8-30-2017

Application of Gaussian Process Priors on Bayesian Regression

Abhishek Bishoyi
abhishek.bishoyi@uconn.edu

Application of Gaussian Process Priors on Bayesian Regression

Abhishek Bishoyi, Ph.D.
University of Connecticut, 2017

ABSTRACT

This dissertation aims at introducing Gaussian process priors on the regression to capture features of dataset more adequately. Three different types of problems occur often in the regression. 1) For the dataset with missing covariates in the semiparametric regression, we utilize Gaussian process priors on the nonparametric component of the regression function to perform imputations of missing covariates. For the Bayesian inference of parameters we specify objective priors on the Gaussian process parameters. Posterior propriety of the model under the objective priors is also demonstrated. 2) For modeling binary and ordinal data, we propose a flexible nonparametric regression model that combines flexible power link function with a Gaussian process prior on the latent regression function. We develop an efficient sampling algorithm for posterior inference and prove the posterior consistency of the proposed model. 3) In the high dimensional dataset, the estimation of regression coefficients especially when the covariates are highly multicollinear is very challenging. Therefore, we develop a model by using structured spike an slab prior on regression coefficients. Prior information of similarity between

Abhishek Bishoyi - University of Connecticut, 2017

covariates can be encoded into the covariance structure of Gaussian process which can be used to induce sparsity. Hyperparameters of Gaussian process can used to control different sparsity pattern. Superiority of the proposed model is demonstrated using various simulation studies and real data examples.

Application of Gaussian Process Priors on Bayesian Regression

Abhishek Bishoyi

B.Sc., M.Sc. Statistics and informatics, Indian Institute of Technology, Kharagpur, India, 2013

M.S., Statistics, University of Connecticut, CT, USA, 2016

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2017

APPROVAL PAGE

Doctor of Philosophy Dissertation

Application of Gaussian Process Priors on Bayesian Regression

Presented by

Abhishek Bishoyi, B.Sc. Statistics and Informatics, M.Sc. Statistics

Major Co-Advisor _____

Dr. Xiaojing Wang

Major Co-Advisor _____

Dr. Dipak K. Dey

Associate Advisor _____

Dr. Ming-Hui Chen

Associate Advisor _____

Dr. Ofer Harel

University of Connecticut

2017

Acknowledgements

First of all I would like to express my deepest gratitude to both of my advisors Dr. Xiaojing Wang and Dr. Dipak K. Dey for their constant guidance and support. Dr. Xiaojing Wang introduced me to an interesting and challenging research field and gave me freedom and excellent atmosphere to do research. I owe special thanks to her for being extremely patient with me and spending tremendous amount of time mentoring me. I am very grateful of Dr. Dey for his constant support through out my research. His wisdom and never ending encouragement have given me confidence to recognize my goals, and the courage to pursue them. I am truly indebted to both of my advisors for giving me the opportunity to complete my PhD work under their advisement.

I would like to thank my advisory committee members: Dr. Ming-Hui Chen and Dr. Ofer Harel for their expertise and feedback. To Dr. Chen, I owe a debt of gratitude for giving me opportunities to work on various consulting projects. Your passion towards research makes me enthusiastic going forward in my research career. To Dr. Harel, thank you for examining my thesis and being a part of my dissertation committee. I am grateful to all the members of Department of Statistics for their support.

Lastly, none of this would have been possible without the support of my parents and friends through both successes and failures. I cannot thank them enough for their love and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In Bayesian regression, the approach is to learn how the mean response is affected by several covariates within the context of Bayesian inference, where probability distributions are used to encode one's prior information about parameters. Unlike frequentist estimation methods, where parameters are assumed to be unknown but fixed, in Bayesian statistics, the uncertainty about the parameters are quantified using probability so that the unknown parameters are considered as random variables.

In real world, data can be complex. Analyzing a complex dataset requires robust and flexible models that can infer features of the dataset as adequately as possible. The parametric model structure is expressed with a finite number of parameters. This limits the complexity of the model even when the complexity of data is unbounded. As a result, the parametric model becomes inadequate for solving complex real world problems. This motivates us to adopt nonparametric/semiparametric methods for regression. In nonparametric regression, the objective is to find relationship between response and covariates without assuming the parametric form of regression function. It offers more flexible way to model the effects of covariates on the response compared to parametric

models, while parametric models have more restrictive conditions on the mean function. Nonparametric regression is a rapidly growing and exciting field. When both responses and covariates are fully observed, the relevant theories and methods are well developed as described in Takezawa (2005). Many competing methods are available for nonparametric regression, including kernel-based methods, regression splines, smoothing splines, and wavelet and Fourier series expansions.

For Bayesian methods, nonparametric regression (and classification) problems are via elicitation of priors on the mean function. Dirichlet process models are very popular methods for Bayesian nonparametrics. It is a distribution over distributions, i.e. each draw from a Dirichlet process is itself a distribution. The nonparametric nature of the Dirichlet process makes it an ideal candidate in Bayesian clustering problems when the number of clusters are unknown. Gaussian process (GP) models are acknowledged as another popular tool for nonparametric regression. In contrast to Dirichlet process prior, GP prior is a flexible and tractable prior over continuous functions, useful for solving regression and classification problems. The usage of GP models is widespread in spatial models, in the analysis of computer experiments and time series, in machine learning and so on (Rasmussen and Williams, 2006). One of the major advantage of using Gaussian process is that it is fully determined by its mean and covariance function. The kind of structure that can be modeled by Gaussian process is determined by its covariance structure. Neal (1996) have shown that many Bayesian regression models based on neural networks converge to Gaussian process in the limit of infinite number of hidden

units.

1.1 Gaussian Process

A Gaussian process is a collection of random variables for which any finite number of those variables has a joint Gaussian distribution (Rasmussen and Williams, 2006). Formally, we denote a GP with correlation function $k(\cdot,\cdot \mid \ell)$ and mean function $\mu(.)$ as $GP(\mu(.), \sigma_z^2 k(\cdot,\cdot \mid \ell))$ and assume a stochastic real valued function $g(\cdot) \sim GP(\mu(.), \sigma_z^2 k(\cdot,\cdot \mid \ell))$. Given any finite n distinct input vector $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \Re^k$, $[g(\mathbf{x}_1), \cdots, g(\mathbf{x}_n)]'$ will follow a multivariate Gaussian distribution with mean vector $[\mu(\mathbf{x}_1), \cdots, \mu(\mathbf{x}_n)]'$ and covariance matrix Σ, with each (i,j)th entry of Σ, i.e., $(\Sigma)_{ij} = \sigma_z^2 k(\mathbf{x}_i, \mathbf{x}_j \mid \ell)$, is determined by the covariance between the points $g(\mathbf{x}_i)$ and $g(\mathbf{x}_j)$ for $i, j = 1, \cdots, n$. In our research, we considered only isotropic correlation kernel, i.e., $k(\mathbf{x}_i, \mathbf{x}_j \mid \ell) = \Psi_\ell(||\mathbf{x}_i - \mathbf{x}_j||)$ for some isotropic correlation function Ψ_ℓ and $||.||$ denotes Euclidean distance. A common choice of isotropic correlation functions is the squared exponential kernel (also known as Gaussian kernel), that is,

$$(\Sigma)_{ij} = \sigma_z^2 k(\mathbf{x}_i, \mathbf{x}_j \mid \ell) = \sigma_z^2 \exp\left(-\sum_{d=1}^{k} \frac{(x_{i,d} - x_{j,d})^2}{2\ell_d^2}\right), \tag{1.1.1}$$

where σ_z^2 and $\{\ell_d\}_{d=1}^k$ are hyperparameters of the GP prior. The scaling parameter σ_z^2 controls the variation of the response surface and the length-scale parameter ℓ_d guides the smoothness of sample paths. Sample paths are smoother with larger length-scale. A

GP with covariance kernel given in equation (1.1.1) supports a large class of functions with various shapes.

GP can be used as a prior distribution for unknown regression function. Intuitively, one can think of a function $g(.) : \Re^k \to \Re$ drawn from a Gaussian process prior as an extremely high-dimensional vector drawn from an extremely high-dimensional multivariate Gaussian distribution. Here, each dimension of Gaussian distribution corresponds to an input $\mathbf{x}_i$, and the corresponding component of the random vector represents the value $g(\mathbf{x}_i)$. Flexible regression models based on Gaussian process provide efficient ways of model learning and carrying out inference.

1.2 Binary or Ordinal Response Data

In many scientific fields, the response variable is not a numerical value. Instead, the response variable is simply a designation of two or more possible outcomes, for example, success or failure, alive or dead. One of the important research question in such type of data is finding out how the expected probability of belonging to one of the category is related to a set of covariates, and aims at better predictions of future outcomes. Generalized linear models (GLMs) are most commonly used methods to model such type of data. Consider a random binary response y_i measured with covariate $\mathbf{x}_i \in \Re^k$, for $i = 1, \cdots, n$. To analyze such binomial response we usually use generalized linear model, where we model the latent probability of "success" through a link function (McCullagh

and Nelder, 1989), that is:

$$P(y_i = 1) = p(\mathbf{x}_i).$$

Traditionally, a parametric approach to specify $p(\mathbf{x})$ is taken using $p(\mathbf{x}_i) = H(\mathbf{x}_i'\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is an unknown parameter vector and H is a cumulative distribution function (cdf), called the link function. The logit, probit and Student-t link functions are the most frequently used link function in generalized linear model (Madsen and Thyregod, 2010).

A critical issue with modeling of binary and ordinal response data using generalized linear models is the choice of link functions. Czado and Santner (1992) have shown that mis-specification of link function leads to a significant bias in estimation of regression parameters and the mean response probability. Moreover, the commonly used link functions like logit, probit and Student-t lack flexibility. Most of them are symmetric links in the sense that they assume that the latent probability of a binomial response will approach towards 0 with the same rate as it approaches to 1. In another word, the probability density function (pdf) that corresponds to the inverse cumulative distribution function (cdf) of the link function is symmetric. In many cases, this may not be a reasonable assumption (Li et al., 2016). A commonly used asymmetric link function is the complementary loglog (cloglog) function. However, it has a fixed negative skewness, which restricts the ability of data to allow for positive skewness. Moreover, misspecification in the link function may lead to an increase in mean squared error of the estimated

probability as well as a substantial bias in estimating the regression parameters and the mean response probability (Czado and Santner, 1992). Therefore, it is very essential to introduce flexible link functions to create more flexible and robust regression model to increase the predictive power of the model.

To overcome this issue, Jiang et al. (2013) proposed a general class of flexible link functions based on symmetric link functions and its mirror reflection. Suppose F_0^{-1} is a symmetric link function. That is, the latent probability of binomial response approaches 0 with the same rate as it approaches 1. The symmetric power link family is then given by

$$F(x,r) = F_0^r\left(\frac{x}{r}\right)\mathbf{I}_{(0,1]}(r) + \left(1 - F_0^{\frac{1}{r}}\left(-rx\right)\right)\mathbf{I}_{(1,+\infty)}(r), \tag{1.2.1}$$

where $\mathbf{I}_c(x)$ is indicator function taking value 1 if $x \in c$, zero otherwise. $r \in (0,\infty)$ is skewness parameter of the link function. The intuition is to utilize the fact that $F_0^r(x)$ is a valid cdf and it achieves flexible left skewness when $r < 1$, while the same property holds for its mirror reflection $1 - F_0^{\frac{1}{r}}(-x)$ with skewness being in opposite direction. By combining the two, greater flexibility in skewness can be achieved. Moreover, when the skewness parameter $r = 1$, $F(x,r)$ is same as baseline link function $F_0(x)$. So, the baseline link function is a special case of $F(x,r)$.

Another critical issue with modeling of binary and ordinal response data is the choice of latent regression function. Limiting the latent regression function to simple linear or parametric form can be restrictive in modeling binary or ordinal data. Frlich (2006)

has shown that mis-specification of parametric model can lead to inconsistent estimates. To overcome this issue, there have been various research work on flexible binary regression model by using Bayesian semiparametric or nonparametric methods. The basic difference among these methods is about where these nonparametric prior are imposed. Newton et al. (1996) have proposed flexible nonparametric model by imposing Dirichlet process prior on the link function. One drawback of this model is the parametric assumption of latent regression function. Another line of reseach is to impose Gaussian process prior on the latent regression function to model it nonparametrically. Rasmussen and Williams (2006) and Choudhuri et al. (2007) have worked on this. Li et al. (2016) have investigated the effect of different link functions on GP binary model. However, a much less investigated problem is the effect of flexible power link functions on GP binary and ordinal model. Another less investigated aspect is the effect of missing data in Bayesian nonparametric models.

1.3 Missing Data

Missing data arise in various experimental settings, including survey, clinical trials, environmental studies. To decide how to deal with missing data, it is important to know why they are missing. As known, for missing data, there are three basic classification based on relationship between missing data mechanism and the missing and observed values (Little and Rubin, 2002). When the nonresponse is not related to any values of

variable, the missing data mechanism is called missing completely at random (MCAR). So, one can think of the observed values as essentially a random sample of the full data set. Therefore, complete case analysis gives the same results as the full data set would have. Unfortunately, most missing data are not MCAR. Missing at random (MAR) is less restrictive than MCAR. The assumption of MAR is that missingness depends only on the observed values. Both MAR and MCAR can be grouped into ignorable missing data mechanism. A much more relaxed assumption is missing not at random (MNAR), where the missing data mechanism depends on data that are missing. Such type of missing data mechanism is also called non-ignorable missing data mechanism. For example, if individuals with higher incomes are less likely to reveal about themselves on a survey than are individuals with lower incomes, the missing data mechanism for income is non-ignorable. Missing data is a serious problem in almost all statistical problems. Things become more difficult when predictors have missing values.

1.4 Motivation

This dissertation is focused on developing flexible model to capture the characteristics of different types of datasets more accurately. We developed three novel nonparametric and semiparametric models to address special structure of datasets. They are respectively focused on semiparametric regression models on missing data, flexible nonparametric binary and ordinal regression models, and variable selection in high dimensional data.

There is very limited literature on either nonparametric or semiparametric models for missing covariates data. Missing data are problematic because classical statistical methods require a value for each variable in model. When the dataset is incomplete, more sophisticated methods are required to deal with this. A useful reference for general parametric statistical inferences with missing data has been comprehensively discussed in Little and Rubin (2002). Although there is a huge literature in regression models with missing data, there have not been much work in the paradigm of regression when the nonprametric component has presented missingness in covariates. In Chapter 2, we have discussed some interesting deficiencies of existing methods, which serve as motivations for development of our proposed model.

In the binary and ordinal regression framework, we have discussed in previous sections that in addition to the structure of latent regression function, the choice of link function also plays a major role in estimation of model parameters. Considerable work has been done in developing flexible model with application to the binary response data. However, those models cannot be used to handle ordinal response data. To construct a more flexible class of nonparametric binary/ordinal regression model, in Chapter 3 we employ the link function discussed in equation (1.2.1) in Bayesian nonparametric framework. We also explored the performance of nonparametric ordinal regression model under various choices of flexible link functions.

Next, we focused on the use of Bayesian nonparametric methods to address variable selection problem in high dimensional dataset. Estimation of regression coefficients

can be a challenging task in high dimensional data when the covariates are grouped. Various methods have been developed to estimate the sparse regression coefficients by partitioning the set of covariates beforehand. But, all of them assumed that the group structures are known before estimation. Therefore, in Chapter 4 we have developed a novel approach to handle sparsity under by enforcing prior information about covariates structure via Gaussian process kernel.

Finally, we have explored the extension of semiparametric regression model with non-ignorable missing covariates. To the best of our knowledge Bayesian nonparametric or semiparametric regression model using Gaussian process prior have never been explored for covariates with non-ignorable misssingness. This motivates us to eventually propose a Bayesian model based method as part of future work in Chapter 5.

1.5 Thesis Outline

In Chapter 2, we introduce a semiparametric regression model in the presence of missing covariates for nonparametric components under a Bayesian framework. We propose an imputation method to solve this issue and perform our analysis using Bayesian inference, where we specify the objective priors on the parameters of Gaussian process models. In Chapter 3, we develop a flexible Bayesian nonparametric binary/ordinal regression model and construct an efficient sampling algorithm for posterior inference. Posterior consistency of the model is also discussed. In Chapter 4 we develop a novel approach

to estimate sparse regression coefficients by grouping the covariates nonparametrically. Finally in Chapter 5, we summarize the findings of all three chapters and then suggest future directions of our models.

Chapter 2

Learning Semiparametric Regression with Missing Covariates Using Gaussian Processes Models

2.1 Introduction

Nonparametric regression offers more flexible way to model the effects of covariates on the response compared to parametric models, which have more restrictive conditions on the mean function. Many competing methods are available for nonparametric regression without missingness, including kernel-based methods, regression splines, smoothing splines, and wavelet and Fourier series expansions. But one of the drawback of nonparametric regression models is the difficulty to interpret its parameters in contrast to parametric regression. Thus, various efforts have been addressed on semiparametric models, which balance the interpretation of parametric models and flexibility of nonparametric models.

However, there is very limited literature on either nonparametric or semiparametric models for missing covariates data. One common approach for nonparametric modeling is splines, such as using basis function representations for the mean function (e.g., Denison (2002)). Yau and Kohn (2003) used thin plate splines to allow the mean and variance to change with covariates. In certain applications, this structure may be overly restrictive due to the specific splines used in their model. However, model estimation using regression splines become more challenging when covariates have missingness. Faes et al. (2011) developed a nonparametric model based on spline basis functions, where covariates are missing. They carried out inference using variational Bayes approximations and showed that variational Bayes approximations (c.f., Beal (2003)) produces multimodality in the posterior distributions in the case of missing covariates when we do not have one-to-one mapping for the latent function. Therefore, in the appearance of missing covariates, to estimate nonparametric models with splines using Markov chain Monte Carlo (MCMC) sampling methods will become very computationally intensive.

In Bayesian nonparametric paradigm, there are few literatures which address missing data problem. Wang et al. (2010) developed a classification model to handle incomplete inputs, where they extended the finite Quadratically Gated Mixture of Experts (QGME) developed by Liao et al. (2007) to an infinite QGME via a Dirichlet process prior. Since the MCMC-based analysis for this model suffers from huge computational costs, Wang et al. (2010) implemented approximate inference via the variational Bayesian method. Recently, Zhang et al. (2016) proposed an approach to solve unsupervised

learning for clustering with missing data. They used infinite Dirichlet process mixture model to automatically determine the number of clusters. They assumed missing data as latent variables and obtained their posterior distributions using the variational Bayesian expectation maximization algorithm. However, the computation burden on all these Dirichlet Process models is heavy. Hence, the inference is carried out using approximate methods like variational Bayes and so far. The current literature for missing predictors of Dirichlet process models is only focused on clustering problems other than regression.

Gaussian process (GP) models are another important method for nonparametric regression. For the properties of GP models, one can refer to Van Der Vaart and Wellner (1996), Adler (1990), Cramér and Leadbetter (2013) and Rasmussen and Williams (2006). Choi and Schervish (2007) showed assigning GP priors to the unknown regression function with normality assumption on the residuals would lead to a consistent estimator for the regression function. However, for GP models, the case of missing inputs has received little attention, due to the challenge of propagating the input uncertainty through the nonlinear GP mapping. Only recently, there are several studies focusing on GP models with observed inputs subject to some measurement uncertainty (Quiñonero-Candela and Roweis (2003), Girard and Murray-Smith (2003) and Damianou and Lawrence (2015)). They often developed a two-stage procedure for such GP models either using variational Bayesian methods or an expectation-maximization procedure, where in the first stage they estimated the model parameters only for complete

case and then in the second step they alternately updated model parameters and adjusted estimates of missing input points. However, the situation to deal with noisy inputs due to measurement uncertainty will be quite different than the situation where the inputs are missing.

Therefore, in this chapter, we consider the scenario when an input of GP models is subject to MCAR or MAR for the purpose of filling in the gap of missing data for GP models in the literature. To avoid the risk of introducing modeling biases in parametric regression models as well as the existing drawbacks of nonparametric regression models (such as the curse of dimensionality, the difficulty of interpretation and lack of extrapolation capability), we will consider semiparametric regression models in our study to balance the interpretation of linear models and flexibility of nonparametric models. Specifically, we will use the partial linear model, the most commonly used semiparametric regression model (c.f., Engle et al. (1986), Ruppert et al. (2003), Härdle and Liang (2007) and references therein). A GP prior will be assigned to such semiparametric regression model with specifying the mean function of the GP for certain linear parametric forms. Further, we will impute the missing covariate of nonparametric component via a Bayesian hierarchical model, which will be a key for us to develop the covariance function of the GP prior.

To complete the prior specification of GP models, we need to elicit the priors on the hyperparameters of a GP, which controls the smoothness and variabilities of a GP model. However, it is often difficult to specify subjective information over hyperparameters of

a GP model. Thus, we will consider to use noninformative priors. But as mentioned in Berger et al. (2001) and thereafter, assigning noninformative priors such as commonly used constant priors and independent Jeffrey's priors for hyperparameters both fail to yield proper posteriors. Instead, they recommended the "exact" reference prior for GP models without appearing white noise. Ren et al. (2012) extended the "exact" reference prior in the case when there are white noises for GP models and showed the posterior propriety. We further prove that under some mild conditions, the posterior propriety of the "exact" reference prior will still hold in the presence of ignorable missing covariates.

The format of this Chapter is as follows. In Section 2.2, we outline the setting of semiparametric regressions in a Bayesian hierarchical modeling framework. Section 2.3 will focus on the discussion of sampling methods to estimate model parameters and deriving posterior predictive distribution. In addition, we show the posterior propriety of GP hyperparameters under the "exact" reference prior. Then, we perform simulation studies in Section 2.4 to validate our proposed method. In Section 2.5, we present two real world applications to compare our proposed model with some competitive models. Finally, in Section 2.6, we draw conclusion and point out some future direction.

2.2 Semiparametric Regression Models with Ignorable Missing Covariates

The task of finding a good function estimation from given data, receives a lot of attention not only in the statistics literature but also in the neural network and machine learning communities. One of the popular approach for nonparametric Bayesian regression model is using a GP prior in modeling the unknown underlying function with nonlinear and nonparametric structures. GP model admits a much richer latent structure than that of a parametric model which restricts to certain fixed parametric structure, thus GP model will potentially better approximate the true response function. In addition, we can specify a parametric structure for the mean of a GP prior, which will make the GP model enjoy the combined merits of the easy interpretation as parametric models and flexibility as the nonparametric models. In this section, we are going to propose our semiparametric regression model in a Bayesian framework to handle missing data.

The semiparametric regression model that we consider is given by

$$y_i = \mathbf{z}_i'\boldsymbol{\beta} + g(x_i) + \epsilon_i, \tag{2.2.1}$$

for $i = 1, 2, \cdots, n$. Here, $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_p]'$ is a $q \times 1$ vector of coefficients of fully observed covariates $\mathbf{z}_i = [1, z_{i1}, \cdots, z_{ip}]'$ and further, define $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_n]' \in \Re^{n \times q}$, where $q = p + 1$. We assume that $p << n$ and denote $\mathbf{y} = [y_1, \cdots, y_n]'$. Here, $g(\cdot)$ is

the unknown function, x_i's $\in \Re$ are the observed inputs (subject to missing) and ϵ_i's are random errors. The errors ϵ_i's are conveniently assumed to be independent and identically distributed normal random variables with mean 0 and variance σ_ϵ^2, with $0 < \sigma_\epsilon^2 < \infty$. In the absence of covariates $\mathbf{Z}$, our model can be reduced to a nonparametric model as:

$$y_i = g(x_i) + \epsilon_i.$$

To estimate unknown function $g(\cdot)$, we are going to introduce a GP prior on $g(\cdot)$. We will consider zero mean GP to avoid confounding of the mean parameter of a GP prior and coefficients $\boldsymbol{\beta}$ in Model (2.2.1). As known, a GP is a collection of random variables for which any finite number of those variables has a joint Gaussian distribution (Rasmussen and Williams (2006)). Formally, we denote a zero mean GP with correlation function $k(\cdot, \cdot \mid \ell)$ as $GP(0, \sigma_z^2 k(\cdot, \cdot \mid \ell))$ and assume a stochastic real valued function $g(\cdot) \sim GP(0, \sigma_z^2 k(\cdot, \cdot \mid \ell))$. Given any finite n distinct inputs $x_1, \cdots, x_n \in \Re$, $[g(x_1), \cdots, g(x_n)]'$ will follow a multivariate Gaussian distribution with zero mean vector and covariance matrix Σ, with each (i, j)th entry of Σ, i.e., $(\Sigma)_{ij} = \sigma_z^2 k(x_i, x_j \mid \ell)$, is determined by the covariance between the points x_i and x_j for $i, j = 1, \cdots, n$. In this chapter, we considered only isotropic correlation kernel, i.e., $k(x_i, x_j \mid \ell) = \Psi_\ell(||x_i - x_j||)$ for some isotropic correlation function Ψ_ℓ and $||.||$ denotes Euclidean distance. A common choice of isotropic correlation functions is the squared exponential kernel (also known

as Gaussian kernel), that is,

$$(\Sigma)_{ij} = \sigma_z^2 k(x_i, x_j \mid \ell) = \sigma_z^2 \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right),$$

where σ_z^2 and ℓ are hyperparameters of the GP prior. The scaling parameter σ_z^2 controls the variation of the response surface and the length-scale parameter ℓ guides the smoothness of sample paths. In this Chapter, we also consider other power exponential correlation and Matérn class of correlation functions.

