Bayesian Analysis (2020) **15**, Number 1, pp. 215–239

Learning Semiparametric Regression with Missing Covariates Using Gaussian Process Models

Abhishek Bishoyi[†], Xiaojing Wang[‡*], and Dipak K. Dey[§]

Abstract. Missing data often appear as a practical problem while applying classical models in the statistical analysis. In this paper, we consider a semiparametric regression model in the presence of missing covariates for nonparametric components under a Bayesian framework. As it is known that Gaussian processes are a popular tool in nonparametric regression because of their flexibility and the fact that much of the ensuing computation is parametric Gaussian computation. However, in the absence of covariates, the most frequently used covariance functions of a Gaussian process will not be well defined. 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. Several simulations are conducted to illustrate effectiveness of our proposed method and further, our method is exemplified via two real datasets, one through Langmuir equation, commonly used in pharmacokinetic models, and another through Auto-mpg data taken from the StatLib library.

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35.

Keywords: Gaussian processes, missing at random, missing covariates, nonparametric regression, semiparametric regression.

1 Introduction

In nonparametric regression, the objective is to find relationships between response and covariates without assuming the parametric form of a regression function. Nonparametric regression is a rapidly growing and exciting field. It offers a more flexible way to model the effects of covariates on the response compared to parametric models, which often have more restrictive conditions on the mean function. Many competing methods are available for nonparametric regression, including kernel-based methods, regression splines, smoothing splines, and wavelet and Fourier series expansions. When both responses and covariates are fully observed, the relevant theories and methods are well developed as described in Takezawa (2005). But a drawback of nonparametric regression models lies in its ability of interpretability in contrast to parametric regression models. Thus, various efforts have been addressed on semiparametric models, which balance the interpretation of parametric models and flexibility of nonparametric models. However,

*Corresponding Author.
[†]Selective Insurance, 40 Wantage Ave, Branchville, NJ, 07890, USA, abhishek.bishoyi@uconn.edu
[‡]Department of Statistics, University of Connecticut, Storrs, CT, 06250, USA, xiaojing.wang@uconn.edu
[§]Department of Statistics, University of Connecticut, Storrs, CT, 06250, USA, dipak.dey@uconn.edu

 https://doi.org/10.1214/18-BA1136

there is limited literature on either nonparametric or semiparametric models for missing covariates appearing in the nonparametric components of a regression.

For missing data, there are three basic classifications. If missingness does not depend on either observed or missing values, the data are called missing completely at random (MCAR). While the assumption of missing at random (MAR) is that missingness depends only on the observed values. The MAR is less restrictive than the MCAR. A much more relaxed assumption is missing not at random (MNAR), where the missingness depends on the data that are missing. A compressively review of general parametric statistical inferences with missing data has been discussed in Little and Rubin (2002).

When we come up with nonparametric modeling, one common approach 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, the structure may be overly restrictive due to the specific splines used in the 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 (cf., Beal (2003)) and showed that in the case of missing covariates, variational Bayes approximations produce multimodality in the posterior distributions where the one-to-one mapping does not exist for the unknown function.

In the Bayesian framework, nonparametric regression (or nonparametric classification) problems become the elicitation of the suitable priors on the mean function. Dirichlet process models are very popular methods for Bayesian nonparametric. 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 Markov chain Monte Carlo (MCMC) based analysis for this model suffers from huge computational costs, Wang et al. (2010) implemented approximate inference via the variational Bayesian approach. Recently, Zhang et al. (2016) proposed an infinite Dirichlet process mixture model to solve unsupervised learning for clustering with missing data. They assumed missing data as latent variables and obtained their posterior distributions using the variational Bayesian expectation maximization algorithm. Often, the computation burden is heavy on all these Dirichlet Process models above. Hence, the inference is carried out using approximate methods like variational Bayes and others. Moreover, the current literature for missing predictors in Dirichlet process models is only focused on clustering problems other than regression.

Gaussian process (GP) models are acknowledged as another popular tool for nonparametric regression. The usage of GP models has been widespread in spatial models, in the analysis of computer experiments and time series, in machine learning and so on (Rasmussen and Williams (2006)). For the properties of GP models, one can refer to Adler (1990), Van Der Vaart and Wellner (1996), Rasmussen and Williams (2006) and Cramér and Leadbetter (2013). Further, with normality assumption on the residuals, Choi and Schervish (2007) have shown assigning GP priors to the unknown regression function 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 appear several studies focusing on GP models with inputs subject to some measurement uncertainty (Girard and Murray-Smith (2003), Quiñonero-Candela and Roweis (2003) and Damianou and Lawrence (2015)). They often developed a two-stage procedure for estimating such GP models either using variational Bayesian methods or expectation maximization procedures, wherein the first stage, they estimated the model parameters only for complete cases 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 from the situation where the inputs are completely missing.

Therefore, in this paper, we consider the scenario when an input of GP models is subject to MCAR or MAR as for the purpose to fill 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 difficulty of interpretation and lack of extrapolation capability), we will consider semiparametric regression models in our study. Specifically, we will use the partially linear model, the most commonly used semiparametric regression model (cf., Engle et al. (1986), Ruppert et al. (2003), Härdle and Liang (2007) and references therein). A GP prior will be assigned to nonparametric components for this semiparametric regression model. Further, we will impute the missing covariate of the nonparametric component via a Bayesian hierarchical model, which will be a key for us to recover 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. Those hyperparameters often control the smoothness and variation of a GP. However, it is often difficult to specify subjective information over hyperparameters of a GP model. Thus, we will consider using noninformative priors. In Berger et al. (2001), they mentioned assigning noninformative priors such as commonly used constant priors and independent Jeffrey's priors for hyperparameters of a GP both fail to yield proper posteriors. Instead, they recommended the 'exact' reference prior for GP models when there is no white noise. Ren et al. (2012) extended the 'exact' reference prior in the case when we have white noises for the responses and showed the posterior is propriety under such prior. We further prove that under some mild conditions, the posterior propriety of the GP under the 'exact' reference prior will still hold in the presence of ignorable missing covariates. In addition, we have conducted a simulation study to compare the results from the 'exact' reference prior with certain weakly informative priors applied to hyperparameters of a GP.

The format of the paper is organized as below. In Section 2, we outline the setting of semiparametric regressions in a Bayesian hierarchical modeling framework. Section 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 our model under the "exact" reference prior for GP hyperparameters. Then, in Section 4, we perform several simulation studies to validate our proposed method and compare with existing methods. We present two real-world applications in Section 5 to show the advantage of our proposed model over some competitive models. Finally, in Section 6, we draw the conclusion and point out some future direction.

2 Semiparametric Regression Models with Ignorable Missing Covariates

The task of finding a good function estimation from a given dataset 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 approaches for Bayesian nonparametric regression 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, where the latter one restricts to certain fixed parametric structure. Thus, the GP model will potentially better approximate the true response function. In this section, we are going to propose our semiparametric regression model in a Bayesian framework to handle missing data, where we will use a GP model to estimate the nonparametric component.

The semiparametric regression model that we consider is given by

$$y_i = \mathbf{z}_i'\boldsymbol{\beta} + g(x_i) + \epsilon_i, \tag{2.1}$$

for $i = 1, 2, \cdots, n$. Here, $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_p]'$ is a $q \times 1$ vector of coefficients for 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. Without loss of generality, we assume $\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. In the absence of covariates $\mathbf{z}_i$'s, our model is reduced to a nonparametric model $y_i = g(x_i) + \epsilon_i$.

