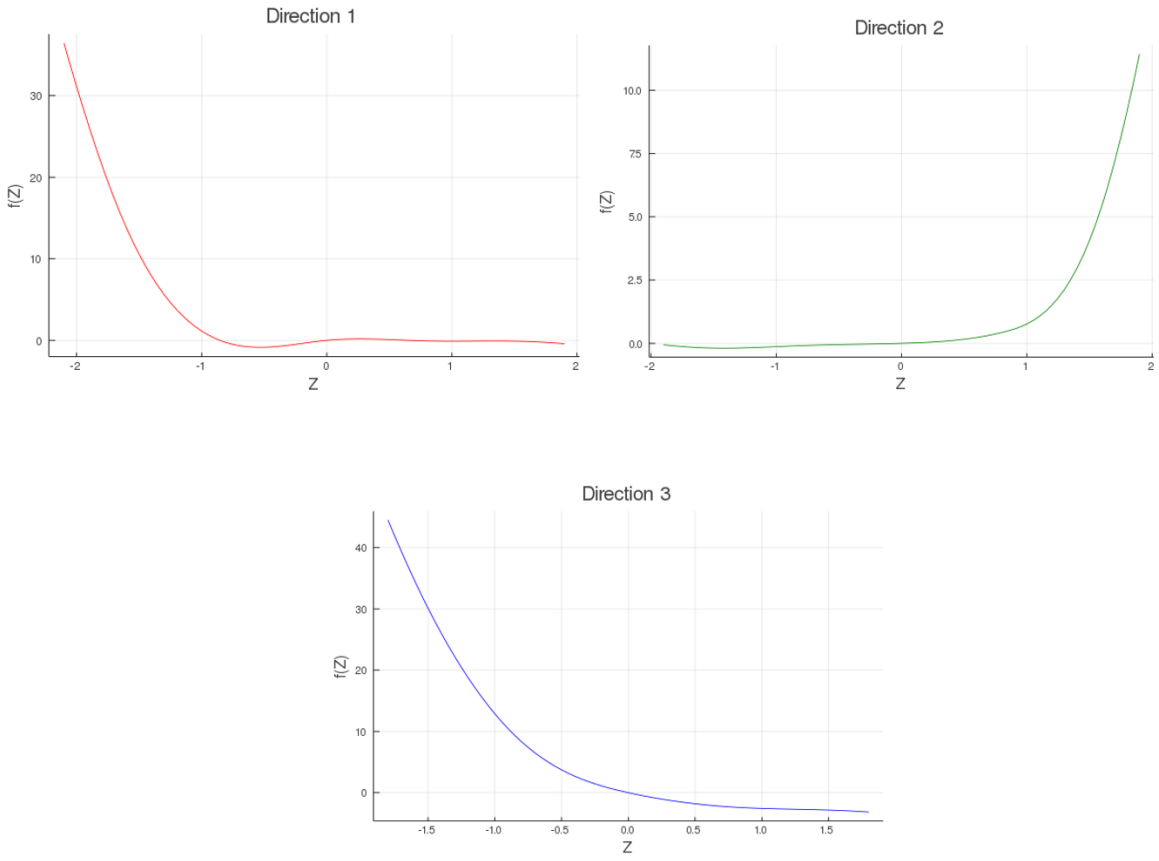

Figure 6.3: Estimated ridge functions for Yacht Hydrodynamics Data by 3-dir BMIAM

methods based on the testing set. We replicate these steps 100 times to calculate the average cross-validated classification error. The results are shown in Table.

Table 6.5: Comparison of cross-validated classification error of different methods for the Wholesale Customers Data Set.

	MIM Categorical	k-NN	SVM	SVM non-linear	LDA
Classification error	0.256	0.285	0.515	0.282	0.290

According to Table 6.5, the proposed method has the smallest classification error among all the competitors. Compared with the second best approach SVM non-linear, the cross-validated classification error of MIM for categorical method is 9% smaller.