In this chapter, we consider the input x_i's in Model (2.2.1) are subject to missing. This may happen because respondents in a survey refuse to fill in certain items, or recorders fail to observe an input due to unknown mistakes in an experimental process or others. Denote $\mathbf{x} = [x_1, \cdots, x_n]'$ with x_i's $\in \Re$ and presume that $\mathbf{x} \sim f(\mathbf{x} \mid \boldsymbol{\omega})$, where $\boldsymbol{\omega}$ are some unknown parameters. Without loss of generality, we write $\mathbf{x} = (\mathbf{x}^{obs}, \mathbf{x}^{mis})$, where $\mathbf{x}_{obs}$ denotes the observed values and $\mathbf{x}_{mis}$ denotes missing values. Suppose m out of n covariates x_i's are missing, i.e., $\mathbf{x}^{mis} \in \Re^m$ and $\mathbf{x}^{obs} \in \Re^{n-m}$. For imputation of missing covariates under Bayesian framework, we need a probabilistic model to estimate the missing x_i's. Let us denote for $i = 1, \cdots, n$, R_i is a binary random variable with success probability π_i and use R_i to indicate whether x_i is observed or not ($R_i = 1$ if x_i is missing and 0 otherwise). For $i = 1, \cdots, n$, R_i is a binary random variable with success probability π_i. Then, we define $R = [R_1, \cdots, R_n]'$ a $n \times 1$ vector of missingness indicator. Bayesian inference for the parameters of regression models will differ according

to the dependence of the distribution of R_i on the data. Here, we consider following two missingness mechanisms:

(1) $\pi_i = P(R_i = 1 \mid y_i, x_i, \mathbf{z}_i) = p$ for some constant $0 < p < 1$. In this case, x_i's are said to be missing completely at random (MCAR) (c.f., Little and Rubin (2002)) and the missingness mechanism is independent of the data.

(2) $\pi_i = P(R_i = 1 \mid y_i, x_i, \mathbf{z}_i) = h(y_i, \mathbf{z}_i)$. The function $h(y_i, \mathbf{z}_i)$, called the conditional probability of observing the predictor given the response, defines the missing mechanism and is in general unknown. In this case, missing data mechanism depends on the observed y_i's but not on the x_i's and $\mathbf{z}_i$'s (c.f., Little and Rubin (2002)).

With these specified missingness mechanisms, then one of the keys to make the statistical inference for Model (2.2.1) is to estimate parameters $\ell, \sigma_z^2, \sigma_\epsilon^2$, and $\boldsymbol{\beta}$ based on marginal likelihood where we integrate out latent function $g(\cdot)$ in the likelihood, i.e.,

$$f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \ell, \sigma_z^2, \sigma_\epsilon^2, \boldsymbol{\beta}) = \mathcal{N}_n(\mathbf{Z}\boldsymbol{\beta}, \sigma_z^2\mathbf{G}).$$

Here, $\mathcal{N}_n(\cdot, \cdot)$ indicates a n-dimensional multivariate normal distribution with $\mathbf{Z}\boldsymbol{\beta}$ being its mean and $\sigma_z^2\mathbf{G}$ being its covariance, where $\mathbf{G} = \eta\mathbf{I}_n + \mathbf{K}$ and $\eta = \sigma_\epsilon^2/\sigma_z^2$ is the variance component of the noise-to-signal ratio. Notice that $\mathbf{K}$ is $n \times n$ isotropic correlation matrix with each (i, j)th entry $k_{ij} = \Psi_\ell(||x_i - x_j||) = (\Sigma)_{ij}/\sigma_z^2$ depending only on ℓ. Throughout this chapter, we will interchange the usage of the notation $\mathbf{K}$ and $\mathbf{K}(\ell)$ to represent the correlation matrix of a GP whenever it is necessary. To simplify the notation, let us

define $\Theta = (\ell, \sigma_z^2, \eta, \boldsymbol{\beta}')'$. Then, the likelihood of $\Theta, \boldsymbol{\omega}$ given the observed data $\mathbf{y}$, $\mathbf{x}^{obs}$, R and $\mathbf{Z}$ for Model (2.2.1) is:

$$\begin{aligned}\mathcal{L}(\Theta, \boldsymbol{\omega} \mid R, \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z}) = \int_{\mathbf{x}^{mis}} \left(\prod_{i=1}^{n} f(R_i \mid y_i, x_i, \mathbf{z}_i, \phi) \right) f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \Theta) \\ \times f(\mathbf{x} \mid \boldsymbol{\omega}) \; d\mathbf{x}^{mis}.\end{aligned} \tag{2.2.2}$$

We assume the parameter in missing data mechanism ϕ is independent of parameters $\{\Theta, \boldsymbol{\omega}\}$. Under the two specified missingness mechanisms, $f(R_i \mid y_i, x_i, \mathbf{z}_i, \phi)$ will not have any effect on estimation of parameters Θ and imputation values of missing $\mathbf{x}^{mis}$ if . Thus, when we derive the posterior distribution of parameters Θ and missing values $\mathbf{x}^{mis}$, we can ignore the first term on the right side of the likelihood (2.2.2). Further, if we assign a prior on $\boldsymbol{\omega}$ as $\pi(\boldsymbol{\omega})$ then we can integrate out nuisance hyperparameters $\boldsymbol{\omega}$ in Equation (5.2.3). Let us define $\pi(\mathbf{x}) = \int_{\boldsymbol{\omega}} f(\mathbf{x} \mid \boldsymbol{\omega}) \times \pi(\boldsymbol{\omega}) d\boldsymbol{\omega}$ as marginal prior on $\mathbf{x}$ after integrating out nuisance parameter $\boldsymbol{\omega}$ and we can factorize $\pi(\mathbf{x}) = \pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}) \times \pi(\mathbf{x}^{obs})$. Then, the likelihood of Θ given the data $\{\mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z}\}$ is given by:

$$\begin{aligned}\mathcal{L}(\Theta \mid \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z}) &= \int_{\mathbf{x}^{mis}} \int_{\boldsymbol{\omega}} f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \Theta) f(\mathbf{x} \mid \boldsymbol{\omega}) \times \pi(\boldsymbol{\omega}) \; d\boldsymbol{\omega} d\mathbf{x}^{mis} \\ &\propto \int_{\mathbf{x}^{mis}} f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \Theta) \times \pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}) d\mathbf{x}^{mis}.\end{aligned} \tag{2.2.3}$$

To utilize Bayesian methods to perform the inference on Model (2.2.1), we need to consider the specification of priors on the unkonwn parameters Θ in order to derive their posterior distributions. One common approach is to use proper priors on Θ, assigned

subjectively or abstracting information from previous data. One of the advantages of proper priors is that they can always achieve propriety of posterior distribution. However, the subjective elicitation of GP hyperparameters (i.e., ℓ, η and σ_z^2) is difficult due to the hard interpretation of their meanings in practice. Therefore, we resort to specify the priors of GP hyperparameters non-informatively. But if we use the conventional noninformative priors, Berger et al. (2001) showed that those priors do not yield proper posterior. Thus, they derived an exact reference prior under the case without the noise variance (i.e., $\sigma_\epsilon^2 = 0$ in our case). Ren et al. (2012) further examined the effect of noise variance and derived an "exact" reference prior under this situation $\sigma_\epsilon^2 \neq 0$. In this chapter, we aim to extend the posterior propriety of this reference prior in the case for missing data for the GP models and we are going to use the "exact" reference priors for unknown GP hyperparameters.

2.3 Posterior Propriety and Posterior Inference

In the Section 2.3.1, we discuss the posterior propriety with the "exact" reference prior. Then, in Section 2.3.3, we specify MCMC procedure to carry out Bayesian inference of parameters. Section 2.3.4 will be discussed about how to estimate new observations from our proposed model.

2.3.1 Posterior Propriety with the "Exact" Reference Prior

In this subsection, we aim to prove the posterior propriety of our GP models with the "exact" reference prior under the situation when the inputs of GP models are missing. Following the discussion of Ren et al. (2012), the "exact" reference prior of hyperparameters of GP, i.e., (ℓ, η, σ_z^2) are based on their Fisher information matrix, which is derived from integrating $\boldsymbol{\beta}$ out using a flat prior in the likelihood of Θ below provided that all data are observed,

$$\mathcal{L}_*(\Theta \mid \mathbf{y}, \mathbf{x}, \mathbf{Z}) \propto \left(\frac{1}{\sigma_z^2}\right)^{n/2} |\mathbf{G}|^{-1/2} \exp\left\{-\frac{1}{2\sigma_z^2}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{G}^{-1}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})\right\}. \tag{2.3.1}$$

Here, $\mathcal{L}_*(\cdot)$ with a subscript '$*$' denoting the assumption that $\mathbf{x}$ is fully observed in this expression. The Fisher information matrix derived from the integrated likelihood of (ℓ, η, σ_z^2) in (2.3.1) is given by

$$I^*(\ell, \eta, \sigma_z^2) = \frac{1}{2}\begin{pmatrix} tr\{\mathbf{R}_G\frac{\partial}{\partial\ell}\mathbf{K}\}^2 & tr\{\mathbf{R}_G^2\frac{\partial}{\partial\ell}\mathbf{K}\} & \frac{1}{\sigma_z^2}tr\{\mathbf{R}_G\frac{\partial}{\partial\ell}\mathbf{K}\} \\ tr\{\mathbf{R}_G^2\frac{\partial}{\partial\ell}\mathbf{K}\} & tr(\mathbf{R}_G^2) & \frac{1}{\sigma_z^2}tr(\mathbf{R}_G) \\ \frac{1}{\sigma_z^2}tr\{\mathbf{R}_G\frac{\partial}{\partial\ell}\mathbf{K}\} & \frac{1}{\sigma_z^2}tr(\mathbf{R}_G) & \frac{n-q}{(\sigma_z^2)^2} \end{pmatrix}, \tag{2.3.2}$$

where $\mathbf{R}_G = \mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{G}^{-1}$, $tr(\cdot)$ is the notation for trace and $\partial\mathbf{K}/\partial l$ indicates the first-order partial derivative of $\mathbf{K}$ with respect to ℓ. Applying the derivation

of the "exact" reference prior from Ren et al. (2012), a non-informative prior for Θ is

$$\pi^R(\Theta) = \pi^R(\ell, \eta, \sigma_z^2, \boldsymbol{\beta}) \propto \frac{1}{\sigma_z^2}\sqrt{|I^*(\ell, \eta, 1)|}, \tag{2.3.3}$$

where $I^*(\ell, \eta, 1)$ implies that we use $\sigma_z^2 = 1$ in Equation (2.3.2). In fact, the non-informative prior of $\pi^R(\ell, \eta, \sigma_z^2, \boldsymbol{\beta})$ can be rewritten as $\pi^R(\ell, \eta, \sigma_z^2, \boldsymbol{\beta}) = \pi(\boldsymbol{\beta})\pi(\sigma_z^2)\pi_*^R(\ell, \eta)$, where $\pi(\boldsymbol{\beta}) \propto 1$, $\pi(\sigma_z^2) \propto 1/\sigma_z^2$ and $\pi_*^R(\ell, \eta) \propto \sqrt{|\Sigma_*(\ell, \eta, 1)|}$.

Then, to show the posterior propriety of Θ using the "exact" reference prior (2.3.3) under the missing data framework for our Model (2.2.1), we only need to show the integration of the joint posterior distributions of Θ and $\mathbf{x}^{mis}$ below

$$\begin{aligned} f(\Theta, \mathbf{x}^{mis} \mid \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z}) &\propto \left(\frac{1}{\sigma_z^2}\right)^{n/2} |\mathbf{G}|^{-1/2} \exp\left\{-\frac{1}{2\sigma_z^2}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{G}^{-1}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})\right\} \\ &\times \pi^R(\Theta)\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}), \end{aligned} \tag{2.3.4}$$

is finite over the domain of Θ and $\mathbf{x}^{mis}$, where in (2.3.4), $\pi^R(\Theta)$ is the reference prior defined in (2.3.3) and $\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})$ is the prior distribution for $\mathbf{x}^{mis}$ given the observed $\mathbf{x}^{obs}$, which depends on the marginal distribution of $\pi(\mathbf{x})$.

To verify the propriety of the joint posterior (2.3.4), first, let us integrate out $\boldsymbol{\beta}$ and σ_z^2 from this joint distribution, which yields

$$f(\ell, \eta, \mathbf{x}^{mis} \mid \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z}) \propto |\mathbf{G}|^{-1/2}|\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2}(S^2)^{-(n-q)/2}\pi_*^R(\ell, \eta)\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}),$$

where $S^2 = (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})'\mathbf{G}^{-1}(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})$ and $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{G}^{-1}\mathbf{y}$. Using the Condition A1 to Condition A4 in Appendix A.1, Ren et al. (2012) have proved that the integrated likelihood $\mathcal{L}_{**}(\mathbf{x}^{mis})$ is finite, that is:

$$0 < \mathcal{L}_{**}(\mathbf{x}^{mis}) = \int_{\ell}\int_{\eta} |\mathbf{G}|^{-1/2}|\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2}(S^2)^{-(n-q)/2}\pi_*^R(\ell,\eta)d\ell d\eta < \infty, \qquad (2.3.5)$$

for a given $\mathbf{x}^{mis}$. Therefore, in the presence of ignorable missingness in covariates, to show the joint posterior distribution of $(\Theta, \mathbf{x}^{mis})$ is proper, we only need to verify that

$$0 < \int_{\ell}\int_{\eta}\int_{\mathbf{x}^{mis}} |\mathbf{G}|^{-1/2}|\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2}(S^2)^{-(n-q)/2}\pi_*^R(\ell,\eta)\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})d\ell d\eta d\mathbf{x}^{mis} < \infty.$$

By using the result (2.3.5), this is equivalent to prove that

$$0 < \int_{\mathbf{x}^{mis}} \mathcal{L}_{**}(\mathbf{x}^{mis})\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})d\mathbf{x}^{mis} < \infty. \qquad (2.3.6)$$

Following the derivation of Ren et al. (2012) and the conditions they stated (see details in Appendix A.1), $\mathcal{L}_{**}(\mathbf{x}^{mis})$ is finite in (2.3.5) and thus bounded. Therefore, if we add additional condition below, i.e.,

$$\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}) \text{ is a proper density,} \qquad (A5)$$

then (2.3.6) will be finite.

The Condition A5 is easy to achieve. For example, if we specify a proper prior on $\mathbf{x}$, often the conditional distribution of $\mathbf{x}^{mis}$ given $\mathbf{x}^{obs}$ will be proper as well. Without loss of generality, for the discussion throughout this chapter, we will assume the covariates x_i's for the unknown function $g(\cdot)$ in Model (2.2.1) follow $x_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu_x, \sigma_x^2)$ and further, presume the prior of hyperparameter μ_x and σ_x^2 to be $\pi(\mu_x, \sigma_x^2) \propto 1/\sigma_x^2$. Integrating out μ_x and σ_x^2, the conditional marginal prior for $\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}) = t_\nu(\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}^{obs}, (\kappa/(n - m - 1))\mathbf{A}_{11}^{-1})$, where $\kappa = \mathbf{x}^{obs'}\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)\mathbf{x}^{obs}$, $\mathbf{A}_{11} = \mathbf{I}_m - \mathbf{J}_m/n$, $\mathbf{A}_{12} = \mathbf{A}_{21}' = -\mathbf{J}_{m\times k}/n$ and $\mathbf{A}_{22} = \mathbf{I}_k - \mathbf{J}_k/n$ with $\mathbf{I}$ being the identity matrix and $\mathbf{J}$ being the unit matrix. The derivation of $\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})$ is postponed to Appendix A.2.

The Condition A1 in Appendix A.1 ensures that the correlation function will decrease to zero as the distance between two points goes to infinity. The Condition A2 in Appendix A.1 ensures $\ell \to \infty$, a Taylor expansion of the correlation function will follow. For the power exponential kernel and Matérn kernel we used in our chapter for the correlation matrix of the GP model, it will automatically satisfy the Conditions A1 and A2. Also, Conditions A1 and A2 are also applicable to other popular correlation matrices including the spherical kernel, rational quadratic kernel and other isotropic kernels. From our discussion, under the Conditions A1-A4 and together with the additional condition A5, we can easily establish the posterior propriety of $(\Theta, \mathbf{x}^{mis})$ in our model.

2.3.2 Posterior Consistency

We will follow the proof given in Choi and Schervish (2007) to prove posterior consistency of our model under ignorable missingness in covariates.

Suppose that x_i has density Q. We can rewrite our model as:

$$\begin{aligned} y_i &= g(x_i) + \epsilon_i \\ g(\mathbf{x}) &\sim GP(\mathbf{Z}\boldsymbol{\beta}, \Sigma) \\ \epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2) \ , \ x_i &\overset{iid}{\sim} Q \\ \sigma_\epsilon^2 &\sim \pi(\sigma_\epsilon^2) \end{aligned} \tag{2.3.7}$$

Here, we assume that covariate x_i takes value in a compact set T. Without loss of generality, we assume that $T = [0, 1]$. Here, we will prove posterior consistency of parameter $(g(.), \sigma_\epsilon^2)$. Let the true value of parameter be denoted as $(g_0(.), \sigma_{\epsilon 0}^2)$. Let the Π be the prior on $(g(.), \sigma_\epsilon^2)$ induced by GP prior on $g(.)$ and a proper prior $\pi(\sigma_\epsilon^2)$. Since, the missing data mechanism is ignorable, that is it does not depend on x_i's, we can ignore it for Bayesian computation and also proving posterior consistency. Now, to prove posterior consistency we have to show that all the conditions mentioned in Theorem 1 in Choi and Schervish (2007) are satisfied. $y_i \mid x_i$ has density f_i with respect to Lebesgue measure λ, which is the normal density with mean $g(x_i)$ and variance σ_ϵ^2. Then f_i is the joint density of (x_i, y_i) with respect to $\nu = Q \times \lambda$. Also denote f_0 as the true density obtained from using the parameter $(g_0(.), \sigma_{\epsilon 0}^2)$. We define the Hellinger

neighborhood of $(g_0(.), \sigma^2_{\epsilon 0})$, $H_\epsilon = \{(g(), \sigma^2_\epsilon) : d_H(f, f_0) < \epsilon\}$, $\epsilon > 0$. Here, $d_H(.,.)$ is the Hellinger distance between the two distributions and is defined as following:

$$d_H(f, f_0) = \int \left[\sqrt{f} - \sqrt{f_0}\right] d\nu(x, y).$$

Note that the model mentioned in equation (2.3.7) is the same as the model studied by Choi and Schervish (2007) under the assumption that the covariates are sampled from a probability density Q. Choi and Schervish (2007) have proved that posterior probability of joint neighborhoods defined above converge almost surely to 1, then it follows that the posterior probability of marginal neighborhoods converge almost surely to 1. Choi and Schervish (2007) have shown that additional assumptions on prior on covariance hyperparameters is required for posterior consistency to hold. Li et al. (2016) have proved that under inverse gamma prior on both ℓ and σ^2_z, the posterior consistency will hold. Hence, for every $\epsilon > 0$,

$$\Pi\{H_\epsilon|(x_i, y_i), i = 1, \cdots, n\} \to 1 \text{ a.s. } P_{f_0}.$$

2.3.3 Bayesian Computation and Sampling Schemes

Since the joint posterior distribution of $(\Theta, \mathbf{x}^{mis})$ in (2.3.4) is proper, we will rely on this joint posterior to make inference for our proposed Model (2.2.1) with missing input x_i's. However, this joint posterior does not have a closed form, thus, we shall resort to MCMC

sampling scheme to draw samples of unknown parameters to make inference. There are two key steps in developing the MCMC scheme. First, we draw the missing values $\mathbf{x}^{mis}$ provided that the unknown parameters Θ in the Model (2.2.1) is known and treat the values drawn for $\mathbf{x}^{mis}$ as their imputed values. Second, we sample Θ based on observed $\mathbf{x}^{obs}$ and imputed $\mathbf{x}^{mis}$. This alternative process simulating missing data and parameter creates a Markov chain that eventually stabilizes to the joint posterior distribution of parameters and missing covariates in (5.2.5). The detailed steps of MCMC schemes are described below.

Step 1: draw $\mathbf{x}^{mis}$ from its posterior conditional distribution:

$$\begin{aligned} f(\mathbf{x}^{mis} \mid \ell, \eta, \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z}) &\propto \sqrt{|I^*(\ell,\eta,1)|} \times |\mathbf{G}|^{-1/2} |\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2} \\ &\times \left(\frac{(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})'\mathbf{G}^{-1}(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})}{2} \right)^{-(n-q)/2} \times \pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}), \end{aligned}$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{G}^{-1}\mathbf{y}$ and the term $I^*(\ell, \eta, 1)$ is defined in Equation (2.3.3). Since the conditional posterior distribution of $\mathbf{x}^{mis}$ do not have a closed form, we consider to use a Metropolis-Hastings algorithm to impute new values of $\mathbf{x}^{mis}$.

Step 2: Given the imputed values $\mathbf{x}^{mis}$, we sample the value of ℓ using the posterior conditional distribution given by:

$$\begin{aligned} f(\ell \mid \eta, \mathbf{y}, \mathbf{x}, \mathbf{Z}) &\propto \sqrt{|I^*(\ell,\eta,1)|} \times |\mathbf{G}|^{-1/2} |\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2} \\ &\times \left(\frac{(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})'\mathbf{G}^{-1}(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})}{2} \right)^{-(n-q)/2}. \end{aligned}$$

It is easy to see that the posterior conditional distribution of ℓ is not closed form, thus we use slice sampling (c.f., Neal (2003)) to draw samples of ℓ from its posterior conditional distribution.

Step 3: Provided that $\mathbf{x}^{mis}$ and ℓ are known, we will sample η from its posterior conditional distribution:

$$\begin{aligned} f(\eta \mid \ell, \mathbf{y}, \mathbf{x}, \mathbf{Z}) &\propto \sqrt{|I^*(\ell, \eta, 1)|} \times |\mathbf{G}|^{-1/2} |\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2} \\ &\times \left(\frac{(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})'\mathbf{G}^{-1}(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})}{2} \right)^{-(n-q)/2}. \end{aligned}$$

The posterior conditional distribution of η does not have a closed form either and we will also use the slice sampling algorithm to draw samples of η from its posterior conditional distribution.

Step 4: When $\mathbf{x}^{mis}$, ℓ and η are known, we will draw σ_z^2 from its posterior conditional distribution:

$$f(\sigma_z^2 \mid \ell, \eta, \mathbf{y}, \mathbf{x}, \mathbf{Z}) \propto (\sigma_z^2)^{-(a+1)} \exp(-b/\sigma_z^2),$$

which is an inverse gamma distribution with the shape parameter $a = (n-q)/2$ and the rate parameter $b = (\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})'\mathbf{G}^{-1}(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}})/2$.

Step 5: Given $\mathbf{x}^{\mathbf{mis}}$, ℓ, η and σ_z^2, then we can sample $\boldsymbol{\beta}$ from its posterior conditional distribution, which is a q-dimensional multivariate normal distribution with mean vector $\widehat{\boldsymbol{\beta}}$ and covariance matrix $\sigma_z^2 \left(\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}\right)^{-1}$.

Once we give the initiate values for $\ell, \eta, \boldsymbol{\beta}, \sigma_z^2$, and $\mathbf{x}^{mis}$, then the Bayesian computation is done by running MCMC algorithms from *Step 1* through *Step 5* until the MCMC has converged. To evaluate the convergence of the MCMC chains, we run the MCMC chains with 10 different starting values of parameters. The Gelman-Rubin potential scale reduction factor (c.f., Brooks and Gelman (1998)) are found to be very close to 1 at most after 25,000 iterations of MCMC runs in our simulations and examples for every parameter needed to estimate in the Model (2.2.1). We also evaluate the convergence by informally looking at trace plots and we find the MCMC chains are mixing well after 25,000 iterations in our simulations and examples. After MCMC samples are converged, the statistical inferences are straightforward by utilizing the MCMC samples. For example, a posterior median estimate and 95% credible interval for the unknown function $g(\cdot)$ can be formed from the median, 2.5%, and 97.5% empirical quantiles of the corresponding MCMC realizations, respectively.

2.3.4 Posterior Predictive Distribution

In Subsection 2.3.3, we have developed a MCMC algorithm to impute the missing covariates under ignorable missing mechanism as well as to estimate the unknown parameters in Model (2.2.1) simultaneously. However, often in the study, one of our goals is to predict responses using Model (2.2.1) when new observations of covariates comes, while, other purpose might be using future observations to assess the performance of our proposed models in comparison to other competitive models. For these reasons, in this

subsection, we are going to derive the posterior predictive distribution of $\mathbf{y}^{new}$ when we observe new covariates in Model (2.2.1).

Let us presume that the n observations $\{x_i, y_i, \mathbf{z}_i\}_{i=1}^{n}$ are training data points and $\left\{x_j^{test}, y_j^{test}, \mathbf{z}_j^{test}\right\}_{j=1}^{t}$ are t test points, where x_i^{test}'s and $\mathbf{z}_i^{test}$'s are observed new covariates with $\mathbf{z}_i^{test} = [1, z_{i1}^{test}, \cdots, z_{ip}^{test}]'$, while y_i^{test}'s are unknown and needed to predict. To estimate y_i^{test}'s under the new observations x_i^{test}'s and $\mathbf{z}_i^{test}$, from Bayesian perspective, we shall first derive the posterior predictive distribution for y_i^{test}'s given the observed y_i's and observed covariates.

