entropy

MDPI

Article

On a Robust MaxEnt Process Regression Model with Sample-Selection

Hea-Jung Kim *, Mihyang Bae and Daehwa Jin

Department of Statistics, Dongguk University-Seoul, Pil-Dong 3Ga, Chung-Gu, Seoul 100-715, Korea; system89@naver.com (M.B.); daehwajin@gmail.com (D.J.)

* Correspondence: kim3hj@dongguk.edu; Tel.: +82-2-2260-3221

Received: 2 February 2018; Accepted: 7 April 2018; Published: 9 April 2018

Abstract: In a regression analysis, a sample-selection bias arises when a dependent variable is partially observed as a result of the sample selection. This study introduces a Maximum Entropy (MaxEnt) process regression model that assumes a MaxEnt prior distribution for its nonparametric regression function and finds that the MaxEnt process regression model includes the well-known Gaussian process regression (GPR) model as a special case. Then, this special MaxEnt process regression model, i.e., the GPR model, is generalized to obtain a robust sample-selection Gaussian process regression (RSGPR) model that deals with non-normal data in the sample selection. Various properties of the RSGPR model are established, including the stochastic representation, distributional hierarchy, and magnitude of the sample-selection bias. These properties are used in the paper to develop a hierarchical Bayesian methodology to estimate the model. This involves a simple and computationally feasible Markov chain Monte Carlo algorithm that avoids analytical or numerical derivatives of the log-likelihood function of the model. The performance of the RSGPR model in terms of the sample-selection bias correction, robustness to non-normality, and prediction, is demonstrated through results in simulations that attest to its good finite-sample performance.

Keywords: Gaussian process model; hierarchical Bayesian methodology; robust sample-selection MaxEnt process regression model; Markov chain Monte Carlo; sample-selection bias; bias correction

MSC: 62G08; 62F15

1. Introduction

The Bayesian nonparametric method is a powerful approach for regression problems when the shape of the underlying regression function is unknown, the function may be difficult to evaluate analytically, or other requirements such as design costs may complicate the process of information acquisition process. Bayesian orthogonal basis expansion regression, spline smoothing regression, wavelet regression, and Gaussian process regression (GPR) are powerful nonparametric Bayesian approaches that address these regression problems. These regression techniques have been extensively used in fields such as psychology, data science, engineering, neuroscience, and fishery [1–5].

Sample selection (or incidental truncation) in regression analysis is known to often arise in a wide variety of practical problems and standard analysis of data with sample selection leads to biased results because the selected sample represents only a subset of a full population; see [6–8]. Regression analysis also has problems regarding sensitivity to outliers and departures from the normality of the dependent variable (see [9,10]). Thus, when one implements nonparametric Bayesian regression with non-normal data with sample selection, the selection mechanism and non-normality of the data must be jointly modeled with the Bayesian nonparametric regression model to correct the sample-selection bias and to implement a robust statistical inference. In this regard, several estimation

procedures have been considered in the literature to produce robust linear regression models that are subject to sample selection including for instance [6,7], for frequentist methods, and [8,9,11] for Bayesian methods. See [9,12] to obtain robust Bayesian sample-selection models other than the regression model. In addition, no studies have generalized a nonparametric regression model to deal with non-normal data with the sample selection.

The objective of this paper is to introduce the Maximum Entropy (MaxEnt) process regression model as a new Bayesian nonparametric regression model and to then generalize this model to propose a robust sample-selection Bayesian nonparametric regression model along with its inferential methodology. The MaxEnt process regression model is obtained by assuming a MaxEnt prior distribution for its nonparametric regression function, and it includes the GPR model as a special case. This provides a relationship between the MaxEnt nonparametric regression approach and the rationale to conduct a Gaussian regression analysis. This study focuses on the GPR model as a special MaxEnt process regression model and a powerful analysis model towards nonparametric regression problems. Then, the GPR model is generalized to obtain a robust sample-selection Gaussian process regression (RSGPR) model. This RSGPR model extends the GPR model to account for the sample selection scheme, and it is robust when the data are heavy-tailed or contain outliers.

The RSGPR model consists of two components. The first is a robust GPR model that determines the level of the dependent variable of interest and the second is an equation that describes the selection mechanism that determines whether we have observed the dependent variable or not. The sample-selection bias arises when these two components are correlated and must be modeled jointly. A Bayesian hierarchical methodology is developed here to estimate the RSGPR model. This methodology relies on a stochastic representation technique (see, e.g., [13]) to set up the Bayesian hierarchy of the RSGPR model, and it has three attractive features. First, given the likelihood function of the model, the posterior of its parameters does not belong to any well-known parametric family, but the methodology uses a simple Markov chain Monte Carlo (MCMC) algorithm that does not resort to generating random draws from the complex posterior. Second, the output of the algorithm not only provides a Bayesian analogue of confidence intervals for the regression function, but it also readily gives an indication of the presence (or absence) of the sample-selection bias. Third, if there is prior information, such as restrictions on the regression function, such information can be incorporated easily through a prior distribution.

The remainder of this study is organized as follows. Section 2 introduces the MaxEnt process regression model that strictly includes the GPR model. Then, this section formulates the RSGPR model that is obtained by incorporating the GPR model with a class of scale mixtures of normal errors as well as a selection model comprising a class of scale mixtures of the probit sample selection equations. Properties of this RSGPR model are studied including the exact distribution of a selected observation, a stochastic representation, a distributional hierarchy, and the magnitude of the sample-selection bias. In Section 3, we construct a Bayesian hierarchical model for inference in the RSGPR model by exploiting the stochastic representation and distributional hierarchy. Then we develop a Bayesian estimation methodology based on the hierarchical model to provide a simple estimation procedure for the RSGPR model. We further construct a computationally feasible MCMC algorithm through a Bayesian hierarchical approach. Section 4 examines the finite-sample performance of the method through a limited but informative simulation. This numerical illustration shows the usefulness of the RSGPR model for the Gaussian process regression analysis of non-normal data with the sample selection. The study then concludes with a discussion in Section 5. Proofs and additional details are provided in the Appendix A.

2. Robust Sample-Selection GPR Model

2.1. MaxEnt Process Regression Model

Consider the following nonparametric regression model,

$$\boldsymbol{y}_n = \boldsymbol{\eta}_n(\mathbf{x}) + \boldsymbol{\epsilon}_n, \tag{1}$$

where $\boldsymbol{y}_n = (y_1, \ldots, y_n)^\top$ is $n \times 1$ vector of responses, $y_i = \eta(\boldsymbol{x}_i) + \epsilon_i$, $\boldsymbol{\eta}_n(\mathbf{x}) = (\eta(\boldsymbol{x}_1), \ldots, \eta(\boldsymbol{x}_n))^\top$ is $n \times 1$ vector of regression function values satisfying $\eta(\boldsymbol{x}_i) = E[y_i|\boldsymbol{x}_i]$, $i = 1, \ldots, n$, $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ is the $n \times p$ design matrix, and $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)^\top$ is a $n \times 1$ vector of i.i.d. random noises with zero mean vector. In the basic model structure of (1), the parametric form of the regression function $\boldsymbol{\eta}_n(\mathbf{x})$ is not assumed, but $\boldsymbol{\eta}_n(\mathbf{x})$ is assumed to have specific types of functional structure. For example, $\boldsymbol{\eta}_n(\mathbf{x})$ can be represented with a Fourier series [14], splines [15], kernels [16] and others.

In the Bayesian nonparametric regression, we assume that the regression function (or signal term) $\boldsymbol{\eta}_n(\mathbf{x})$ is a random function that follows a particular distribution. This distribution is subjective in the sense that the distribution reflects our uncertain prior information regarding the function. Sometimes, we have a situation in which partial prior information on $\boldsymbol{\eta}_n(\mathbf{x})$ is available, outside of which it is desirable to use a prior that is as non-informative as possible. In this situation, Boltzmann's maximum entropy theorem (see, e.g., [17]) yields a maximum entropy prior $\pi_{max}(\boldsymbol{\eta}_n(\mathbf{x}))$ that is an exponential form and maximizes the entropy,

$$\mathrm{H}(\pi) = -\int_{\mathsf{R}^n} \pi(\boldsymbol{\eta}_n(\mathbf{x})) \log \pi(\boldsymbol{\eta}_n(\mathbf{x})) \mathrm{d}\boldsymbol{\eta}_n(\mathbf{x}),$$

in the presence of partial information for various moment functions of $\boldsymbol{\eta}_n(\mathbf{x})$. In a special case where we only have partial prior information about the mean vector and covariance matrix functions of $\boldsymbol{\eta}_n(\mathbf{x})$ of the Bayesian nonparametric regression model (1), Boltzmann's maximum entropy theorem yields the following prior distribution of $\boldsymbol{\eta}_n(\mathbf{x})$.