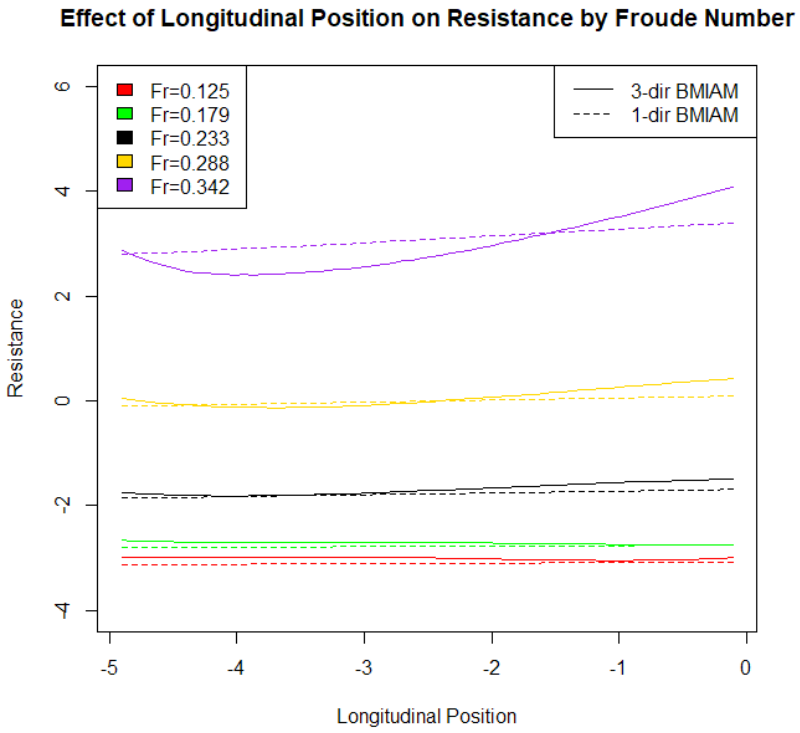

Figure 6.4: Marginal effect of Longitudinal Position on Resistance by Froude.

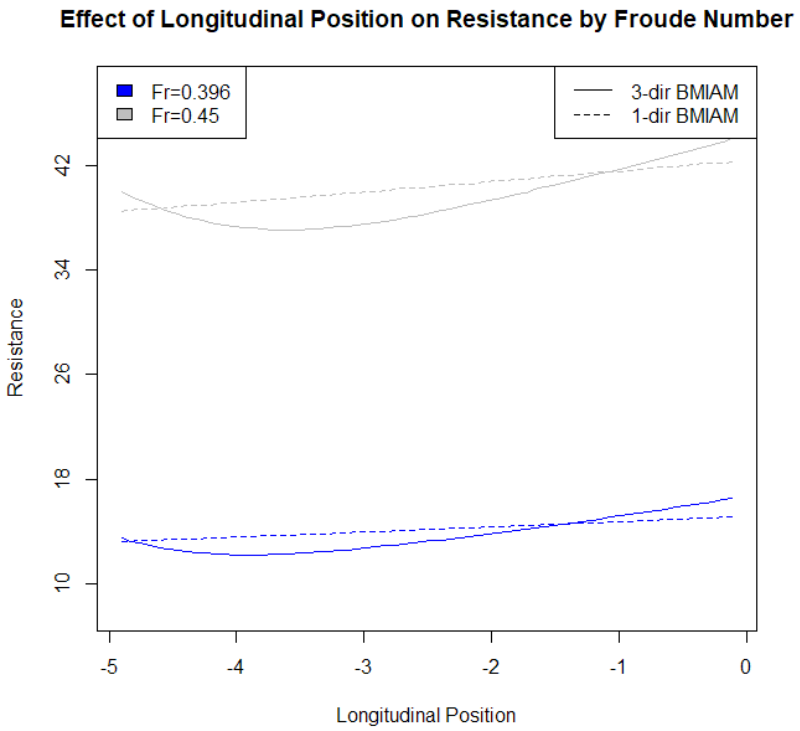

Figure 6.5: Marginal effect of Longitudinal Position on Resistance by Froude.

CHAPTER 7

DISCUSSION

In this paper, we propose two Bayesian multi-index regression models. One is for continuous response variable, and the other one is for categorical response variable. When moving from the linear model to the index model, the family of the models has been enlarged substantially and it could be used to model many complicated problems.

In the Bayesian multi-index additive model, the ridge functions are approximated by B-spline, and the index parameters are parametrized by polar coordinates. The number of knots and directions are selected by a modified BIC approach. For the first time in the Bayesian multi-index additive model framework, we provide theoretical guarantees in posterior consistency. Simulation studies show that the proposed methods have advantages both in estimating the central space and predicting the response when compared with the competing methods. In the application on the Yacht Hydrodynamics Data, the proposed method shows good prediction accuracy and identifies the predictor that contribute the most to the response variable.