In addition, denote $\mathbf{y}^{test} = (y_1^{test}, \cdots, y_t^{test})'$, $\mathbf{x}^{test} = (x_1^{test}, \cdots, x_n^{test})'$, and $\mathbf{Z}^{test} = [\mathbf{z}_1^{test}, \cdots, \mathbf{z}_t^{test}]'$. Then, the posterior predictive distribution of $\mathbf{y}^{test}$ given $\mathbf{y}$ and other observed covariates can be written as to integrate out all the unknown parameters Θ and missing values $\mathbf{x}^{mis}$ from the posterior conditional distribution of $\mathbf{y}^{test}$ below provided that $\mathbf{x}^{test}, \mathbf{Z}^{test}, \mathbf{x}^{mis}, \Theta$, $\mathbf{x}^{obs}$, $\mathbf{y}$ and $\mathbf{Z}$ are known,

$$\begin{aligned} f(\mathbf{y}^{test}|\mathbf{x}^{test}, \mathbf{Z}^{test}, \mathbf{x}^{obs}, \mathbf{Z}, \mathbf{y}) &= \int\int f(\mathbf{y}^{test} \mid \mathbf{x}^{test}, \mathbf{Z}^{test}, \mathbf{x}^{mis}, \Theta, \mathbf{x}^{obs}, \mathbf{y}, \mathbf{Z}) \\ &\times f(\Theta, \mathbf{x}^{mis} \mid \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z}) d\Theta d\mathbf{x}^{mis}, \end{aligned} \tag{2.3.8}$$

where $f(\Theta, \mathbf{x}^{mis} \mid \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z})$ is the joint posterior distribution of $(\Theta, \mathbf{x}^{mis})$ derived in

(5.2.5) and $f(\mathbf{y}^{test} \mid \mathbf{x}^{test}, \mathbf{Z}^{test}, \mathbf{x}^{mis}, \Theta, \mathbf{x}^{obs}, \mathbf{y}, \mathbf{Z})$ is following a multivariate normal distribution, i.e.,

$$f(\mathbf{y}^{test} \mid \mathbf{x}^{test}, \mathbf{Z}^{test}, \mathbf{x}^{mis}, \Theta, \mathbf{x}^{obs}, \mathbf{y}, \mathbf{Z}) = \mathcal{N}\left(\bar{f}_{test}, \mathrm{Cov}(f_{test})\right), \tag{2.3.9}$$

where $\bar{f}_{test} = \mathbf{Z}^{test}\boldsymbol{\beta} + \Sigma_{(\mathbf{x}^{test},\mathbf{x})}(\sigma_z^2\mathbf{G})^{-1}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})$ and $\mathrm{Cov}(f_{test}) = \Sigma_{(\mathbf{x}^{test},\mathbf{x}^{test})} - \Sigma_{(\mathbf{x}^{test},\mathbf{x})}(\sigma_z^2\mathbf{G})^{-1}\Sigma_{(\mathbf{x},\mathbf{x}^{test})}$. Notice that $\Sigma_{(\mathbf{x},\mathbf{x}^{test})} = \Sigma'_{(\mathbf{x},\mathbf{x}^{test})}$ is a $n \times t$ matrix and its (i,j)th element $(\Sigma_{(\mathbf{x},\mathbf{x}^{test})})_{i,j} = \sigma_z^2\mathbf{K}(x_i, x_j^{test})$, where x_i is a training point for $i = 1, \cdots, n$ and x_j^{test} is a test point for $j = 1, \cdots, t$.

Let M be total number iterations of MCMC samples after burn-in period. Then, to generate a random sample $\mathbf{y}^{test}$ from its posterior predictive distribution in (2.3.8), it involves two major iterative steps, that is, for $i = 1, \cdots, M$,

1. draw $(\Theta, \mathbf{x}^{mis})$ from $f(\Theta, \mathbf{x}^{mis} \mid \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z})$, where the detailed steps are described in Subsection 2.3.3.

2. after given the values of $(\Theta, \mathbf{x}^{mis})$ at the i-th iteration, we sample the i-th iteration values of $\mathbf{y}^{test}$ from

$$\mathbf{y}^{test} \sim f(\mathbf{y}^{test} \mid \mathbf{x}^{test}, \mathbf{Z}^{test}, (\mathbf{x}^{mis})^{(i)}, \Theta^{(i)}, \mathbf{x}^{obs}, \mathbf{y}, \mathbf{Z}) = \mathcal{N}\left(\bar{f}_{test}^{(i)}, \mathrm{Cov}(f_{test}^{(i)})\right),$$

where $\bar{f}_{test}^{(i)} = \mathbf{Z}^{test}\boldsymbol{\beta}^{(i)} + \Sigma_{(\mathbf{x}^{test},\mathbf{x}^{(i)})}(\sigma_z^{2(i)}\mathbf{G}^{(i)})^{-1}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}^{(i)})$, $\mathrm{Cov}(f_{test}) = \Sigma_{(\mathbf{x}^{test},\mathbf{x}^{test})} - \Sigma_{(\mathbf{x}^{test},\mathbf{x}^{(i)})}(\sigma_z^{2(i)}\mathbf{G}^{(i)})^{-1}\Sigma_{(\mathbf{x}^{(i)},\mathbf{x}^{test})}$ and noticing $\mathbf{x}^{(i)} = (\mathbf{x}^{obs}, (\mathbf{x}^{mis})^{(i)})'$.

Then, the posterior median estimate of $\mathbf{y}^{test}$ can be easily calculated by using $\mathbf{y}^{test} = \sum_{i=1}^{M} \mathbf{y}_{(i)}^{test}/M$.

2.4 Simulation Examples

In this section, we design some simulation examples to validate the inference procedure proposed in Section 2.3 and compare the benefits by imputing the missing values in Model (2.2.1) instead of using complete data only. Further, we conduct some experiments to analyze the sensitivity of misspecification of correlation functions for GP priors assigned to $g(\cdot)$ in Model (2.2.1).

2.4.1 Simulation I

Let us consider the semiparametric regression model (2.2.1) with the following specification,

$$y_i = \beta_0 + \beta_1 z_i + g(x_i) + \epsilon_i,\ i = 1, \cdots, n, \tag{2.4.1}$$

where $\beta_0 = -10$, $\beta_1 = 20$, $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 10)$, $z_i \overset{i.i.d.}{\sim} \mathcal{N}(1, 5)$, $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 0.4$ and $n = 75$. Moreover, we assume $g(x_i)$ has a GP prior with the mean function centered at 0 and the correlation function being squared exponential correlation function, that is

$$(\Sigma)_{ij} = \sigma_z^2 \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right),$$

where the values of hyperparameters are $\ell = 10$, $\sigma_z^2 = 2$. Thus, $\eta = \sigma_z^2/\sigma_\epsilon^2 = 0.2$. In order to test the performance of our proposed method, we randomly select 50 data points out of 75 generated data points from (2.4.1) to be training datasets, while the rest 25 data points are left for the assessment of the prediction power for the model. Next, we create an average of 10%, 25% and 40% missingness of covariates x_i's in $g(x_i)$ for training data points according to the procedure described below, that is, we randomly generate the missing indicator from

$$R_i \sim Bin(1, p_i), \quad \text{with } p_i = \frac{\exp(b_0 + b_1 y_i)}{1 + \exp(b_0 + b_1 y_i)}, \tag{2.4.2}$$

where $R_i = 1$ indicates x_i is missing for the ith subject, $R_i = 0$ otherwise. We fix $b_1 = -0.1$ in (2.4.2) and then in each simulation run, we solve the value of b_0 to make the average missing probability of p_i's over 50 training points equals to 0.1, 0.25 and 0.4, respectively, in three different scenarios of percentages for missingness.

After the data were generated, we employ the MCMC sampling methods developed in Subsection 2.3.3 to estimate model parameters and impute missing values of x_i's. We apply the reference prior discussed in Subsection 2.3.1 for the unknown parameters $\ell, \eta, \beta_0, \beta_1$ and σ_z^2 in Model (2.4.1). Further, we assume the covariates x_i's follow $\pi(x_i) \overset{i.i.d.}{\sim} \mathcal{N}(\mu_x, \sigma_x^2)$ with the hyperprior on μ_x and σ_x^2 being $\pi(\mu_x, \sigma_x^2) \propto 1/\sigma_x^2$. Using the derivation in Appendix A.2, we know that the conditional prior distribution of $\mathbf{x}^{mis}$ given $\mathbf{x}^{obs}$ will follow a multivariate t-distribution.

For each simulated data, we run the MCMC for 100,000 iterations, where the first 50,000 draws are discarded as a burn-in phase and every 10th values of MCMC samples are stored to reduce level of correlation between successive values of the chain. For each different scenario of missing percentage, we repeat the entire simulation procedure described above for 50 times using different random seeds. Then, we compare the parameters estimated in Model (2.4.1) using our proposed methods (PM) with the naive method using only complete cases (CC) (i.e., fitting Model (2.4.1) using only those data points where covariate values x_i's are observed).

The comparison of the two methods is shown in Table 1, where we have compared their predicted mean squared error (PMSE) of y_i's for testing points respectively as well as their estimated bias of the parameters in Model (2.4.1) relative to the truth. Notice in Table 1, the bias of the parameters are calculated using the absolute distance between posterior median estimates of the parameters and their corresponding true values in simulations. From Table 1, it is clear to see that using our proposed method to impute the missing covariates x_i's, we are able to predict the testing points with better accuracy than using the naive method in all three different levels of missingness. Moreover, when the missing rate is higher, the posterior median estimates of the hyperparameters of GP prior as well as the parametric coefficients in Model (2.4.1) have relative lower biases by using our proposed method than using the naive method.

Table 1: Comparison between our proposed model and the complete case analysis for Model (2.4.1)

	10 % Missing		25 % Missing		40 % Missing	
	PM	CC	PM	CC	PM	CC
PMSE for y	1.7945	1.9152	2.4017	2.9412	2.7488	4.4022
Bias for ℓ	0.2333	0.2138	0.1966	0.3131	0.4182	0.5661
Bias for η	0.0204	0.0156	0.0139	0.0342	0.0252	0.0374
Bias for σ_z^2	1.1991	1.0514	1.5437	1.6178	2.6262	2.9159
Bias for β_0	1.0241	0.9004	1.2804	1.8320	1.6101	1.7255
Bias for β_1	0.1169	0.1420	0.0775	0.2365	0.2427	0.8512

2.4.2 Simulation II

In this subsection, we have designed several simulation experiments to test the performance of our proposed method under misspecification of correlation functions for the GP prior assigned to $g(\cdot)$ in Model (2.2.1). Here, we consider three types of covariance functions for the GP prior, which are commonly used in spatial statistics and machine learning field, i.e.,

1. Squared Exponential (SE) Covariance Function:

$$(\Sigma)_{ij} = \sigma_z^2 \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right),$$

2. γ-exponential (γ-E) Covariance Function:

$$(\Sigma)_{ij} = \sigma_z^2 \exp\left(-\frac{|x_i - x_j|^\gamma}{\ell}\right),$$

where $|x|$ is the absolute value of x and $0 < \gamma \leq 2$. In the simulation, we choose $\gamma = 1$.

3. Mátern Class (MC) of Covariance Functions:

$$(\Sigma)_{ij} = \sigma_z^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu|x_i - x_j|}}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu|x_i - x_j|}}{\ell} \right),$$

with positive parameters ν and ℓ, where $K_\nu(\cdot)$ is a modified Bessel function. The most interesting cases for Mátern class of covariance functions are $\nu = 3/2$ and $\nu = 5/2$ (abbreviations are $MC_{3/2}$ and $MC_{5/2}$ in Table 2) , that is:

$$\begin{aligned}
(\Sigma)_{ij} &= \sigma_z^2 \left(1 + \frac{\sqrt{3}|x_i - x_j|}{\ell} \right) \exp\left(-\frac{\sqrt{3}|x_i - x_j|}{\ell} \right), \text{ for } \nu = 3/2, \\
(\Sigma)_{ij} &= \sigma_z^2 \left(1 + \frac{\sqrt{5}|x_i - x_j|}{\ell} + \frac{5(x_i - x_j)^2}{3\ell^2} \right) \exp\left(-\frac{\sqrt{5}|x_i - x_j|}{\ell} \right), \text{ for } \nu = 5/2.
\end{aligned}$$

In our simulation, we have considered both the choice of $\nu = 3/2$ and $\nu = 5/2$.

For each choice of covariance functions above and each missing covariate percentages (i.e., 10%, 25% and 40%), we apply Model (2.4.1) to generate 10 different sets of data using different random seeds. The simulation procedure is the same as described in Subsection 2.4.1 with only changing the covariance function specified for the GP prior assigned to $g(\cdot)$ in Model (2.4.1) and the missingness is created using the same missing at random mechanism explained in (2.4.2). Therefore, we will have $10 \times 4 \times 3 = 120$ datasets in total.

In Table 2, we assess the performance of our proposed methods under misspecification of covariance functions for the GP prior assigned to $g(\cdot)$ in Model (2.4.1). We use the mean squared error of imputed missing values of x_i's (MSEx), the predicted mean squared error (PMSE) of testing points y_i's and deviance information criterion (DIC) (Spiegelhalter et al., 2002) to evaluate the performance and test the goodness of fit for the difference choices of covariance functions under different datasets. The values of DIC (in fact, due to the complication of integrating out $\mathbf{x}^{mis}$ in the likelihood, we use the conditional DIC defined in Celeux et al. (2006)) are easily obtained from MCMC samples. Notice that for model comparison, we can define the deviance as

$$D(\Theta, \mathbf{x}^{mis}) = -2\log(f(\mathbf{y} \mid \Theta, \mathbf{x}^{mis}, \mathbf{x}^{obs}, \mathbf{Z})),$$

where $f(\mathbf{y} \mid \Theta, \mathbf{x}^{mis}, \mathbf{x}^{obs}, \mathbf{Z})$ is the conditional likelihood of $\mathbf{y}$. Then, apply the original definition of DIC to this conditional distribution, which leads to

$$DIC = -2\mathrm{E}_{\Theta,\mathbf{x}^{mis}}[D(\Theta, \mathbf{x}^{mis}) \mid \mathbf{y}] + 2\log f(\mathbf{y} \mid \widetilde{\Theta}, \widetilde{\mathbf{x}}^{mis}, \mathbf{x}^{obs}, \mathbf{Z}),$$

where $\mathrm{E}_{\Theta,\mathbf{x}^{mis}}(\cdot)$ implies taking expectation respect to the joint posterior distribution of Θ and $\mathbf{x}^{mis}$, which can easily be approximated using an MCMC run by taking the sample mean of the simulated values of $D(\Theta, \mathbf{x}^{mis})$ and we choose $\widetilde{\Theta}$ and $\widetilde{\mathbf{x}}^{mis}$ as their posterior medians in our study. Every value listed in Table 2 has already been averaged

over the 10 different datasets. The smaller the values of MSEx, PMSE and DIC are, the better the fit of the corresponding covariance models are. In Table 2, the number with the bold blue color indicates the smallest value we picked is the same choice as the true covariance kernel used to generate data, while the number with the bold black color indicates the true covariance kernel used to generate data does not have the smallest value and instead the covariance kernel in the fitted model with the the red color number yield comparatively smaller value. In Table 2, we could see the values of MSEx, PMSE and DIC are not substantial differences among different choices of covariance functions. Most frequently, the true kernels will yield the smallest values of MSEx, PMSE and DIC in the corresponding categories.

Since the SE covariance function has lots of good properties and supports a large class of functions with various shapes, we further examine on the performance of using SE covariance functions when the other covariance kernels are true. The detailed results were summarized in Table 3, where those numbers are computed via (using DIC values as an example),

$$ratio = \frac{|DIC_{SE} - DIC_{True}|}{DIC_{True}},$$

where $|\cdot|$ represents the absolute value, DIC_{SE} is the DIC values using SE covariance function in the model fit, while DIC_{True} is the DIC values employing the true generated covariance function in the model fit, and DIC values can be replaced by MSEx and PMSE values. From Table 3, we could see the relative changes of MSEx, PMSE and

DIC values of using SE covariance function in comparison to using the true kernel is relative small. Thus, it shows that the performance of our model using SE covariance under misspecification of covariance kernel is kind of robust. Then, in our application, we will choose to work with SE covariance.

2.5 Application

Since our approach has successfully applied to the simulated data and recovered the true values of parameters well, we will employ our methods to two applications. According to our investigation in the previous section for the relative robustness of misspecification of covariance functions in GP prior, we are going to use SE covariance function throughout the applications.

2.5.1 Application I

First, we are going to apply our methodology and evaluate our algorithm in Adsorption Isotherm data for R-113 refrigerant vapors on BPL activated carbon at 298 Kelvin obtained from Mahle et al. (1994). BPL activated carbon is a virgin granular activated carbon designed for use in gas phase applications. It can be reactivated for reuse which eliminates disposal problem. One of the usage of BPL activated carbon is gas purification and solvent recovery. R-113 is 1,1,2-Trichloro-1,2,2-Trifluoroethane, which is a colorless to water white, non-flammable liquid with a slight, ether like odor at high concentrations.

		MSEx				PMSE				DIC			
Fitted Kernel \ Generated Kernel		SE	γ-E	$MC_{3/2}$	$MC_{5/2}$	SE	γ-E	$MC_{3/2}$	$MC_{5/2}$	SE	γ-E	$MC_{3/2}$	$MC_{5/2}$
Missing=10%	SE	**2.8417**	2.1737	2.3339	2.2603	**2.5973**	3.0808	1.2561	1.2603	**195.7971**	184.9669	197.4832	182.1793
	γ-E	2.9858	**2.1384**	2.2346	2.1599	2.9378	**2.9076**	1.2839	0.9114	201.2749	**185.4313**	196.0822	183.2887
	$MC_{3/2}$	3.1829	3.2448	**1.9577**	1.2992	3.0104	2.8244	**1.3273**	1.4481	201.1927	188.3942	**195.3640**	180. 2765
	$MC_{5/2}$	3.0099	3.3451	2.0662	**1.1414**	2.7812	3.1223	1.3999	**1.0139**	201.2321	187.8190	198.7922	**179.1275**
Missing=25%	SE	**1.2250**	2.6233	1.3349	2.7791	**1.3685**	1.3451	1.8227	1.8981	**202.0202**	200.8871	195.8263	226.2590
	γ-E	2.1508	**2.4224**	1.3257	2.9911	1.3546	**1.2621**	1.8520	2.1911	210.0052	**197.7755**	192.6172	222.4992
	$MC_{3/2}$	1.1652	2.7468	**1.4660**	2.0455	1.4607	1.2866	**1.8086**	1.8252	204.7253	210.6224	**192.2594**	221.1930
	$MC_{5/2}$	1.7727	3.5419	1.3431	**1.8781**	1.4199	1.2276	1.9909	**1.6766**	211.0411	209.1989	193.4752	**216.2747**
Missing=40%	SE	**1.6222**	4.0737	1.4891	2.4141	**1.3298**	1.2728	3.3702	2.8902	**210.3894**	205.1100	195.9898	205.8788
	γ-E	1.7767	**4.1026**	2.4091	2.1879	1.1891	**1.2559**	3.6011	2.4740	210.9872	**203.0577**	199.0233	214.2484
	$MC_{3/2}$	2.1971	5.0167	**0.9716**	1.6614	1.3232	1.2691	**2.9741**	2.0111	210.0123	209.6501	**193.1111**	203.4333
	$MC_{5/2}$	2.7811	5.1779	1.1163	**2.4223**	1.2134	1.3443	2.9115	**2.0010**	214.1919	210.7721	193. 4595	**203.5159**

Table 2: The comparison of the usage of different covariance kernels based on MSEx, PMSE and DIC

		MSEx			PMSE			DIC		
Fitted Kernel \ Generated Kernel		γ-E	$MC_{3/2}$	$MC_{5/2}$	γ-E	$MC_{3/2}$	$MC_{5/2}$	γ-E	$MC_{3/2}$	$MC_{5/2}$
Missing=10%	SE	0.0165	0.1921	0.9803	0.0596	0.0536	0.2430	0.0025	0.0108	0.0170
Missing=25%	SE	0.0829	0.0894	0.4797	0.0658	0.0078	0.1321	0.0157	0.0186	0.0462
Missing=40%	SE	0.0070	0.5256	0.0034	0.0135	0.1332	0.4444	0.0101	0.0149	0.0116

Table 3: The sensitivity analysis of using squared exponential kernels based on MSEx, MSEy and DIC

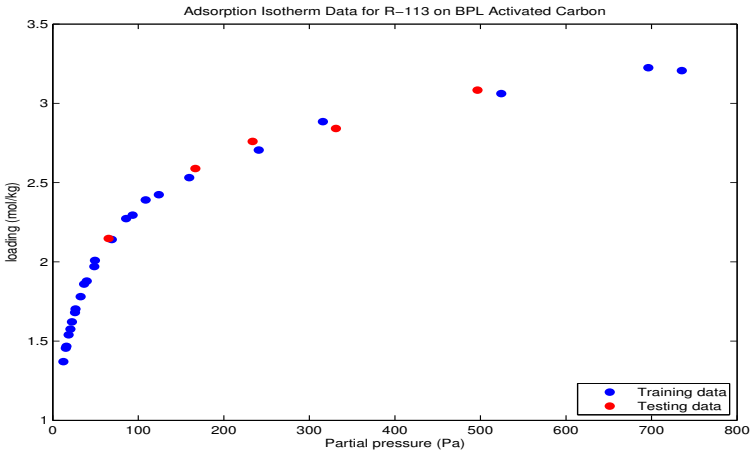

Figure 1: Scatterplot of Adsorption Isotherm Data

It has been used as a cold degreasing agent, dry cleaning solvent, refrigerant, blowing agent, chemical intermediate and drying agent. The data we considered contains 29 observations. We partitioned the data into training and test dataset containing 24 and 5 observations respectively. Figure 1 shows the plot of 24 training data as blue colors and 5 test points as red colors for Adsorption Isotherm data in Mahle et al. (1994).

Adsorption is usually described through isotherms, that is, the amount of adsorbate on the adsorbent (i.e., loading in Figure 1) as a function of its pressure (defined as partial pressure in Figure 1). It is clear that the loading has non-decreasing relationship with the partial pressure from Figure 1. The Langmuir equation, defined in Langmuir (1918)

is one of the most popular models that correlates the amount of adsorbed gases y on plane surfaces of glass, mica, and platinum with the equilibrium aqueous concentration x through a nonlinear function given by

$$y_i = \frac{\alpha\beta x_i}{1+\alpha x_i} + \epsilon_i,\ i = 1, \cdots, n \tag{2.5.1}$$

where $\alpha > 0$, $\beta > 0$, n is the total number of observations and ϵ_i takes account of random measurement errors with the assumption that $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. This formula is the most commonly used isotherm equation because of its simplicity and its ability to fit a variety of adsorption data. In our dataset, y_i in Equation (2.5.1) presents loading (mol/kg), while x_i corresponds to partial pressure (pa) and $n = 24$. However, some of the assumptions used to derive Equation (2.5.1) are seldom all true. Moreover, accuracy of the data collected during the experimental procedure may be affected due to various reasons like equipment failure, data entry error and etc. Thus, in the presence of missing or inaccurate data, the inference based on Langmuir equation may be invalid.

When the data is fully observed and accurate, Dey et al. (1997) proposed a model

$$y_i = \alpha + \beta \log(x_i) + \epsilon_i,\ i = 1, \cdots, n \tag{2.5.2}$$

to be a competitive model with the Langmuir equation, where n is the total number of observations and $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ is a random error. There is no constraints on the values

of parameters α and β in Equation (2.5.2). However, their defined model is merely based on the approximation of the geometric representation of the data generated from the Langmuir equation to ease the computation.

In this part, we compare our proposed model with the model specified in Equation (2.5.2) as well as with the Langmuir equation (2.5.1) using the Adsorption Isotherm Data for R-113 on BPL activated carbon at 298 K obtained from Mahle et al. (1994). We evaluated the accuracy of all the three models for missing imputation using mean squared errors criteria. Let us name our proposed model described in Section 3 as Model 1, and the models described in Equation (2.5.1) and (2.5.2) as Model 2 and Model 3, respectively. From Figure 1, the domain of x, i.e., the partial pressure is always positive. Thus, to make the imputation for the missing covariates of in all three models more efficient, we consider to use the truncated normal prior on covariates, which are truncated at zero on the left. Specifically, we assume $\pi(x_i \mid \mu_x, \sigma_x^2) \propto \mathcal{N}_+(\mu_x, \sigma_x^2)$, where $\mathcal{N}_+(\cdot, \cdot)$ indicates a normal distribution $\mathcal{N}(\cdot, \cdot)$ truncated by the left. For the priors on the hyperparameters μ_x and σ_x^2, we use the same non-informative priors as before, that is, $\pi(\mu_x \mid \sigma_x^2) \propto 1$ and $\pi(\sigma_x^2) \propto 1/\sigma_x^2$. Details about imputation scheme for Model 2 and Model 3 are postponed to Appendix A.3 and Appendix A.4, while the imputation scheme for Model 1 is similar as we discussed in Section 2.3 by merely changing the priors on x_i's.

We artificially create missingness in the covariates using ignorable missing mechanism to compare imputed missing covariate with the true value based on mean squared errors

(MSEx) as well as their predicted mean squared errors (PMSE). We use Equation (2.4.2) to yield the ignorable missingness for the covariate x_i (i.e., the training observations of the partial pressure in Figure 1). We produce the missingness via Equation (2.4.2) with three different percentages, i.e., 10%, 25% and 40%, each of which we repeat the generation 50 times. Thus, for each percentage, we average the values of MSEx and PMSE over 50 times for Model 1, Model 2 and Model 3, a summary of which is given in Table 4. On the basis of prediction of y_i's using imputed x_i's, Model 1 (GP model) and Model 3 (log model) both are able to predict very accurately. However, based on imputation of missing x_i's, even though Model 3 (log model) performs slightly better than Model 1 (GP model), both of them are able to impute far better than Langmuir model. Thus, in comparison to Model 1 and Model 3, Model 2 (Langmuir equation) performed very poorly in the criteria of PMSE and MSEx.

From Table 4, the performance of Model 1 (GP model) is comparable to Model 3 (log model) and much better than Model 2(Langmuir equation). Although Model 3 (log model) is the best among the three models, it has no theory foundation in adsorption isotherm data and it is just approximation to Langmuir model from experimental data. Therefore, Model 3 (log model) will have high risk of misspecification in real application. While, Model 1 (GP model) has nonparametric nature in its fit, thus it will be more flexible in regressing adsorption isotherm data and avoiding misspecification. Hence, our Model 1(GP model) will be a better choice for the analysis of adsorption isotherm data in comparison to the Langmuir model when we have missing covariates.