To estimate unknown function $g(\cdot)$, we are going to introduce a GP prior on $g(\cdot)$. We will consider a zero mean GP to avoid confounding of the mean parameter of a GP prior with coefficients $\boldsymbol{\beta}$ in Model (2.1). Let us denote $g(\cdot) \sim GP(0, \sigma_z^2 k(\cdot, \cdot \mid \ell))$, where $k(\cdot, \cdot \mid \ell)$ is the correlation function, and σ_z^2 and ℓ are hyperparameters of the GP. Then, 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 (i, j)th entry of Σ, i.e., $(\Sigma)_{ij} = \sigma_z^2 k_{ij} = \sigma_z^2 k(x_i, x_j \mid \ell)$, for $i, j = 1, \cdots, n$. In this paper, we only considered isotropic correlation kernel, that is, $k_{ij} = k(x_i, x_j \mid \ell) = \Psi_\ell(||x_i - x_j||)$ for some isotropic correlation function Ψ_ℓ and $||.||$ denote Euclidean distance. A common choice of isotropic correlation functions is the squared exponential kernel (also known as Gaussian kernel), that is,

$$(\Sigma)_{ij} = \sigma_z^2 k(x_i, x_j \mid \ell) = \sigma_z^2 \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right), \tag{2.2}$$

where 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 Subsection 4.2, we will also discuss about other power exponential correlation and Matérn class of correlation functions.

Let us consider the input x_i's in Model (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 $\mathbf{x} = [x_1, \cdots, x_n]'$ and presume that $\mathbf{x} \sim f(\mathbf{x} \mid \boldsymbol{\omega})$, where $\boldsymbol{\omega}$ are some unknown parameters. Without loss of generality, we can partition $\mathbf{x}$ into $\left\{\mathbf{x}^{mis}, \mathbf{x}^{obs}\right\}$, where $\mathbf{x}^{obs}$

collects the fully observed values, while $\mathbf{x}^{mis}$ denotes the observations absent of x 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 model the missing x_i's. For $i = 1, \cdots, n$, let R_i be 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). Then, we define $\mathbf{R} = [R_1, \cdots, R_n]'$ as a $n \times 1$ vector of missingness indicator. Here, we consider the 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 missing completely at random (MCAR) and the missingness mechanism is independent of the data.

(2) $\pi_i = P(R_i = 1 \mid y_i, x_i, \mathbf{z}_i) = P(R_i = 1 \mid y_i, \mathbf{z}_i)$ defines missingness is ignorable.

With these specified missingness mechanisms, our goal is to make the statistical inference for Model (2.1). We estimate parameters $\ell, \sigma_z^2, \sigma_\epsilon^2$, and $\boldsymbol{\beta}$ based on marginal likelihood, where we integrate out the unknown 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 $\mathbf{G} = \eta\mathbf{I}_n + \mathbf{K}$, $\eta = \sigma_\epsilon^2/\sigma_z^2$ is the variance component of the noise-to-signal ratio, and $\mathbf{K}$ is $n \times n$ isotropic correlation matrix with (i, j)th entry being k_{ij} and depending only on ℓ. Throughout the paper, 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, define $\Theta = (\ell, \sigma_z^2, \eta, \boldsymbol{\beta}')'$. Given the observed data $\mathcal{D} = \{\mathbf{R}, \mathbf{y}, \mathbf{x}^{obs}, \mathbf{Z}\}$, the likelihood of $\Theta, \boldsymbol{\omega}$ for Model (2.1) is:

$$\mathcal{L}(\Theta, \boldsymbol{\omega} \mid \mathcal{D}) = \int_{\mathbf{x}^{mis}} \left(\prod_{i=1}^{n} P(R_i \mid y_i, x_i, \mathbf{z}_i) \right) f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \Theta) f(\mathbf{x} \mid \boldsymbol{\omega}) \, d\mathbf{x}^{mis}. \tag{2.3}$$

Under the two specified missingness mechanisms, $P(R_i \mid y_i, x_i, \mathbf{z}_i)$ will not have any effect on estimation of parameters Θ and imputation values of missing $\mathbf{x}^{mis}$. 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 (2.3). Further, if we assign a prior on $\boldsymbol{\omega}$ as $\pi(\boldsymbol{\omega})$ then we can integrate out nuisance hyperparameters $\boldsymbol{\omega}$ in Equation (2.3). Define $\pi(\mathbf{x}) = \int_{\boldsymbol{\omega}} f(\mathbf{x} \mid \boldsymbol{\omega}) \pi(\boldsymbol{\omega}) d\boldsymbol{\omega}$ as the marginal prior on $\mathbf{x}$ and factorize $\pi(\mathbf{x}) = \pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}) \pi(\mathbf{x}^{obs})$. Then, given the data $\mathcal{D}$, the likelihood of Θ is:

$$\begin{aligned} \mathcal{L}(\Theta \mid \mathcal{D}) &\propto \int_{\mathbf{x}^{mis}} \int_{\boldsymbol{\omega}} f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \Theta) f(\mathbf{x} \mid \boldsymbol{\omega}) \times \pi(\boldsymbol{\omega}) \, d\boldsymbol{\omega} d\mathbf{x}^{mis} \\ &\propto \int_{\mathbf{x}^{mis}} f(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \Theta) \times \pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}) d\mathbf{x}^{mis}. \end{aligned} \tag{2.4}$$

To utilize Bayesian methods to perform the inference on Model (2.1), we need to specify priors on the unknown parameters Θ. One common approach is to use proper priors on Θ, assigning subjective priors or abstracting information from previous data. One advantage of proper priors is that they can always achieve posterior propriety. 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 noninformatively. But Berger et al. (2001) showed that the conventional noninformative priors lead to improper posterior. Thus, they derived an exact reference prior under the case without the noise (i.e., $\sigma_\epsilon^2 = 0$ in our case). Ren et al. (2012) further examined the effect of noise and derived the "exact" reference prior under the situation $\sigma_\epsilon^2 \neq 0$. In this paper, we aim to extend the posterior propriety of this reference prior in the case of missing data for the GP models.

3 Posterior Propriety and Posterior Inference

In the Section 3.1, we discuss the posterior propriety with the "exact" reference prior. Then, in Section 3.2, we specify MCMC procedure to carry out Bayesian inference of parameters in Model (2.1). Section 3.3 will be discussed about how to estimate new observations from our proposed model.

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 priors of (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 $\mathbf{x}^{mis}$ is known,

$$\mathcal{L}_*(\Theta \mid \mathbf{y}, \mathbf{x}^{mis}, \mathbf{x}^{obs}, \mathbf{Z}) \propto \left(\frac{1}{\sigma_z^2}\right)^{n/2} |\mathbf{G}|^{-1/2} \exp\left\{-\frac{1}{2\sigma_z^2}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{G}^{-1}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})\right\}. \tag{3.1}$$

Here, $\mathcal{L}_*(\cdot)$ with a subscript '$*$' indicating 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 (3.1) is given by

$$I^*(\ell, \eta, \sigma_z^2 \mid \mathbf{x}^{mis}, \mathbf{x}^{obs}) = \frac{1}{2}\begin{pmatrix} tr\{\mathbf{R}_G\frac{\partial}{\partial\ell}\mathbf{K}\}^2 & tr\{\mathbf{R}_G^2\frac{\partial}{\partial\ell}\mathbf{K}\} & \frac{1}{\sigma_z^2}tr\{\mathbf{R}_G\frac{\partial}{\partial\ell}\mathbf{K}\} \\ tr\{\mathbf{R}_G^2\frac{\partial}{\partial\ell}\mathbf{K}\} & tr(\mathbf{R}_G^2) & \frac{1}{\sigma_z^2}tr(\mathbf{R}_G) \\ \frac{1}{\sigma_z^2}tr\{\mathbf{R}_G\frac{\partial}{\partial\ell}\mathbf{K}\} & \frac{1}{\sigma_z^2}tr(\mathbf{R}_G) & \frac{n-q}{(\sigma_z^2)^2} \end{pmatrix}, \tag{3.2}$$