In the Bayesian multi-index models for Categorical Response Variable, the link functions for the auxiliary variables are approximated by B-spline, and the index parameters are parametrized by polar coordinates. The number of knots is selected by a modified BIC approach. Simulation studies show that the proposed methods have a accurate estimation of index parameters. Additionally, simulation studies also indicate that the proposed methods have advantages in prediction (classification)

when compared with some competing methods. In the application on the Wholesale Customers Data, the proposed method shows good prediction performance compared with some competing methods.

Currently, the proposed methods can only deal with single response for each observation only. We intend to extend our Bayesian methods to the multi-index model with multiple response variable for each observation. We leave all these into the future research.

BIBLIOGRAPHY

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679.

Antoniadis, A., Grégoire, G., & Mckeague, I. W. (2004). Bayesian estimation in single-index models. *Statistica Sinica*, *14*(4), 1147–1164.

Bhattacharya, A., Pati, D., & Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics*, *47*(1), 39 – 66.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Cassotti, M., Ballabio, D., Todeschini, R., & Consonni, V. (2015). A similarity-based qsar model for predicting acute toxicity towards the fathead minnow (pimephales promelas). *SAR and QSAR in environmental research*, *26*(3), 217-243.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Choi, T., Shi, J. Q., & Wang, B. (2011). A gaussian process regression approach to a single-index model. *Journal of Nonparametric Statistics*, *23*(1), 21-36.

Cook, D. R. (1998). *Regression graphics*. John Wiley & Sons, Inc., New York. (Ideas for studying regressions through graphics, A Wiley-Interscience Publication)

Cook, D. R. (2000). SAVE: a method for dimension reduction and graphics in regression. *Communications in statistics-Theory and methods*, *29*(9-10), 2109–2121.

Cook, D. R., Li, B., & Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika*, *94*(3), 569–584.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, 1–26.

Dhara, K., Lipsitz, S., Pati, D., & Sinha, D. (2020). A new Bayesian single index model with or without covariates missing at random. *Bayesian Analysis*, *15*(3), 759 – 780.

Friedman, J., & Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, *C-23*(9), 881-890. doi: 10.1109/T-C.1974.224051

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J. H., Grosse, E., & Stuetzle, W. (1983). Multidimensional additive spline approximation. *SIAM Journal on Scientific and Statistical Computing*, *4*(2), 291-301.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*(376), 817–823.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gerritsma, J., Onnink, R., & Versluis, A. (1981). Geometry, resistance and stability of the delft systematic yacht hull series. *In International Shipbuilding Progress*, *28*, 276–297.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *41*(2), 190–195.

Hsing, T., & Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, *20*(2), 1040–1061.

Huang, J. Z., & Yang, L. (2004). Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *66*(2), 463–477.

Lang, S., & Brezger, A. (2004). Bayesian p-splines. *Journal of Computational and Graphical Statistics*, *13*(1), 183–212.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*(414), 316–327.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, *87*(420), 1025–1039.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Lin, Q., Zhao, Z., & Liu, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics*, *46*(2), 580 – 610.

Lin, Q., Zhao, Z., & Liu, J. S. (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, *114*(528), 1726-1739.

McGee, G., Wilson, A., Webster, T. F., & Coull, B. A. (2021). *Bayesian multiple index models for environmental mixtures.*

Ni, L., Cook, D. R., & Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. *Biometrika*, *92*(1), 242–247.

O'Brien, S. M., & Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, *60*(3), 739–746.

Park, C. G., Vannucci, M., & Hart, J. D. (2005). Bayesian methods for wavelet series in single-index models. *Journal of Computational and Graphical Statistics*, *14*(4), 770–794.

Reich, B. J., Bondell, H. D., & Li, L. (2011). Sufficient dimension reduction via bayesian mixture modeling. *Biometrics*, *67*(3), 886–895.

Roosen, C. B., & Hastie, T. J. (1994). Automatic smoothing spline projection pursuit. *Journal of Computational and Graphical Statistics*, *3*(3), 235–248.

Sun, Y., & Jiang, H. (2020). *Bayesian variable selection for single index logistic model.* arXiv. Retrieved from `https://arxiv.org/abs/2012.06199` doi: 10 .48550/ARXIV.2012.06199

Wei, R., Reich, B. J., Hoppin, J. A., & Ghosal, S. (2020). Sparse bayesian additive nonparametric regression with application to health effects of pesticides mixtures. *Statistica Sinica*, *30*, 55–79.