Table 4: Comparison of Model 1, Model 2 and Model 3 on PMSE and MSEx

	Model 1(GP)	Model 2(Langmuir)	Model 3(log)
PMSE (10% missing)	0.0021	0.0123	0.0019
PMSE (25% missing)	0.0023	0.0124	0.0020
PMSE (40% missing)	0.0024	0.0127	0.0023
MSEx (10% missing)	3.1718	224.1800	3.0866
MSEx (25% missing)	6.8122	272.2199	5.6542
MSEx (40% missing)	10.2698	301.6536	12.6118

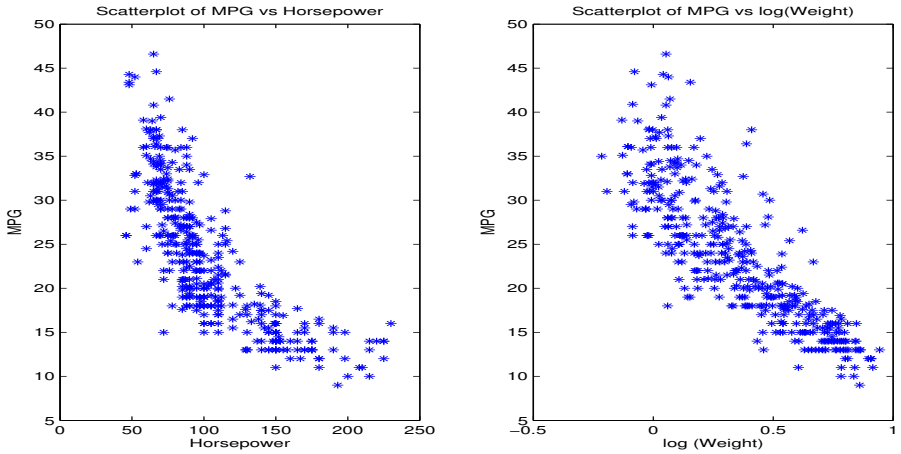

Figure 2: Scatterplot of MPG vs Horsepower and MPG vs log(Weight)

2.5.2 Application II

In this subsection, we are going to use our method on Auto-mpg data. This dataset is from the StatLib library maintaining by Carnegie Mellon University and previously was used in the 1983 American Statistical Association Exposition. This data is also available in the Statistics and Machine Learning Toolbox in MATLAB with a filename called "carbig.mat". One of its application goal is to predict the fuel consumption in miles per gallon (mpg) using the weight and horsepower of a car. In this data, it

contains 398 instances and we have 6 missing values in the horsepower attribute and it is reasonable to consider that missingness in the horsepower attribute is ignorable.

A common approach to model the fuel consumption for this data is to apply the linear regression technique. Our initial study shows that there is a nonlinear relationship between mpg and horsepower, but there is a linear relationship between mpg and the natural logarithm of the weight (denote as log(weight)) of the car. Both of these phenomena can be clearly seen from Figure 2. We randomly sample 30, 60 and 90 instances, respectively from the original data and each sample will include those 6 missing observations, which miss the horsepower attribute. We repeat such random draws for 50 times of each 3 cases of instances and we consider the rest of the observations in the data as test points.

We employed our GP semiparametric model as well as the linear regression on the three cases of instances for the randomly sampled observations. Specifically, our GP semiparametric model is fitted using the linear structure for the natural logarithm of the weight (in tons) and the nonparametric structure for the horsepower, that is:

$$y_i = \beta_0 + \beta_1 z_i + g(x_i) + \epsilon_i,\ i = 1, \cdots, n.$$

Similarly, for the linear regression, we use the horsepower and the natural logarithm of

the weight as predictor variables and the mpg as the response variable, i.e.,

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + \epsilon_i,\ i = 1, \cdots, n.$$

In both regressions above, y_i corresponds to the mpg, x_i is the horsepower attribute, z_i indicates the natural logarithm of the weight of the car, ϵ_i is the random error with the assumption that $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ and n is the number of instances we consider. From Figure 2, it is natural to assume that the values of the horsepower attribute is nonnegative, thus, we assume a truncated normal prior on x_i's, i.e., $\pi(x_i \mid \mu_x, \sigma_x^2) \propto \mathcal{N}_+(\mu_x, \sigma_x^2)$. All the other priors of unknowns are also the same as Subsection 2.5.1. We compare the performance of both models based on the predicted mean squared errors (PMSE), which is the differences between their predictive MPG values relative to the truth in the test sets of the data. To get rid of randomness, we have averaged PMSE over 50 draws for each case of instances.

Table 5 shows the results for both models in the scenarios of imputation (IM) and complete cases (CC). As expected, with the increase in the number of training data, the PMSE decreases for both models. However, our GP semiparametric model is able to perform better than the linear model in either scenarios. In addition, our proposed methods to impute the missing covariates in the GP semiparametric model are dominant in the performance of PMSE in comparison to the complete cases analysis using GP semiparametric model in all three cases of instances. Thus, our proposed GP semiparametric

model is superior in the analysis of the Auto-mpg data.

Table 5: The comparison of our GP semiparametric model and the linear model using the PMSE criteria

	GP Semiparametric Model		Linear Model	
	IM	CC	IM	CC
PMSE dataset 1 (n=30)	21.4647	21.8333	23.4285	23.2001
PMSE dataset 2 (n=60)	17.2350	17.7081	18.1333	19.0001
PMSE dataset 3 (n=90)	16.7127	16.8739	17.8999	17.9111

2.6 Discussion

In this chapter, we have considered the problem of imputation of missing covariates for the nonparametric part in a semiparametric regression under Bayesian framework. In the absence of parametric regression part, our semiparametric model can be reduced to the nonparametric regression setting. Our proposed procedure permits us to model nonparametric as well as semiparametric regression in the presence of missing covariate by imposing a GP prior on the unknown regression function and to employ appropriate missing imputation schemes to handle the missing covariates.

Besides, from the two application data, we demonstrated that our proposed method is able to perform better than the competitive parametric methods when there are missing covariates in the data and we are not certain about the parametric relationship between the response and the predictor variables. Thus, our method will be particularly appealing for analyzing the data where the covariates are subject to ignorable missingness and the

relationship between the response and the covariates is unclear.

Throughout the chapter, we assume the missing data mechanism is ignorable, we have extended our proposed procedure for non-ignorable missing mechanism. Future directions of this work is discussed in more details in Chapter 5.

Chapter 3

Flexible Symmetric Power Link Functions in Nonparametric Ordinal Regression with Gaussian Process Priors

3.1 Introduction

Consider a random binary response y_i measured with covariate $\mathbf{x}_i$, for $i = 1, \cdots, n$. To analyze such binomial response we usually use generalized linear model (GLM), where we model the latent probability of "success" through a link function (McCullagh and Nelder, 1989), that is:

$$P(y_i = 1) = p(\mathbf{x}_i).$$

Traditionally, a parametric approach to the specification of $p(x)$ is taken using $p(x_i) = H(\mathbf{x}_i\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is unknown parameter vector and H is a cumulative distribution

function (cdf), called the link function. The logit, p‘robit and Student-t link functions are the most frequently used link function in GLM. Despite the wide use to these link functions, the inference is sensitive to the choice of link functions. Moreover, these link functions lack flexibility. Most of them are symmetric links in the sense that they assume that the latent probability of a binomial response will approach towards 0 with the same rate as it approaches to 1. In another words, the probability density function (pdf) that corresponds to the inverse cumulative distribution function (cdf) of the link function is symmetric. In many cases, this may not be a reasonable assumption. A commonly used asymmetric link function is the complementary loglog (cloglog) function. However, it has a fixed negative skewness, which restricts the ability of data to allow for positive skewness. To overcome this issue, several authors have proposed models to introduce flexibility into the link functions. Unfortunately, most of these proposed link functions suffer from issues like improper posteriors or bounded range of the skewness. Recently, Jiang et al. (2013) proposed a general class of symmetric power link function by introducing a power parameter into the cdf corresponding to a symmetric link function and its mirror reflection. By doing so, greater flexibility in skewness can be achieved in both positive and negative directions. However, the drawback of their approach is that they have assumed the commonly used latent regression functions have parametric forms. This parametric assumption may not be appropriate for many data, as shown by Li et al. (2016). Limiting the latent regression function to a simple linear or parametric form is clearly restrictive in modeling the binary data. Parametric modeling often leads

to inconsistent estimates when the model is mis-specified (Frlich, 2006). Moreover, misspecification in the link function leads to an increase in mean squared error of the estimated probability as well as a substantial bias in estimating the regression parameters and the mean response probability (Czado and Santner, 1992). In an effort to create more flexible binary regression model, Li et al. (2016) used Gaussian process prior on the latent regression function, while they used generalized extreme value (GEV) link functions. However, the end points of GEV distribution depends on parameter values. Smith (2003) showed that it is unlikely to obtain maximum likelihood estimators when the skewness parameter $\xi < -1$ because the log likelihood is J-shaped, which shows there can no consistent maximum likelihood estimator. Moreover, the commonly used link function, logit and probit, are not special case of GEV link. Moreover, Li et al. (2016) did not compare the GEV link function with other flexible link functions like the one proposed in Jiang et al. (2013), and also their model cannot handle ordinal response data. Wang and Dey (2011) introduced a flexible skewed link function for modeling ordinal response data with covariates based on the generalized extreme value (GEV) distribution. However, they assumed the latent regression function as linear.

We would distinguish our work from the previous semiparametric and nonparametric approaches as well as extend our model to handle ordinal response data. Our contribution is to investigate appropriate link function in a GP binary/ordinal regression model. In this Chapter, we will propose a new class of flexible binary link regression models which combines the symmetric power link function proposed by Jiang et al. (2013) with

a Gaussian process prior on the latent structure similar to that employed in Li et al. (2016). This Chapter is organized as follows. In section 3.2, we present the proposed model for binary response data as well as ordinal response data. In section 3.3, we specify the priors on hyperparameters and develop an efficient sampling algorithm for posterior inference, and discuss model selection criteria for comparing our model with competing models. Section 3.4 discusses findings from simulated datasets by comparing the proposed model with various alternative models. In Section 3.5, we will analyze two real data application as motivating examples and conclude in Section 3.6.

3.2 GP-Power link model

In the Section 3.2.1, we will propose our nonparametric model by combining the flexible power link functions with GP prior to achieve double flexibility to handle binary response data. In Section 3.2.2, we will present the results on posterior consistency of the model. In Section 3.2.3, we will extend our model for ordinal response data.

3.2.1 GP-Power Regression Model

Let us denote the observed data as $\mathcal{D} = \{\mathbf{X}, y\}$, where $\mathbf{X}$ is $n \times k$ matrix of covariates and y is $n \times 1$ vector of binary responses. So, y_i takes $\{0,1\}$ values and the index i $(i = 1, \cdots, n)$ refers to observations in the sample. We will follow the work of Albert and Chib (1993) by assuming the binary data outcomes as arising from an underlying latent

variable threshold-crossing framework. In particular, the model is setup by assuming latent random variable h_i depends on covariates $\mathbf{x}_i$ through the model

$$h_i = w(\mathbf{x}_i) + \epsilon_i, \tag{3.2.1}$$

where, w(.) is a latent regression function and $\epsilon_i \sim F$, where F is a cumulative distribution function (cdf) and we assume $E(\epsilon_i \mid w(\mathbf{x}_i)) = 0$. The outcome y_i arises according to

$$\begin{aligned} &y_i = 0 \text{ if } -\infty < h_i < 0, \text{ and} \\ &y_i = 1 \text{ if } 0 < h_i < \infty, \end{aligned} \tag{3.2.2}$$

So, given $w(.)$, $\mathbf{x}_i$,

$$P(y_i = 0) = 1 - P(y_i = 1) = F(-w(\mathbf{x}_i)), \tag{3.2.3}$$

that is, the success probability, p_i, of binary response variable is $1 - F(-w(\mathbf{x}_i))$. So, if ϵ_i are independent $N(0, 1)$, normal distribution with mean 0 and variance 1, p_i will reduce to $p_i = \Phi(w(\mathbf{x}_i))$. The likelihood function of the model can be written as:

$$L(\boldsymbol{\gamma}, \mathbf{w} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \left[\{1 - F(-w(\mathbf{x}_i))\}^{y_i} \times \{F(-w(\mathbf{x}_i))\}^{(1-y_i)} \right]. \tag{3.2.4}$$

A key component of this model is the specification of the link function F. But, the commonly used probit, logit, etc. lack flexibility in skewness. Wang and Dey (2010)

showed that the symmetric link has an inferior performance when the data structure requires skewed response probability function. So, to address this issue Jiang et al. (2013) introduced the symmetric power link family given by:

$$F(x,r) = F_0^r\left(\frac{x}{r}\right)\mathbf{I}_{(0,1]}(r) + \left(1 - F_0^{\frac{1}{r}}\left(-rx\right)\right)\mathbf{I}_{(1,+\infty)}(r), \tag{3.2.5}$$

where $\mathbf{I}_c(x)$ is indicator function taking value 1 if $x \in c$, zero otherwise. The intuition is to utilize the fact that $F_0^r(x)$ is a valid cdf and it achieves flexible left skewness when $r < 1$, while the same property holds for its mirror reflextion $1 - F_0^{\frac{1}{r}}(-x)$ with skewness being in opposite direction. By combining the two, greater flexibility in skewness can be achieved. Moreover, when the skewness parameter $r = 1$, $F(x,r)$ is same as baseline link function $F_0(x)$. So, the baseline link function is a special case of flexible power link family, $F(x,r)$. Jiang et al. (2013) have studied the skewness behavior of symmetric power link function under various choices of F_0 and showed that using logistic cdf for F_0, the resulting logit power link function, $F(.,r)$, can achieve entire range of skewness.

Another key component of this model is the choice of latent regression function. From equation (3.2.3) it is clear that if we assume the latent regression function is linear, that is, $w(\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$, $P(y_i \leq j)$ will only evolve in a monotonic trend. However, in many scenarios, as discussed in Li et al. (2016), such assumption can be inappropriate. To overcome this issue we will model the latent regression function, $w(.)$, nonparametrically by assuming GP prior on $w(.)$.

A random, real-valued function $w(.)$ is said to follow a GP denoted as $GP(\mu(\mathbf{x}), R(.,.))$ with mean function $\mu(\mathbf{x})$ and covariance kernel $R(\mathbf{x}_i, \mathbf{x}_j)$ if given any finite n distinct vectors, $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^k$, $\mathbf{w} = (w(\mathbf{x}_1), \cdots, w(\mathbf{x}_n))'$ follows a multivariate Gaussian distribution with mean vector $(\mu(\mathbf{x}_1), \cdots, \mu(\mathbf{x}_n))'$ and covariance matrix Σ. That is, $(w_1, \cdots, w_n)' \sim N((\mu(\mathbf{x}_1), \cdots, \mu(\mathbf{x}_n))', \Sigma)$, where $w_i = w(\mathbf{x}_i)$, and $\Sigma_{i,j} = R(\mathbf{x}_i, \mathbf{x}_j)$. A common choice of kernel R is the squared exponential kernel,

$$(\Sigma_\theta)_{i,j} = \mathcal{C}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_z^2 \; exp\left[-\sum_{d=1}^{k} \left\{ (x_{i,d} - x_{j,d})^2 / \ell_d^2 \right\} \right] \tag{3.2.6}$$

with a set of hyperparameters $\theta = \{\sigma_z^2, \{\ell_d\}_{d=1}^k\}$. The scaling parameter σ_z^2 controls the vertical scale of variation of the response function and the length-scalse parameters $\ell_1, \cdots, \ell_k$ control the smoothness of sample paths (Rasmussen and Williams, 2006). A GP with covariance structure (1) supports a large class of functions with various shapes. So, we will use GP with covariance structure (1) as a prior over latent regression function, that is, $\mathbf{w} = (w(\mathbf{x}_1), \cdots, w(\mathbf{x}_n))' \sim GP(\mathbf{x}_i\boldsymbol{\beta}, \Sigma_\theta)$.

By combining symmetric power link function given in equation (3.2.5) with GP prior we can achieve double level of flexibility. That is, flexibility both in skewness and nonparametric latent regression function. We propose our GP-Power binary regression

model as:

$$\begin{aligned} y_i &= \mathbf{I}(h_i > 0) \\ h_i &= w(\mathbf{x}_i) + \epsilon_i, \\ \epsilon_i &\sim F(x, r), \\ \mathbf{w} \mid \mathbf{X}, \theta &= (w_i, \cdots, w_n)' \sim GP(\mathbf{x}_i\boldsymbol{\beta}, \Sigma_\theta). \end{aligned} \tag{3.2.7}$$

Here, $\mathbf{I}(h_i > 0)$ is indicator function and takes value 1 if $h_i > 0$ and 0 otherwise. We employ Bayesian computation to obtain inference of the GP-Power model. Each positive GP hyperparameters is given a independent diffused inverse-gamma prior with a large variance. For skewness parameter r of symmetric power link function, a diffuse gamma prior with large variance is placed. We will use normal distribution with large variance as prior on $\boldsymbol{\beta}$. In the next section we will discuss the posterior consistency of our model for binary response data under these priors. In the section 3, we will discuss the details of the priors and algorithm we used to obtain inference.

3.2.2 Posterior consistency

We will follow the proof given in Ghosal and Roy (2006) to prove the posterior consistency of our model with the choice of flexible power link function and kernel given in equation (3.2.5) and (3.2.6), respectively. Ghosal and Roy (2006) proved posterior consistency of Gaussian process prior in binary regression. In their model, they used

fixed and symmetric link function. But our model has uncertainty in the link function because of prior specified on the skewness parameter.

We can rewrite the model (3.2.7) as:

$$Y_i \mid p_i \overset{ind}{\sim} Bin(1, p_i),\ i = 1, \cdots, n \tag{3.2.8}$$

$$p_i \equiv \mathbb{E}(Y_i = 1 \mid \mathbf{x}_i) = 1 - F(-w(\mathbf{x}_i), r) \tag{3.2.9}$$

$$\mathbf{w} \mid \mathbf{X}, \theta = (w_i, \cdots, w_n)' \sim GP(\mathbf{x}_i\boldsymbol{\beta}, \Sigma_\theta). \tag{3.2.10}$$

$F^{-1}(.,.)$, is the link function as given in equation (3.2.5). Let $p_0(x)$ be the true response probability function. The corresponding density is $f_0(x, y) = p_0(x)^y(1-p_0(x))^{1-y}$. Let Π be a prior on $p(x)$ induced by GP prior on $w(x)$ with covariance kernel given in equation (3.2.6). We will prove posterior consistency in the following two cases:

Case 1 : Covariate comes from random design, that is, $x \sim Q$ for some distribution Q. Then we can define the joint density of x and y with respect to the product of Q and counting measure on $\{0, 1\}$, say, $\aleph$, as $f(x, y) = p(x)^y(1 - p(x))^{1-y}Q$. So, the observations $\{x_i, y_i\}$, for $i = 1, \cdots, n$ are i.i.d with distribution $f(x, y)$.

Case 2 : Covariate values arise from a fixed design. Then we can define density of y with respect to counting measure on $\{0, 1\}$, say, $\aleph$, as $f(y \mid x) = p(x)^y(1 - p(x))^{1-y}$. We can see that the observations $\{y_i\}$, for $i = 1, \cdots, n$ are independent and nonidentically distributed with distribution given by $f(y \mid x)$.

We need the following definitions:

Definition 1.(Ghosal et al., 1999) A strong or L_1-neighborhood of f_0 is a set containing a set of the form $U = \{f \in \mathcal{F} : \int |f - f_0| < \epsilon\}$, where, $\mathcal{F}$ is set of all densities.

Definition 2:(Ghosal et al., 1999) Let x be from a random design. Then a prior Π is said to be strongly consistent at f_0, if with P_{f_0}-probability 1,

$$\Pi\left(U_\epsilon \mid (y_1, x_1), \cdots, (y_n, x_n)\right) \to 1 \text{ as } n \to \infty,$$

where, U_ϵ is strong neighborhood of f_0.

Definition 3: For any $f_0 \in \mathcal{F}$, we denote the Kullback-Leibler (K-L) neighborhood $\{f : \int f_0 \log(f_0/f) < \epsilon\}$ by $K_\epsilon(f_0)$. And we say that f_0 is in $K - L$ support of Π if $\Pi(K_\epsilon(f_0)) > 0 \ \forall \epsilon > 0$.

Definition 4:(Ghosal et al., 1999) Let $\mathcal{G} \subset \mathcal{F}$. For $\delta > 0$, the L_1- metric entropy $J(\delta, \mathcal{G})$ is defined as the logarithm of the minimum of all k such that there exist $f_1, f_2, \cdots, f_k$ in $\mathcal{F}$ with the property $\mathcal{G} \subset \cup_{i=1}^{k} \{f : \int |f - f_i| < \delta\}$.

Case 1: We will use the following theorem:

Theorem 1: (Ghosal et al., 1999) Let Π be a prior on $\mathcal{F}$. Suppose $f_0 \in \mathcal{F}$ is in the $K-L$ support of Π and let $U_\epsilon = \{f \in \mathcal{F} : \int |f - f_0| < \epsilon\}$. If there is a $\delta < \epsilon/4$, c_1, $c_2 > 0$, $\beta < \epsilon^2/8$ and $\mathcal{F}_n \subset \mathcal{F}$ such that, for all large n and any fixed ϵ:

1. $\Pi(\mathcal{F}_n^c) < c_1 \exp(-nc_1)$, and

2. $J(\delta, \mathcal{F}_n) < n\beta$,

then $\Pi(U_\epsilon \mid X_1, X_2, \cdots, X_n) \to 1$ a.s. P_{f_0}

Observe that, $\int f_0 \log(f_0/f) d\aleph dQ = \int p_0 \log(p_0/p) dQ + \int (1-p_0)\log((1-p_0)/(1-p)) dQ$. So, to verify $K-L$ condition of Theorem 1 we need to prove $\Pi\{f : \int f_0 \log(f_0/f) d\aleph dQ < \epsilon\} > 0 \ \forall \epsilon > 0$, or equivalently,

$$\Pi\left\{p : \int p_0 \log\frac{p_0}{p} dQ + \int (1-p_0)\log\frac{(1-p_0)}{(1-p)} dQ < \epsilon\right\} > 0 \ \forall \epsilon > 0. \tag{3.2.11}$$

From the lemma 5 in Ghosal and Roy (2006), it can be shown that:

$$\int p_0 \log\frac{p_0}{p} dQ + \int (1-p_0)\log\frac{(1-p_0)}{(1-p)} dQ \leq \sup_{x\in\chi} |p(x) - p_0(x)|^2. \tag{3.2.12}$$

Hence, to prove equation (3.2.11), it suffices to show that $\Pi(p : \sup_{xin\chi}(p(x) - p_0(x)) < \epsilon) > 0$ for every $\epsilon > 0$. $p_0(x) = 1 - F(-w_0(x), r)$ and $F(., r)$ is bounded and differentiable everywhere for any value of r. Hence, it is Lipschitz continuous. So, to prove $K - L$ condition of Theorem 1, it is enough to show that

$$\Pi(w : \sup_{x\in\chi} |w(x) - w_0(x)| < \epsilon) > 0 \ \text{ for every } \epsilon > 0. \tag{3.2.13}$$

Now, for any $p_0(x)$ continuous function, we can easily find a continuous function $w_0(x)$ on the bounded set χ and some parameter r_0 such that $p_0(x) = 1 - F(-w_0(x), r)$.

For example, if we take $r \leq 1$ and F_0 as logistic distribution:

$$p_0(x) = 1 - \left(\frac{1}{1 + exp(w_0(x)/r)} \right)^r \Leftrightarrow w_0(x) = r \log \left(\frac{1}{(1-p_0)^{(1/r)} - 1} \right).$$

So, we can use Theorem 4 in Ghosal and Roy (2006) to show that $\Pi(w : \sup_{x \in \chi} |w(x) - w_0(x)| < \epsilon) > 0$ for every $\epsilon > 0$.

To prove the condition (i) and (ii) in theorem 1, we need to show that there exists sieves $\mathcal{F}_n$ which is a subset of the space of all the $f(x)$ such that, entropy number of $\mathcal{F}_n$, that is, $J(\delta, \mathcal{F}_n)$ is of order $O(n)$ and $\Pi(\mathcal{F}_n^c)$ is exponentially small. We take the $\mathcal{F}_n$ same as that of the model in Ghosal and Roy (2006):

$$\mathcal{F}_n = \{f(x,y) = p(x)^y (1 - p(x))^{(1-y)} : p \in \Theta_n\}, \tag{3.2.14}$$

where $\Theta_n = \{p(x) = H(w(x)) : |D^m w| \leq M_n, m \leq \alpha\}$. $D^m w = \partial^p w / \partial x_1^{i_1} \cdots \partial x_p^{i_p}$ with $i_1 + \cdots + i_p = m$, M_n is some sequence of real numbers depending on n and α is some constant.

Ghosal and Roy (2006) has shown that the Assumption (G) in Ghosal and Roy (2006) will hold by choosing priors on α and $\{\ell_d\}_{d=1}^k$ which have very thin tail and covariance kernel with sufficiently large derivative, that is, α large enough. Note that squared-exponential kernel is smooth so we can take α as large as needed. Hence, by choosing inverse-gamma as prior on σ_z^2 and $\{\ell_i\}_{i=1}^d$, and taking alpha large enough, we can satisfy Assumption (G) in Ghosal and Roy (2006). So, we can take $M_n = bn^{\alpha/d}$ with

sufficiently small constant $b > 0$, we can ensure that $J(\delta, \mathcal{F}_n) < n\beta$. Hence, condition (ii) of Theorem 1 is satisfied.

Finally using lemma 1 in Ghosal and Roy (2006) we can prove condition (i) in Theorem 1.

Case 2: For the fixed design we will follow the approach of Ghosal and Roy (2006). To prove the $K - L$ condition for the prior we can use argument given above and testing condition holds with the test constructed in Equation (5.8) in Ghosal and Roy (2006).