Lemma 1. *Let $n \times 1$ regression function vector $\boldsymbol{\eta}_n(\boldsymbol{x})$ have a prior distribution on R^n whose partial information on the mean and covariance functions are $m(\boldsymbol{x}) = (m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_n))^\top$ and $K(\boldsymbol{x}) = \{\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$, respectively. Then the maximum entropy prior of $\boldsymbol{\eta}_n(\boldsymbol{x})$ is*

$$\pi_{max}(\boldsymbol{\eta}_n(\boldsymbol{x})) = (2\pi)^{-n/2} |K(\boldsymbol{x})|^{-1/2} \exp\Big\{ -\frac{1}{2} (\boldsymbol{\eta}_n(\boldsymbol{x}) - m(\boldsymbol{x}))^\top K(\boldsymbol{x})^{-1} (\boldsymbol{\eta}_n(\boldsymbol{x}) - m(\boldsymbol{x})) \Big\} \tag{2}$$

for $\boldsymbol{\eta}_n(\boldsymbol{x}) \in \mathsf{R}^n$. This is a density of $\mathcal{GP}(m(\boldsymbol{x}), K(\boldsymbol{x}))$, a Gaussian process defined by the mean function $m(\boldsymbol{x})$ and the covariance function $K(\boldsymbol{x})$.

Note that the Gaussian process $\mathcal{GP}\big(m(\mathbf{x}), K(\mathbf{x})\big)$ defines a collection of random functions wherein any finite subset of the process has multivariate normal (Gaussian) distribution. From now on, we will write the Gaussian process as $\boldsymbol{\eta}_n(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), K(\mathbf{x})\big)$. The only restriction on the Gaussian process is that the covariance function $K(\mathbf{x})$ must be an $n \times n$ positive definite symmetric (pds) matrix. If $K(\mathbf{x})$ is not a pds matrix, then the corresponding value of $\mathrm{H}(\pi_{max}) = p(1 + \log(2\pi))/2 + \log(|K(\mathbf{x})|)/2$ will not be defined (see, e.g., [18]). As a result, this paper introduces yet another Bayesian nonparametric regression model by combining the regression model (1) and the MaxEnt prior in Lemma 1 and introducing a normal regression error distribution. The model is named as a "MaxEnt process regression model" and defined by

$$\begin{aligned} \boldsymbol{y}_n &= \boldsymbol{\eta}_n(\mathbf{x}) + \boldsymbol{\epsilon}_n,\ \boldsymbol{\epsilon}_n \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathrm{I}_n), \\ \sigma^2 &\sim \mathcal{IG}(\nu_1, \nu_2), \\ \boldsymbol{\eta}_n(\mathbf{x}) &\sim \mathcal{GP}\big(m(\mathbf{x}), K(\mathbf{x})\big), \end{aligned} \tag{3}$$

where $\mathcal{N}_n(\mathbf{0}, \sigma^2 \mathrm{I}_n)$ is an n-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 \mathrm{I}_n$, $\mathcal{IG}(\nu_1, \nu_2)$ denotes an inverse gamma distribution with shape parameter ν_1 and scale parameter ν_2, $\boldsymbol{\eta}_n(\mathbf{x})$ is independent of σ^2 and $\boldsymbol{\epsilon}_n$, the mean function $m(\boldsymbol{x}_i)$ reflects the expected function value at input $\boldsymbol{x}_i$, i.e., $m(\boldsymbol{x}_i) = \mathrm{E}[\eta(\boldsymbol{x}_i)]$, and the covariance function $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ models the dependence between the function values at different input points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, i.e., $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{E}[(\eta(\boldsymbol{x}_i) - m(\boldsymbol{x}_i))(\eta(\boldsymbol{x}_j) - m(\boldsymbol{x}_j))]$, $i, j = 1, \ldots, n$. See [19] for choice of an appropriate covariance function based on assumptions such as smoothness and likely patterns to be expected in the data. A commonly used isotropic covariance function in practice is the squared exponential covariance function given by

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = Cov(\eta(\boldsymbol{x}_i), \eta(\boldsymbol{x}_j)) = u_0 \exp\Big\{ -\frac{w_0}{2} (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top (\boldsymbol{x}_i - \boldsymbol{x}_j) \Big\}, \tag{4}$$

where u_0 and w_0 are hyperparameters and which are relevant for the shape of MaxEnt process regression. Here u_0 stands for global scale of the covariance matrix $K(\mathbf{x})$ and w_0 stands for smoothing parameter, respectively.

We can easily see that the MaxEnt process regression model (5) is the same as the GPR model considered by [19,20]. This proves the following corollary and can hence be used as an information theoretic justification for using the GPR model as a Bayesian approach for a nonparametric regression analysis.

Corollary 1. *Suppose* $\mathrm{E}[\boldsymbol{\eta}_n(\boldsymbol{x})] = m(\boldsymbol{x})$ *and* $\mathrm{Cov}(\boldsymbol{\eta}_n(\boldsymbol{x})) = K(\boldsymbol{x})$ *are all the prior information on* $\boldsymbol{\eta}_n(\boldsymbol{x})$ *for the Bayesian nonparametric regression model* (1)*. Then MaxEnt prior distribution of* $\boldsymbol{\eta}_n(\boldsymbol{x})$ *is* $\mathcal{GP}\big(m(\boldsymbol{x}), K(\boldsymbol{x})\big)$ *which defines the GPR model.*

According to Corollary 1, we shall denote the MaxEnt process regression model by the GPR model. When there is no functional constraint in the GPR model, then the prior specification in the model (3) can be used, and posterior inference can be performed without difficulty. It is seen, from [20], that the conditional posterior distribution of $\boldsymbol{\eta}_n(\mathbf{x})$ is normal with the mean and covariance given by

$$\begin{aligned} \mathrm{E}[\boldsymbol{\eta}_n(\mathbf{x}) | (\boldsymbol{y}_n, \mathbf{x}), \sigma^2] &= K(\mathbf{x})(K(\mathbf{x}) + \sigma^2 \mathrm{I}_n)^{-1} \boldsymbol{y}_n + \sigma^2 \mathrm{I}_n (K(\mathbf{x}) + \sigma^2 \mathrm{I}_n)^{-1} m(\mathbf{x}), \\ \mathrm{Var}[\boldsymbol{\eta}_n(\mathbf{x}) | (\boldsymbol{y}_n, \mathbf{x}), \sigma^2] &= (K(\mathbf{x})^{-1} + \sigma^{-2} \mathrm{I}_n)^{-1} = \sigma^2 K(\mathbf{x})(K(\mathbf{x}) + \sigma^2 \mathrm{I}_n)^{-1}. \end{aligned} \tag{5}$$

However, in the GPR analysis, a sample-selection scheme often applies to the response variable that results in missing not at random (MNAR) observations on the variable. In this case, the regression analysis using only the selected cases will lead to biased results (see, e.g., [6–8]). This study provides a bias correction procedure for the GPR analysis with MNAR data generated from the sample-selection scheme. For the analysis, we propose a robust sample-selection GPR (RSGPR) model based on a family of scale mixtures of normal (SMN) distributions (see [21,22] for details). This approach reflects the MNAR mechanism as well as its robustness against departures from the normality assumption (see, e.g., [11,12]), and proposes a robust GPR model to analyze the partially observed sample-selection data.

2.2. Proposed Model

We propose the RSGPR model through the following steps. First, we modify the GPR model (3) by incorporating the SMN error distribution for a robust GPR analysis. Then, we connect the robust GPR model directly to a sample-selection model by introducing some latent variables to explain the partially observed sample-selection data. To model the sample-selection mechanism, we need to introduce some notation for the partially observed data.