Xia, Y., Tong, H., Li, W. K., & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *64*(3), 363–410.

Yuan, M. (2011). On the identifiability of additive index models. *Statistica Sinica*, *21*(4), 1901–1911.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of bruno de finetti* (p. 233-243). North-Holland: Elsevier.

Zhu, L., Miao, B., & Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, *101*(474), 640–643.

Zhu, L.-X., & Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, *5*(2), 727–736.

APPENDIX

PROOF OF THEOREM 1

Proof. Let $g(\boldsymbol{X}_i) = \sum_{d=1}^{D} g_d(\boldsymbol{X}_i)$ be the regression function, where $g_d(\cdot)$, $d = 1, ..., D$ are the individual link functions including the indexes.

$$\begin{aligned} Y_i &= g(\boldsymbol{X}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, ..., n, && (.1) \\ g(\boldsymbol{X}_i) &= \sum_{d=1}^{D} \boldsymbol{B}_d \boldsymbol{\eta}_d + \delta_i \stackrel{\text{def}}{=} \boldsymbol{B}\boldsymbol{\eta} + \delta_i. && (.2) \end{aligned}$$

For the theoretical derivation of posterior risk bounds, we will use the fractional likelihood and results from Bhattacharya et al. (2019). A fractional likelihood is obtained by raising the likelihood to a fractional power v. An examination of the fractional likelihood $\mathcal{L}_{n,v}(g)$ and the corresponding posterior $p_{n,v}(g)$ under (.1) yields

$$p_{n,v}(g) = \frac{\left\{\mathcal{N}_n\left(\boldsymbol{y}; \widetilde{\boldsymbol{g}}, \sigma^2 \mathbf{I}_n\right)\right\}^v p(g)}{\int \left\{\mathcal{N}_n\left(\boldsymbol{y}; \widetilde{\boldsymbol{g}}, \sigma^2 \mathbf{I}_n\right)\right\}^v p(g)} = \frac{\mathcal{N}_n\left(\boldsymbol{y}; \widetilde{\boldsymbol{g}}, \psi^2 \mathbf{I}_n\right) p(g)}{\int \mathcal{N}_n\left(\boldsymbol{y}; \widetilde{\boldsymbol{g}}, \psi^2 \mathbf{I}_n\right) p(dg)},$$

where $p(g)$ is the prior distribution of g and $\mathcal{N}_n(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the n-multivariate normal density evaluated at y with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Here, $\boldsymbol{y} = (y_1, \ldots, y_n)'$ with y_i being the realization of Y_i, $\widetilde{\boldsymbol{g}} = \left(g\left(\boldsymbol{X}_1\right), \ldots, g\left(\boldsymbol{X}_n\right)\right)'$ and $\psi = \sigma/\sqrt{v}$. Hence, the fractional posterior for (.1) is essentially a standard posterior with a different variance parameter in the likelihood. In the following, we state the assumptions on the true data generation mechanism and the prior.

Denote $\Pi(A)$ to be the probability of A under the prior (2.2-5), (2.2-6), and (2.2-8).

To find the posterior convergence, we need a lower bound for $\Pi\left(B_n\left(\widetilde{\boldsymbol{g}}_0, \varepsilon; \widetilde{\boldsymbol{g}}_0\right)\right)$.

$$B_n\left(\widetilde{\boldsymbol{g}}_0, \varepsilon; \widetilde{\boldsymbol{g}}_0\right) = \left\{\widetilde{\boldsymbol{g}} \in \widetilde{\boldsymbol{g}} : E_{\widetilde{\boldsymbol{g}}_0}\left(\ell_n\right) \leq n\varepsilon^2, \mathrm{Var}_{\widetilde{\boldsymbol{g}}_0}\left(\ell_n\right) \leq n\varepsilon^2\right\}$$

where

$$\ell_n(\boldsymbol{Y}) = \log \frac{p^{(n)}_{\widetilde{\boldsymbol{g}}_0}}{p^{(n)}_{\widetilde{\boldsymbol{g}}}}(\boldsymbol{Y})$$