3.2.3 GP-Power ordinal regression

One can view the binary regression problem as a special case of ordinal regression problem. Let us denote the observed data as $\mathcal{D} = \{\mathbf{X}, y\}$, where $\mathbf{X}$ is $n \times k$ matrix of covariates and y is $n \times 1$ vector of ordinal responses. So, y_i takes one of the J categories, j, where $j = 0, \cdots, J-1$, and the index i $(i = 1, \cdots, n)$ refers to observations in the sample.

Again, we will follow the work of Albert and Chib (1993) by assuming the ordinal data outcomes as arising from n independent latent variable, $\mathbf{h} = (h_1, \cdots, h_n)$, such that:

$$y_i = j \text{ if } \gamma_j < h_i < \gamma_{j+1}, \tag{3.2.15}$$

where, $-\infty = \gamma_0 < \gamma_1 = 0, \gamma_2 < \cdots < \gamma_{J-1} < \gamma_J = \infty$ are cutpoint parameters that determine the discretization of the data into J ordered categories. To make sure all

parameters are identifiable we fix $\gamma_1 = 0$. We setup the model is setup by assuming latent random variable h_i depends on covariates $\mathbf{x}_i$ through the model

$$h_i = w(\mathbf{x}_i) + \epsilon_i, \tag{3.2.16}$$

where, w(.) is a latent regression function and $\epsilon_i \sim F(., r)$ where, $F(., r)$ is symmetric power link function as defined in equation (3.2.5) and we assume $E(\epsilon_i \mid w(\mathbf{x}_i)) = 0$. So, the likelihood function of the model can be written as:

$$L(\boldsymbol{\gamma}, \mathbf{w} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \prod_{j=0}^{J-1} \left[F(\gamma_{j+1} - w(\mathbf{x}_i)) - F(\gamma_j - w(\mathbf{x}_i))\right]^{\mathbf{I}(y_i=j)}, \tag{3.2.17}$$

where $\mathbf{I}(y_i = j)$ is the indicator function which takes value 1 if $y_i = j$ and 0 otherwise. And, given $w(.)$, $\mathbf{x}_i$, and $\boldsymbol{\gamma} = (\gamma_2, \cdots, \gamma_{J-1})'$,

$$P(y_i = j) = F(\gamma_{j+1} - w(\mathbf{x}_i)) - F(\gamma_j - w(\mathbf{x}_i)), \text{ and} \tag{3.2.18}$$

$$P(y_i \leq j) = F(\gamma_{j+1} - w(\mathbf{x}_i)). \tag{3.2.19}$$

Here, if we take $w(\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$, $P(y_i \leq j)$ will evolve monotonically for every $j = 0, \cdots, J-1$. This assumption can be inappropriate for many situations. Thus, we

propose our GP-Power ordinal regression model as:

$$\begin{aligned} y_i &= j \text{ if } \gamma_j < h_i < \gamma_{j+1}, \\ h_i &= w(\mathbf{x}_i) + \epsilon_i, \\ \epsilon_i &\sim F(x, r), \\ \mathbf{w} \mid \mathbf{X}, \theta &= (w_i, \cdots, w_n)' \sim GP(\mathbf{x}_i\boldsymbol{\beta}, \Sigma_\theta). \end{aligned} \tag{3.2.20}$$

One can easily see that, when we have only two classes, that is $J = 1$, the model (3.2.20) will reduce to the model (3.2.7).

3.3 Prior Specification and Posterior Inference

In the previous section, we saw that binary regression can be thought of as a special case of ordinal regression. So, in this section we will develop sampling algorithm for GP-power ordinal regression. In the section 3.3.1, we specify priors over parameters θ, r, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ and in section 3.3.2 we will discuss model un-identifiability issue. We will carry out posterior inference in section 3.3.3. We will also discuss the model comparison criterion in section 3.3.4.

3.3.1 Prior Specifications

GP hyperparameters, $\theta = \left\{\sigma_z^2, \{\ell\}_1^d\right\}$ are given inverse-gamma priors. In particular, $\pi(\sigma_z^2) \propto (1/\sigma_z^2)^{\alpha+1} exp(-\beta/\sigma_z^2) \times \mathbf{I}(\sigma_z^2 > 0)$, and $\pi(\ell_i) \propto (1/\ell_i)^{\alpha+1} exp(-\beta/\ell_i) \times \mathbf{I}(\ell_i >$

0), for $i = 1, \cdots, d$. We also assume parameters θ are independent of each other, that is, $\pi(\theta) = \pi(\sigma_z^2) \prod_{i=1}^{d} \pi(\ell_i)$. Prior on r is assumed to be gamma. So, $\pi(r) \propto r^{a-1} exp(-r/b) \times \mathbf{I}(r > 0)$. We will assume a flat prior on $\boldsymbol{\gamma}$ and independent N(0,100) prior on $\boldsymbol{\beta}$. We use $\alpha = 2.01$, $\beta = 1.01$, $a = 0.01$, and $b = 100$. These choices of hyperparameters will result in a diffused priors with mean = 1 and variance = 100.

3.3.2 Model Unidentifiability

Let $\zeta = \{\mathbf{w}, r\}$, which consists of latent regression function $\mathbf{w}$ and skewness parameter r. In the appendix C we have shown that this parameter vector is not identifiable. That is, there exists two sets of parameters ζ and $\tilde{\zeta}$, such that $L(\zeta|data) = L(\tilde{\zeta}|data)$. Xie and Carlin (2006) have shown that model unidentifiability does not imply that there is no Bayesian learning of the parameter. One can resolve this issue in Bayesian model by incorporating prior information. Through various simulation studies we have shown that by placing a weakly informative priors, the posterior distributions concentrate around true parameter values.

3.3.3 Posterior Inference

Consider the likelihood of GP-Power ordinal model:

$$L(\boldsymbol{\gamma}, \mathbf{w} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \prod_{j=0}^{J-1} \left[F(\gamma_{j-1} - w(\mathbf{x}_i)) - F(\gamma_j - w(\mathbf{x}_i)) \right]^{\mathbf{I}(y_i = j)}.$$

Assuming the priors defined in section 3.1, we obtain the joint distribution of $\boldsymbol{\gamma}$, $w(.)$, r, θ as

$$\pi(\boldsymbol{\gamma}, w(.), r, \theta \mid \mathbf{y}, \mathbf{X}) \propto \prod_{i=1}^{n} \prod_{j=0}^{J-1} \left[F(\gamma_{j-1} - w(\mathbf{x}_i)) - F(\gamma_j - w(\mathbf{x}_i))\right]^{\mathbf{I}(y_i=j)} \times N_n(\mathbf{w}; \mathbf{X}\boldsymbol{\beta}, \Sigma_\theta) \times N_d(\boldsymbol{\beta}; \mathbf{0}, 100 \times \mathbf{I}_d) \times \pi(\theta) \times \pi(r). \tag{3.3.1}$$

Recall that, for $i = 1, \cdots, n$, $y_i = j$ if $\gamma_j < h_i < \gamma_{j+1}$ and $h_i = w(\mathbf{x}_i) + \epsilon_i$. So, the posterior conditional distribution of h_i given $w(.)$, $\boldsymbol{\gamma}$, r, $\mathbf{x}_i$ for $i = 1, \cdots, n$ is given by:

$$h_i \mid w(.), \boldsymbol{\gamma}, r, \mathbf{x}_i \propto F((h_i - w(\mathbf{x}_i)), r)\ \mathrm{I}(\gamma_{j-1} < h_i < \gamma_j), \tag{3.3.2}$$

where, $F(., r)$ is flexible power cdf as given in equation (3.2.5), and is easy to sample using inverse cdf method, that is, generate $u \sim uniform(0, 1)$, then, $h_i = F^{-1}(u, r) + w(\mathbf{x}_i)$. Also, for $j = 2, \cdots, J-1$, the posterior conditional distribution of γ_j has cloesd form and is given by

$$\gamma_j \mid \gamma_{k \neq j}, \mathbf{h}, \mathbf{y} \sim \text{Uniform}\left[\max\left\{\max(h_i : y_i = j), \gamma_{j-1}\right\}, \min\left\{\min(h_i : y_i = j+1), \gamma_{j+1}\right\}\right]. \tag{3.3.3}$$

Sampling of $w(.)$, r, and θ using "standard" Metropolis-Hastings (MH) algorithm are not only computationally expensive but also converge slowly and mix poorly. One reason for such poor performance is that it is difficult to design proposals for this high

dimensional $\mathbf{w}$ that lead to reasonable acceptance rate. Efficient sampling algorithm is available when GP ordinal regression uses probit link function. Therefore, for sampling using the flexible power link function we will extend the surrogate slice sampling (SSLS) algorithm developed by Murray and Adams (2010). To our best understanding, our sampling method is the first successful attempt at sampling a GP ordinal regression model with a flexible power link function.

Surrogate data model. Following the definition of surrogate data in Murray and Adams (2010), we define the surrogate variables $\mathbf{g} = (g_1, \cdots, g_n)'$ as $g_i = w_i + z_i$, that is, a noisy version of gaussian process. Here, z_i's are normally distributed errors. So, conditional distribution is $\mathbf{g}|\mathbf{w}, \theta \sim N(\mathbf{g}; \mathbf{w}, S_\theta)$,where S_θ is the diagonal noise covariance matrix which is often chosen to be $S_\theta = \mathbf{c}\mathbf{I}_n$ and the vector $\mathbf{c}$ can be set by hand to a fixed value or can also be obtained by individually matching variance from a Gaussian fit obtained from Laplace approximation to the posterior of each latent variable. Details about fixing the value of $\mathbf{c}$ is exemplified in Appendix B using few of the baseline link functions.

Now, integrating out the $\mathbf{w}$ using GP prior, the marginal distribution of $\mathbf{g}$ is given by, $P(\mathbf{g} \mid \theta, \boldsymbol{\beta}) = N(\mathbf{g}; \mathbf{X}\boldsymbol{\beta}, \Sigma_\theta + S_\theta)$. So, the posterior distribution of $\mathbf{w}$ conditional on $\mathbf{g}$ and θ is then given by $\mathbf{w} \mid \mathbf{g}, \theta \sim N(\mathbf{w}; \mathbf{m}_{\theta,g,\beta}, R_\theta)$, where $R_\theta = S_\theta - S_\theta(S_\theta + \Sigma_\theta)^{-1}S_\theta$, and $\mathbf{m}_{\theta,g,\beta} = R_\theta(S_\theta^{-1}\mathbf{g} + \Sigma_\theta^{-1}\mathbf{X}\boldsymbol{\beta})$. So, sampling of latent function $\mathbf{w}$ can be done using the surrogate data $\mathbf{g}$ and GP hyperparameters θ.

One issue with this auxiliary model is that $\mathbf{w}$ are highly informative about the hyper-parameters, and it significantly limits the ability to update θ and r with fixed $\mathbf{w}$ for any Markov chain. In view of that, Murray and Adams (2010) proposed to further reparameterize this auxiliary model to deal with sampling from strongly coupled variables. A Cholesky decomposition of R_θ gives $R_\theta = L_\theta L'_\theta$, where L_θ is the lower triangular matrix. During sampling, a draw is first obtained from a multivariate Gaussian $\boldsymbol{\eta} \sim N(\boldsymbol{\eta}; 0, \mathbf{I}_n)$ and then we can calculate the latent variables $\mathbf{w}$ by

$$\mathbf{w} = L_\theta \boldsymbol{\eta} + m_{\theta,g,\beta}. \tag{3.3.4}$$

Such a reparameterization helps in sampling of θ using a fixed $\boldsymbol{\eta} = L_\theta^{-1}(\mathbf{w} - m_{\theta,g,\beta})$, instead of fixed $\mathbf{w}$. Note that $\boldsymbol{\eta}$ is independent of θ. So, the posterior conditional distribution of θ, r, and $\boldsymbol{\beta}$ given $\mathbf{g}$, $\boldsymbol{\eta}$ is:

$$\pi(\theta, r, \boldsymbol{\beta}|\mathbf{g}, \boldsymbol{\eta}, \mathbf{X}, \mathbf{y}) \propto \prod_{i=1}^{n} \mathcal{L}(\theta, \eta, r, \mathbf{g}) \times N_n(\mathbf{g}; \mathbf{X}\boldsymbol{\beta}, \Sigma_\theta + S_\theta) \times \pi(\theta) \times \pi(r) \times \pi(\boldsymbol{\beta}). \tag{3.3.5}$$

Here the likelihood function $\mathcal{L}(.)$ is obtained from using $\mathbf{g}$ and $\boldsymbol{\eta}$ instead of $\mathbf{w}$ using equation (3.3.4). Now, based on the surogate data model proposed by Murray and Adams (2010) and the flexible power link function we will propose a slice sampling algorithm, SSLS-power, to jointly updata all the parameters of interest. This is a very robust algorithm and has a free parameter σ which is not required to be carefully tuned.

The SSLS-power algorithm is described in Appendix A.

Prediction. For the prediction of new point $\tilde{y}$ at test point $\tilde{\mathbf{x}}$, we first need to estimate the $w(\mathbf{x})$. The conditional on $w(\mathbf{x})$, the estimate of latent regression function at test point $\tilde{\mathbf{x}}$ can be obtained after integrating out prior on $\boldsymbol{\beta}$ using $N(\mathbf{b}, B)$. So,

$$w(\tilde{\mathbf{x}} \mid \mathbf{X}), \theta \sim N(\mathbf{w}_*, cov(\mathbf{w}_*)),$$

where, $\mathbf{w}_* = \tilde{\mathbf{x}}\tilde{\boldsymbol{\beta}} + \Sigma_{\theta,21}\Sigma_{\theta,11}^{-1}(w(\mathbf{X}) - \mathbf{X}\tilde{\boldsymbol{\beta}})$ and $cov(\mathbf{w}_*) = \Sigma_* + R'(B^{-1} + \mathbf{X}'\Sigma_{\theta,11}^{-1}\mathbf{X})^{-1}R$, $\tilde{\boldsymbol{\beta}} = (B^{-1} + \mathbf{X}'\Sigma_{\theta,11}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\Sigma_{\theta,11}^{-1}w(\mathbf{X}) + B^{-1}\mathbf{b})$, $R = \Sigma_{\theta,22} - \Sigma_{\theta,21}\Sigma_{\theta,11}^{-1}\Sigma_{\theta,12}$, $\Sigma_{\theta,11} = var(w(\mathbf{X}))$, $\Sigma_{\theta,12} = \Sigma'_{\theta,21} = cov(w(\mathbf{X}), w(\tilde{\mathbf{x}}))$, and $\Sigma_{\theta,22} = var(w(\tilde{\mathbf{x}}))$.

Let us denote N as the number of MCMC chains remaining after burn-in and thinning and M as the number of times we make prediction for each MCMC sample. So, the predictive probability for each category j and for each MCMC sample (i) can be obtained as:

$$\hat{p}_{(i),j} = P_{(i)}(\tilde{y} = j|\mathbf{y}, \mathbf{X}) = \sum_{k=1}^{M} \left[F(\gamma_{j+1}^{(i)} - w^{(k)}(\tilde{\mathbf{x}}|\mathbf{X}, \theta^{(i)})) - F(\gamma_j - w^{(k)}(\tilde{\mathbf{x}}|\mathbf{X}, \theta^{(i)})) \right] /M.$$

A rule of thumb can be applied to obtain the prediction of $\tilde{y}$ based on the estimated predictive probability. For example, if $\{\bar{p}_0, \cdots, \bar{p}_{J-1}\}$ are estimated as mean of predictive probability for each category, then $y_i = j$ if $\bar{p}_j$ is the maximum predictive probability among all categories.

3.3.4 Model comparison criterion

We use deviance information criteria (DIC) as defined in Spiegelhalter et al. (2002). Let Θ denote the set of all parameters contained in the model under consideration. The deviance is defined as $D(\mathbf{y}, \Theta) = -2 \log (p(\mathbf{y} \mid \Theta))$, that is, -2 times log likelihood. Here, $p(\mathbf{y} \mid \Theta)$ is same as the likelihood defined in equation (3.2.17). Then DIC $= 2\hat{D}_{avg}(\mathbf{y}) - D_{\hat{\Theta}}(\mathbf{y})$, where $\hat{D}_{avg}(\mathbf{y}) = \sum_{i=1}^{S} \{D(\mathbf{y}, \Theta_s)\} / S$, θ_s is the s^{th} sampling value of Θ, and $D_{\hat{\Theta}}(\mathbf{y}) = D(\mathbf{y}, \hat{\Theta})$, where, $\hat{\Theta}$ is the mean of the samples from posterior distribution.

We also include prediction error (PE) as a measure for model selection criterion. In a sample of size $i = 1, \cdots, n$, if $\{p_{0,i}, \cdots, p_{J-1,i}\}$ are estimated predictive probabilities for each category and define $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ such that $\mathbf{y}_i$ is a sparse J dimensional vector with j^{th} element as 1 if the response $y_i = j$. Then $PE = \left(\sum_{i=1}^{n} \sum_{j=0}^{J-1} (y_{j,i} - p_{j,i})^2 \right) / (nJ)$.

3.4 Simulations

In this section we perform various simulation studies to validate and test the performance of our proposed model for ordinal regression under various choice of baseline link function, F_0. We compare the proposed model with logit and probit as baseline link functions. We also compare the proposed model with GEV link for ordinal data. The

GEV link is defined as:

$$F(w_i,\xi) = 1 - GEV(-w_i,\xi) = \begin{cases} 1 - exp\left\{-(1-\xi w_i)_+^{-1/\xi}\right\}, & \xi \neq 0 \\ 1 - exp\left\{-exp(w_i)\right\}, & \xi = 0, \end{cases} \tag{3.4.1}$$

where $GEV(x;\xi)$ represents the cumulative probability at x for the GEV distribution with parameters $\phi = (\mu = 0, \sigma = 1, \xi)$. Datasets are generated using two values of skewness parameter of the proposed model with logit and probit as baseline link functions and also GEV link function with GP as prior on latent regression function. We generated data for two types of response variable: binary (J=2) and ordinal (J=6).

The data is generated by first taking the covariate generated from independent normal distribution with mean 0 and variance 10, $x_i \sim N(0, 10)$, $i = 1, \cdots, n$. The sample size n of the dataset is taken as 100. The latent regression function $\mathbf{g}$ is generated from multivariate Gaussian distribution with 0 mean vector and squared exponential covariance matrix (from equation 3.2.6) with parameters $\ell = 1$ and $\sigma_z^2 = 1$. Now, ϵ_i is generated using three different flexible link functions as discussed before. Two values of skewness parameter are considered, $r = 2$ and 0.5, and $\xi = -5$ and 0.5. The response y_i is obtained by first calculating h_i by $h_i = w_i + \epsilon_i$, setting the cutoff points $\boldsymbol{\gamma}$, and taking $y_i = j$ if $\gamma_j < h_i \leq \gamma_{j+1}$. We fit all the three models in these 12 types of scenario. This experiment is repeated 20 times. The result of the simulation study is presented in Table 6 and Table 7. There are 100,000 MCMC samplings but we only used 4,000 iterations, obtained from every 20^{th} iteration, to compute all quantities of interest, using

a burn-in of 20,000 iterations.

3.5 Application

In this section we present two real data applications to illustrate the flexibility of our model. In Section 5.1, we use a pre-clinical study data to demonstrate the flexibility of the model in binary response data. In Section 5.2, we will present a ordinal response data application for the proposed model.

3.5.1 Experiment on Attention Paradigm

This example studies the attention paradigm experiment performed on a monkey. Details of this experiment is described in Smith et al. (2009). In this experiment the goal is to determine if deep brain stimulation (DBS) allows the monkey to recover pre-fatigue level of performance once the performance has decreased as a result of spontaneous fatigue. The monkey performed 1250 trials. In each trial, the monkey had to perform a task and the monkey is rewarded if successful in completion of task. Stimulation was applied during four periods across tirals 300-364, 498-598,700-799 and 1000-1099. A total of 741 out of 1250 are 1's.

We fit our proposed model under logit link as baseline link function. We also fit GP-GEV model and GP-logit model. Table 8 shows the model comparison results for different link functions. This shows that GP-power logit is a better fit in comparison to

Binary response (J=2)			
	Fitted		
Generated	GP-power logit	GP-power probit	GP-GEV
GP-power logit($r = 0.5$)	DIC = 138.2299 PE=0.0011 Mean ℓ=1.1129 Mean σ_z^2=0.9088 Mean r=0.4652	DIC = 140.4174 PE=0.0019 Mean ℓ=0.8001 Mean σ_z^2= 0.8223 Mean r=0.6119	DIC =139.1190 PE=0.0015 Mean ℓ=1.2368 Mean σ_z^2= 1.3211 Mean ξ=-0.3568
GP-power logit($r = 2$)	DIC = 142.6225 PE=0.0015 Mean ℓ=0.8999 Mean σ_z^2=0.9266 Mean r=1.8901	DIC = 147.7641 PE=0.0021 Mean ℓ=0.8721 Mean σ_z^2= 1.1295 Mean r=1.5300	DIC =144.2000 PE=0.0020 Mean ℓ=1.1249 Mean σ_z^2= 1.2655 Mean ξ=0.4110
GP-power probit($r = 0.5$)	DIC = 151.3347 PE=0.0016 Mean ℓ=1.1086 Mean σ_z^2=1.5257 Mean r=0.3277	DIC = 149.0332 PE=0.0015 Mean ℓ=0.9151 Mean σ_z^2= 1.2332 Mean r=0.5912	DIC =152.2220 PE=0.0016 Mean ℓ=1.1100 Mean σ_z^2= 1.4217 Mean ξ=-0.4900
GP-power probit($r = 2$)	DIC = 145.0001 PE=0.0020 Mean ℓ=1.4001 Mean σ_z^2=1.3965 Mean r=2.2526	DIC = 141.0222 PE=0.0015 Mean ℓ=0.8766 Mean σ_z^2= 0.8014 Mean r=2.3301	DIC =143.9091 PE=0.0018 Mean ℓ=1.5550 Mean σ_z^2= 1.4208 Mean ξ= 0.5238
GP-GEV($\xi = 0.5$)	DIC = 140.0654 PE=0.0015 Mean ℓ=1.9568 Mean σ_z^2=1.6733 Mean r=0.3374	DIC = 143.1199 PE=0.0016 Mean ℓ=1.8233 Mean σ_z^2= 1.9911 Mean r=0.7198	DIC = 139.8851 PE=0.0013 Mean ℓ=1.5995 Mean σ_z^2= 1.5901 Mean ξ=-0.6377
GP-GEV($\xi = -0.5$)	DIC = 148.3747 PE=0.0101 Mean ℓ=1.8732 Mean σ_z^2=1.8988 Mean r=1.7100	DIC = 147.6771 PE=0.0111 Mean ℓ=1.5666 Mean σ_z^2= 2.0120 Mean r=1.8988	DIC =147.9000 PE=0.0089 Mean ℓ=1.3221 Mean σ_z^2= 1.5777 Mean ξ=0.4041

Table 6: Comparison between GP-Splogit, GP-Spprobit, and GP-GEV model for Binary response

Ordinal response (J=6)			
	Fitted		
Generated	GP-power logit	GP-power probit	GP-GEV
GP-power logit($r = 0.5$)	DIC = 327.1567 PE=0.1033 Mean ℓ=1.1567 Mean σ_z^2=1.1324 Mean r=0.6519	DIC = 383.9001 PE=0.1191 Mean ℓ=1.8511 Mean σ_z^2= 1.1229 Mean r=0.2210	DIC =367.8811 PE=0.1200 Mean ℓ=1.8811 Mean σ_z^2= 1.4501 Mean ξ=-0.1006
GP-power logit($r = 2$)	DIC = 433.6889 PE=0.1131 Mean ℓ=0.8788 Mean σ_z^2=1.8755 Mean r=2.3899	DIC = 493.3389 PE=0.1677 Mean ℓ=1.7662 Mean σ_z^2= 1.9991 Mean r=3.1011	DIC =400.1567 PE=0.1990 Mean ℓ=1.8977 Mean σ_z^2= 1.8891 Mean ξ=0.3300
GP-power probit($r = 0.5$)	DIC = 461.2100 PE=0.1908 Mean ℓ=0.7666 Mean σ_z^2=1.9013 Mean r=0.3900	DIC = 443.2765 PE=0.1287 Mean ℓ=0.8912 Mean σ_z^2= 1.3133 Mean r=0.6122	DIC =511.0911 PE=0.1799 Mean ℓ=1.4311 Mean σ_z^2= 1.7566 Mean ξ=-0.5112
GP-power probit($r = 2$)	DIC = 512.0551 PE=0.1865 Mean ℓ=1.8013 Mean σ_z^2=0.8977 Mean r=2.5781	DIC = 500.9876 PE=0.1298 Mean ℓ=1.0955 Mean σ_z^2= 1.1244 Mean r=1.8915	DIC =591.0914 PE=0.1900 Mean ℓ=1.4331 Mean σ_z^2= 1.1988 Mean ξ= 0.3499
GP-GEV($\xi = 0.5$)	DIC = 410.7641 PE=0.1498 Mean ℓ=1.6751 Mean σ_z^2=2.0199 Mean r=0.2981	DIC = 421.3471 PE=0.1822 Mean ℓ=1.9100 Mean σ_z^2= 1.8842 Mean r=0.6911	DIC = 389.1001 PE=0.1643 Mean ℓ=1.6744 Mean σ_z^2= 1.3542 Mean ξ=-0.4233
GP-GEV($\xi = -0.5$)	DIC = 481.2260 PE=0.1814 Mean ℓ=1.8732 Mean σ_z^2=0.8821 Mean r=3.0098	DIC = 471.0941 PE=0.1873 Mean ℓ=1.5666 Mean σ_z^2= 1.6723 Mean r=2.9842	DIC =412.8699 PE=0.1709 Mean ℓ=1.3221 Mean σ_z^2= 1.4301 Mean ξ=0.3982

Table 7: Comparison between GP-Splogit, GP-Spprobit, and GP-GEV model for Ordinal response

GP-GEV. Estimates of skewness parameters of power-logit link function and GEV link function suggest that it is left skewed.