Let s_i be a binary variable that takes on value 1 if y_i of subject i is observed using the sample-selection scheme, and 0 if that of the subject is not observed using the same scheme. Then,

we introduce the following RSGPR model to represent the regression equation of the observable variable y_i :

$$\begin{aligned} y_i &= \begin{cases} \eta(\boldsymbol{x}_i) + \epsilon_i & \text{for } s_i = \mathrm{I}(z_i \geq 0), \\ \text{missing} & \text{for } s_i = \mathrm{I}(z_i < 0), \end{cases} \\ z_i &= \boldsymbol{v}_i^\top \boldsymbol{\gamma} + \varepsilon_i, \ \ i = 1, \ldots, n, \ \ \begin{pmatrix} \epsilon_i \\ \varepsilon_i \end{pmatrix} \overset{iid}{\sim} \mathcal{SMN}_2(\mathbf{0}, \Sigma, \delta, G), \\ \boldsymbol{\eta}_n(\mathbf{x}) &\sim \mathcal{GP}\big(m(\mathbf{x}), K(\mathbf{x})\big), \end{aligned} \tag{6}$$

where $\mathrm{I}(\cdot)$ is an indicator function, $\mathcal{SMN}_2(\mathbf{0}, \Sigma, \delta, G)$, a scale mixture of bivariate normal distributions with mixture function $\delta(\omega)$ and mixing variable $\omega \sim G$. Here $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ and $\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$ are parameters to be elicited by using the priors $p_0(\boldsymbol{\gamma})$ and $p_0(\Sigma)$.

Without loss of generality, we assume that the single sample selection scheme $s_i = I(z_i \geq 0)$ is applied to a random sample of n observations (y_i's) associated with the model (1) and gives only the first n_1 observed values of y_i's out of the n ($n > n_1$) observations according to the sample-selection scheme. Thus, the overall available data information of the RSGPR model consists of the set of s_i binary values and the n_1-tuples of observations $(y_i, \boldsymbol{v}_i)$ corresponding to individuals with $s_i = 1$, while $\boldsymbol{v}_i$ for those with $s_i = 0$. The purpose of this study is to estimate the regression function $\boldsymbol{\eta}_n(\mathbf{x})$ based on partially observed data (i.e., sample-selection data) with size n_1.

For fixed $\eta(\boldsymbol{x}_i)$, the density of the RSGPR model (6) is composed of a continuous component $h(y_i|s_i = 1)$ and a discrete component $p(s_i)$. The discrete component is

$$p(s_i) = [\bar{F}(C_i; 0, 1)]^{s_i} \, [1 - \bar{F}(C_i; 0, 1)]^{1 - s_i}, \tag{7}$$

where $\bar{F}(C_i; d, \tau) = \int_0^\infty \Phi(C_i; d, \delta(\omega)\tau) dG(\omega)$ with a selection interval $C_i = (\alpha_i, \infty)$, $\alpha_i = -\boldsymbol{v}_i^\top \boldsymbol{\gamma}$, and $\Phi(C_i; d, \kappa(\omega)\tau) = \int_{C_i} \phi(x; d, \delta(\omega)\tau) dx$ denotes the probability of the interval C_i under the $\mathcal{N}(d, \delta(\omega)\tau)$ distribution with the density $\phi(x; d, \delta(\omega)\tau)$. The continuous component is a density of $[y_i|\eta(\boldsymbol{x}_i), s_i = 1] \overset{d}{=} [y_i|\eta(\boldsymbol{x}_i), \ \varepsilon_i \in C_i]$ for $i = 1, \ldots, n_1$. This density is given by

$$h(y_i|s_i = 1) = \frac{\int_0^\infty \phi(y_i; \ \eta(\boldsymbol{x}_i), \delta(\omega)\sigma^2) \Phi\big(C_i; \theta_{\varepsilon_i|y_i}, \delta(\omega)\sigma^2_{\varepsilon_i|y_i}\big) dG(\omega)}{\bar{F}(C_i; 0, 1)}, \ \ y_i \in \mathsf{R}, \tag{8}$$

where $\theta_{\varepsilon_i|y_i} = \rho[y_i - \eta(\boldsymbol{x}_i)]/\sigma$, and $\sigma^2_{\varepsilon_i|y_i} = 1 - \rho^2$. This distribution is essentially a member of the class of skew-scale mixtures of normal (skew-SMN) distributions discussed by [13,23,24]. We will denote the distribution law of $[y_i|\eta(\boldsymbol{x}_i), s_i = 1]$ with density (8) by skew-$\mathcal{SMN}(C_i; \boldsymbol{\theta}_i, \Sigma, \delta, G)$, where $\boldsymbol{\theta}_i = (\eta(\boldsymbol{x}_i), 0)^\top$. The following lemma is useful to generate the partially observed y_i's and indicate the difference between the RSGPR model (6) and the GPR model (3).

Lemma 2. *For a given value of $\eta(\boldsymbol{x}_i)$, the selected observation $[y_i|\eta(\boldsymbol{x}_i), s_i = 1]$ for the RSGPR model can be represented by the following two-stages of distributional hierarchy:*

$$\begin{aligned} [y_i|\omega, \eta(\boldsymbol{x}_i), s_i = 1] &\overset{d}{=} \eta(\boldsymbol{x}_i) + \rho\sigma Z_{C_i} + \sigma(1 - \rho^2)^{1/2} U_i, \ \ i = 1, \ldots, n_1, \\ \omega &\sim G(\omega), \end{aligned} \tag{9}$$

where $U_i \overset{iid}{\sim} \mathcal{N}(0, \delta(\omega))$ and $Z_{C_i} \overset{ind}{\sim} \mathcal{TN}_{C_i}(0, \delta(\omega))$ are independent conditionally on ω, and $\mathcal{TN}_{C_i}(0, \delta(\omega))$ denotes a $\mathcal{N}(0, \delta(\omega))$ distribution truncated to the interval $C_i = (\alpha_i, \infty)$.

Lemma 2 shows that the RSGPR model applies to relax the classic assumption of the underlying normality as well as to reflect the sample-selection scheme. This lemma also indicates that the partially

observed data y_i's does not represent a random sample from the GPR model generating y_i's, even after controlling for the regression function $\eta(\boldsymbol{x}_i)$. If we want to apply a GPR analysis to the partially observed sample-selection data, a fitted model should be the RSGPR model. The RSGPR model changes depending on the choice of the distribution of ω and its function $\delta(\omega)$. In the special case wherein the distribution of ω degenerates at $\delta(\omega) = 1$, the RSGPR model produces a sample-selection Gaussian process normal error regression (SGPR_N) model. When we choose $\omega \sim \mathcal{G}(\nu/2, \nu/2)$, a gamma distribution with mean 1 and $\delta(\omega_i) = 1/\omega_i$, the model becomes a sample-selection Gaussian process t_ν error regression (SGPR_{t_ν}) model, allowing to regulate the tail distribution of the model by means of the degrees of freedom. We also see that the RSGPR model strictly includes the GPR model because the latter is obtained by setting $\rho = 0$. For the remainder of this study, we use the symbols in the preceding sections with the same definitions.

2.3. The Sample-Selection Bias

As indicated by the density (8) and Lemma 2 the selected observations $[y_i|s_i = 1]$'s do not represent a random sample from the GPR model generating y_i's, but they are missing not at random (MNAR) [25] inducing a sample-selection bias. The following results on the sample-selection bias are noted in the Bayesian estimation of the GPR model with the partially observed data.

Lemma 3. *Given the RSGPR model (6), a stochastic representation of conditional posterior distribution of the regression function* $\boldsymbol{\eta}_{n_1} = (\eta(\boldsymbol{x}_1), \ldots, \eta(\boldsymbol{x}_{n_1}))^\top$ *is*

$$[\boldsymbol{\eta}_{n_1}|\boldsymbol{y}, \omega, \Psi] \quad \stackrel{d}{=} \quad \boldsymbol{\theta}_1 + \Gamma\Omega_2^{-1}\boldsymbol{W}_1^{\boldsymbol{C}_\beta} + \boldsymbol{W}_2, \tag{10}$$

where $\Psi = \{\sigma^2, \rho, \gamma\}$, $\boldsymbol{W}_1 \sim \mathcal{N}_{n_1}(\boldsymbol{0}, \Omega_2)$ *and* $\boldsymbol{W}_2 \sim \mathcal{N}_{n_1}(\boldsymbol{0}, \Omega_1 - \Gamma\Omega_2^{-1}\Gamma^\top)$ *are independent random vectors,* $\boldsymbol{W}_1^{\boldsymbol{C}_\beta} \stackrel{d}{=} [\boldsymbol{W}_1|\boldsymbol{W}_1 \in \boldsymbol{C}_\beta]$, $\boldsymbol{C}_\beta = \cap_{i=1}^{n_1}\{w_{1i};\ \beta_i \le w_{1i} \le \infty\}$, $\boldsymbol{W}_1 = (w_{11}, \ldots, w_{1n_1})^\top$, $\beta_i = (\alpha_i - \theta_{2i})/\sqrt{\delta(\omega)}$, $\boldsymbol{\theta}_1 = K_{11}(\boldsymbol{x})^{-1}H^{-1}\boldsymbol{y}_{n_1} + \delta(\omega)\sigma^2H^{-1}m_1(\boldsymbol{x})$, $\boldsymbol{\theta}_2 = (\theta_{21}, \ldots, \theta_{2n_1})^\top = \rho(\boldsymbol{y}_{n_1} - m_1(\boldsymbol{x}))/\sigma$, $\Gamma = -\rho\Omega_1/\sigma$, $H = \left(K_{11}(\boldsymbol{x}) + \delta(\omega)\sigma^2 I_{n_1}\right)$, $\Omega_1 = \delta(\omega)\sigma^2K_{11}(\boldsymbol{x})H^{-1}$, $\Omega_2 = (1-\rho^2)I_{n_1} + \rho^2\Omega_1/\sigma^2$, $\boldsymbol{y}_{n_1} = (y_1, \ldots, y_{n_1})^\top$ *be an* $n_1 \times 1$ *observed vector,* $m_1(\boldsymbol{x}) = (m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_{n_1}))^\top$, *and* $K_{11}(\boldsymbol{x})$ *is the first* $n_1 \times n_1$ *diagonal sub-matrix of* $K(\boldsymbol{x})$.