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 noninformative prior for Θ is

$$\pi^{Rf}(\Theta \mid \mathbf{x}^{mis}, \mathbf{x}^{obs}) = \pi^{Rf}(\ell, \eta, \sigma_z^2, \boldsymbol{\beta} \mid \mathbf{x}^{mis}, \mathbf{x}^{obs}) \propto \frac{1}{\sigma_z^2}\sqrt{|I^*(\ell, \eta, 1 \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})|}, \tag{3.3}$$

where $I^*(\ell, \eta, 1 \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})$ implies that we use $\sigma_z^2 = 1$ in Equation (3.2). In fact, the non-informative prior of $\pi^{Rf}(\ell, \eta, \sigma_z^2, \boldsymbol{\beta} \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})$ can be rewritten as $\pi^{Rf}(\ell, \eta, \sigma_z^2, \boldsymbol{\beta} \mid \mathbf{x}^{mis}, \mathbf{x}^{obs}) = \pi(\boldsymbol{\beta})\pi(\sigma_z^2)\pi_*^{Rf}(\ell, \eta \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})$, where $\pi(\boldsymbol{\beta}) \propto 1$, $\pi(\sigma_z^2) \propto 1/\sigma_z^2$ and $\pi_*^{Rf}(\ell, \eta \mid \mathbf{x}^{mis}, \mathbf{x}^{obs}) \propto \sqrt{|I^*(\ell, \eta, 1 \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})|}$.

Then, to show the posterior propriety of Θ using the "exact" reference prior (3.3) under the missing data framework for our Model (2.1), we only need to prove the integration of the joint posterior distributions of Θ and $\mathbf{x}^{mis}$ below

$$\begin{aligned}\pi(\Theta, \mathbf{x}^{mis} \mid \mathcal{D}) &\propto \left(\frac{1}{\sigma_z^2}\right)^{n/2} |\mathbf{G}|^{-1/2} \exp\left\{-\frac{1}{2\sigma_z^2}(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta})'\mathbf{G}^{-1}(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta})\right\} \\ &\times \pi^{Rf}(\Theta \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}), \end{aligned} \tag{3.4}$$

is finite over the domain of Θ and $\mathbf{x}^{mis}$. Here, in (3.4), $\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})$ is the prior 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 (3.4), first, let us integrate out $\boldsymbol{\beta}$ and σ_z^2 from this joint distribution, which yields

$$\begin{aligned}\pi(\ell, \eta, \mathbf{x}^{mis} \mid \mathcal{D}) &\propto |\mathbf{G}|^{-1/2}|\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2}(S^2)^{-(n-q)/2} \\ &\times \pi^{Rf}(\ell, \eta \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}),\end{aligned}$$

where $S^2 = (\mathbf{y}-\mathbf{Z}\widehat{\boldsymbol{\beta}})'\mathbf{G}^{-1}(\mathbf{y}-\mathbf{Z}\widehat{\boldsymbol{\beta}})$ and $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{G}^{-1}\mathbf{y}$. 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

$$\begin{aligned}0 < \int_\ell \int_\eta \int_{\mathbf{x}^{mis}} &\Big\{|\mathbf{G}|^{-1/2}|\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2}(S^2)^{-(n-q)/2} \\ &\times \pi^{Rf}(\ell, \eta \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})\Big\}\, d\ell d\eta d\mathbf{x}^{mis} < \infty.\end{aligned}$$

Using the Condition A1 to Condition A4 in Supplementary S.1 (Bishoyi et al., 2019), Ren et al. (2012) have proved that the integrated likelihood $\mathcal{L}_{**}(\mathbf{x}^{mis})$ is finite, that is:

$$\begin{aligned}0 &< \mathcal{L}_{**}(\mathbf{x}^{mis}) \\ &= \int_\ell \int_\eta |\mathbf{G}|^{-1/2}|\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2}(S^2)^{-(n-q)/2}\pi^{Rf}(\ell, \eta \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})d\ell d\eta < \infty, \end{aligned} \tag{3.5}$$

for a given $\mathbf{x}^{mis}$. Thus, $\mathcal{L}_{**}(\mathbf{x}^{mis})$ is bounded. Then, this is equivalent to prove that

$$0 < \int_{\mathbf{x}^{mis}} \mathcal{L}_{**}(\mathbf{x}^{mis})\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})d\mathbf{x}^{mis} \leq \int_{\mathbf{x}^{mis}} C\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}) < \infty, \tag{3.6}$$

where C is some bounded constant. Therefore, if we add additional condition below,

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

then (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 the paper, we will assume the covariates x_i's for the unknown function $g(\cdot)$ in Model (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 $\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})$ follows a multivariate t distribution (see details in Equation (S.1) and its derivation in Supplementary S.2 (Bishoyi et al., 2019)). Moreover, in Subsection 4.3, we analyze the sensitivity of the prior chosen on the missing covariates x_i's that satisfies Condition A5.

From our discussion, using Conditions A1–A4 in Supplementary S.1 (Bishoyi et al., 2019) with additional Condition A5, we can easily establish the posterior propriety of $(\Theta, \mathbf{x}^{mis})$ in Model (2.1). The Condition A1 ensures that the correlation function will decrease to zero as the distance between two points goes to infinity, while the Condition A2 ensures $\ell \to \infty$, a Taylor expansion of the correlation function will follow. The commonly used correlation matrix of the GP model such as the power exponential kernel, Matérn kernel, spherical kernel, rational quadratic kernel, and other isotropic kernels will often automatically satisfy the Conditions A1 and A2.

3.2 Bayesian Computation and Sampling Schemes

Since the joint posterior distribution of $(\Theta, \mathbf{x}^{mis})$ in (3.4) is proper, we will rely on this joint posterior to make inference for our proposed Model (2.1) when some of input x_i's are missing.

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 an inference. There are two key steps in developing the MCMC scheme. First, we draw the missing values $\mathbf{x}^{mis}$ provided that Θ 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 $\mathbf{x}^{mis}$. These alternative iterations create a Markov chain that eventually stabilizes to the joint posterior distribution of parameters and missing covariates in (3.4). The detailed steps of MCMC schemes are described below.

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

$$\begin{aligned}\pi(\mathbf{x}^{mis} \mid \ell, \eta, \mathcal{D}) \quad &\propto \quad \sqrt{|I^*(\ell, \eta, 1 \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})|} \times |\mathbf{G}|^{-1/2} |\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2} \\ &\times \quad (S^2)^{-(n-q)/2} \times \pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs}),\end{aligned}$$

where the term $I^*(\ell, \eta, 1 \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})$ is defined in Equation (3.3) and $\pi(\mathbf{x}^{mis} \mid \mathbf{x}^{obs})$ is derived in Equation (S.1). Since the conditional posterior distribution of $\mathbf{x}^{mis}$ do not have a closed form, we use a Metropolis-Hastings algorithm to impute 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:

$$\pi(\ell \mid \eta, \mathcal{D}, \mathbf{x}^{mis}) \quad \propto \quad \sqrt{|I^*(\ell, \eta, 1 \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})|} \times |\mathbf{G}|^{-1/2} |\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2} \times (S^2)^{-(n-q)/2}.$$

The posterior conditional distribution of ℓ is not closed form, thus we use slice sampling (cf., Neal (2003)) to draw samples of ℓ from its posterior conditional distribution.

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

$$\pi(\eta \mid \ell, \mathbf{x}^{mis}, \mathcal{D}) \quad \propto \quad \sqrt{|I^*(\ell, \eta, 1 \mid \mathbf{x}^{mis}, \mathbf{x}^{obs})|} \times |\mathbf{G}|^{-1/2} |\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}|^{-1/2} \times (S^2)^{-(n-q)/2}.$$

Also, the posterior conditional distribution of η does not have a closed form and we will 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 draw σ_z^2 from its posterior conditional distribution:

$$[\sigma_z^2 \mid \ell, \eta, \mathbf{x}^{mis}, \mathcal{D}] \sim IG((n-q)/2, S^2/2),$$

where IG indicates an inverse gamma distribution with the shape parameter $(n-q)/2$ and the rate parameter $S^2/2$.