Figure 3 shows the comparison between GP models under the above choices of link functions. The shaded area denotes the duration of trials when the DBS was switched ON. The observed probability estimates denoted by solid dots are empirically obtained by calculating percentage of 1's in 40 consecutive trials. In all the GP models the estimates follow a similar pattern. At the beginning the probability estimates was close to 0.7 and decreases significantly after the third ON. In the fourth ON period the performance increased significantly.

GP-power logit	GP-logit	GP-GEV
DIC = 1595.2366	DIC = 1601.4744	DIC = 1597.2391
PE=0.1391	PE=0.1622	PE=0.1411
r=0.5611	-	ξ=-0.7988

Table 8: Model comparison for attention paradigm example.

3.5.2 Patient Satisfaction Data Application

To illustrate flexibility of our proposed model, we use the patient satisfaction dataset obtained from MINITAB. The objective of this study is to know how each factors influence patient satisfaction. Relevant predictors include age and proximity to office. Both the predictors are continuous variables. "How likely a patient is to return" is used as a response variable. The categories in the response variable have a natural order: unlikely, somewhat likely, very likely. So, the response variable is ordinal. The dataset consists

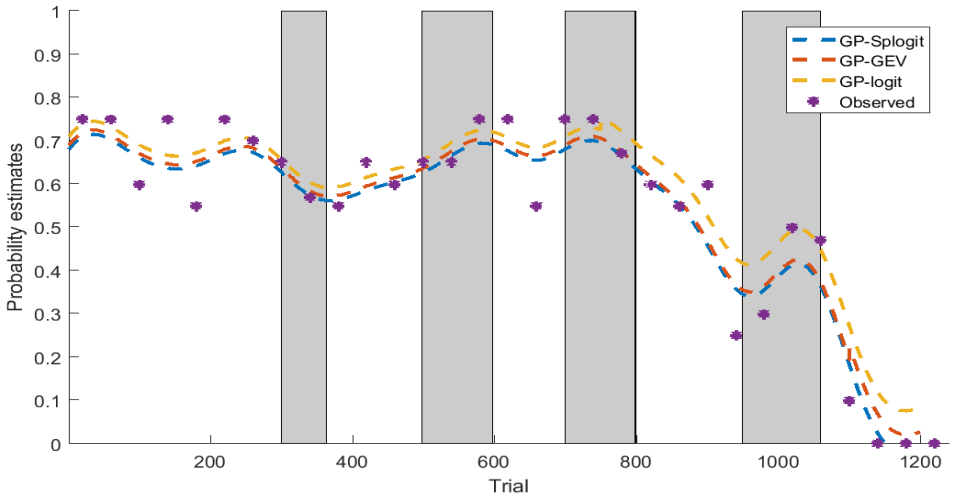

Figure 3: Probability estimates for the experiment on attention paradigm example. Posterior mean of predictive probabilities of GP model under power-logit, power-probit and GEV as link functions. The shaded grey area denotes the duration of trial when the DBS was switched ON.

of 73 observations. We encode $y_i = 0$ if "unlikely", $= 1$ if "somewhat likely", and $= 2$ if "very likely".

We fit GP-power model with logit as baseline link function, and also fit GP-GEV and GP-logit model. The results are presented in Table 9. The model selection criterion shows that GP-power logit and GP-GEV model are better than GP-logit. The probability estimates are closer to the observed values in GP-power and GP-GEV model in comparison to GP-logit model. This shows that flexible link functions like GEV and flexible power are able to adjust to the data in a better way by choosing a suitable skewness parameter in the model.

Variables	GP-power logit	GP-logit	GP-GEV	
DIC	2615.2411	2929.3987	2641.1509	
PE	0.1812	0.1922	0.1814	
Skewness	r=2.0510	-	ξ=0.2521	
				Observed
P(Y=0)	0.1490	0.1485	0.1520	0.1507
P(Y=1)	0.2701	0.2689	0.2712	0.2603
P(Y=2)	0.5869	0.5991	0.5892	0.5890

Table 9: Model comparison for patient satisfaction data.

3.6 Discussion

In this Chapter we propose a family of flexible nonparametric binary and ordinal regression models. The flexibility in the links is important to avoid link misspecification. Usually, a skewed link could be more appropriate when there exists an extremely uneven distribution of observations across different categories. However, the determination of link function is much more complicated than simply just counting the observations in each categories. Further, the incorrect decision on choice of link function could have much severe consequences. Thus, in such a situation it is important to study the link functions which can handle skewness in both the directions. Both flexible power link and GEV link models can handle all directions of skewness.

Chapter 4

Variable Selection Using Gaussian Process Prior

4.1 Introduction

Linear inverse problem occurs frequently, whenever we need to infer unobserved features of interest from the quantities that we can measure. Typically, a linear model describing the relationship between the features and measure quantities is of the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1.1}$$

where, $\mathbf{y} \in \Re^n$ is a measurement vector, $\mathbf{X} \in \Re^{n\times k}$ is a measurement matrix, $\boldsymbol{\beta} \in \Re^k$ is the vector of features, and $\boldsymbol{\epsilon} \in \Re^n$ is a vector of noise. Estimation of feature $\boldsymbol{\beta}$ can be a challenging task when the dimension k is comparable to or exceeds n. In such situations, the ordinary least square (OLS) estimator is not well behaved and no longer unique due to the singularity of the design matrix. Problems of this type are very common in signal processing, genetic research, neuroscience, and machine learning. For example, a genome

wide association study looks at millions of single nucleotide polymorphisms (SNPs) to identify several relevant genes to a certain characteristic.

A typical solution for this problem is sparse modeling. The main assumption is that the k-dimensional vector $\boldsymbol{\beta}$ is sparse with many components being exactly zero thereby eliminating irrelevant predictors from the models. This idea has been used as a motivation for various models (Tibshirani, 2011; Zou, 2006; Fan and Li, 2001) that performs estimation as well as variable selection simultaneously, by proposing a penalized loss function to estimate $\boldsymbol{\beta}$ as follows:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} [L\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}\} + \text{Pe}(\boldsymbol{\beta}, \lambda)],$$

where, $L(.)$ is a loss function and Pe(.) is a penalty function with a tuning parameter λ. Even though the penalized likelihood method produces nice estimates of regression coefficients, one major issue is difficulty in obtaining standard errors of the estimator for small sample size (Kyung et al., 2010). To overcome this issue, various Bayesian methods have been developed in sparse high-dimensional problem by treating $\boldsymbol{\beta}$ as a random variable and thus the uncertainty can be explained using its posterior distribution. In the Bayesian framework, sparsity can be induced by enforcing sparsity promoting priors on the $\boldsymbol{\beta}$. For example, Park and Casella (2008) showed that LASSO estimate is equivalent to posterior mode obtained by placing independent Laplace priors on β_i's. Mitchell and Beauchamp (1988), George and McCulloch (1993) and Chipman (1996) proposed

Bernoulli-Gaussian prior (spike and slab prior) for variable selection. This involves designing a mixture of two priors, one that is very peaky and another that is very broad. Mathematically, one can define spike and slab prior on β_i in following way:

$$\pi(\beta_i|z_i,) = (1 - z_i)\delta(\beta_i) + z_i N(\beta_i; 0, \sigma^2), \tag{4.1.2}$$

which is a mixture of a Dirac Delta function, $\delta(.)$, and a Gaussian distribution. z_i controls how likely β_i is nonzero, and therefore, it takes the role of a complexity parameter controlling the size of the model. Typical default choice of prior on z_i include the i.i.d. Bernoulli prior with success probability p, with a uniform prior on p.

One of the drawbacks of these methods is their inability to take into account any prior knowledge of the structure of sparsity pattern. In many applications, for example, in compressive sensing (Huang et al., 2009) the sparse regression coefficients need not be randomly distributed but have certain pattern.

A few methods (Simon et al., 2013; Jacob et al., 2009) have been proposed to model group sparsity by partitioning the set of covariates beforehand. In Bayesian framework, group sparsity is achieved by placing group spike and slab prior (Hernández-Lobato et al., 2013). Let $\mathcal{G}$ be a partition of set of variables into G groups, then the form of

group spike and slab prior can be written as:

$$\pi(\boldsymbol{\beta}|\mathbf{z}) = \prod_{g=1}^{G}[(1-z_g)\delta(\boldsymbol{\beta}_g) + z_g N(\boldsymbol{\beta}_g; 0, \sigma^2\mathbf{I}_g)]$$
$$\pi(\mathbf{z}) = \prod_{g=1}^{G} \text{Bernoulli } (p_g).$$

Here, z_g indicates whether the variables in a group are active or not. Andersen et al. (2014b) introduced another approach to encode prior belief of sparsity structure by using a structured spike and slab prior and inducing a structured sparsity by using Gaussian Process (GP) prior on spike and slab probabilities. However, in their model they have failed to incorporate prior information from covariates. They have also used expectation propagation approach for estimation and have failed to estimate GP hyperparameters. Their method assumes that prior knowledge on GP hyperparameters are known. However, GP hyperparameters are very hard to interpret in practice. Therefore, we have developed a novel approach by using covariates in GP kernel to encode prior sparsity pattern via spike and slab prior. Moreover, we performed Bayesian inference using Markov chain Monte Carlo (MCMC) methods.

This Chapter is organized as follows. In Section 4.2, we describe our model, and in Section 4.3 we discuss the Bayesian inference scheme based on MCMC for the proposed model. In Section 4.4, we perform simulations to validate our model. In Section 4.5, we discuss the extensions and future works of our model.

4.2 The Proposed Method

Consider the linear regression:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.2.1}$$

where, $\mathbf{y} \in \Re^n$ is the response vector, $\mathbf{X} \in \Re^{n\times k}$ is the design matrix and the columns of $\mathbf{X}$ have scaled to unit ℓ_2-norm. $\boldsymbol{\beta} \in \Re^k$ is the vector of coefficients, and $\boldsymbol{\epsilon} \in \Re^n$ is the error vector. Also consider $n < k$. In a Bayesian approach, model specification is completed by specifying priors on parameters $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. Prior on ϵ_i is assumed to be iid normal with mean 0 and variance σ^2_ϵ. We also assume the following form of spike and slab prior on $\boldsymbol{\beta}$

$$\begin{aligned} \pi(\beta_i \mid z_i) &= (1 - z_i)\delta(\beta_i) + z_i N(\beta_i \mid 0, \sigma^2_\epsilon/\tau^2), \\ z_i &\sim \text{Bernoulli}(p_i), \\ F^{-1}(p_i) &= g(\mathbf{X}_{[,i]}) \end{aligned} \tag{4.2.2}$$

for $i = 1, \cdots, k$. Here, F^{-1} is a link function and g(.) is a function which can be used to encode information covariates. $\mathbf{X}_{[,i]}$ represents i^{th} column vector of design matrix $\mathbf{X}$. We assume that prior on $\beta_i \mid z_i$ is independent for $i = 1, \cdots, n$. So, $\pi(\boldsymbol{\beta}|\mathbf{z}) = \prod_{i=1}^n \pi(\beta_i|z_i)$, where $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_k]'$ and $\mathbf{z} = [z_1, \cdots, z_n]'$. We will also assume prior on β_0 is $N(\beta_0 \mid 0,$ where $\sigma^2_\epsilon/\tau^2)$. σ^2_ϵ/τ^2 is the variance parameter and F is any inverse link function. So, $F(.)$ can be normal cumulative distribution function, $\Phi(.)$.

In some situations it is desirable that the two regression coefficients to have similar

values if their corresponding covariates are close to each other. We can encode such informations via the covariance kernel of Gaussian process (GP) prior on the function $g(.)$. So, $\mathbf{g} = \left(g(\mathbf{X}_{[,1]}), \cdots, g(\mathbf{X}_{[,k]})\right)' \sim GP\left(\boldsymbol{\mu}, \Sigma\right)$ where, Σ is $k \times k$ matrix with

$$(\Sigma)_{ij} = \sigma_z^2 \psi\left(\frac{||\mathbf{X}_{[,i]} - \mathbf{X}_{[,j]}||}{\ell}\right) \tag{4.2.3}$$

and $\psi(.)$ is some isotropic correlation function, and $\boldsymbol{\mu}$ is mean function. For example, ψ can be the squared exponential kernel. Then,

$$(\Sigma)_{i,j} = \sigma_z^2 exp\left(-\frac{1}{2}\left(\mathbf{X}_{[,i]} - \mathbf{X}_{[,j]}\right)' M \left(\mathbf{X}_{[,i]} - \mathbf{X}_{[,j]}\right)\right), \tag{4.2.4}$$

where $M = \ell^{-2}\mathbf{I}_k$ is a symmetric matrix. So, if the i^{th} and j^{th} column are close to each other distance wise, that is, $\left(\mathbf{X}_{[,i]} - \mathbf{X}_{[,j]}\right)' M \left(\mathbf{X}_{[,i]} - \mathbf{X}_{[,j]}\right)$ is small then $g_i(= g(\mathbf{X}_{[,i]}))$ and $g_j(= g(\mathbf{X}_{[,j]}))$ will be close to each other and hence, the values of β_i and β_j will be more likely to be close to each other. Hence, ℓ here controls the sparsity patterns.

Using the probit link function, that is, $F(.) = \Phi(.)$, the marginal prior probability of i^{th} regression coefficient being nonzero is :

$$p(z_i = 1) = \int p(z_i = 1 \mid g_i) p(g_i \mid \mu_i, \Sigma_{ii})\, dg_i = \Phi\left(\frac{\mu_i}{\sqrt{1 + \sigma_z^2}}\right). \tag{4.2.5}$$

The derivation of above result is given in Chapter 3.9 of Rasmussen and Williams (2006). If we have $\mu_i = 0$, then, probability that $z_i = 1$ is 0.5 for all $i = 1, \cdots, k$. Now if we

assume $\boldsymbol{\mu} = \mu\mathbf{1}$, where, μ is a scalar, then $p(z_i = 1)$ is a function of μ and σ_z^2 and it will be same for all $i = 1, \cdots, k$. So, the parameter vector $\{\mu, \sigma_z^2\}$ is unidentifiable. That is one can find two distinct values of vector $\{\mu, \sigma_z^2\}$ such that we have same sparsity level as well as same sparsity pattern. To avoid this model unidentifiability for the inference of parameters, we will fix σ_z^2 to a constant value.

4.3 Bayesian inference

In this section, we will design an efficient sampling method for the inference of parameters of interest. Consider the likelihood of regression model defined in equation (4.2.1).

$$\mathcal{L}(\boldsymbol{\beta}, \sigma_\epsilon^2) \propto \left(\frac{1}{\sigma_\epsilon^2}\right)^{n/2} exp\left\{-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma_\epsilon^2}\right\} \tag{4.3.1}$$

Here, we will use the probit link function, $F(.) = \Phi(.)$, and it can be easily generalized to other link functions. The joint posterior distribution of interest becomes:

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma_\epsilon^2, \mathbf{z}, \mathbf{g}, \ell, \boldsymbol{\mu} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\mu}) \propto & \left(\frac{1}{\sigma_\epsilon^2}\right)^{n/2} exp\left\{-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma_\epsilon^2}\right\} \times \pi(\sigma_\epsilon^2) \\ & \times \prod_{\imath=1}^{k} \left[(1 - z_i)\delta(\beta_i) + z_i \mathcal{N}(\beta_i | 0, \sigma_\epsilon^2/\tau^2)\right] \times N(\beta_0 | 0, \sigma_\epsilon^2/\tau^2) \\ & \times \prod_{i=1}^{k} \left[\Phi(g(\mathbf{X}_{[,i]}))^{z_i}(1 - \Phi(g(\mathbf{X}_{[,i]})))^{1-z_i}\right] \\ & \times |\Sigma|^{-1/2} exp\left\{(\mathbf{g} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{g} - \boldsymbol{\mu})\right\} \times \pi(\ell) \times \pi(\boldsymbol{\mu}), \end{aligned} \tag{4.3.2}$$

where, $\pi(\sigma_\epsilon^2)$, $\pi(\ell)$, and $\pi(\boldsymbol{\mu})$ are priors on σ_ϵ^2, ℓ, and $\boldsymbol{\mu}$, respectively. The true posterior distribution of the parameters of interest is not analytically tractable. So, for posterior inference we have to implement an efficient sampling algorithm.

Recall that our parameters of interest are $\Theta = \{\boldsymbol{\beta}, \sigma_\epsilon^2, \mathbf{z}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \ell\}$. Note that carrying out Gibbs sampling by alternatively sampling from these parameters will not create an irreducible Markov chain. This is because when z_i is zero, β_i is also zero. Thus, this leads to a Markov chain with absorbing states. We can overcome this issue by first integrating out the $\boldsymbol{\beta}$ from the likelihood and then sampling $\mathbf{z}$ from the integrated likelihood. We take prior on σ_ϵ^2 as inverge-gamma with shape and rate parameter as a_0 and b_0, respectively. Let s be the number of covariates corresponding to nonzero β_i's. Then, the posterior conditional distribution of $\mathbf{z}$ is:

$$
\begin{aligned}
p(\mathbf{z} \mid \mathbf{y}, \tau^2, \gamma) =& \Bigg[\int_{\sigma_\epsilon^2} \int_{\boldsymbol{\beta}} N(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma_\epsilon^2 \mathbf{I}_n) \times f(\boldsymbol{\beta} \mid \mathbf{z}, \sigma_\epsilon^2, \tau^2) \times \pi(\sigma_\epsilon^2)\ d\boldsymbol{\beta}\ d\sigma_\epsilon^2 \Bigg] \\
& \times \prod_{i=1}^{k} \left[\Phi(g(\mathbf{X}_{[,i]}))^{z_i} (1 - \Phi(g(\mathbf{X}_{[,i]})))^{1-z_i} \right] \\
=& \frac{1}{(2\pi)^{n/2}} \frac{|\mathbf{B}_n^\delta| (\tau^2)^{(s+1)/2} \Gamma(d_n) b_0^{a_0}}{\Gamma(a_0) (\mathbf{D}_n^\delta)^{d_n}} \times \prod_{i=1}^{k} \left[\Phi(g(\mathbf{X}_{[,i]}))^{z_i} (1 - \Phi(g(\mathbf{X}_{[,i]})))^{1-z_i} \right]
\end{aligned}
\tag{4.3.3}
$$

Here,

$$\begin{aligned}
\mathbf{B}_n^{\delta} =& \left(\left(\mathbf{X}^{\delta}\right)' \mathbf{X}^{\delta} + \tau^2 \mathbf{I}_{(s+1)} \right)^{-1} \\
\mathbf{b}_n^{\delta} =& \mathbf{B}_n^{\delta} \left(\mathbf{X}^{\delta}\right)' \mathbf{y} \\
d_n =& a_0 + n/2 \\
\mathbf{D}_n^{\delta} =& b_0 + \left(\mathbf{y}'\mathbf{y} - \left(\mathbf{b}_n^{\delta}\right)' \left(\mathbf{B}_n^{\delta}\right)^{-1} \mathbf{b}_n^{\delta} \right).
\end{aligned}$$

Here, $\mathbf{X}^{\delta}$ represents the design matrix which contains only the columns corresponding to the nonzero β_i's. So, $\mathbf{X}^{\delta}$ is a matrix with dimension $n \times (s+1)$. Sampling of $\mathbf{z}$ can be done by updating the components of $\mathbf{z}$ one at a time in a random order from the posterior conditional distribution of $p(z_i \mid \mathbf{z}_{-i}, \mathbf{y}, \tau^2, \mathbf{g})$. Here, $\mathbf{z}_{-i}$ is the vector of $\mathbf{z}$ without the component z_i. Let ν_i is $p(z_i = 1 \mid \mathbf{z}_{-i}, \mathbf{y}, \tau^2, g)$ and ν_i^* is $p(z_i = 0 \mid \mathbf{z}_{-i}, \mathbf{y}, \tau^2, \gamma)$. Then, z_i can be sampled from Bernoulli distribution with success probability $\nu_i/(\nu_i + \nu_i^*)$.

The posterior conditional distribution of σ_ϵ^2 can be obtained as Inverse gamma distribution with shape and rate parameter as $(a_0 + n/2)$ and r_*, respectively. Here,

$$r_* = \left(b_0 + \frac{1}{2} \left(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}^{\delta} \left(\left(\mathbf{X}^{\delta}\right)' \mathbf{X}^{\delta} + \tau^2 \right)^{-1} \left(\mathbf{X}^{\delta}\right)' \mathbf{y} \right) \right). \tag{4.3.4}$$

Next, we can sample the block of nonzero β, $\boldsymbol{\beta}^{\delta}$, from its posterior conditional ditribution

given by $N_s(m_b, \sigma_\epsilon^2 v_b)$, with covariance matrix and mean vector given by:

$$v_b = \left(\left(\mathbf{X}^\delta \right)' \mathbf{X}^\delta + \tau^2 \right)^{-1} \tag{4.3.5}$$

$$m_b = v_b \left(\mathbf{X}^\delta \right)' \mathbf{y}. \tag{4.3.6}$$

To sample $\mathbf{g}$, ℓ, and $\boldsymbol{\mu}$ under probit link we will follow the latent variable approach described in Choudhuri et al. (2007). Let us define new latent variable $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_k)'$ such that conditional on g_i, the η_i's follow independent normal distribution with mean $g(\mathbf{X}_{[,i]})$ and variance 1. Also assume that z_i's are function of these η_i's with $z_i = I(\eta_i > 0)$, where $I(.)$ is the indicator function. Hence, conditional on g_i, z_i's are independent Bernoulli random variables with success probability $\Phi(g(\mathbf{X}_{[,i]}))$, thus leading to probit link. Now we can find the conditional distribution of $\mathbf{g}, \boldsymbol{\eta}$ given $\mathbf{z}$:

$$\begin{aligned} f(\mathbf{g}, \boldsymbol{\eta} | \mathbf{z}, \ell, \sigma_z^2, \boldsymbol{\mu}) \propto & \left(\frac{1}{|\Sigma|^{1/2}} \right) \exp\left\{ -(\mathbf{g} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{g} - \boldsymbol{\mu}) \right\} \\ & \times \prod_{i=1}^{k} \{ 1(\eta_i > 0) 1(z_i = 1) + 1(\eta_i < 0) 1(z_i = 0) \} \\ & \times \exp\left\{ -\frac{(\boldsymbol{\eta} - \mathbf{g})' (\boldsymbol{\eta} - \mathbf{g})}{2} \right\} \end{aligned} \tag{4.3.7}$$

Thus, $\mathbf{g} | \eta, \mathbf{z}, \ell, \sigma_z^2, \boldsymbol{\mu} \sim N(\mu^*, \Sigma^*)$, where covariance matrix $\Sigma^* = (I_k + \Sigma^{-1})^{-1}$ and mean vector $\mu^* = \Sigma^* (\eta - \boldsymbol{\mu}) + \boldsymbol{\mu}$. When k is large, computation of Σ^{-1} must be avoided as Σ will be near-singular matrix. To overcome this issue we will use spectral decomposition of Σ. Consider spectral decomposition $\Sigma = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}'$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_k)$ is a

diagonal matrix whose elements are the eigenvalues in descending order of magnitude and $\mathbf{U}$ is a matrix of eigenvectors. Let us partition Σ as:

$$\Sigma = [\mathbf{U}_m \mathbf{U}_{k-m}] \begin{pmatrix} \mathbf{\Lambda}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{k-m} \end{pmatrix} [\mathbf{U}_m \mathbf{U}_{k-m}]'. \tag{4.3.8}$$

Let $||.||_2$ and $||.||_F$ denote the spectral and Frobenius norms, respectively. Then, the best rank m approximation to Σ, with respect to $||.||_2$ and $||.||_F$, is $\Sigma_m = \mathbf{U}_m \mathbf{\Lambda}_m \mathbf{U}'_m$ (Stewart, 1993). We obtained rank m by finding the m^* such that $\inf_{m^*} \frac{\sum_{i=1}^{m^*} \lambda_i^2}{\sum_{i=1}^{k} \lambda_i^2} > 0.9$, that is, Σ_m captures 90% of the signal.

Posterior conditional distribution of $\boldsymbol{\eta}$ is given by:

$$\eta_i | \boldsymbol{\gamma}, \mathbf{z} \overset{ind}{\sim} \begin{cases} N(g(\mathbf{x}_{[,i]}), 1) | \eta_i > 0, & \text{if } z_i = 1 \\ N(g(\mathbf{x}_{[,i]}), 1) | \eta_i < 0, & \text{if } z_i = 0. \end{cases} \tag{4.3.9}$$

We assume inverse gamma $\mathcal{G}^{-1}(a, b)$ prior on ℓ, i.e, prior on ℓ has density $\pi(\xi; a, b) = \mathcal{G}^{-1}(a, b) \propto (\xi)^{a+1} exp\{-b/\xi\} 1(\xi > 0)$. Let $a = a_\ell$ and $b = b_\ell$ be parameters of the prior on ℓ. We put prior on $\mu \mid \sigma_z^2 \sim N(0, V_\beta)$. Then the posterior conditional distribution of μ is given by $N(m^*, V^*)$, with $V^* = (\mathbf{1}'\Sigma^{-1}\mathbf{1} + V_\beta^{-1})^{-1}$ and $m^* = V^* \mathbf{1}'\Sigma^{-1}\eta$. Then, the

posterior conditional distributions of ℓ are given by:

$$f(\ell \mid \mathbf{g}) \propto |\Sigma|^{-1/2} exp\left(-\frac{1}{2}(\mathbf{g}-\mu\mathbf{1})'\Sigma^{-1}(\mathbf{g}-\mu\mathbf{1})\right) \times (\ell)^{-a_\ell-1} exp\{-b_\ell/\ell\}1(\ell>0), \tag{4.3.10}$$

respectively. Sampling from $f(\ell \mid \mathbf{g})$ is performed using independent Metropolis-Hastings algorithm by taking proposal as $\mathcal{G}^{-1}(a_\ell, b_\ell)$.