As shown in Lemma 3, if we use the partially observed y_i's to estimate the GPR model, the existence of the truncated normal distribution term (i.e., $\mathbf{W}_1^{\boldsymbol{C}_\beta}$) in Equation (10) induces a biased estimation of the regression function. Note that the distribution becomes normal (i.e., $\mathbf{W}_1 \sim \mathcal{N}_{n_1}(\mathbf{0}, \Omega_2)$) in the GPR model for the case where y_i's are fully observed. Therefore, the usual estimation of the regression function based on the GPR model will produce inconsistent results when $\rho \neq 0$. This clearly reveals that sample-selection bias occurs in Bayes estimation of the regression function $\boldsymbol{\eta}_{n_1}$. The magnitude of this bias is as follows.

Corollary 2. *Instead of the* SGPR_N, *if the GPR model is used for estimating* $\boldsymbol{\eta}_{n_1}$ *based on observed data* $\boldsymbol{y}_{n_1}$ *then a sample-selection bias occurs in its conditional posterior mean. This bias is*

$$\mathrm{E}[\boldsymbol{\eta}_{n_1}|\boldsymbol{y}_{n_1}, \Psi] - \mathrm{E}[\boldsymbol{\eta}_{n_1}|\boldsymbol{y}_{n_1}, \sigma^2] = -\left(\frac{\rho}{\sigma}I_{n_1} + \frac{\sigma(1-\rho^2)}{\rho}\Omega_1^{*-1}\right)^{-1}\boldsymbol{\xi},$$

where $\boldsymbol{\xi} = \mathrm{E}[\mathbf{W}_1^{\boldsymbol{C}_\beta}] = (\xi_1, \ldots, \xi_{n_1})^\top$, $\xi_i = \omega_{ii}^*\phi(\beta_i; 0, \omega_{ii}^*)/[1 - \Phi(\beta_i/\sqrt{\omega_{ii}^*})]$, ω_{ii}^* *denotes the i-th diagonal element of* Ω_2^*, $\Omega_1^* = \Omega_1\big|_{\delta(\omega)=1}$, *and* $\Omega_2^* = \Omega_2\big|_{\delta(\omega)=1}$.

The sample-selection bias in calculating the marginal effect (or propensity) of a predictor can be also expected.

Corollary 3. *Suppose that $v_{ki} = x_{ki}$, where v_{ki} and x_{ki} are k-th element of $\boldsymbol{v}_i$ and $\boldsymbol{x}_i$, then difference in the marginal effect of the predictor x_{ki} on the selected observation y_i between the RSGPR model and the GPR model is*

$$\gamma_k \rho \sigma E_\omega \left[\frac{1}{\delta(\omega)} \left(\delta_2(\boldsymbol{v}_i^\top \boldsymbol{\gamma}, \omega) - \delta_1(\boldsymbol{v}_i^\top \boldsymbol{\gamma}, \omega)^2 \right) \right], \tag{11}$$

where γ_k is the k-th element of $\boldsymbol{\gamma}$,

$$\begin{aligned} b_1(\boldsymbol{v}_i^\top \boldsymbol{\gamma}, \omega) &= \delta(\omega)\phi(\alpha_i; 0, \delta(\omega)) / \left[1 - \Phi\left(\alpha_i / \sqrt{\delta(\omega)} \right) \right], \\ b_2(\boldsymbol{v}_i^\top \boldsymbol{\gamma}, \omega) &= \alpha_i \delta(\omega)\phi(\alpha_i; 0, \delta(\omega)) / \left[1 - \Phi\left(\alpha_i / \sqrt{\delta(\omega)} \right) \right], \end{aligned}$$

$\alpha_i = -\boldsymbol{v}_i^\top \boldsymbol{\gamma}$, and E_ω denotes the expectation is taken with respect to the distribution of $\omega \sim G(\omega)$.

To compare the SGPR_N model with the *GPR* model, various values of the sample-selection bias associated with the posterior mean (see, Corollary 2) and the difference in the marginal effect of the *k*-th predictor (see, Corollary 3) were calculated and are depicted in Figure 1. For the calculation, we set $\sigma = 1$, $\gamma_k = 1$, and $K_{11}(\mathbf{x}) = 0.5 I_{n_1} + 0.5 \mathbf{1}_{n_1} \mathbf{1}_{n_1}^\top / n_1$, an intra-class covariance matrix, where $\mathbf{1}_{n_1}$ is an $n_1 \times 1$ summing vector whose elements are all one. The left panel in Figure 1 is a graph of the sample-selection bias for different values of β_i and ρ. This graph shows the values of the first element of the bias vector given in Corollary 2. From the graph, we see that the sample-selection bias occurs in the GPR analysis with sample selection, and its magnitude becomes larger as the values of $|\rho|$ or β_i become larger. The sign of the bias is opposite to that of ρ. The right panel shows a graph of the difference in the marginal effect (defined by Equation (11)) as a function of α_i and ρ. This graph shows that the absolute value of the difference increases rapidly as α_i tends to have a large value, and this difference tends to be larger as the absolute value of ρ becomes larger. Furthermore, the signs of the difference and ρ are different, which is expected for the case where $\gamma_k > 0$. These panels imply that an inconsistent nonparametric regression analysis is unavoidable, provided that the GPR model is fitted to the partially observed sample-selection data. Instead, the proposed RSGPR model should be used to correct the sample-selection bias and to estimate the true marginal effect of each predictor in the regression analysis.

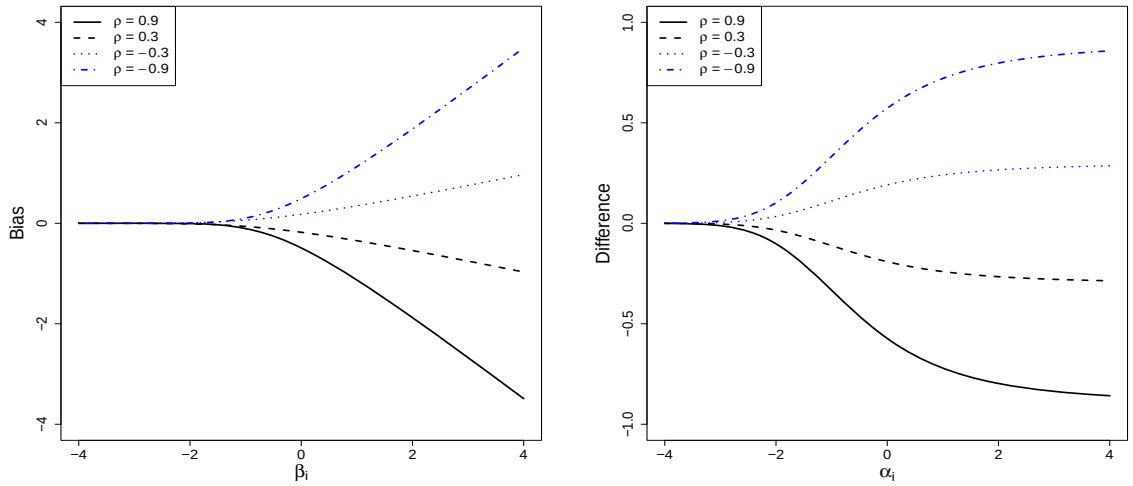

Figure 1. Graphs of the sample-selection bias and the difference in marginal effect of the *k*-th predictor.