Step 5: Given $\mathbf{x}^{mis}$, ℓ, η and σ_z^2, then we can sample $\boldsymbol{\beta} \sim \mathcal{N}_q(\widehat{\boldsymbol{\beta}}, \sigma_z^2 \left(\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}\right)^{-1})$.

Once we give the initial 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 chain has converged. To evaluate the convergence of the MCMC chains, we run the MCMC chains with 10 different starting values for the unknown parameters. The Gelman-Rubin potential scale reduction factor (cf., Brooks and Gelman (1998)) is found to be very close to 1 at most after 25,000 iterations of MCMC runs in our simulations and examples for all unknown parameters in the model. 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.

3.3 Posterior Predictive Distribution

In Subsection 3.2, 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 (2.1) simultaneously. However, often in the study, one of our goals is to predict responses using Model (2.1) when new observations of covariates come; while, another 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 (2.1).

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_j^{test}'s and $\mathbf{z}_j^{test}$'s are observed new covariates with $\mathbf{z}_j^{test} = [1, z_{j1}^{test}, \cdots, z_{jp}^{test}]'$, while y_i^{test}'s are unknown and needed to predict. To estimate y_j^{test}'s under the new observations x_j^{test}'s and $\mathbf{z}_j^{test}$, from Bayesian perspective, we shall first derive the posterior predictive distribution for y_j^{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 $\mathbf{x}^{mis}$, that is,

$$\begin{aligned}\pi(\mathbf{y}^{test}|\mathbf{x}^{test},\mathbf{Z}^{test},\mathcal{D}) &= \int\int f(\mathbf{y}^{test}\mid \mathbf{x}^{test},\mathbf{Z}^{test},\mathbf{x}^{mis},\Theta,\mathcal{D}) \\ &\times \pi(\Theta,\mathbf{x}^{mis}\mid\mathcal{D})d\Theta d\mathbf{x}^{mis}, \end{aligned} \tag{3.7}$$

where $\pi(\Theta,\mathbf{x}^{mis}\mid\mathcal{D})$ is the joint posterior distribution of $(\Theta,\mathbf{x}^{mis})$ derived in (3.4) and $f(\mathbf{y}^{test}\mid\mathbf{x}^{test},\mathbf{Z}^{test},\mathbf{x}^{mis},\Theta,\mathcal{D})$ is following a multivariate normal distribution, i.e.,

$$f(\mathbf{y}^{test}\mid\mathbf{x}^{test},\mathbf{Z}^{test},\mathbf{x}^{mis},\Theta,\mathcal{D}) = \mathcal{N}_t\left(\bar{f}_{test},\mathrm{Cov}(f_{test})\right), \tag{3.8}$$

with $\bar{f}_{test} = \mathbf{Z}^{test}\boldsymbol{\beta} + \Sigma_{(\mathbf{x}^{test},\mathbf{x})}(\sigma_z^2\mathbf{G})^{-1}(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta})$ 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 $\Sigma_{(\mathbf{x},\mathbf{x}^{test})} = \Sigma'_{(\mathbf{x}^{test},\mathbf{x})}$ is a $n\times t$ matrix and its (i,j)th element $(\Sigma_{(\mathbf{x},\mathbf{x}^{test})})_{i,j} = \sigma_z^2 k(x_i,x_j^{test}\mid\ell)$, 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 a total number of iterations for MCMC samples after burn-in period. Then, to generate a random sample $\mathbf{y}^{test}$ from its posterior predictive distribution in (3.7), it involves two major iterative steps, that is, for $i=1,\cdots,M$,

Step 1 Draw $(\Theta^{(i)},(\mathbf{x}^{mis})^{(i)})$ from $\pi(\Theta,\mathbf{x}^{mis}\mid\mathcal{D})$, where the detailed steps are described in Subsection 3.2.

Step 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

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

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 note $\mathbf{x}^{(i)} = (\mathbf{x}^{obs},(\mathbf{x}^{mis})^{(i)})'$.

Then, $\widehat{\mathbf{y}}^{test} = \sum_{i=1}^{M}\mathbf{y}^{test(i)}/M$ is the value of the posterior median estimate of $\mathbf{y}^{test}$.

4 Simulation Examples

In Subsection 4.1, we design some simulation examples to validate the inference procedure proposed in Section 3 and compare the benefits by imputing the missing values in Model (2.1) instead of using complete data only. Also, we compare the results of using our reference priors versus the weakly informative priors for the hyperparameters in the GP prior. Moreover, we empiricially investigage the posterior consistency of our proposed models under the 'exact' reference priors. In Subsection 4.2, we conduct some experiments to evaluate the sensitivity of misspecification of correlation functions for GP priors assigned to $g(\cdot)$ in Model (2.1). Further, we analyze the prior choices assigned for the missing covariates in Subsection 4.3. In the meantime, we perform a simulation study to compare our proposed method with other competitive methods in a nonparametric regression when the covariates are missing.

4.1 Simulation I

Consider the semiparametric regression model (2.1) with the following specification,

$$\begin{aligned} y_i &= \beta_0 + \beta_1 z_i + g(x_i) + \epsilon_i,\ i = 1, \cdots, n, \\ g(x_i) &\sim GP\left(0, \sigma_z^2 \exp\left(-\frac{(x_i - \cdot)^2}{2\ell^2}\right)\right), \end{aligned} \tag{4.1}$$

where $\beta_0 = -1$, $\beta_1 = 2$, $\ell = 2$, $\sigma_z^2 = 1$, $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 10)$, $z_i \overset{i.i.d.}{\sim} \mathcal{N}(1, 5)$, $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 0.4$ and $n = 120$. Thus, $\eta = \sigma_\epsilon^2/\sigma_z^2 = 0.4$. To test the performance of our proposed method, we randomly select 100 data points out of 120 generated data points from (4.1) to be training datasets, while the rest 20 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 the training data according to the procedure described below. That is, we randomly generate the missing indicator from

$$R_i \sim Bernoulli(p_i), \quad \text{with } p_i = \frac{\exp(b_0 + b_1 y_i)}{1 + \exp(b_0 + b_1 y_i)}, \tag{4.2}$$

where $R_i = 1$ indicates x_i is missing for the ith subject, $R_i = 0$ otherwise. We fix $b_1 = -0.1$ in (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 100 training points equals to 0.1, 0.25 and 0.4, respectively.

After the data were generated, we employ the MCMC sampling scheme developed in Subsection 3.2 to estimate model parameters and impute missing values of x_i's. We assume $\pi(x_i) \overset{i.i.d.}{\sim} \mathcal{N}(\mu_x, \sigma_x^2)$ 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 Supplementary S.2 (Bishoyi et al., 2019), we know that the conditional prior distribution of $\mathbf{x}^{mis}$ given $\mathbf{x}^{obs}$ will follow a multivariate t-distribution. Although we advocate to use the reference prior discussed in Subsection 3.1 for the unknown parameters $\ell, \eta, \beta_0, \beta_1$ and σ_z^2 in Model (4.1), we have also performed a comparison with other two popular priors used for our setting in practice. For the three priors considered, the information containing in the priors is gradually increasing as below:

1. Reference Prior (named *Prior 1*): We use the priors proposed in Subsection 3.1.
2. Vaguely Informative Prior (named *Prior 2*): According to the range of x_i's we generate, we assign a uniform prior on $[-20, 20]$ to ℓ. Besides, we presume an inverse gamma prior with mean 1 and variance 100 for σ_z^2 and $\log(\eta) \sim \mathcal{N}(0, 100)$. The priors for all other parameters are specified the same as *Prior 1*.
3. Weakly Informative Prior (named *Prior 3*): We assign the following priors for ℓ, η, β_0 and β_1: $\ell \sim \mathcal{N}_+(0, 10)$, $\log(\eta) \sim N(0, 2)$, $\beta_0 \sim \mathcal{N}(0, 10)$, $\beta_1 \sim \mathcal{N}(0, 10)$, where $\mathcal{N}_+(\cdot, \cdot)$ indicates the normal distribution is truncated on its left. Further, we assume σ_z^2 follows an inverse gamma prior with mean 1 and variance 10.