4.4 Simulalation study

In this section we performed a simulation study to validate our model. In this study, we fit our model to a dataset simulated from known values of parameters and fit our model to assess whether the model can precisely recover the parameter of interest. We also generated a dataset using Andersen et al. (2014a) model with known values of parameters and fit their model to see how precisely their model can recover the parameters of interest.

We simulate 10 datasets from each model, model 1, that is, the proposed model and model 2 as described in Andersen et al. (2014a). We generate various instances of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ by first generating ϵ from *i.i.d.* $N(0, 4)$, and $\mathbf{X} \in \Re^{n\times 200}$ is generated from *i.i.d.* Gaussian. $\boldsymbol{\beta}$ is generated from prior described in equation (4.2.2) with $\ell = 1$, and $\tau^2 = 1$. We used two covariance kernel. One of them is as proposed in our model, that is $(\Sigma)_{i,j} = \sigma_z^2 exp\left(-\frac{||\mathbf{X}_{[,i]}-\mathbf{X}_{[,j]}||^2}{2\ell^2}\right)$ and the other one is $(\Sigma)_{i,j} = \sigma_z^2 exp\left(-\frac{(i-j)^2}{2\ell^2}\right)$. We fixed the degree of undersampling n/k to 0.5, with fixed sparsity of 0.25 by taking appropriate

values of μ and fixing the $\sigma_z^2 = 1$. To compare the performance of models, we calculate two summary measures. The first one is the Normalized Mean Square Error (NMSE) between the true $\boldsymbol{\beta}$ and the estimated $\hat{\boldsymbol{\beta}}$, which is given by:

$$\text{NMSE} = \frac{||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_2}{||\boldsymbol{\beta}||_2}. \tag{4.4.1}$$

The other measure we consider here is the F-measure, which we will use to measure the accuracy in estimating the sparsity pattern. F-measure is given by:

$$\text{F-measure} = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{4.4.2}$$

where, the precision measure provides what percentage of the tuples that is classifier labeled as positive is acutually positive. Mathematically, precision = True positive/(True positive + False positive) . Recall measures what percentage of positive tuples did the classifier labeled as positive. Mathematically, recall = True positive/(True positive + False negative).

	Model 1	Model 2
Bias ℓ	0.00227	0.01426
Bias μ	0.044	0.0071
NMSE	1.0113	1.0995

Table 10: Parameter estimation accuracy

Results in Table 10 show that we can estimate both the models parameters with very

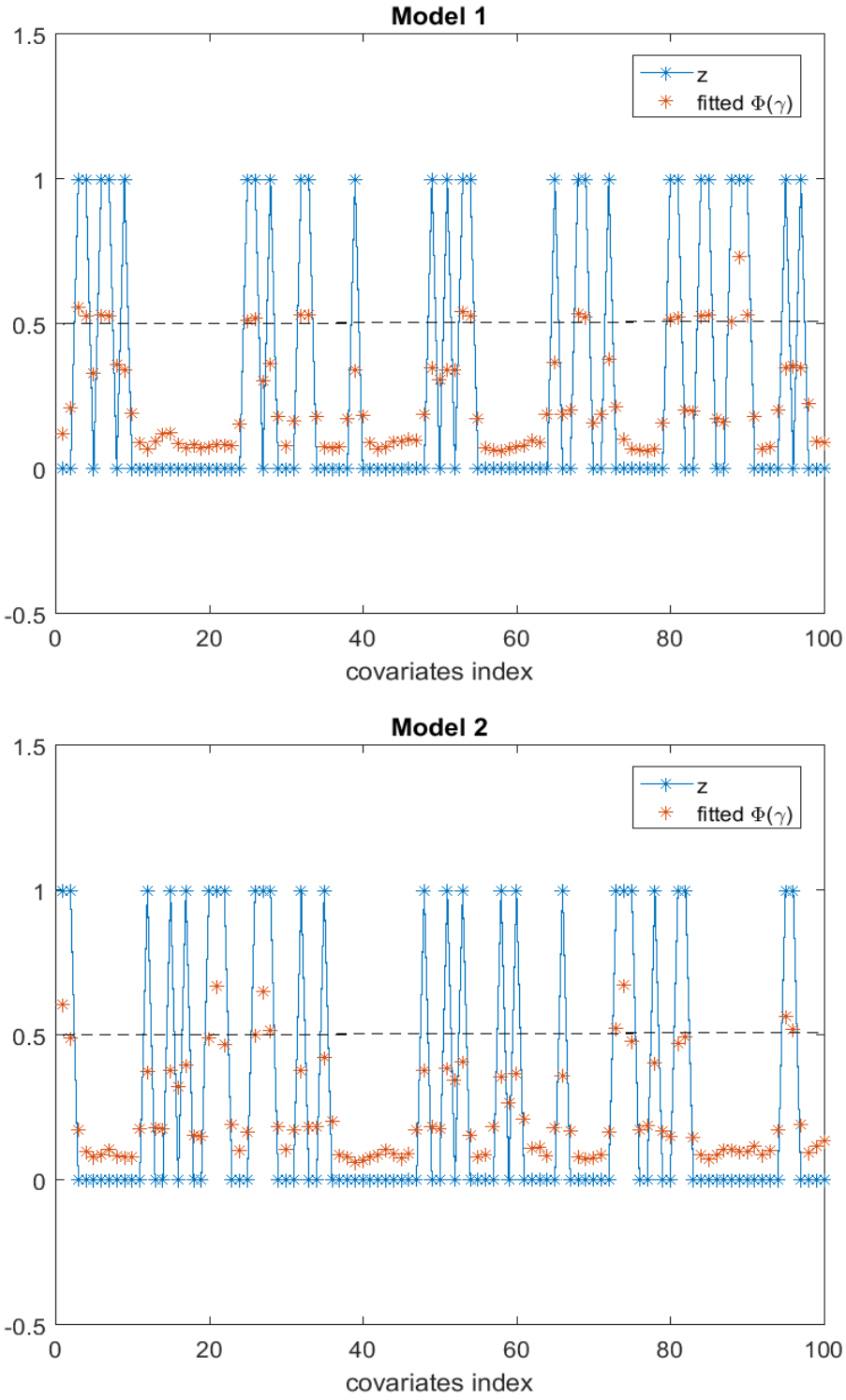

Figure 4: Sparsity structure

small bias. Figure 4 shows the sparsity structure generate from both the models in one of the instance. From the figure we can see that both the models are able to recover the sparsity pattern. Hence, both the models are able to precisely estimate the parameters of interest when the data is generated from the respective models.

4.5 Discussions

In this chapter we have introduced a novel Bayesian approach to perform variable selection by utilizing covariates values to gain prior information about sparsity patterns. The proposed model enforces the prior belief that regression coefficient values will be close to each other if the corresponding predictors are close to each other. We performed a simulation study to show that the model parameters are identifiable and can be estimated precisely using MCMC. We need to perform simulation studies to compare the proposed model with LARS algorithm (Efron et al., 2004), the model proposed in Andersen et al. (2014a), and the "oracle least square estimator" that knows the true support of the solutions. More simulation studies needed to understand the performance of the model under various degree of n/k ratio using a simulated dataset.

Chapter 5

Conclusions and Future Works

5.1 Concluding Remarks

In this dissertation we discussed some key applications of Gaussian process prior in several research areas. In Chapter 2, we consider the application of Gaussian process to deal with missing covariates for nonparametric regression problem. To deal with missing covariates for the nonparametric regression is often difficult. Especially, when we assign a GP prior on the unknown nonparametric function, the missing covariates will cause the problem to establish the covariance function in the GP prior. Our proposed method is the first one in solving this problem for the GP prior and it has kept the flexibility of GP prior in the computation for the nonparametric/semiparametric modeling from Bayesian perspective. Moreover, we have proved the posterior propriety under the 'exact' reference prior for the hyperparameters of GP prior in the appearance of missing covariates for the nonparametric part in the model. Further, we have showed that our model can perform better than the naive method to use complete cases only in the presence of ignorable missing covariates for the proposed semiparametric regression.

In Chapter 3, we explored the application of Gaussian process prior to increase

the flexibility in ordinal and binary regression models. By using a family of flexible power link function and Gaussian process prior on the latent regression function, we achieved double level of flexibility. Through various simulation studies and applications, we showed that the performance of this Bayesian nonparametric model using flexible power link function obtained using logistic as baseline link function is comparable to that of using Generalized extreme value link function. However, the performance of our model is significantly better than using linear latent regression function.

In Chapter 4, we explored the application of Gaussian process prior to encode information of similarity among covariates in variable selection problem. We showed that incorporating the observed covariates information in the spike and slab prior via Gaussian process we can estimate regression coefficients more accurately.

5.2 Extensions of Chapter 2

In Chapter 2, we assume the missing data mechanism is ignorable, we can extend our proposed procedure for non-ignorable missing mechanism. Consider the semiparametric regression model given by

$$y_i = \mathbf{z}_i'\boldsymbol{\beta} + g(\mathbf{x}_i) + \epsilon_i \tag{5.2.1}$$

for $i = 1, 2, \cdots, n$. $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_p]'$ is a $p \times 1$ vector of coefficients of fully observed covariates $\mathbf{Z} = [\mathbf{1}_p, \mathbf{z}_1 \cdots, \mathbf{z}_n]' \in \Re^{n \times p}$ and ϵ_i's are random errors. Assume $p << n$. $g(\cdot)$ is unknown nonlinear function. We assumed x_i's are covariates that are susceptible to

missing not at random. Non-ignorable missingness or missing not at random is a much more relaxed assumption in comparison to MAR, where the missing data mechanism depends on data that are missing. So, the probability whether the covariate value is missing or observed can be modeled in the following ways:

$$\pi_{ij} = P(R_{ij} = 1 \mid y_i, x_{ij}, \mathbf{z}_i) = h(y_i, \mathbf{z}_i, x_i).$$

In this case, missing data mechanism depends on the x_i's making it missing not at random case. Let $\mathcal{D} = \{(R_i, y_i, x_i, \mathbf{z}_i) : i = 1, \cdots, n\}$ denote complete data. To ease the inference of parameters $\ell, \sigma_z^2, \sigma_\epsilon^2$, and $\boldsymbol{\beta}$, it is important to obtain marginal likelihood given the parameters, that is to integrate out latent function $g(\cdot)$ in the likelihood, and we have

$$f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \ell, \sigma_z^2, \sigma_\epsilon^2, \boldsymbol{\beta}) = N(\mathbf{Z}\boldsymbol{\beta}, \sigma_z^2\mathbf{G}), \tag{5.2.2}$$

where $N_n(\cdot, \cdot)$ indicates a n-dimension multivariate normal distribution with $\mathbf{Z}\boldsymbol{\beta}$ being its mean and $\sigma_z^2\mathbf{G}$ being its covariance, $\mathbf{G} = \eta\mathbf{I}_n + \mathbf{K}$ and $\eta = \sigma_\epsilon^2/\sigma_z^2$ is the variance component of the noise-to-signal ratio. Here, $\mathbf{K}$ is $n \times n$ isotropic correlation matrix as defined in Section 2.2. We will interchange the usage of the notation $\mathbf{K}$ and $\mathbf{K}(\ell)$ to represent this correlation matrix throughout the paper when it is necessary. Let us define $\Theta = (\ell, \sigma_z^2, \eta, \boldsymbol{\beta}')'$. Now, we form a likelihood of $\Theta, \boldsymbol{\omega}$ given the observed data $\mathbf{y}$, $\mathbf{x}^{obs}$, R

and $\mathbf{Z}$ for Model (5.2.1): Model parameters we need to estimate are $g(.), \Theta, \sigma^2_\epsilon, \boldsymbol{\beta}, x^{mis}_{ij}, \boldsymbol{\omega}$.

$$\begin{aligned}\mathcal{L}(\Theta, \boldsymbol{\omega}, \phi \mid \mathbf{R}, \mathbf{y}, x^{obs}, \mathbf{Z}) &= \int_{\mathbf{x}^{mis}} \int_{g(X)} \Big(\prod f(R_{ij} \mid y_i, x_{ij}, z_{ij} \mid \phi)\Big) \times f(\mathbf{y} \mid g(\mathbf{X}), \mathbf{Z}) \\ &\qquad \times f(g(\mathbf{X}) \mid \Theta) \times f(\mathbf{X} \mid \boldsymbol{\omega})\ dg(X)\ d\mathbf{x}^{mis}. \\ &= \int_{\mathbf{x}^{mis}} \Big(\prod f(R_{ij} \mid y_i, x_{ij}, z_{ij} \mid \phi)\Big) \times f(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, \Theta) \\ &\qquad \times f(\mathbf{X} \mid \boldsymbol{\omega})\ d\mathbf{x}^{mis}.\end{aligned} \tag{5.2.3}$$

where, $f(\mathbf{y} \mid \mathbf{X}, \mathbf{Z}, \Theta) = \mathcal{N}_n(\mathbf{Z}\boldsymbol{\beta}, \sigma^2_z \mathbf{G})$ and $\mathbf{G} = \eta \mathbf{I}_n + \mathbf{K}$ and $\eta = \sigma^2_\epsilon / \sigma^2_z$ (noise-to-signal ratio).

$$(K)_{a,b} = (\Sigma)_{a,b} / \sigma^2_z \tag{5.2.4}$$

$\mathbf{K}$ is the correlation matrix, where, $\mathbf{K}_{i,j}$ is correlation between 'i' and 'j' observation.

Now, to complete Bayesian specification, we need to put prior on the GP hyperparameters to derive their posterior distributions. We will use the prior defined in equation (2.3.3). So, the joint posterior distribution for the parameters, which is given by:

$$\begin{aligned}f(\Delta, \mathbf{x}^{mis}, \phi, \boldsymbol{\omega} \mid \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z}) &= \Big(\prod f(R_{ij} \mid y_i, x_{ij}, z_{ij} \mid \phi)\Big) \\ &\qquad \left(\frac{1}{\sigma^2_z}\right)^{n/2} |\mathbf{G}|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2_z}(\mathbf{y} - \mathbf{Z}'\boldsymbol{\beta})' \mathbf{G}^{-1} (\mathbf{y} - \mathbf{Z}'\boldsymbol{\beta})\right\} \\ &\qquad \times \pi^R(\boldsymbol{\Delta} \mid \mathbf{X}) \times p(\mathbf{X} \mid \boldsymbol{\omega}) \times p(\phi) \times p(\boldsymbol{\omega}).\end{aligned} \tag{5.2.5}$$

We implement Gibbs sampling to draw samples from posterior distribution of $\Delta, \mathbf{x}^{mis}, \phi, \mathbf{\Omega}$. Posterior conditional distribution of $\mathbf{x}^{mis}, \ell$ and η do not have standard format. Sampling from their posterior conditional distribution is carried out by slice sampling method (Neal, 2003).

To validate our model, we have performed few simulation studies. We generated data using known values of parameters and fit our model to assess whether the model is able to precisely recover the parameters of interest. Future work include simulation studies to assess the performance of the model when the covariance structure is mis-specified, and also finding the appropriate application for the proposed model.

5.3 Future Works

In Chapter 3, we have explored the effects of flexible link function on GP binary and ordinal regression model. We can extend this model to handle longitudinal binary/ordinal response data. To the best of our knowledge Bayesian nonparametric model with flexible link functions have never been explored for longitudinal binary or ordinal response data.

To begin with, let $y_i(t)$ denote binary or ordinal response obtained at time $t \in [0, T]$ for subject i. Consider the predictors are related to the response variable via following

model:

$$\begin{aligned} y_i(t) &\sim Bin(p_i(t)) \\ p_i(t) &= F(g(x_i(t)), r) \end{aligned} \tag{5.3.1}$$

In practice, we often do not have any specific knowledge to specify the functional forms of $g_i = g(x_i(.))$, but some shape properties of mean structure may be known (e.g. non-decreasing nature of growth curves). So, Gaussian process prior can be used on $g_i(.)$ with appropriate mean structure to model this latent regression structure. We would also like to explore the longitudinal data subject to missing and censored observations.

In Chapter 4, we proposed a novel variable selection method where prior information about the covariates can be utilized to better estimate of parameters. In this model, the choice of covariance structure of GP plays a very crucial role to encode covariate information. For example, if we use squared-exponential covariance kernel as defined in equation (4.2.4), then in that case, if i^{th} and j^{th} column are close to each other distance wise, the values of β_i and β_j will be more likely to be close to each other. Therefore, it is important to assess the fitness of our model on a simulated dataset under various covariance structure, to better address their role in enforcing prior information about the covariates.

Appendix A

Posterior Propriety and Inference of Semiparametric Regression Model with Missing Covariates using GP Models

In this Chapter, we will provide some conditions required to prove posterior propriety of our model in Chapter 2 and also derived posterior conditional distribution of missing covariates. We also derived the Gibbs sampling algorithm for inference of Langmuir model and log model.

A.1 Posterior Propriety Conditions

In this subsection, we restate about the four conditions used in Ren et al. (2012) to prove posterior propriety of Θ in Equation (2.3.5) and we also utilize these conditions for proving the posterior propriety for the joint posterior distribution of Θ and $\mathbf{x}^{mis}$ in

(5.2.5). The four conditions are:

A1. Suppose $c_\ell(d)$ is a continous function of $\ell > 0$, for every $d \geq 0$ such that $c_\ell(d) = c^0(d/\ell)$, where $c^0(\cdot)$ is a correlation function satisfying $\lim_{u\to 0} c^0(u) = 0$.

A2. There exists a nonsingular and symmetric matrix $\mathbf{D}$ satisfying $\mathbf{1}_n'\mathbf{D}^{-1}\mathbf{1} \neq 0$, a fixed matrix $\mathbf{D}^*$, nondecreasing and differentiable function $\nu(\ell) > 0$, differentiable function $w(\ell)$, and a differentiable matrix $\mathbf{R}(\ell)$, so that as $\ell \to \infty$

$$\begin{aligned} \mathbf{K}(\ell) &= \mathbf{K}^*(\ell) + \nu(\ell)w(\ell)\{\mathbf{D}^* + \mathbf{R}(\ell)\}, \\ \mathbf{K}^*(\ell) &= \mathbf{1}_n\mathbf{1}_n' + \nu(\ell)\mathbf{D}, \end{aligned}$$

and

$$\nu(\ell) \to 0,\ w(\ell) \to 0,\ \frac{w'(\ell)\nu(\ell)}{\nu'(\ell)} \to 0,\ \|\mathbf{R}(\ell)\|_\infty \to 0,\ \text{and}\ \frac{\|\frac{\partial}{\partial\ell}\mathbf{R}(\ell)\|_\infty w(\ell)}{w'(\ell)} \to 0,$$

where $\|A\|_\infty = max_{i,j}|a_{i,j}|$.

A3. $[tr\{(\partial/\partial\ell)\mathbf{K}(\ell)\}^2]^{1/2}$ is integrable at zero for all ℓ.

A4. There exists a constant $b > 0$ and $c > 0$, such that

$$\mid w'(\ell) \mid \leq c\left|\frac{\partial}{\partial\ell}|log\nu(\ell)|^{-b}\right| \text{ as } \ell \to \infty.$$

A.2 Derivation of the Conditional Distribution of $\mathbf{x}^{mis}$ Given the observed $\mathbf{x}^{obs}$

Following our prior assumption about the hyperparamters μ_x and σ_x^2, i.e., $\pi(\mu_x, \sigma_x^2) \propto 1/\sigma_x^2$, and assume $k = n - m$, we have

$$\begin{aligned}
\pi(\mathbf{x}) &\propto \int\int \left(\frac{1}{\sigma_x^2}\right)^{n/2} \left(\frac{1}{\sigma_x^2}\right) \exp\left\{\frac{-1}{2\sigma_x^2}\left((\mathbf{x}-\mu_x\mathbf{1}_n)'(\mathbf{x}-\mu_x\mathbf{1}_n)\right)\right\} d\mu_x d\sigma_x^2 \\
&\propto \int \left(\frac{1}{\sigma_x^2}\right)^{(n-1)/2+1} \exp\left\{\frac{-1}{2\sigma_x^2}\left(\mathbf{x}'(\mathbf{I}_n - \mathbf{J}_n/n)\mathbf{x}\right)\right\} d\sigma_x^2 \\
&\propto \left(\frac{2}{\mathbf{x}'\left(\mathbf{I}_n - \mathbf{J}_n/n\right)\mathbf{x}}\right)^{(n-1)/2} \propto \left(\mathbf{x}'\mathbf{A}\mathbf{x}\right)^{-(n-1)/2},
\end{aligned}$$

where we denote $\mathbf{A} = \mathbf{I}_n - \mathbf{J}_n/n$. Then, we partition the matrix $\mathbf{A}$ as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_m - \mathbf{J}_m/n & -\mathbf{J}_{m\times k}/n \\ -\mathbf{J}_{k\times m}/n & \mathbf{I}_k - \mathbf{J}_k/n \end{pmatrix},$$

and notice that $\mathbf{A}_{12} = \mathbf{A}_{21}'$. After some algebra and let $\kappa = \mathbf{x}^{obs'}\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)\mathbf{x}^{obs}$, we can write

$$\begin{aligned}
& \left(\mathbf{x}'\mathbf{A}\mathbf{x}\right)^{-(n-1)/2} \\
= & \left[\left(\mathbf{x}^{mis} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}^{obs}\right)'\mathbf{A}_{11}\left(\mathbf{x}^{mis} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}^{obs}\right) + \kappa)\right]^{-(n-1)/2}, \\
\propto & \left[1 + \frac{1}{(n-m-1)}\left(\mathbf{x}^{mis} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}^{obs}\right)'\left((n-m-1)\mathbf{A}_{11}/\kappa\right)\left(\mathbf{x}^{mis} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}^{obs}\right)\right]^{-\xi},
\end{aligned}$$

where $\xi = (\nu + m)/2$ and $\nu = (n - m - 1)$. Then, it is easy to derive that the conditional distribution of $\mathbf{x}^{mis}$ provided that $\mathbf{x}^{obs}$ is known, that is,

$$\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}) = t_{\nu}\left(\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}^{obs}, (\kappa/(n-m-1))\mathbf{A}_{11}^{-1}\right), \tag{A.0.1}$$

where $t_{\nu}(\cdot,\cdot)$ indicates $\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})$ follows a multivariate t-distribution, ν is the degrees of freedom for the multivariate t-distribution, $\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}^{obs}$ is the mean and $(\kappa/(n-m-1))\mathbf{A}_{11}^{-1}$ is the covariance matrix for the multivariate t-distribution, respectively.

A.3 Model 2 (Langmuir Equation) Estimation

Recall that Model 2 (Langmuir equation) is

$$y_i = \frac{\alpha\beta x_i}{(1+\alpha x_i)} + \epsilon_i,\ i = 1, \cdots, n$$

where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{\epsilon}^2)$, $\alpha > 0$ and $\beta > 0$. To facilitate the Bayesian inference on Model 2 (Langmuir equation), we further assign priors on α and β as $\pi(\alpha) \propto \mathbf{1}(\alpha > 0)$ and $\pi(\beta) \propto \mathbf{1}(\beta > 0)$ with $\mathbf{1}(\cdot)$ being indicator functions. Let us presume the first m of n x_i's are missing. Following the assumption that $x_i \sim N_{+}(\mu_x, \sigma_x^2)$, $\pi(\sigma_{\epsilon}^2) \propto 1/\sigma_{\epsilon}^2$, $\pi(\mu_x \mid \sigma_x^2) \propto 1$ and $\pi(\sigma_x^2) \propto 1/\sigma_x^2$, then the joint distribution of unknown parameters in

Model 2 is

$$\begin{aligned}\pi(\alpha,\beta,\sigma_\epsilon^2,\mathbf{x}^{mis}\mid\mathbf{y},\mathbf{x}^{obs}) &\propto \left(\frac{1}{\sigma_\epsilon^2}\right)^{n/2+1}\exp\left(-\sum_{i=1}^{n}\frac{(y_i-\alpha\beta x_i/(1+\alpha x_i)^2}{2\sigma_\epsilon^2}\right)\\ &\times \mathbf{1}(\alpha>0)\mathbf{1}(\beta>0)\pi(\mathbf{x}^{mis}\mid\mathbf{x}^{obs})\mathbf{1}(x_i^{mis}\geq 0, i=1,\cdots,m),\end{aligned}$$

where $\pi(\mathbf{x}^{mis}\mid\mathbf{x}^{obs})$ follows the multivariate t-distribution derived in (A.0.1). Since $\pi(\alpha,\beta,\sigma_\epsilon^2,\mathbf{x}^{mis}\mid\mathbf{y},\mathbf{x}^{obs})$ is not in the closed form, we will resort to MCMC sampling scheme to draw the unknown parameters from this joint distribution. The key steps are to sample the unknown parameters from their corresponding conditional posterior distributions in an iterative manner from Step 1 to Step 4 until their MCMC samples are convergent. The conditional posterior distributions are below:

$$\begin{aligned}&\textit{Step 1}: \pi(\alpha\mid\beta,\sigma_\epsilon^2,\mathbf{x}^{mis},\mathbf{y},\mathbf{x}^{obs})\propto\exp\left(-\sum_{i=1}^{n}\frac{(y_i-\alpha\beta x_i/(1+\alpha x_i))^2}{2\sigma_\epsilon^2}\right)\mathbf{1}(\alpha>0),\\ &\textit{Step 2}: \pi(\beta\mid\alpha,\sigma_\epsilon^2,\mathbf{x}^{mis},\mathbf{y},\mathbf{x}^{obs})=\mathcal{N}_+\left(\frac{\sum_{i=1}^{n}\frac{x_i}{1+\alpha x_i}}{\alpha\sum_{i=1}^{n}\frac{x_i^2}{(1+\alpha x_i)^2}},\frac{\sigma_\epsilon^2}{\alpha^2\sum_{i=1}^{n}\frac{x_i^2}{(1+\alpha x_i)^2}}\right),\\ &\textit{Step 3}: \pi(\sigma_\epsilon^2\mid\alpha,\beta,\mathbf{x}^{mis},\mathbf{y},\mathbf{x}^{obs})=IG\left(n/2,\frac{\sum_{i=1}^{n}\left(y_i-\alpha\beta x_i/(1+\alpha x_i)\right)^2}{2}\right),\\ &\textit{Step 4}: \pi(\mathbf{x}^{mis}\mid\alpha,\beta,\sigma_\epsilon^2,\mathbf{y},\mathbf{x}^{obs})\propto\exp\left(-\sum_{i=1}^{m}\frac{(y_i-\alpha\beta x_i^{mis}/(1+\alpha x_i^{mis}))^2}{2\sigma_\epsilon^2}\right)\\ &\qquad\qquad\times\pi(\mathbf{x}^{mis}|\mathbf{x}^{obs})\mathbf{1}(x_i^{mis}\geq 0, i=1,\cdots,m).\end{aligned}$$

Noticing that sampling from the conditional posterior distribution of $\mathbf{x}^{mis}$ and α are done using slice sampling algorithm (Neal (2003)).