3. Bayesian Hierarchical Methodology

3.1. Hierarchical Representation of the RSGPR Model

Let us revisit the RSGPR model (6) in Section 2.2. From Equations (7) and (8), we see that the log-likelihood function of the RSGPR model based on the partially observed n-tuples of observations $(y_i, \boldsymbol{x}_i, \boldsymbol{v}_i, s_i)$ is

$$l(\boldsymbol{\eta}_{n_1}, \boldsymbol{\gamma}, \rho, \sigma^2) = \sum_{i=1}^{n} \Big[s_i \Big\{ \ln \bar{F}(C_i; 0, 1) + \ln h(y_i | s_i = 1) \Big\} + (1 - s_i) \ln \Big\{ 1 - \bar{F}(C_i; 0, 1) \Big\} \Big]. \quad (12)$$

This is a complex function for the Bayesian estimation of the parameters ($\boldsymbol{\eta}_{n_1}$ and Ψ) of the RSGPR model. Instead, the following hierarchical representation of the RSGPR model is useful for a simple estimation of the parameters.

First, the likelihood function in Equation (12) can be represented by the following distributional hierarchy.

Theorem 1. *For the n-pairs of independent observations, (y_i, s_i), generated from the RSGPR model defined by Equation (6), their distribution can be written by the following Bayesian hierarchical model:*

$$\begin{aligned}
[y_i | \omega_i, z_{C_i}, s_i = 1] &\sim \mathcal{N}(\eta(\boldsymbol{x}_i) + \zeta z_{C_i}, \delta(\omega_i)\tau^2), \\
p(s_i | z_i, \omega_i) &= I(z_i \geq 0) I(s_i = 1) + I(z_i < 0) I(s_i = 0), \\
[z_i | \omega_i] &\sim \mathcal{N}(\boldsymbol{v}_i^\top \boldsymbol{\gamma}, \delta(\omega_i)), \\
\omega_i &\sim G(\omega_i), \ \ i = 1, \ldots, n, \\
\boldsymbol{\eta}_{n_1} &\sim \mathcal{N}_{n_1}(m_1(\boldsymbol{x}), K_{11}(\boldsymbol{x})), \\
[\zeta | \tau^2] &\sim \mathcal{N}(\theta_0, \sigma_0 \tau^2), \\
\tau^2 &\sim \mathcal{IG}(c, d), \\
\boldsymbol{\gamma} &\sim \mathcal{N}_q(\boldsymbol{\gamma}_0, \Omega_0),
\end{aligned}$$

where $z_{C_i} = z_i - \boldsymbol{v}_i^\top \boldsymbol{\gamma}$, $\zeta = \rho\sigma$, $\tau^2 = \sigma^2(1 - \rho^2)$, $\mathcal{IG}(c, d)$ denotes an inverse gamma distribution with the p.d.f. $IG(\tau^2; c, d) = d^c \tau^{-2(c+1)} e^{-d/\tau^2} / \Gamma(c)$, and $G(\cdot)$ is a distribution function of the scale mixing variable ω.

When the prior information on ξ, τ^2, and $\boldsymbol{\gamma}$ is not available, a convenient strategy of avoiding improper posterior distribution is to use proper priors with their hyperparameters fixed as appropriate quantity to reflect the diffuseness of the priors (i.e., limiting non-informative priors). For this convenience, the prior distributions in Theorem are used to elicit the prior distributions of ξ, τ^2, and $\boldsymbol{\gamma}$. All hyperparameters that appeared in the prior distributions of the Bayesian hierarchical model are assumed to be given from the prior information of previous studies or other sources.

3.2. Full Conditional Posteriors

Let $\mathbf{y}_{n_1} = (y_1, \ldots, y_{n_1})^\top$, $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$, and $\mathbf{s} = (s_1, \ldots, s_n)^\top$ be observed. Further suppose that $\mathbf{z} = (z_1, \ldots, z_n)^\top$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ are the latent observation vector and the scale mixing vector, respectively. Then, based on the RSGPR model, we obtained joint posterior distribution of $\Theta = \{\boldsymbol{\eta}_{n_1}, \tau^2, \zeta, \boldsymbol{\gamma}, \mathbf{z}, \boldsymbol{\omega}\}$ given the observed data set $\mathcal{D}_n = \{\mathbf{y}_{n_1}, \mathbf{V}, \mathbf{s}\}$:

$$\begin{aligned} p(\Theta \mid \mathcal{D}_n) \quad &\propto \quad \prod_{i=1}^{n_1} \phi(y_i; \eta(\boldsymbol{x}_i) + \zeta z_{C_i}, \delta(\omega_i)\tau^2) \\ &\times \quad \prod_{i=1}^{n} \left[p(s_i | z_i, \omega_i) \phi(z_i; \boldsymbol{v}_i^\top \boldsymbol{\gamma}, \delta(\omega_i)) g(\omega_i) \right] \\ &\times \quad IG(\tau^2; c, d) \phi(\zeta; \theta_0, \sigma_0 \tau^2) \phi_{n_1}(\boldsymbol{\eta}_{n_1}; \mathrm{m}_1(\mathbf{x}), K_{11}(\mathbf{x})) \phi_q(\boldsymbol{\gamma}; \boldsymbol{\gamma}_0, \Omega_0), \end{aligned} \tag{13}$$

where $g(\cdot)$ is the p.d.f. of the scale mixing variable ω. Note that the joint posterior in (13) is not simplified in an analytic form of the known density and is thus intractable for posterior inference. Instead, we derive conditional posterior distribution of each parameter in Θ in an explicit form, which will be useful for posterior inference by using a Markov chain Monte Carlo (MCMC) method.

Given the joint posterior distribution (13), we can obtain the following posterior distributions whose derivations are provided in Appendix A:

(1) The full conditional posterior distribution of $\boldsymbol{\eta}_{n_1}$ is given by

$$\left[\boldsymbol{\eta}_{n_1} \mid \Theta_{\setminus \boldsymbol{\eta}_{n_1}}, \mathcal{D}_n \right] \sim \mathcal{N}_{n_1}(\boldsymbol{\theta}_{\boldsymbol{\eta}_{n_1}}, \Sigma_{\boldsymbol{\eta}_{n_1}}), \tag{14}$$

where $\boldsymbol{\theta}_{\boldsymbol{\eta}_{n_1}} = \Sigma_{\boldsymbol{\eta}_{n_1}} \left(K_{11}(\mathbf{x})^{-1} m_1(\mathbf{x}) + D_1(\delta(\boldsymbol{w}))^{-1} (\mathbf{y}_{n_1} - \zeta \mathbf{z}_C)/\tau^2 \right)$, $\Sigma_{\boldsymbol{\eta}_{n_1}} = \left(K_{11}(\mathbf{x})^{-1} + D_1(\delta(\boldsymbol{w}))^{-1}/\tau^2 \right)^{-1}$, $D_1(\delta(\boldsymbol{\omega})) = \mathrm{diag}\{\delta(\omega_1), \ldots, \delta(\omega_{n_1})\}$, $\mathbf{z}_C = (z_{C_1}, \ldots, z_{C_{n_1}})^\top$, and $z_{C_i} = z_i - \boldsymbol{v}_i^\top \boldsymbol{\gamma}$.

(2) The full conditional posterior distribution of τ^2 is an inverse Gamma distribution:

$$\left[\tau^2 \mid \Theta_{\setminus \tau^2}, \mathcal{D}_n \right] \sim \mathcal{IG}\left(c + \frac{n_1 + 1}{2},\ d + \frac{1}{2} \sum_{i=1}^{n_1} \frac{(y_i - \eta(\boldsymbol{x}_i) - \zeta z_{C_i})^2}{\delta(\omega_i)} + \frac{(\zeta - \theta_0)^2}{2\sigma_0} \right). \tag{15}$$

(3) The full conditional posterior distribution of ζ is a normal distribution:

$$\left[\zeta \mid \Theta_{\setminus \zeta}, \mathcal{D}_n \right] \sim \mathcal{N}(\theta_\zeta, \sigma_\zeta^2), \tag{16}$$

where

$$\theta_\zeta = \frac{\theta_0/\sigma_0 + \sum_{i=1}^{n_1} (y_i - \eta(\boldsymbol{x}_i)) z_{C_i}/\delta(\omega_i)}{1/\sigma_0 + \sum_{i=1}^{n_1} z_{C_i}^2/\delta(\omega_i)} \text{ and } \sigma_\zeta^2 = \left(\frac{1}{\sigma_0 \tau^2} + \frac{\sum_{i=1}^{n_1} z_{C_i}^2}{\delta(\omega_i)\tau^2} \right)^{-1}.$$