and $E_{\widetilde{\boldsymbol{g}}_0}$ and $\mathrm{var}_{\widetilde{\boldsymbol{g}}_0}$ denote expectations and variances under $\boldsymbol{Y} \sim \mathcal{N}\left(\widetilde{\boldsymbol{g}}_0, \sigma^2 \mathbf{I}_n\right)$ Simplifying the log-likelihood ratio, one obtains,

$$\ell_n(\boldsymbol{Y}) = \frac{1}{2\sigma^2}\left[\left\|\widetilde{\boldsymbol{g}} - \widetilde{\boldsymbol{g}}_0\right\|^2 + 2\left\langle \boldsymbol{Y} - \widetilde{\boldsymbol{g}}_0, \widetilde{\boldsymbol{g}}_0 - \widetilde{\boldsymbol{g}}\right\rangle\right]$$

Then,

$$E_{\widetilde{\boldsymbol{g}}_0}\left(\ell_n\right) = \frac{1}{2\sigma^2}\left\|\widetilde{\boldsymbol{g}} - \widetilde{\boldsymbol{g}}_0\right\|^2$$

and

$$\mathrm{Var}_{\widetilde{\boldsymbol{g}}_0}\left(\ell_n\right) = \frac{1}{\sigma^4} E_{\widetilde{\boldsymbol{g}}_0}\left\langle \boldsymbol{Y} - \widetilde{\boldsymbol{g}}_0, \widetilde{\boldsymbol{g}}_0 - \widetilde{\boldsymbol{g}}\right\rangle^2 = \frac{1}{\sigma^4} E\left(\widetilde{\boldsymbol{g}} - \widetilde{\boldsymbol{g}}_0\right)'\left(\boldsymbol{Y} - \widetilde{\boldsymbol{g}}_0\right)\left(\boldsymbol{Y} - \widetilde{\boldsymbol{g}}_0\right)'\left(\widetilde{\boldsymbol{g}} - \widetilde{\boldsymbol{g}}_0\right) = \frac{1}{\sigma^4}\left\|\widetilde{\boldsymbol{g}} - \widetilde{\boldsymbol{g}}_0\right\|^2$$

Let $\lambda = min\{2\sigma^2, \sigma^4\}$, then it follows that

$$B_n\left(\widetilde{\boldsymbol{g}}_0, \varepsilon; \widetilde{\boldsymbol{g}}_0\right) \leq \left\{\widetilde{g} \in \Theta : \left\|\widetilde{\boldsymbol{g}} - \widetilde{\boldsymbol{g}}_0\right\|^2 \leq n\lambda\epsilon^2\right\}$$

For simplicity, we assume λ is known, and without loss of generality, equals one.

Therefore, we need a lower bound for $\Pi\left(\|\widetilde{\boldsymbol{g}}-\widetilde{\boldsymbol{g}}_0\|^2<n\epsilon^2\right)$.

$$
\begin{aligned}
&\Pi\left(\|\widetilde{\boldsymbol{g}}-\widetilde{\boldsymbol{g}}_0\|^2<n\epsilon^2\right)\\
\geq\ &\Pi\left(\|\widetilde{\boldsymbol{g}}-\widetilde{\boldsymbol{g}}_0\|_\infty<\epsilon\right)\\
\geq\ &\Pi\left\{\|\widetilde{\boldsymbol{g}}(\boldsymbol{X}\boldsymbol{\beta})-\widetilde{\boldsymbol{g}}_0(\boldsymbol{X}\boldsymbol{\beta})\|_\infty<\frac{\epsilon}{2}\right\}\Pi\left\{\left\|\widetilde{\boldsymbol{g}}_0(\boldsymbol{X}\boldsymbol{\beta})-\widetilde{\boldsymbol{g}}_0(\boldsymbol{X}\boldsymbol{\beta}^0)\right\|_\infty<\frac{\epsilon}{2}\right\}
\end{aligned}
$$

Notice that, the response Y can be written as:

$$
\boldsymbol{Y}=\sum_{d=1}^{D}\boldsymbol{B}_d(\boldsymbol{X}\beta_d^0)\boldsymbol{\eta}_d^0+\boldsymbol{\delta}+\boldsymbol{\epsilon}
$$