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

		10 % Missing		25 % Missing		40 % Missing	
		PM	CC	PM	CC	PM	CC
PMSE for y	*Prior 1*	0.0201	0.0211	0.0305	0.0419	0.0579	0.0891
	Prior 2	0.0195	0.0215	0.0276	0.0394	0.0566	0.0883
	Prior 3	0.0173	0.0190	0.0193	0.0317	0.0409	0.0662
Bias for ℓ	*Prior 1*	0.0491	0.0492	0.0525	0.0812	0.0837	0.1219
	Prior 2	0.0487	0.0502	0.0526	0.0714	0.0829	0.1154
	Prior 3	0.0266	0.0289	0.0355	0.0481	0.0593	0.0901
Bias for η	*Prior 1*	0.0035	0.0037	0.0042	0.0066	0.0103	0.0213
	Prior 2	0.0029	0.0039	0.0040	0.0063	0.0095	0.0187
	Prior 3	0.0011	0.0018	0.0026	0.0044	0.0066	0.0105
Bias for σ_z^2	*Prior 1*	0.0623	0.0652	0.0801	0.0961	0.1591	0.1949
	Prior 2	0.0601	0.0626	0.0788	0.0915	0.1451	0.1896
	Prior 3	0.0312	0.0315	0.0671	0.0888	0.1077	0.1354
Bias for β_0	*Prior 1*	0.0255	0.0253	0.0349	0.0467	0.0643	0.0810
	Prior 2	0.0210	0.0250	0.0298	0.0312	0.0487	0.0729
	Prior 3	0.0199	0.0205	0.0247	0.0289	0.0455	0.0712
Bias for β_1	*Prior 1*	0.0415	0.0430	0.0511	0.0701	0.0805	0.1061
	Prior 2	0.0387	0.0399	0.0481	0.0515	0.0772	0.0950
	Prior 3	0.0223	0.0271	0.0364	0.0419	0.0600	0.0798

Table 1: Comparison between our proposed method and the naive method via three different priors for Model (4.1).

are stored to reduce the 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. The total computation time costs 11 hours to run on a Xeon E5-2690 CPU with 2.60 GHz frequency, 128 GB RAM and 24 cores. Then, under the scenarios of three different priors mentioned above, we compare the parameters estimated in Model (4.1) using our proposed methods (PM) to the naive method. In the naive method, we only use the complete cases (CC) (i.e., those data points where covariate values x_i's are observed) to fit Model (4.1).

A comparison of these two methods with three different priors is shown in Table 1. We have compared the predicted mean squared error (PMSE) of y_i's for test points and the bias of the estimated parameters in Model (4.1). Here, we define PMSE $= \sum_{i=1}^{t}(y_i^{test} - \hat{y}_i^{test})^2/t$, where $\{y_i^{test}, x_i^{test}\}_{i=1}^{t}$ are test points and $\hat{y}_i^{test}$'s are posterior median estimates of predicted values at input x_i^{test}'s. $\hat{y}_i^{test}$'s are computed via the procedure described in Subsection 3.3. Notice in Table 1, the bias of the estimated parameters are calculated using the absolute distance between posterior median estimates of the parameters and their corresponding true values in the simulation.

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 test 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 (4.1) have relative lower biases by using

		10% Missing	20% Missing	40% Missing
Bias ℓ	$n = 100$	0.0491	0.0525	0.0837
	$n = 500$	0.0489	0.0511	0.0811
	$n = 1000$	0.0415	0.0493	0.0742
Bias η	$n = 100$	0.0035	0.0042	0.0103
	$n = 500$	0.0031	0.0038	0.0095
	$n = 1000$	0.0028	0.0030	0.0089
Bias σ_z^2	$n = 100$	0.0623	0.0801	0.1591
	$n = 500$	0.0616	0.0799	0.1533
	$n = 1000$	0.0598	0.0723	0.1488
Bias β_0	$n = 100$	0.0255	0.0349	0.0643
	$n = 500$	0.0200	0.0295	0.0601
	$n = 1000$	0.0192	0.0266	0.0584
Bias β_1	$n = 100$	0.0415	0.0511	0.0805
	$n = 500$	0.0380	0.0497	0.0790
	$n = 1000$	0.0295	0.0431	0.0742

Table 2: A simulation study of posterior consistency of our proposed methods under the 'exact' reference prior.

our proposed method than using the naive method. Generally speaking, the weakly informative prior will yield less absolute bias and smaller PMSE than the other two priors. It makes sense as if the weakly informative prior can provide extra information to pinpoint the right range of the target parameters. However, if we are lacking of such information, by using our proposed reference prior in the analysis, our reference prior is still doing a good job for the inference and prediction in comparison to the vaguely informative prior shown in Table 1 and is not much different than the weakly informative prior. However, an advantage for the practitioners in the usage of our proposed priors is that they can automatically run our program without knowing how to elicit the priors, whereas they are expected to obtain the comparable results as they do have some weakly information on the prior beliefs.

Another important topic is to explore the posterior consistency of our proposed models under the 'exact' reference priors. To empirically investigate this problem, we consider three different sample sizes: $n = 100$, 500, and 1000 for simulating data from Model (4.1). We illustrate the bias for each parameter averaging over 50 datasets that use different random seeds to simulate the data from Model (4.1). Notice that in Table 2, the row of $n = 100$ is the same as the row of Prior 1 in Table 1. From Table 2, it is obviously the bias of all parameters become smaller when the sample size gets larger. This pattern indicates that the posterior consistency empirically holds for our proposed models under the 'exact' reference prior.

4.2 Simulation II

In this subsection, we design 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 (2.1). Here, we consider three different types of covariance functions, commonly used in the spatial statistics and machine learning field.

1. Squared Exponential (SE) Covariance Function, see details in Equation (2.2).
2. γ-exponential (γ-E) Covariance Function:

$$(\Sigma)_{ij} = \sigma_z^2 \exp\left(-\frac{|x_i - x_j|^\gamma}{\ell}\right), \text{ with } 0 < \gamma \leq 2,$$

   where $|x|$ is the absolute value of x. 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 $\nu = 5/2$ (denote them as $MC_{3/2}$ and $MC_{5/2}$ in Table 3 and Table S), that is:

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

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

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

For each choice of covariance functions (i.e., 4 choices) and each missing percentages for the covariate x_i's (i.e., 10%, 25% and 40%), we apply Model (4.1) to generate 10 different sets of data using different random seeds. Thus, we have a total of $10 \times 4 \times 3 = 120$ datasets. Here, we choose $\beta_0 = -10$, $\beta_1 = 20$, $\ell = 10$ and $\sigma_z^2 = 2$ in Model (4.1), while all other settings are the same as Subsection 4.1. The data is generated in the same way as described in Subsection 4.1 with only changing the covariance function of the GP prior assigned to $g(\cdot)$ in Model (4.1).

We use the mean squared error of imputed missing values of x_i's (MSEx), PMSE of test 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 different choices of covariance functions. 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) for computing DIC. 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

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

		10% Missing	20% Missing	40% Missing
	F / T	SE	SE	SE
MSEx	γ-E	0.0165	0.0829	0.0071
	$MC_{3/2}$	0.1922	0.0894	0.5326
	$MC_{5/2}$	0.9803	0.4797	0.0034
PMSE	γ-E	0.0596	0.0658	0.0135
	$MC_{3/2}$	0.0536	0.0078	0.1332
	$MC_{5/2}$	0.2430	0.1321	0.4444
DIC	γ-E	0.0025	0.0157	0.0101
	$MC_{3/2}$	0.0108	0.0186	0.0149
	$MC_{5/2}$	0.0170	0.0462	0.0116

Table 3: The sensitivity analysis of using squared exponential kernels based on MSEx, PMSE and DIC. 'T' represents the true kernel used in generating the data and 'F' indicates the kernel applied in fitting the data.