(4) The full conditional posterior distributions of z_i's are independent and their distributions are given by

$$\left[z_i | \Theta_{\setminus z_i}, \mathcal{D}_n \right] \overset{ind}{\sim} \begin{cases} \mathcal{TN}_{(-\infty,0)}(\boldsymbol{v}_i^\top \boldsymbol{\gamma}, \delta(\omega_i)) & \text{if } s_i = 0, \\ \mathcal{TN}_{(0,\infty)}(\theta_{z_i}, \sigma_{z_i}^2) & \text{if } s_i = 1 \end{cases} \tag{17}$$

for $i = 1, \ldots, n$, where

$$\theta_{z_i} = \boldsymbol{v}_i^\top \boldsymbol{\gamma} + \frac{\zeta(y_i - \eta(\boldsymbol{x}_i))}{\zeta^2 + \tau^2} \text{ and } \sigma_{z_i}^2 = \frac{\delta(\omega_i)\tau^2}{\zeta^2 + \tau^2}.$$

(5) The full conditional posterior density of $\boldsymbol{\gamma}$ is:

$$\left[\boldsymbol{\gamma} \mid \Theta_{\setminus \boldsymbol{\gamma}}, \mathcal{D}_n \right] \propto \mathcal{N}_q(\boldsymbol{\theta}_\gamma, \Sigma_\gamma), \tag{18}$$

where $\boldsymbol{\theta}_\gamma = \Sigma_\gamma\Big(\sum_{i=1}^{n}\frac{1}{\delta(\omega_i)}z_i\boldsymbol{v}_i + \sum_{i=1}^{n_1}\frac{\zeta}{\tau^2\delta(\omega_i)}(\zeta z_i + \eta(\boldsymbol{x}_i) - y_i)\boldsymbol{v}_i + \Omega_0^{-1}\gamma_0\Big)$ and $\Sigma_\gamma = \big(\Omega_0^{-1} + \sum_{i=1}^{n}\frac{1}{\delta(\omega_i)}\boldsymbol{v}_i\boldsymbol{v}_i^\top + \sum_{i=1}^{n_1}\frac{\zeta^2}{\tau^2\delta(\omega_i)}\boldsymbol{v}_i\boldsymbol{v}_i^\top\big)^{-1}$.

(6) The full conditional posterior densities of ω_i's are independent and they are given by

$$\begin{aligned} p(\omega_i|\Theta_{\setminus\omega_i}, \mathcal{D}_n) &\propto \phi(y_i; \eta(\boldsymbol{x}_i) + \zeta z_{C_i}, \delta(\omega_i)\tau^2)\phi(z_i; \boldsymbol{v}_i^\top\gamma, \delta(\omega_i))g(\omega_i)\mathrm{I}(i \le n_1) \\ &+ \phi(z_i; \boldsymbol{v}_i^\top\gamma, \delta(\omega_i))g(\omega_i)\mathrm{I}(i > n_1). \end{aligned} \tag{19}$$

3.3. Markov Chain Monte Carlo Method

The MCMC scheme, working with the full conditional distributions of the parameters in Θ, is not complicated to implement. A routine Gibbs sampler can be used to generate posterior samples of $\boldsymbol{\eta}_{n_1}$, τ^2, ζ, z_i, and γ based on each of their full conditional posterior distributions obtained in Section 3.2. In posterior sampling of ω_i's, Metropolis–Hastings (M–H) within the Gibbs algorithm can be applied because their conditional posterior densities may not have explicit form of known distribution as in Equation (19). For Gibbs sampling, one should note the following points:

(1) Given the initial values of $\Theta^{(0)}$, the implementation of the Gibbs sampler involves R iterative sampling from each of the full conditional posterior distributions obtained in Equation (14) through Equation (19).

(2) Gibbs samples of ρ and σ^2 can be obtained by using those of $\zeta = \rho\sigma$ and $\tau^2 = \sigma^2(1-\rho^2)$.

(3) If ω_i degenerates at $\delta(\omega_i) = 1$, the RSGPR model can be reduced to the SGPR_N model. In this case, the MCMC procedure excludes the Gibbs sampling of ω_i's by using the posterior distribution (19).

(4) For various distributions of mixing variable ω_i and mixing functions $\delta(\omega_i)$ of the SMN distributions such as t_ν, *logit*, *stable*, *slash*, and *exponential power* models (see, e.g., [21,22]).

(5) When $\omega_i \overset{iid}{\sim} \mathcal{G}(\nu/2, \nu/2)$ and $\delta(\omega_i) = 1/\omega_i$, the RSGPR model becomes the SGPR_{t_ν} model. For generating ω_i's, we may use the following posteriors

$$\omega_i \overset{ind}{\sim} \begin{cases} \mathcal{G}\Big(\frac{\nu+2}{2}, \frac{\nu+z_{C_i}^2}{2} + \frac{(y_i-\eta(\boldsymbol{x}_i)-\zeta z_{C_i})^2}{2\zeta^2}\Big) & \text{for } i \le n_1, \\ \mathcal{G}\Big(\frac{\nu+1}{2}, \frac{\nu+z_{C_i}^2}{2}\Big) & \text{for } i > n_1, \end{cases} \tag{20}$$

where $z_{C_i} = z_i - \boldsymbol{v}_i^\top\gamma$. Except for the SGPR_N and SGPR_{t_ν}, we need to adopt the Metropolis–Hastings algorithm within the Gibbs sampler because the conditional posterior density of ω_i does not have explicit form of known distribution. See [26,27] for the algorithm for sampling ω_i from the posterior density.

(6) When the squared exponential covariance function $K(\mathbf{x})$ in Equation (4) is chosen with unknown hyperparameters u_0 and w_0, we need to elicit the priors of u_0 and w_0 for the full Bayes methods based on the MCMC method. The priors considered by [28] can be used for this assessment as follows. The prior distributions are a conjugate $u_0 \sim \mathcal{IG}(a, b)$ and $w_0 \sim \mathcal{HC}(c, d)$. Here $\mathcal{HC}(c, d)$ denotes the half-Cauchy distribution with the p.d.f. $HC(w_0; c, d)$, location parameter c, and scale parameter d. See [28], for compatibility with $w_0 \sim \mathcal{HC}(c, d)$ to elicit the prior information on w_0.

(7) Full conditional posterior distributions of u_0 and w_0 are

$$[u_0|\Theta, w_0, \mathcal{D}] \sim \mathcal{IG}(a^*, b^*) \text{ and } p(w_0|\Theta, u_0, \mathcal{D}) \propto \phi_{n_1}(\boldsymbol{\eta}_{n_1}; m_1(\mathbf{x}), K_{11}(\mathbf{x}))HC(w_0; c, d),$$

where $a^* = a + n_1/2$ and $b^* = b + u_0(\boldsymbol{\eta}_{n_1} - m_1(\mathbf{x}))^\top K_{11}(\mathbf{x})^{-1}(\boldsymbol{\eta}_{n_1} - m_1(\mathbf{x}))$. Note that the conditional posterior density of w_0 does not have explicit form of known distribution. This implies the use of the Metropolis–Hastings algorithm within the Gibbs sampler to generate w_0 from the posterior density.

(8) After obtaining the Gibbs samples of Θ, we can use them for Monte Carlo estimation of regression function $\boldsymbol{\eta}_{n_2}$ and missing observations $\mathbf{y}_{n_2}$. They can be also used for predicting regression functions and y_i's evaluated at new predictors (see, e.g., [26]).