Then we have,

$$
\begin{aligned}
&\Pi\left\{\|\widetilde{\boldsymbol{g}}(\boldsymbol{X}\boldsymbol{\beta})-\widetilde{\boldsymbol{g}}_0(\boldsymbol{X}\boldsymbol{\beta})\|_\infty<\frac{\epsilon}{2}\right\}\\
=\ &\Pi\left\{\left\|\sum_{d=1}^{D}\boldsymbol{B}_d(\boldsymbol{X}\boldsymbol{\beta}_d)(\boldsymbol{\eta}_d-\boldsymbol{\eta}_d^0)+\boldsymbol{\delta}\right\|_\infty<\frac{\epsilon}{2}\right\}\\
\geq\ &\Pi\left\{\left\|\sum_{d=1}^{D}\boldsymbol{B}_d(\boldsymbol{X}\boldsymbol{\beta}_d)(\boldsymbol{\eta}_d-\boldsymbol{\eta}_d^0)\right\|_\infty<\frac{\epsilon}{2}-\|\boldsymbol{\delta}\|_\infty\right\}
\end{aligned}
$$

According to assumption **A1**, and **A3**,

$$
\begin{aligned}
&\Pi\left\{\left\|\sum_{d=1}^{D}\boldsymbol{B}_d(\boldsymbol{X}\boldsymbol{\beta}_d)(\boldsymbol{\eta}_d-\boldsymbol{\eta}_d^0)\right\|_\infty<\frac{\epsilon}{2}-\|\boldsymbol{\delta}\|_\infty\right\}\\
\geq\ &\Pi\left\{\left\|\boldsymbol{\eta}_d-\boldsymbol{\eta}_d^0\right\|_\infty<\frac{t\epsilon}{\sqrt{m}D}\right\}\\
\geq\ &\Pi\left\{\|\boldsymbol{\eta}_d\|_\infty\in(\left\|\boldsymbol{\eta}_d^0\right\|_\infty-\frac{t\epsilon}{\sqrt{m}D},\left\|\boldsymbol{\eta}_d^0\right\|_\infty+\frac{t\epsilon}{\sqrt{m}D})\right\}\\
\geq\ &\frac{t\epsilon}{\sqrt{m}D}\inf_{\|\boldsymbol{\eta}_d\|_\infty\leq W}\Pi(\boldsymbol{\eta}_d)
\end{aligned}
$$

for some constant t.

Then, we need to find a lower bound of $\inf_{\|\boldsymbol{\eta_d}\|_\infty \leq W} \Pi(\boldsymbol{\eta}_d) < n\epsilon^2$

$$\begin{aligned}
\inf_{\|\boldsymbol{\eta}_d\|_\infty \leq W} \Pi(\boldsymbol{\eta}_d) &= \int_0^\infty (2\pi\tau)^{-\frac{m}{2}} e^{-\frac{mW^2}{2\tau}} \Pi(\tau)\, d\tau \\
&= \int_0^\infty (2\pi\tau)^{-\frac{m}{2}} e^{-\frac{mW^2}{2\tau}} \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} \tau^{-(a_\tau+1)} e^{-\frac{b_\tau}{\tau}}\, d\tau \\
&= (2\pi)^{-\frac{m}{2}} \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} \int_0^\infty e^{-\frac{b_\tau + \frac{mW^2}{2}}{\tau}} \tau^{-(a_\tau+\frac{m}{2}+1)}\, d\tau \\
&= (2\pi)^{-\frac{m}{2}} \frac{\Gamma(a_\tau+\frac{m}{2}) b_\tau^{a_\tau}}{\Gamma(a_\tau)(b_\tau+\frac{mW^2}{2})^{a_\tau+\frac{m}{2}}} \int_0^\infty \frac{(b_\tau+\frac{mW^2}{2})^{a_\tau+\frac{m}{2}}}{\Gamma(a_\tau+\frac{m}{2})} e^{-\frac{b_\tau+\frac{mW^2}{2}}{\tau}} \tau^{-(a_\tau+\frac{m}{2}+1)}\, d\tau \\
&= (2\pi)^{-\frac{m}{2}} \frac{\Gamma(a_\tau+\frac{m}{2}) b_\tau^{a_\tau}}{\Gamma(a_\tau)(b_\tau+\frac{mW^2}{2})^{a_\tau+\frac{m}{2}}} \\
&= b_\tau^{a_\tau} (2\pi)^{-\frac{m}{2}} (\frac{m}{2})! (b_\tau + \frac{mW^2}{2})^{-(a_\tau+\frac{m}{2})}
\end{aligned}$$