A.4 Model 3 (Log Model) Estimation

Recall that Model 3 (log model) is

$$y_i = \alpha + \beta \log(x_i) + \epsilon_i,\ i = 1, \cdots, n,$$

where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. To use Bayesian inference on the unknown parameters of Model 3, we assign the priors for α and β be $\pi(\alpha) \propto 1$, $\pi(\beta) \propto 1$ and the priors of all other parameters are the same as specified in Appendix A.3 for Model 2 (Langmuir equation). Without loss of generality, we assume the first m of n x_i's are missing. It is easy to derive that the joint distribution of unknown parameters in Model 3 is

$$\begin{aligned}\pi(\alpha, \beta, \sigma_\epsilon^2, \mathbf{x}^{mis} \mid \mathbf{y}, \mathbf{x}^{obs}) \quad &\propto \quad \left(\frac{1}{\sigma_\epsilon^2}\right)^{n/2+1} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \alpha - \beta log(x_i))^2}{2\sigma_\epsilon^2}\right) \\ &\times \quad \mathbf{1}(\alpha > 0)\mathbf{1}(\beta > 0)\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})\mathbf{1}(x_i^{mis} \geq 0, i = 1, \cdots, m),\end{aligned}$$

where similarly, $\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})$ follows the multivariate t-distribution derived in (A.0.1). Since $\pi(\alpha, \beta, \sigma_\epsilon^2, \mathbf{x}^{mis} \mid \mathbf{y}, \mathbf{x}^{obs})$ is not in the closed form, we will utilize MCMC sampling scheme to draw the unknown parameters from this joint distribution. The key sampling

steps are below:

$$
\begin{aligned}
&\textit{Step 1}: \pi(\alpha \mid \beta, \sigma_\epsilon^2, \mathbf{x}^{mis}, \mathbf{y}, \mathbf{x}^{obs}) = \mathcal{N}\left(\frac{\sum_{i=1}^{n}(y_i - \beta log(x_i))}{n}, \sigma_\epsilon^2/n\right), \\
&\textit{Step 2}: \pi(\beta \mid \alpha, \sigma_\epsilon^2, \mathbf{x}^{mis}, \mathbf{y}, \mathbf{x}^{obs}) = \mathcal{N}\left(\frac{\sum\left(log(x_i)(y_i - \alpha)\right)}{\sum\left(log(x_i)\right)^2}, \frac{\sigma_\epsilon^2}{\sum\left(log(x_i)\right)^2}\right), \\
&\textit{Step 3}: \pi(\sigma_\epsilon^2 \mid \alpha, \beta, \mathbf{x}^{mis}, \mathbf{y}, \mathbf{x}^{obs}) = IG\left(n/2, \frac{\sum\left(y_i - \alpha - \beta log(x_i)\right)^2}{2}\right), \\
&\textit{Step 4}: \pi(\mathbf{x}^{mis} \mid \alpha, \beta, \sigma_\epsilon^2, \mathbf{y}, \mathbf{x}^{obs}) \propto \exp\left(-\sum_{i=1}^{m}\frac{(y_i - \alpha - \beta log(x_i^{mis}))^2}{2\sigma_\epsilon^2}\right) \\
&\qquad\qquad \times \pi(\mathbf{x}^{mis}|\mathbf{x}^{obs})\mathbf{1}(x_i^{mis} \geq 0, i = 1, \cdots, m).
\end{aligned}
$$

From Step 1 to Step 4, we sample the unknown parameters from their corresponding conditional posterior distributions in an iterative manner until their MCMC samples are convergent. Especially for drawing the conditional posterior distribution of $\mathbf{x}^{mis}$, we employ the slice sampling algorithm.

Appendix B

Surrogate Data Slice Sampling Algorithm for Ordinal Regression using Flexible Power Link Function

Consider the GP ordinal regression with flexible power link function for one covariate case. For illustration we will use one covariate even though this algorithm can be easily extended to handle multiple covariates. Denote the observed data as $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x}$ is $n \times 1$ matrix of covariates and y is $n \times 1$ vector of ordinal responses. Here, $y_i \in \{0, 1, \cdots, J-1\}$, and the index i $(i = 1, \cdots, n)$ refers to observations in the sample. Following work of Albert and Chib (1993) by assuming the ordinal data outcomes as arising from n independent latent variable, $\mathbf{h} = (h_1, \cdots, h_n)$, such that $y_i = j$ if $\gamma_j < h_i < \gamma_{j+1}$, where, $-\infty = \gamma_0 < \gamma_1 = 0, \gamma_2 < \cdots < \gamma_{J-1} < \gamma_J = \infty$. Also $h_i = w(x_i) + \epsilon_i$, where w(.) is a latent regression function and $\epsilon_i \sim F(., r)$. So, the likelihood of $\boldsymbol{\gamma}$ and

$\mathbf{w}$ is given by:

$$L(\boldsymbol{\gamma}, \mathbf{w}, r \mid \mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} \prod_{j=0}^{J-1} [F(\gamma_{j+1} - w(x_i), r) - F(\gamma_j - w(x_i), r)]^{\mathbf{I}(y_i=j)} .$$

Note that sampling of $\boldsymbol{\gamma}$, $\mathbf{h}$ can be easily done from their posterior conditional distributions described in Section 3.2. We next describe the SSLS-power sampling algorithm to jointly update ℓ, r, and $\mathbf{w}$ by assuming σ_z^2 has already been updated using SSLS-power algorithm. Given the updated $\boldsymbol{\gamma}$, $\mathbf{h}$, σ_z^2, and $\boldsymbol{\beta}$, let us denote the current states of parameters as $\theta = \{\ell, \sigma_z^2\}$, r, $\mathbf{w}$, $\boldsymbol{\gamma}$, $\mathbf{h}$, σ_z^2, and β; next we will update r and ℓ along with $\mathbf{w}$, and denote updated states as θ^*, r^*, $\mathbf{w}^*$, β. Finally we will update β with $\mathbf{w}$ to the new state β^* and $\mathbf{w}^{**}$.

step 1: Draw surrogate data $\mathbf{g} \sim N(\mathbf{w}, S_\theta)$, where S_θ is the noise covariance matrix;

step 2: Compute implied latent variables: $\eta = L_\theta^{-1}(\mathbf{w} - \mathbf{m}_{\theta,g,\beta})$, where $L_\theta L_\theta' = R_\theta$.

step 3: Randomly center a bracket around current ℓ and r:

$\nu_\ell \sim Uniform(0, \sigma)$,

$\ell_{min} = max(0, \ell - \nu_\ell)$, $\ell_{max} = \ell_{\min} + \sigma$.

$\nu_r \sim Uniform(0, \sigma)$,

$r_{min} = max(0, r - \nu_r)$, $r_{max} = r_{\min} + \sigma$. Let's set $\sigma = 100$.

Using such a large enough σ(say,100) a boundary can be formed whose vertices are outside the posterior conditional density. So, we don't need to properly tune

σ. If σ is chosen small, appropriate stepping out procedure can be implemented to make the vertex outside of the density.

step 4: Compute the log of full conditional posterior distribution of all parameters:

$$\log(y) = \log(\mathcal{L}(\mathbf{w}(\mathbf{g}, \theta), r)) + \log(N(\mathbf{g}; \mathbf{x}\beta, S_\theta + \Sigma_\theta)) + \log(\pi(\ell)) + \log(\pi(r)).$$

step 5: We will now sample ℓ and r from the plane formed by boundaries derived in step 3.

Choose candidate:

$\ell^* \sim Uniform(\ell_{min}, \ell_{max})$ and $r^* \sim Uniform(r_{min}, r_{max})$.

step 6: Compute the updated $\mathbf{w}^* = L_{\theta^*}\eta + \mathbf{m}_{\theta^*,g,\beta}$.

step 7: Compute the updated log of posterior condtional density:

$$\log(y)^* = \log(\mathcal{L}(\mathbf{w}^*(\mathbf{g}, \theta^*), r^*)) + \log(N(\mathbf{g}; \mathbf{x}\beta, S_{\theta^*} + \Sigma_{\theta^*})) + \log(\pi(\ell^*)) + \log(\pi(r^*)).$$

step 8: Check whether sampled point from the interval is within the distribution. Otherwise, adjust the vertices to shrink the boundary and sample again.

Draw $u \sim uniform(0, 1)$.

if $\log(y)^* > \log(y) + \log(u)$

return $\mathbf{w}^*, \ell^*$, and r^*.

break;

else if $\ell^* < \ell$ and $r^* < r$

then, $\ell_{min} = \ell^*$ and $r_{min} = r^*$

else if $\ell^* < \ell$ and $r^* > r$

then, $\ell_{min} = \ell^*$ and $r_{max} = r^*$

else if $\ell^* > \ell$ and $r^* < r$

then, $\ell_{max} = \ell^*$ and $r_{min} = r^*$

else if $\ell^* > \ell$ and $r^* > r$

then, $\ell_{max} = \ell^*$ and $r_{max} = r^*$

Go to step 5 until break.

Now we will sample $\tilde{w}$ and β together.

Choose ellipse for β

step 9: Draw $\phi_\beta \sim Uniform[0, 2\pi]$, and define bracket $[(\phi_\beta)_{min}, (\phi_\beta)_{max}] = [\phi_\beta - 2\pi, \phi_\beta]$.

Choose an ellipse for β, $\nu \sim N(0, 25)$.

step 10: Draw surrogate data $\mathbf{g}^* \sim N(\mathbf{w}^*, S_{\theta^*})$

step 11: Compute implied latent variables: $\eta^* = L_{\theta^*}^{-1}(\mathbf{w}^* - \mathbf{m}_{\theta^*, g^*, \beta})$, where $L_{\theta^*} L'_{\theta^*} = R_{\theta^*}$.

step 12: Compute

$$\log\ (y)^{\#} = \log(\mathcal{L}(\mathbf{w}^*(\mathbf{g}^*, \theta^*), r^*)) + \log(N(\mathbf{g}^*; \mathbf{x}\beta, S_{\theta^*} + \Sigma_{\theta^*})) + \log(\pi(\beta)).$$

step 13: We will now sample β^* from the ellipse

$\beta^* = \beta\ \cos\ \phi_\beta + \nu_\beta\ \sin\phi_\beta$.

step 14: Compute function $\mathbf{w}^{**} = L_{\theta^*}\eta^* + \mathbf{m}_{\theta^*, g^*, \beta^*}$

step 15: Compute

$$\log\ (y)^{\#*} = \log(\mathcal{L}(\mathbf{w}^{**}(\mathbf{g}^*, \theta^*), r^*)) + \log(N(\mathbf{g}^*; \mathbf{x}\beta^*, S_{\theta^*} + \Sigma_{\theta^*})) + \log(\pi(\beta^*)).$$

Check whether sampled point from the hyperplane is within the distribution. Otherwise, adjust the boundaries to shrink the boundary and sample again.

Draw $u^{\#} \sim uniform(0, 1)$.

if $\log(y)^{\#*} > \log(y)^{\#} + \log(u^{\#})$

return $\mathbf{w}^{**}, \beta^*$

break;

else if $\phi_\beta < 0$

$(\phi_\beta)_{min} = \phi_\beta$

else if $\phi_\beta > 0$

$$(\phi_\beta)_{max} = \phi_\beta.$$

step 16: Sample $\phi_\beta \sim Uniform[(\phi_\beta)_{min}, (\phi_\beta)_{max}]$ and go to step 13.

Appendix C

Derivation of the auxiliary noise covariance S_θ using Laplace approximation

The auxiliary noise covariance S_θ is often chosen to be $\mathbf{cI}$. Here, we will discuss the Laplace approximation of Gaussian distribution approach to fix the vector $\mathbf{c}$ by matching the posterior of each observation to its Gaussian fit.

Using the latent regression function $\mathbf{w} = (w_1, \cdots, w_n)$ we can write the likelihood of $\mathcal{L}(\mathbf{w}) = \prod_{i=1}^{n} \mathcal{L}(w_i)$, where $\mathcal{L}(\mathbf{w})$ is same as the GP-ordinal likelihood defined in equation (3.2.17). In order to see how much the likelihood restricts each variable individually, the posterior density for observation i can be written as:

$$Q(w_i \mid \theta, \boldsymbol{\beta}) \propto \mathcal{L}(w_i) \times N(w_i; \mathbf{x}_i\boldsymbol{\beta}, (\Sigma_\theta)_{ii}) \tag{C.0.1}$$

Given a Gaussian fit to the individual posterior in equation (C.0.1) with variance v_i, the auxiliary noise can be set to a level that has the same posterior variance at that

observation:

$$(S_\theta)_{ii} = (v_i^{-1} - (\Sigma_\theta)_{ii}^{-1})^{-1}. \tag{C.0.2}$$

Any negative $(S_\theta)_{ii}$ must be thresholded. Now we derive S_θ considering logit link as baseline link function in GP-ordinal regression. Under, logit-power link:

$$\mathcal{L}(w_i) = \prod_{j=0}^{J-1} [F(\gamma_{j-1} - w(\mathbf{x}_i)) - F(\gamma_j - w(\mathbf{x}_i))]^{\mathbf{I}(y_i=j)}$$

where, $F(x,r) = F_0^r\left(\frac{x}{r}\right)\mathbf{I}_{(0,1]}(r) + \left(1 - F_0^{\frac{1}{r}}(-rx)\right)\mathbf{I}_{(1,+\infty)}(r)$ and $F_0(x) = 1/(1+exp(-x))$. Therefore,

$$Q(w_i \mid \theta, \beta_0) \propto \mathcal{L}(w_i) \times N(w_i; \mathbf{x}_i\boldsymbol{\beta}, (\Sigma_\theta)_{ii}) \tag{C.0.3}$$

To obtain Gaussian fit on Q using Laplace approximation, we first take log of Q. Then, using `fminunc` funcion in MATLAB we obtained the Hessian and $w_{max,i}$. So, $(S_\theta)_{ii} = 1/(1/v_i - (\Sigma_\theta)_{ii}^{-1})$. Here, $1/v_i = (Hessian)_i$. Similarly, changing the baseline link function $F_0(.)$ to $\Phi(.)$, that is, probit link we can obtain S_θ for probit-power link function.

Appendix D

Proof of Model Unidentifiability

Let us consider the case of binary response data. We will first prove that model is unidentifiable for binary response data case. As a result, model will also be unidentifiable for ordinal response data. Let $\zeta = \{r, \mathbf{w}\}$ and $\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ be the observed data. Here, y_i is binary data. Let us assume that $\mathbf{w}$ is known for simplicity. There can be two cases: either $r > 1$ or $0 < r < 1$. Let us assume $0 < r < 1$. One has

$$L(\zeta) = \prod_{i=1}^{n} \left\{ 1 - \left(\frac{1}{1 + exp(w_i/r)} \right)^r \right\}^{y_i} \times \left\{ \left(\frac{1}{1 + exp(w_i/r)} \right)^r \right\}^{1-y_i}.$$

Model will be considered unidentifiable if there exists two sets of parameters, ζ and $\tilde{\zeta}$, such that $L(\zeta) = L(\tilde{\zeta})$. We will use mathematical induction approach to prove this.

When $n = 1$, if $y_i = 0$, then we have

$$\left\{ \left(\frac{1}{1 + exp(w_i/r)} \right)^r \right\} = \left\{ \left(\frac{1}{1 + exp(\tilde{w}_i/\tilde{r})} \right)^{\tilde{r}} \right\}. \tag{D.0.1}$$

Clearly, the above equation is not one-to-one. So, there exists atleast two sets of ζ such that $L(\zeta) = L(\tilde{\zeta})$.

Next, assume that equation (D.0.1) holds for $n-1$ observations. Now, we will prove that equation (D.0.1) also holds for n observations.

$$\prod_{i=1}^{n}\left\{1-\left(\frac{1}{1+exp(w_i/r)}\right)^r\right\}^{y_i}\times\left\{\left(\frac{1}{1+exp(w_i/r)}\right)^r\right\}^{1-y_i}$$
$$=\prod_{i=1}^{n}\left\{1-\left(\frac{1}{1+exp(\tilde{w}_i/\tilde{r})}\right)^{\tilde{r}}\right\}^{y_i}\times\left\{\left(\frac{1}{1+exp(\tilde{w}_i/\tilde{r})}\right)^{\tilde{r}}\right\}^{1-y_i}$$
$$\text{or, } A\left\{1-\left(\frac{1}{1+exp(w_n/r)}\right)^r\right\}^{y_n}\times\left\{\left(\frac{1}{1+exp(w_n/r)}\right)^r\right\}^{1-y_n}$$
$$=B\left\{1-\left(\frac{1}{1+exp(\tilde{w}_n/\tilde{r})}\right)^{\tilde{r}}\right\}^{y_n}\times\left\{\left(\frac{1}{1+exp(\tilde{w}_n/\tilde{r})}\right)^{\tilde{r}}\right\}^{1-y_n},$$

where A and B denote $L(\zeta)$ and $L(\tilde{\zeta})$, respectively for $n-1$ obeservations. If we assume $y_n=0$, then the equation gets satisfied when $A=B$ and $\left\{\left(\frac{1}{1+exp(w_n/r)}\right)^r\right\}=\left\{\left(\frac{1}{1+exp(\tilde{w}_n/\tilde{r})}\right)^{\tilde{r}}\right\}$. From $A=B$ using our assumption for $n-1$ case, we can show $\left\{\left(\frac{1}{1+exp(w_i/r)}\right)^r\right\}=\left\{\left(\frac{1}{1+exp(\tilde{w}_i/\tilde{r})}\right)^{\tilde{r}}\right\}$.

Hence, for any n one can show that there exists two sets of parameters, ζ and $\tilde{\zeta}$, such that $L(\zeta)=L(\tilde{\zeta})$.

Bibliography

R. J. Adler. An introduction to continuity, extrema, and related topics for general gaussian processes. *Lecture Notes-Monograph Series*, 12:i–155, 1990. 14

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993. 55, 64, 108

M. R. Andersen, O. Winther, and L. K. Hansen. Bayesian inference for structured spike and slab priors. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 1745–1753, Cambridge, MA, USA, 2014a. MIT Press. 91, 94

M. R. Andersen, O. Winther, and L. K. Hansen. Bayesian inference for structured spike and slab priors. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1745–1753. Curran Associates, Inc., 2014b. 83

M. J. Beal. *Variational algorithms for approximate Bayesian inference*. University of London London, 2003. 13

J. O. Berger, V. D. Oliveira, and B. Sansó. Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374, 2001. 16, 22

S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998. 31

G. Celeux, F. Forbes, C. P. Robert, D. M. Titterington, et al. Deviance information criteria for missing data models. *Bayesian analysis*, 1(4):651–673, 2006. 39

H. Chipman. Bayesian variable selection with related predictors. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 24(1):17–36, 1996. ISSN 03195724. 81

T. Choi and M. J. Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987, 2007. 14, 27, 28

N. Choudhuri, S. Ghosal, and A. Roy. Nonparametric binary regression using a gaussian process prior. *Statistical Methodology*, 4(2):227–243, 2007. 7, 89

H. Cramér and M. R. Leadbetter. *Stationary and related stochastic processes: Sample function properties and their applications*. Courier Corporation, 2013. 14

C. Czado and T. J. Santner. Orthogonalizing parametric link transformation families in binary regression analysis. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 20(1):51–61, 1992. ISSN 03195724. 5, 6, 54

A. Damianou and N. D. Lawrence. Semi-described and semi-supervised learning with Gaussian processes. *arXiv preprint arXiv:1509.01168*, 2015. 14

D. G. Denison. *Bayesian methods for nonlinear classification and regression*, volume 386. John Wiley & Sons, 2002. 13

D. K. Dey, M.-H. Chen, and H. Chang. Bayesian approach for nonlinear random effects models. *Biometrics*, 53(4):1239–1252, 1997. ISSN 0006341X, 15410420. 44

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 04 2004. 94

R. F. Engle, C. W. J. Granger, J. Rice, and A. Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):310–320, 1986. 15

C. Faes, J. T. Ormerod, and M. P. Wand. Variational bayesian inference for parametric and nonparametric regression with missing data. *Journal of the American Statistical Association*, 106(495):959–971, 2011. 13

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. 81

M. Frlich. Non-parametric regression for binary dependent variables. *Econometrics Journal*, 9(3):511–540, 2006. ISSN 1368-423X. 6, 54

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. 81

S. Ghosal and A. Roy. Posterior consistency of gaussian process prior for nonparametric binary regression. *Ann. Statist.*, 34(5):2413–2429, 10 2006. 59, 62, 63, 64

S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *Ann. Statist.*, 27(1):143–158, 03 1999. 61

A. Girard and R. Murray-Smith. Learning a gaussian process model with uncertain inputs. Technical report, Department of Computing Science, University of Glasgow, 2003. 14

W. Härdle and H. Liang. *Partially Linear Models*, pages 87–103. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-32691-5. 15

D. Hernández-Lobato, J. M. Hernández-Lobato, and P. Dupont. Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14:1891–1945, 2013. 82

J. Huang, X. Huang, and D. Metaxas. Learning with dynamic group sparsity. In *2009 IEEE 12th International Conference on Computer Vision*, pages 64–71, Sept 2009. 82

L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. 82

X. Jiang, D. K. Dey, R. Prunier, A. M. Wilson, and K. E. Holsinger. A new class of flexible link functions with application to species co-occurrence in cape floristic region. *Ann. Appl. Stat.*, 7(4):2180–2204, 12 2013. 6, 53, 54, 57

M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Anal.*, 5(2):369–411, 06 2010. 81

I. Langmuir. The adsorption of gases on plane surfaces of glass, mica and platinum. *Journal of the American Chemical Society*, 40(9):1361–1403, 1918. 43

D. Li, X. Wang, L. Lin, and D. K. Dey. Flexible link functions in nonparametric binary regression with gaussian process priors. *Biometrics*, 72(3):707–719, 2016. ISSN 1541-0420. 5, 7, 28, 53, 54, 55, 57

X. Liao, H. Li, and L. Carin. Quadratically gated mixture of experts for incomplete data classification. In *Proceedings of the 24th International Conference on Machine learning*, pages 553–560. ACM, 2007. 13

R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2002. 7, 9, 20

H. Madsen and P. Thyregod. *Introduction to general and generalized linear models*. CRC Press, 2010. 5

J. J. Mahle, L. C. Buettner, and D. K. Friday. Measurement and correlation of the adsorption equilibria of refrigerant vapors on activated carbon. *Industrial & Engineering Chemistry Research*, 33(2):346–354, 1994. 41, 43, 45

P. McCullagh and J. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989. ISBN 9780412317606. 4, 52

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. ISSN 01621459. 81

I. Murray and R. P. Adams. Slice sampling covariance hyperparameters of latent gaussian models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, pages 1732–1740, USA, 2010. Curran Associates Inc. 69, 70

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. ISBN 0387947248. 2

R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–741, 2003. 30, 99, 105

M. A. Newton, C. Czado, and R. Chappell. Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association*, 91(433):142–153, 1996. ISSN 01621459. 7

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. 81

J. Quiñonero-Candela and S. T. Roweis. Data imputation and robust training with gaussian processes. Technical report, Citeseer, 2003. 14

C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006. 2, 3, 7, 14, 18, 58, 85

C. Ren, D. Sun, and C. He. Objective bayesian analysis for a spatial model with nugget effects. *Journal of Statistical Planning and Inference*, 142(7):1933 – 1946, 2012. 16, 22, 23, 24, 25, 101

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003. 15

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013. 82

A. C. Smith, S. A. Shah, A. E. Hudson, K. P. Purpura, J. D. Victor, E. N. Brown, and N. D. Schiff. A bayesian statistical analysis of behavioral facilitation associated with deep brain stimulation. *Journal of neuroscience methods*, 183(2):267–276, 2009. 74

R. L. Smith. Statistics of extremes, with applications in environment, insurance and finance. *Extreme values in finance, telecommunications and the environment*, pages 1–78, 2003. 54

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. ISSN 1467-9868. 39, 72

G. W. Stewart. On the early history of the singular value decomposition. *SIAM Rev.*, 35(4):551–566, Dec. 1993. ISSN 0036-1445. 90

K. Takezawa. *Introduction to nonparametric regression*, volume 606. John Wiley & Sons, 2005. 2

R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3): 273–282, 2011. ISSN 1467-9868. 81

A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996. 14

C. Wang, X. Liao, L. Carin, and D. B. Dunson. Classification with incomplete data using Dirichlet process priors. *Journal of Machine Learning Research*, 11(Dec): 3269–3311, 2010. 13

X. Wang and D. K. Dey. Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. *Ann. Appl. Stat.*, 4(4):2000–2023, 12 2010. 56

X. Wang and D. K. Dey. Generalized extreme value regression for ordinal response data. *Environmental and ecological statistics*, 18(4):619–634, 2011. 54

Y. Xie and B. Carlin. Measures of bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10):3458–3477, 10 2006. ISSN 0378-3758. 67

P. Yau and R. Kohn. Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, 13(3):191–208, 2003. 13

X. Zhang, S. Song, L. Zhu, K. You, and C. Wu. Unsupervised learning of Dirichlet process mixture models with missing data. *Science China Information Sciences*, 59 (1):1–14, 2016. ISSN 1869-1919. 13

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. 81