3.4. Prediction with Bias Corrected Regression Function

According to the Gaussian (i.e., MaxEnt) process prior, the joint distribution of the training outputs $(\boldsymbol{\eta}_{n_1})$ and test outputs $(\boldsymbol{\eta}_{n_2})$ is

$$\begin{pmatrix} \boldsymbol{\eta}_{n_1} \\ \boldsymbol{\eta}_{n_2} \end{pmatrix} \Big| \mathbf{x} \sim \mathcal{N}_n \left(m(\mathbf{x}) = \begin{pmatrix} m_1(\mathbf{x}) \\ m_2(\mathbf{x}) \end{pmatrix}, K(\mathbf{x}) = \begin{bmatrix} K_{11}(\mathbf{x}) & K_{12}(\mathbf{x}) \\ K_{21}(\mathbf{x}) & K_{22}(\mathbf{x}) \end{bmatrix} \right),$$

where $\boldsymbol{\eta}_n = (\boldsymbol{\eta}_{n_1}^\top, \boldsymbol{\eta}_{n_2}^\top)^\top$, $K_{12}(\mathbf{x})$ denotes the $n_1 \times n_2$ matrix of the covariances evaluated at all pairs of training points $\{\boldsymbol{x}_i | i = 1, \ldots, n_1\}$ and test points $\{\boldsymbol{x}_j | j = n_1 + 1, \ldots, n\}$, and similarly for the other entities $K_{11}(\mathbf{x})$, $K_{21}(\mathbf{x})$, $K_{22}(\mathbf{x})$. The RSGPR framework provides a straightforward way of predicting test outputs based on the relevant test points and the training outputs. Conditioning the joint Gaussian prior distribution on the training observations, we arrive at the predictive distribution for the future (or missing) regression function given by

$$[\boldsymbol{\eta}_{n_2} | \boldsymbol{\eta}_{n_1}, \mathbf{x}] \sim \mathcal{N}_{n_2}(m_2(\mathbf{x}) + K_{21}(\mathbf{x}) K_{11}(\mathbf{x})^{-1}(\boldsymbol{\eta}_{n_1} - m_1(\mathbf{x})), K_{22}(\mathbf{x}) - K_{21}(\mathbf{x}) K_{11}(\mathbf{x})^{-1} K_{12}(\mathbf{x})). \tag{21}$$

The regression function $(\boldsymbol{\eta}_{n_2})$ value can be sampled from the predictive distribution (21) by evaluating the mean and covariance matrix of the distribution. Thus, it can be generated within the preceding MCMC algorithm for estimating the RSGPR model: We can generate $\boldsymbol{\eta}_{n_2}$ and unobserved observation vector $\mathbf{y}_{n_2} = (y_{N_1+1}, \ldots, y_n)^\top$ in the r-th iteration of the algorithm whose Markov chain is augmented by the following conditional distributions.

$$\begin{aligned} [\boldsymbol{\eta}_{n_2}^{(r)} \mid \boldsymbol{\eta}_{n_1}^{(r)}, \mathbf{x}] &\sim \mathcal{N}_{n_2}(m_2(\mathbf{x}) + K_{21}(\mathbf{x}) K_{11}(\mathbf{x})^{-1}(\boldsymbol{\eta}_{n_1}^{(r)} - m_1(\mathbf{x})), K_{22}(\mathbf{x}) - K_{21}(\mathbf{x}) K_{11}(\mathbf{x})^{-1} K_{12}(\mathbf{x})), \\ [\mathbf{y}_{n_2}^{(r)} \mid \Theta^{(r)}, \mathbf{y}_{n_1}, \mathbf{x}] &\sim \mathcal{N}(\boldsymbol{\eta}_{n_2}^{(r)}, \sigma^{2,(r)} D_2(\delta(\boldsymbol{\omega}))^{(r)}), \end{aligned}$$

where $\boldsymbol{\eta}_{n_2}^{(r)} = (\eta(\boldsymbol{x}_{n_1+1})^{(r)}, \ldots, \eta(\boldsymbol{x}_n)^{(r)})^\top$ and $D_2(\delta(\boldsymbol{\omega}))^{(r)} = \mathrm{diag}\{\delta(\omega_{n_1+1})^{(r)}, \ldots, \delta(\omega_n)^{(r)}\}$. Let $\boldsymbol{\eta}_{n_2}^{(1)}, \ldots, \boldsymbol{\eta}_{n_2}^{(R)}$ and $\mathbf{y}_{n_2}^{(1)}, \ldots, \mathbf{y}_{n_2}^{(R)}$ are respective samples generated from R iterations, then bias corrected expected value of $\boldsymbol{\eta}_{n_2}$ and that of posterior predictive distribution of $\mathbf{y}_{n_2}$ can be approximated via Monte Carlo by

$$\hat{\boldsymbol{\eta}}_{n_2} = \mathrm{E}[\boldsymbol{\eta}_{n_2} | \mathbf{x}] \approx \frac{1}{R} \sum_{r=1}^{R} \boldsymbol{\eta}_{n_2}^{(r)} \quad \text{and} \quad \mathrm{E}[\mathbf{y}_{n_2} | \mathbf{y}_{n_1}, \mathbf{x}] \approx \frac{1}{R} \sum_{r=1}^{R} \mathbf{y}_{n_2}^{(r)}.$$

Note that $\mathrm{Cov}(\boldsymbol{\eta}_{n_2} | \mathbf{x}) = K_{22}(\mathbf{x}) - K_{21}(\mathbf{x}) K_{11}(\mathbf{x})^{-1} K_{12}(\mathbf{x})$.

4. Numerical Illustrations

This section presents empirical results of the Bayesian hierarchical RSGPR analysis of non-normal data with the sample selection. We provide results obtained from simulated data applications comparing the performance of the RSGPR model with that of the GPR model. We developed our program written in R (see, e.g., [29]), which is available from the authors upon request.

4.1. Simulation Scheme

In this simulation, we evaluated the finite-sample performance of the RSGPR model by using sample-selection data generated for different sizes. The performance was assessed in terms of sample-selection bias correction and robustness to non-normal model errors. These could be measured by comparing the posterior estimation and prediction results of the RSGPR model with those based on

the GPR model. Specifically, we compared the results obtained from the SGPR_N (or $\text{SGPR}_{t_{10}}$) analysis with the results of the GPR (or $\text{GPR}_{t_{10}}$) analysis based on a partially observed sample-selection data. This study also demonstrated that the SGPR_{t_ν} model is more robust against outliers compared to the SGPR_N model. To evaluate the performance, we generated $M = 300$ sets of partially observed sample-selection data with size $n = 300$ with $n_1 = 150$ (i.e., the missing rate is 0.5) from each of the three models (see details below). The general form of the three models is as follows:

$$\begin{aligned} y_i &= \begin{cases} 50\, x_i + 5\sin(10x_i) + \epsilon_i & \text{for } s_i = 1, \quad x_i \in (0,1), \\ \text{missing} & \text{for } s_i = 0, \quad i = 1,\ldots,n, \end{cases} \\ z_i &= \gamma + \varepsilon_i, \quad \begin{pmatrix} \epsilon_i \\ \varepsilon_i \end{pmatrix} \overset{iid}{\sim} \mathcal{SMN}_2(\mathbf{0}, \Sigma, \delta, G), \end{aligned} \tag{22}$$

where $s_i = \text{I}(z_i \geq 0)$, $\gamma = 0$, $\rho = 0.5$, and $\sigma = 3$.

Model 1 was defined by assuming that the distribution G degenerates at $\omega = 1$. Model 2 was obtained from the model (22) by setting $\delta(\omega) = 1/\omega$ and $G \sim \mathcal{G}(10/2, 10/2)$. Model 3 assumed a mixture of bivariate normal errors instead of the $\mathcal{SMN}_2(\mathbf{0}, \Sigma, \delta, G)$ distribution. Throughout our simulation, the hyper-parameters for the Bayesian hierarchical model in Theorem 1 were chosen to reflect the diffuseness of the priors. To obtain the limiting non-informative priors of ζ, τ^2 and γ, their hyper-parameters were assessed as $\theta_0 = 0$, $\sigma_0 = 10$, $c = 0.001$, $d = 0.001$, $\gamma_0 = 0$, and $\Omega_0 = 10$. Note that when our observational data were augmented through proper prior information, as in this simulation study, the issue to identify the parameters in the RSGPR model disappeared.

In the simulation, we proceeded as follows to estimate the parameters. Using each of $M = 300$ datasets generated from the models (Model 1, Model 2, and Model 3), we fitted the RSGPR and GPR models and applied the proposed Bayesian hierarchical methodology to estimate the parameters of the fitted models by assuming the above prior distributions. To implement the methodology by using each generated dataset, we obtained 15,000 posterior samples from the developed MCMC algorithm (in Section 3) with 5 thinning periods after a burn-in period of 5000 samples. This sampling plan guaranteed a convergence of the chain of the MCMC algorithm. The MCMC method (applied to each of $M = 300$ datasets) gave estimates (or predictions) of the nonparametric regression function ($\eta(x)$) as well as the other parameters of the RSGPR model.