Now, we can bound $\Pi\left\{\|\widetilde{\boldsymbol{g}}(\boldsymbol{X\beta}) - \widetilde{\boldsymbol{g}}_0(\boldsymbol{X\beta})\|_\infty^2 < \frac{n\epsilon^2}{2}\right\}$.

$$\Pi\left\{\|\widetilde{\boldsymbol{g}}(\boldsymbol{X\beta}) - \widetilde{\boldsymbol{g}}_0(\boldsymbol{X\beta})\|_\infty^2 < \frac{n\epsilon^2}{2}\right\} \geq \frac{t\epsilon}{\sqrt{m}D} b_\tau^{a_\tau} (2\pi)^{-\frac{m}{2}} (\frac{m}{2})! (b_\tau + \frac{mW^2}{2})^{-(a_\tau+\frac{m}{2})}$$

To bound the second term, observe that

$$\boldsymbol{g}_0\left(\boldsymbol{X\beta}\right) = \boldsymbol{g}_0\left(\boldsymbol{X\beta}_0\right) + \sum_{d=1}^{D} \sum_{j \leq \lfloor \kappa \rfloor} \left(\frac{(\boldsymbol{X\beta}_d - \boldsymbol{X\beta}_d^0)^j}{j!} \boldsymbol{g}_0^{d(j)}\left(\boldsymbol{X\beta}_d^0\right) + R\left(\boldsymbol{\beta}_d, \boldsymbol{\beta}_d^0, \boldsymbol{X}\right) \right)$$

where $|R\left(\boldsymbol{\beta}_d, \boldsymbol{\beta}_d^0, \boldsymbol{X}\right)| \leq c_1 \|\boldsymbol{X}\| \left\|\boldsymbol{\beta}_d - \boldsymbol{\beta}_d^0\right\|^\kappa$ for some constant $c_1 > 0$. Then $\Pi\left\{\|\boldsymbol{g}_0\left(\boldsymbol{X\beta}\right) - \boldsymbol{g}_0\left(\boldsymbol{X\beta}^0\right)\|_\infty < \frac{1}{2}\epsilon\right\} \geq \Pi\left(\sum_{d=1}^D \left\|\boldsymbol{\beta}_d - \boldsymbol{\beta}_d^0\right\|_2 < c_2\epsilon\right)$ for some constant $c_2 > 0$. Hence, it is enough to find a lower bound to the prior probability corresponding to $\Pi\left(\left\|\boldsymbol{\beta}_d - \boldsymbol{\beta}_d^0\right\|_2 < c_2 \frac{\epsilon}{D}\right)$. This probability is proportional to $(\frac{\epsilon}{D})^{p-1}$, see the supplementary material of Dhara et al. (2020).

Based on these two lower bounds, we have the following bound:

$$\Pi\left(\|\widetilde{\boldsymbol{g}}-\widetilde{\boldsymbol{g}}_0\|_{2,n}^2<n\epsilon^2\right)\geq\frac{t\epsilon}{\sqrt{m}D}{b_\tau}^{a_\tau}(2\pi)^{-\frac{m}{2}}(\frac{m}{2})!(b_\tau+\frac{mE^2}{2})^{-(a_\tau+\frac{m}{2})}(\frac{\epsilon}{D})^{D(p-1)}$$

The rate for posterior convergence can be obtained by solving:

$$\frac{t\epsilon}{\sqrt{m}D}{b_\tau}^{a_\tau}(2\pi)^{-\frac{m}{2}}(\frac{m}{2})!(b_\tau+\frac{mE^2}{2})^{-(a_\tau+\frac{m}{2})}(\frac{\epsilon}{D})^{D(p-1)}\geq e^{-n\epsilon^2}$$

For

$$\begin{aligned}\epsilon^2=&\frac{[log(n)]^{D(p-1)+1}}{n}+\frac{m}{2n}log(2\pi)-\frac{1}{n}log\left(\frac{(\frac{m}{2})!}{\sqrt{m}(b_\tau+\frac{m}{2})^{a_\tau+\frac{m}{2}}}\right)\\=&\frac{1}{n}\left[[log(n)]^{D(p-1)+1}+log\left(\frac{(2\pi)^{\frac{m}{2}}\sqrt{m}(b_\tau+\frac{m}{2})^{a_\tau+\frac{m}{2}}}{(\frac{m}{2})!}\right)\right],\end{aligned}$$

the condition required in Theorem 3.1 by Bhattacharya et al. (2019) is satisfied. □