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 be easily approximated using an MCMC run by taking the sample mean of the simulated values of $D(\Theta, \mathbf{x}^{mis})$; and for $\widetilde{\Theta}$ and $\widetilde{\mathbf{x}}^{mis}$, we choose their posterior medians in our study.

Since the SE covariance function has lots of good properties and supports a large class of functions with various shapes, we want to focus on the performance of using SE covariance functions when the other covariance kernels are true. For the reference, we present a detailed result in Table S of Supplementary S.3 (Bishoyi et al., 2019) to assess the performance of our proposed methods under misspecification of covariance functions for the GP prior in Model (4.1) based on MSEx, PMSE and DIC values. To better compare those values in Table S, we construct Table 3 using numbers computed via (using DIC values as an example),

$$ratio = \frac{|DIC_{SE} - DIC_{True}|}{DIC_{True}}, \tag{4.3}$$

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 that employ the true generated covariance function in the model fit. The DIC values in Equation (4.3) 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 relatively small. Thus, it shows that the performance of our model using SE covariance under misspecification of covariance kernel is kind of robust. Thus, in our application, we will choose to work with SE covariance.

4.3 Simulation III

In this subsection, we first design some simulation examples to assess the sensitivity of the priors assigned to the missing covariates. Then, we perform a simulation

	MSEx			PMSE		
Prior / Missing	Gaussian	Uniform	Cauchy	Gaussian	Uniform	Cauchy
10%	0.0117	0.0118	0.0116	0.0407	0.0411	0.0406
25%	0.0133	0.0135	0.0134	0.0624	0.0639	0.0630
40%	0.0238	0.0240	0.0236	0.1065	0.1109	0.1060

Table 4: The sensitivity analysis of the choice of priors on imputation of missing covariates based on MSEx and PMSE.

study to compare our proposed method with two competitive methods. One using cubic smooth splines, where the missing covariates are imputed through multiple imputation by chained equations (MICE) algorithm (cf., van Buuren and Groothuis-Oudshoorn (2011)). Another one is the method proposed by Faes et al. (2011), who solved the issue of missing covariates in the usage of spline basis functions for nonparametric regression.

Since the focus here is more on the estimation of the function $g(\cdot)$ and its missing covariate, we slightly revise the assumption of Model (4.1) and omit the covariate z_i and the intercept there, i.e.,

$$\begin{aligned} y_i &= g(x_i) + \epsilon_i, \qquad (4.4) \\ \text{where } g(x_i) = x_i^3, \quad x_i &\overset{i.i.d.}{\sim} Uniform[-5,5], \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,2), \end{aligned}$$

for $i = 1, \cdots, n$ and $n = 120$. Out of 120 observations, we randomly select 100 data points for training and use the rest of 20 observations as the test dataset. The percentage of 10%, 25%, and 40% missingness are created among the training points using the missing mechanism shown in Equation (4.2).

First, let us explore the sensitivity of the prior assigned to the missing covariate x_i's. To be specific, we investigate three different priors for x_i's in Model (4.4): 1) Gaussian prior: $\pi(x_i) = \mathcal{N}(0, 100)$; 2) Uniform prior: $\pi(x_i) = \mathbf{1}(\{-5 \le x_i \le 5\})/10$, where $\mathbf{1}(\cdot)$ is an indicator function; and 3) standard Cauchy prior: $\pi(x_i) = 1/[\pi(1 + x_i^2)]$ for $x_i \in \mathcal{R}$. Clearly, the second choice of the prior is identical to the distribution where the covariate x_i's are generated from. Typically, the Cauchy distribution has much heavier tailer and thus it might be expected to tolerate more on the misspecification for the choice of the prior on x_i's.

We have repeated 50 times for the process to generate data from Model (4.4) using different random seeds. The values in Table 4 are averaged over 50 times and calculated by using our proposed method in Subsection 3.2 with changing the corresponding priors specified on the x_i's. From Table 4, it is obvious that the values of MSEx and PMSE are almost the same among three different choices of the priors for x_i's in all three scenarios of missingness. This result suggests that the choice of the prior for the covariate x_i's is insensitive to the inference and prediction in our proposed method provided that the prior elicitation is containing the domain of the covariate x_i's.

Next, we are going to examine the performance of our proposed procedure to estimate the unknown function $g(\cdot)$ in comparison with the other two competitive methods. The three methods compared in Table 5 are the following:

Missing \ Method	MSEx			PMSE		
	M1	*M2*	*M3*	*M1*	*M2*	*M3*
10%	0.0117	0.0118	0.0123	0.0407	0.1205	0.0409
25%	0.0133	0.0133	0.0134	0.0624	0.1400	0.0564
40%	0.0238	0.0258	0.0253	0.1065	0.1542	0.0995

Table 5: The comparison of three methods based on MSEx and PMSE.

- *Method 1 (M1).* For estimating $g(\cdot)$, we assign $g(\cdot) \sim GP(0, \sigma_z^2 k(\cdot, \cdot, \mid \ell))$, where we choose the SE correlation for $k(\cdot, \cdot, \mid \ell)$. We apply the reference prior discussed in Subsection 3.1 for the hyperparameters of this GP prior.
- *Method 2 (M2).* We first perform multiple imputation of the missing covariates x_i's using MICE algorithm, which we call the `MICE` package in R. Next, we fit cubic smoothing spline for $g(\cdot)$ based on the imputed data.
- *Method 3 (M3).* We apply the method proposed in Faes et al. (2011) to impute missing covariates and estimate $g(\cdot)$ simultaneously. In their paper, they used penalized spline with mixed model representation to estimate $g(\cdot)$ (cf., Section 4.1 of Faes et al. (2011)).

We repeated the entire experiment 50 times using different random seeds and in each dataset, we analyze the data with the three methods described above. We assess the performance of these methods using MSEx and PMSE, and the results shown in Table 5 are averaging over these 50 experiments.

From Table 5, we can see that the values of MSEx for all three methods are similar when the missing rates are comparatively lower (i.e., 10% and 25%). It is obviously seen that when the missing percentage is higher (i.e., 40%), our proposed method performs much better than the other two in term of MSEx. Based on PMSE, the method proposed by Faes et al. (2011) and our proposed method have relative similar PMSE values, although Faes et al. (2011)'s method is doing slightly better in higher missing percentages than ours. In term of PMSE, both our method and Faes et al. (2011)'s method are doing much better than the cubic smoothing splines method. The cubic smoothing splines (i.e., *M2*) uses a two-stage estimation procedure, thus it fails to take advantage of the functional relationship between the response and covariates when imputing the missing covariates.

5 Application

Since our approach has successfully applied to the simulated data and recovered the true values of parameters well, we are going to employ our methods to two applications. According to our investigation for the relative robustness of misspecification of covariance functions of GP prior and the sensitivity of the prior choice on the missing covariate x_i's in Section 4, we are going to use SE covariance kernel on the GP prior and assume a normal distribution on the covariate x_i's throughout the applications.

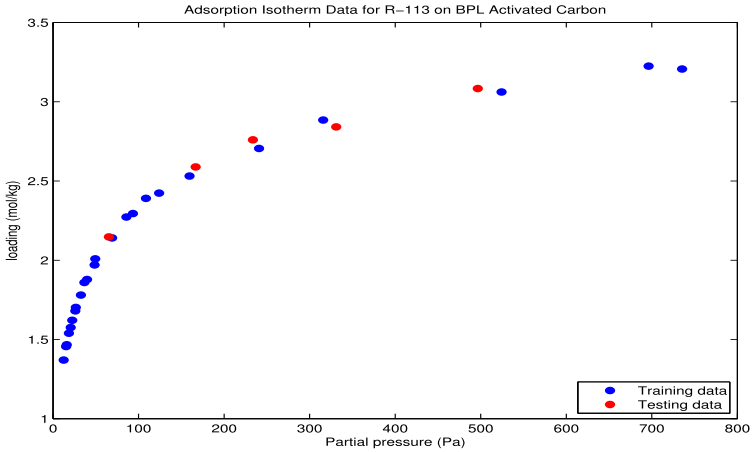

Figure 1: Scatterplot of Adsorption Isotherm Data.

5.1 Application I

First, we are going to apply our proposed method 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 for 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. 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 datasets, containing 24 and 5 observations, respectively. Figure 1 displays the 24 training points as blue colors and the 5 test points as red colors from 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 a 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

$$\text{Langmuir Model}: \quad y_i = \frac{\alpha\beta x_i}{1+\alpha x_i} + \epsilon_i, \ i = 1, \cdots, n, \tag{5.1}$$

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 (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 (5.1) are seldom all true. In addition, the 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

$$\text{Log Model}: \quad y_i = \alpha + \beta \log(x_i) + \epsilon_i, \ i = 1, \cdots, n \tag{5.2}$$

to be a competitive model with the Langmuir equation. There are no constraints on the values of parameters α and β in Equation (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.

Particularly, we use the semiparametric model below

$$\text{GP Model}: \quad y_i = \alpha + g(x_i) + \epsilon_i, \ i = 1, \cdots, n,$$

to compare the performance with the model specified in Equation (5.2) as well as with the Langmuir Equation (5.1) using the Adsorption Isotherm Data. We evaluate the accuracy of missing imputation based on mean squared errors criteria for all three models. From Figure 1, the domain of x, i.e., the partial pressure is always positive. We want to incorporate this information on the prior of x_i's. Thus, we consider a truncated normal prior on covariates x_i's, i.e., $\pi(x_i \mid \mu_x, \sigma_x^2) \propto \mathcal{N}_+(\mu_x, \sigma_x^2)$. For the priors on the hyperparameters μ_x and σ_x^2, we use the same noninformative 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 computation scheme for Langmuir Model and Log Model are postponed to Supplementary S.4 and Supplementary S.5 in (Bishoyi et al., 2019), while the computation scheme for GP Model is similar as discussed in Section 3 by merely changing the prior on x_i's.

We artificially create missingness in the covariates and compare imputed missing covariate with the true value based on MSEx. In addition, we compare the different models based on DIC and PMSE for the test points. We use Equation (4.2) to yield the ignorable missingness for the covariate x_i, where we produce the missingness with three different percentages, i.e., 10%, 25% and 40%. We repeat each generation 50 times. Thus, for each percentage, we average the values of MSEx, DIC, and PMSE over 50 times for Langmuir Model, Log Model, and GP Model, a summary of which is given in Table 6. For prediction of y_i's, Log Model and GP Model both are able to predict very accurately. Similarly, in term of DIC values, Log Model and GP Model are preferred to Langmuir Model. In term of the MSEx for imputation of missing x_i's, Log Model and GP Model are able to impute far better than Langmuir Model. Thus, in comparison to Log Model and GP Model, Langmuir Model performs very poorly in the criteria of PMSE, MSEx, and DIC.

From Table 6, the performance of GP Model is comparable to Log Model and much better than Langmuir Model. Although Log model is the best among the three, it has no theoretical foundation in adsorption isotherm data and it is just approximation

	Missingness	Langmuir Model	Log Model	GP Model
PMSE	10%	0.0123	0.0019	0.0021
	25%	0.0124	0.0020	0.0023
	40%	0.0127	0.0023	0.0024
MSEx	10%	224.1800	3.0866	3.1718
	25%	272.2199	5.6542	6.8122
	40%	301.6536	12.6118	10.2698
DIC	10%	330.6221	309.0777	317.2788
	25%	372.2199	321.2100	339.4221
	40%	401.3221	370.7133	371.8622

Table 6: The comparison of Langmuir Model, Log Model and GP Model on PMSE, MSEx and DIC.

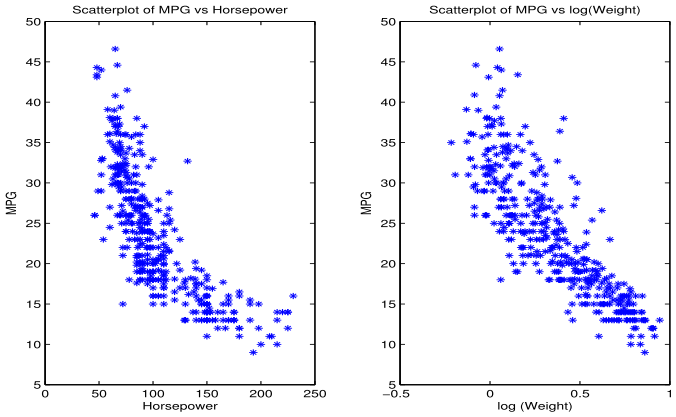

Figure 2: Scatterplot of MPG vs Horsepower and MPG vs log(Weight).

to Langmuir Model from experimental data. Therefore, Log Model will have a high risk of misspecification in real applications. GP Model has nonparametric nature in its fit, thus it will be more flexible in regressing adsorption isotherm data and can avoid misspecification. Hence, our 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.

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

Method \ Size	PMSE			DIC		
	$n = 30$	$n = 60$	$n = 90$	$n = 30$	$n = 60$	$n = 90$
Our Model	21.4647	17.2350	16.7127	416.2000	391.2511	361.1001
Linear Model	23.4285	18.1333	17.8999	459.0211	411.0542	390.2566

Table 7: The comparison of our proposed semiparametric model and the linear model using the PMSE and DIC criteria based on the imputed data.

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) in Figure 2) of the car. Both of these phenomena can be clearly seen from the original data in 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 employ the linear regression and our proposed semiparametric model on the three cases of instances for the randomly sampled observations. Specifically, we use the linear structure for the natural logarithm of the weight (in tons) and the nonparametric structure for the horsepower in our proposed model, that is:

$$y_i = \beta_0 + \beta_1 z_i + g(x_i) + \epsilon_i,\ i = 1, \cdots, n. \tag{5.3}$$

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 the same as Subsection 5.1. We compare the performance of both models based on PMSE and DIC, where their values are averaged over 50 draws for each case of instances.

Table 7 shows the results for our semiparametric model versus the linear model in the scenarios with missing data imputed. As expected, with the increase in the number of training data points, both values of PMSE and DIC have become smaller for these two models. However, our proposed semiparametric model is able to perform better than the linear model in all different sample sizes situations. Thus, our proposed GP

Method \ Sample Size	$n = 30$	$n = 60$	$n = 90$
CC	21.8333	17.7081	16.8739
MICE	21.4512	17.9011	17.2433

Table 8: The comparison of PMSE results for Model (5.3) via using complete cases (CC) and the MICE algorithm in dealing with missing data for Model (5.3).

semiparametric model is superior in the analysis of the Auto-mpg data in comparison to the linear model.

In addition, we build up Table 8, where we compare the PMSE results for Model (5.3) using complete cases and the MICE algorithm to deal with missing data for Model (5.3). When we compare the values of PMSE in the first row from Table 7 with those shown in Table 8, we can see that our proposed method (which is to impute the missing covariates and estimate the unknown parameters in a simultaneous way) performs much better in comparison with MICE imputation algorithms when the sample size is median (i.e., $n = 60$ and $n = 90$); while our method is dominant in the performance of PMSE in comparison to the complete cases analysis in all three cases.

6 Discussion

In this paper, we have considered the problem of imputation of missing covariates for the nonparametric part in a semiparametric regression under the Bayesian framework. In the absence of a parametric regression part, our semiparametric model can be reduced to a nonparametric regression setting. Our proposed procedure permits us to model nonparametric as well as semiparametric regression in the presence of missing covariate.

To deal with missing covariates for the nonparametric regression is often difficult. Especially, when we assign a GP prior to the unknown nonparametric function, the missing covariates will cause the problem to establish the covariance function of the GP prior. Our proposed method is the first one to solve this problem for the GP prior, while it has still kept the flexibility of GPs in the computation for the nonparametric/semiparametric modeling from Bayesian perspective. Also, we have proved the posterior propriety under the "exact" reference prior for the hyperparameters of GP in the appearance of missing covariates. Moreover, we have illustrated that in the presence of ignorable missing covariates for the semiparametric regression model, our proposed method can perform better than the naive method using complete cases and the cubic splines using imputed data. In addition, we have demonstrated our proposed method is at least comparable to the performance of the spline methods proposed by Faes et al. (2011) to deal with missing covariates in nonparametric regression. In the two applications, we have displayed that our proposed method is able to perform better than the competitive parametric methods when there are missing covariates in the data, especially 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. However, we are at least expected that the unknown function $g(\cdot)$ has an one-to-one mapping, otherwise, we might encounter the same multimodality problem in the posterior distribution for the missing covariates as mentioned in Beal (2003) and Faes et al. (2011).

Throughout this paper, we assume the missing data mechanism is ignorable. Sometimes, this assumption is somewhat restricted. Thus, our next goal is to extend our proposed procedure for a non-ignorable missing mechanism.

Supplementary Material

Supplementary Materials for "Learning Semiparametric Regression with Missing Covariates Using Gaussian Process Models" (DOI: 10.1214/18-BA1136SUPP; .pdf). We have restated about the four conditions used in Ren et al. (2012) and the derivation for the Conditional Distribution of $\mathbf{x}^{mis}$ Given $\mathbf{x}^{obs}$ in Section S.1 and Section S.2 of the supplement, respectively. Moreover, we have put the detailed results of MSEx, PMSE and DIC for different covariance kernels in Simulation II of Section 4.2 as Section S.3 of the supplement material. Also, in Section S.4 and Section S.5 of the supplement material, we have included the MCMC sampling scheme for Langmuir model estimation as well as the MCMC sampling scheme for Log Model Estimation for Section 5.2.
See more details in Supplement S (http://doi.org/10.2307/1390675).

References

Adler, R. J. (1990). "An introduction to continuity, extrema, and related topics for general Gaussian processes." *Lecture Notes-Monograph Series*, 12: i–155. MR1088478. 216

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London, London. 216, 237

Berger, J. O., Oliveira, V. D., and Sansó, B. (2001). "Objective Bayesian Analysis of Spatially Correlated Data." *Journal of the American Statistical Association*, 96(456): 1361–1374. MR1946582. doi: https://doi.org/10.1198/016214501753382282. 217, 220

Bishoyi, A., Wang, X., and Dey, D. K. (2019). "Supplementary Materials for "Learning Semiparametric Regression with Missing Covariates Using Gaussian Process Models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1136SUPP. 221, 222, 225, 229, 233

Brooks, S. P. and Gelman, A. (1998). "General methods for monitoring convergence of iterative simulations." *Journal of computational and graphical statistics*, 7(4): 434–455. MR1665662. doi: https://doi.org/10.2307/1390675. 223

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., et al. (2006). "Deviance information criteria for missing data models." *Bayesian analysis*, 1(4): 651–673. MR2282197. doi: https://doi.org/10.1214/06-BA122. 228

Choi, T. and Schervish, M. J. (2007). "On posterior consistency in nonparametric regression problems." *Journal of Multivariate Analysis*, 98(10): 1969–1987. MR2396949. doi: https://doi.org/10.1016/j.jmva.2007.01.004. 216

Cramér, H. and Leadbetter, M. R. (2013). *Stationary and related stochastic processes: Sample function properties and their applications*. Courier Corporation. MR0217860. 216

Damianou, A. and Lawrence, N. D. (2015). "Semi-described and semi-supervised learning with Gaussian processes." *arXiv preprint arXiv:1509.01168*. 217

Denison, D. G. (2002). *Bayesian methods for nonlinear classification and regression*, volume 386. John Wiley & Sons. MR1962778. 216

Dey, D. K., Chen, M.-H., and Chang, H. (1997). "Bayesian Approach for Nonlinear Random Effects Models." *Biometrics*, 53(4): 1239–1252. 233

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). "Semiparametric Estimates of the Relation Between Weather and Electricity Sales." *Journal of the American Statistical Association*, 81(394): 310–320. 217

Faes, C., Ormerod, J. T., and Wand, M. P. (2011). "Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data." *Journal of the American Statistical Association*, 106(495): 959–971. MR2894756. doi: https://doi.org/10.1198/jasa.2011.tm10301. 216, 230, 231, 236, 237

Girard, A. and Murray-Smith, R. (2003). "Learning a Gaussian process model with uncertain inputs." Technical report, Department of Computing Science, University of Glasgow. 217

Härdle, W. and Liang, H. (2007). *Partially Linear Models*, 87–103. Berlin, Heidelberg: Springer, Berlin Heidelberg. 217

Langmuir, I. (1918). "The Adsorption of Gases on Plane Surfaces of Glass, Mica and Platinum." *Journal of the American Chemical Society*, 40(9): 1361–1403. 232

Liao, X., Li, H., and Carin, L. (2007). "Quadratically gated mixture of experts for incomplete data classification." In *Proceedings of the 24th International Conference on Machine learning*, 553–560. ACM. 216

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons. MR1925014. doi: https://doi.org/10.1002/9781119013563. 216

Mahle, J. J., Buettner, L. C., and Friday, D. K. (1994). "Measurement and correlation of the adsorption equilibria of refrigerant vapors on activated carbon." *Industrial & Engineering Chemistry Research*, 33(2): 346–354. 232

Neal, R. M. (2003). "Slice Sampling." *The Annals of Statistics*, 31(3): 705–741. MR1994729. doi: https://doi.org/10.1214/aos/1056562461. 222

Quiñonero-Candela, J. and Roweis, S. T. (2003). "Data imputation and robust training with Gaussian processes." Technical report, Citeseer. 217

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. The MIT Press. MR2514435. 216

Ren, C., Sun, D., and He, C. (2012). "Objective Bayesian analysis for a spatial model with nugget effects." *Journal of Statistical Planning and Inference*, 142(7): 1933–1946. MR2903403. doi: https://doi.org/10.1016/j.jspi.2012.02.034. 217, 220, 221, 237

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. 12. Cambridge University Press. MR1998720. doi: https://doi.org/10.1017/CBO9780511755453. 217

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639. MR1979380. doi: https://doi.org/10.1111/1467-9868.00353. 228

Takezawa, K. (2005). *Introduction to nonparametric regression*, volume 606. John Wiley & Sons. MR2181216. 215

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, 45(3): 1–67. 230

Van Der Vaart, A. W. and Wellner, J. A. (1996). "Weak Convergence." In *Weak Convergence and Empirical Processes*, 16–28. Springer. MR1385671. doi: https://doi.org/10.1007/978-1-4757-2545-2. 216

Wang, C., Liao, X., Carin, L., and Dunson, D. B. (2010). "Classification with incomplete data using Dirichlet process priors." *Journal of Machine Learning Research*, 11(Dec): 3269–3311. MR2756185. 216

Yau, P. and Kohn, R. (2003). "Estimation and variable selection in nonparametric heteroscedastic regression." *Statistics and Computing*, 13(3): 191–208. MR1982474. doi: https://doi.org/10.1023/A:1024293931757. 216

Zhang, X., Song, S., Zhu, L., You, K., and Wu, C. (2016). "Unsupervised learning of Dirichlet process mixture models with missing data." *Science China Information Sciences*, 59(1): 1–14. 216