The variability in the regression function estimates ($\hat{\boldsymbol{\eta}}_{n_1}$) and predictions ($\hat{\boldsymbol{\eta}}_{n_2}$) obtained by using a dataset were then visualized as shown in Figures 2 and 3. These figures compare the estimates (or predictions) of the nonparametric regression function obtained from two models (the RSGPR and GPR models). The black line of each graph in the figures shows the true regression function of the model (22). The red dashed line denotes the posterior mean (or predicted value) of the regression function of the model (22) obtained by using a Bayesian hierarchical RSGPR analysis with the sample-selection data of size $n_1 = 150$, while the blue dashed line depicts that obtained by using a GPR analysis of the sample-selection data. The 97.5th quantile and 2.5th quantile of 3000 posterior samples (predictions) of each regression function ($\eta(x_i)$) in the RSGPR model were also calculated. In each figure, these quantiles were used to draw 95% posterior (or prediction) intervals of $\eta(x_i)$'s by using the gray band. The accuracy of parameter estimates was calculated by using the mean absolute bias (MAB) and the root mean square error (RMSE):

$$\text{MAB} = \frac{1}{M}\sum_{k=1}^{M}\sum_{\ell=1}^{p}|\hat{\theta}_{k\ell} - \theta_\ell| \quad \text{and} \quad \text{RMSE} = \left\{\frac{1}{M}\sum_{k=1}^{M}\sum_{\ell=1}^{p}(\hat{\theta}_{k\ell} - \theta_\ell)^2\right\}^{1/2},$$

where $M = 300$ and $\hat{\theta}_{k\ell}$ is the posterior estimate of ℓ-th element of $p \times 1$ parameter vector $\boldsymbol{\theta}$ in the k-th replication.

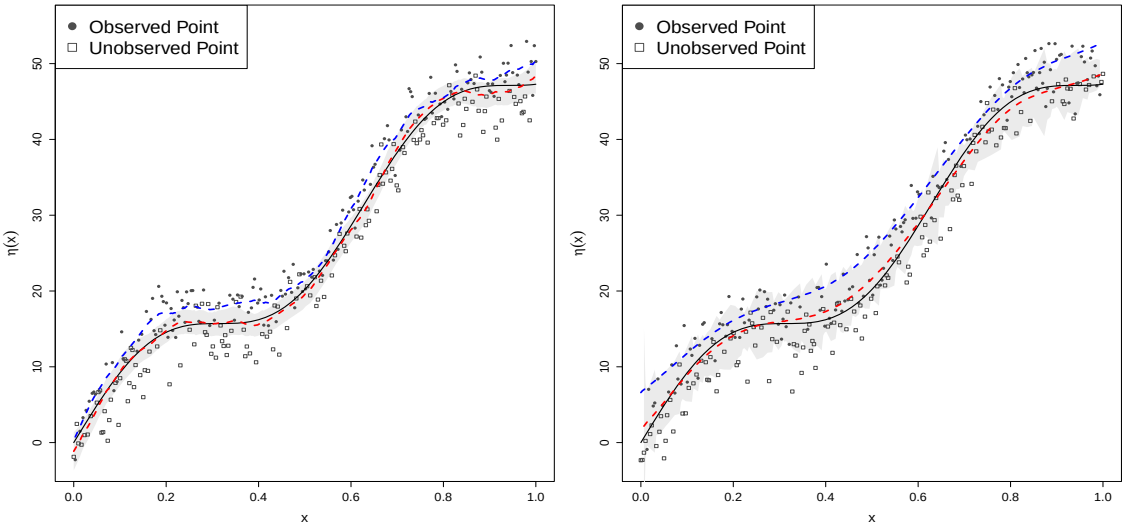

Figure 2. Graphs of estimated regression functions (**left panel**) and predicted regression functions (**right panel**): (**i**) black lines are used for the true regression function; (**ii**) red dashed lines for the robust sample-selection Gaussian process regression (RSGPR) models; (**iii**) blue dashed lines for the Gaussian process regression (GPR) models.

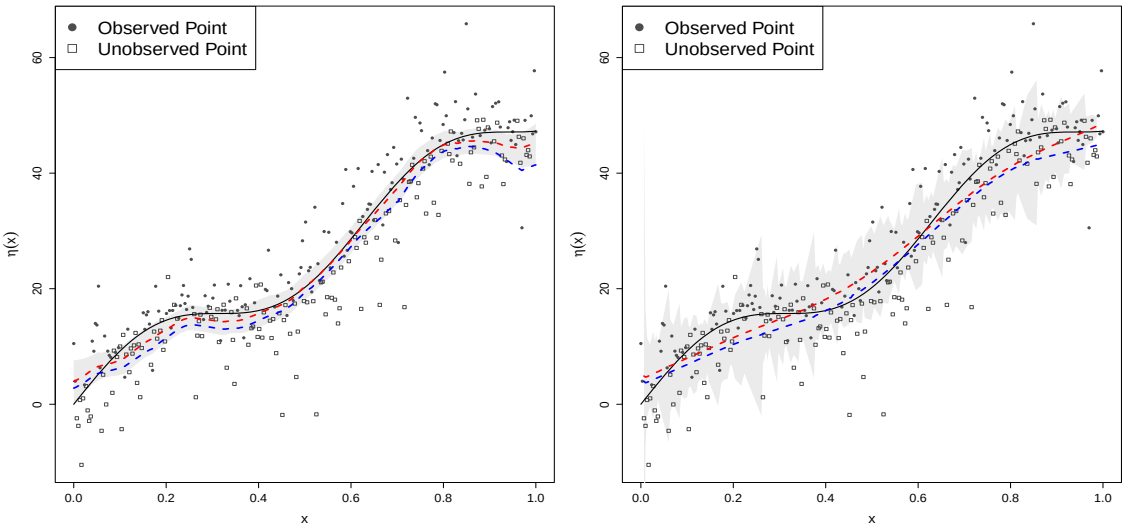

Figure 3. Graphs of regression functions: estimated regression functions (**left panel**) and predicted regression functions (**right panel**).

4.2. Performance of the RSGPR Model

4.2.1. Sample-Selection Data Generated from Model 1

If the distribution G is degenerated at $\omega = 1$, we then obtain the SGPR$_N$ model from the RSGPR model (6). Using each of $M = 300$ datasets generated from Model 1, the proposed Bayesian hierarchical methodology was applied to estimate the parameters of the model. If we set $\rho = 0$, the methodology yielded posterior estimates of the GPR model. The estimation results for the parameter $\boldsymbol{\eta}_{n_1}$ of our primary interest, based on the SGPR$_N$ and GPR models, are shown in the left panel of Figure 2. The left panel provides the following results. First, the posterior estimates of $\boldsymbol{\eta}_{n_1}$ based on the proposed SGPR$_N$ model (red dashed line) are close to their true values (black line), while those based on the GPR model (blue dashed line) tend to have severe sample-selection bias. Second, when the SGPR$_N$ model was used to fit the generated sample-selection dataset, the posterior estimates of regression

function based on the model concentrated true values of the $\eta(x_i)$'s as shown in their 95% posterior intervals (gray band). Third, the difference between the true regression function (black line) and the estimated regression function obtained by using the GPR model (the blue dashed line) confirms Lemma 4, which shows the existence of the sample-selection bias in the GPR regression for data with the sample selection. In summary, the left panel of Figure 2 illustrates the existence of the sample-selection bias in the GPR analysis with the sample-selection data, as discussed in Section 2.3. It also demonstrates the performance of the proposed methodology based on the SGPR_N model to eliminate the sample-selection bias (or inconsistency in estimating the regression function), which could not be achieved by using the GPR model.

The mean of $M = 300$ estimation results for the other parameters were listed in Table 1. As shown in Table 1, the MCMC parameter estimates were close to their true values for the SGPR_N model, while those based on the GPR model were severely biased. In addition, the MAB and RMSE values in the table ensure that the performance of the SGPR_N model is far better than the GPR model when the sample selection data was used for Bayesian nonparametric regression analysis. For both models, the small values of Monte Carlo (MC) error (compared to the RMSE) of each parameter suggests that approximate convergence was reached and the sequence generated from the MCMC samples was well mixed.

Table 1. Posterior Summary.

True Value	Mean	s.d.	SGPR_N Model		MC Error	Mean	s.d.	GPR Model		MC Error
			RMSE	MAB				RMSE	MAB	
$\sigma = 3$	2.831	0.308	0.351	0.426	0.018	2.094	0.104	0.912	0.800	0.002
$\rho = 0.5$	0.380	0.376	0.563	0.287	0.064	NA	NA	NA	NA	NA
			$\text{SGPR}_{t_{10}}$ Model					$\text{GPR}_{t_{10}}$ Model		
$\sigma = 3$	2.880	0.974	0.509	0.515	0.050	2.130	0.109	0.876	0.800	0.003
$\rho = 0.5$	0.435	0.275	0.627	0.422	0.032	NA	NA	NA	NA	NA

s.d.: standard deviation; SGPR_N: sample-selection Gaussian process normal error regression; RMSE: root mean square error; MAB: mean absolute bias; MC: Monte Carlo; GPR: Gaussian process regression.

4.2.2. Data Generated from Model 2

The proposed Bayesian hierarchical methodology for the $\text{SGPR}_{t_{10}}$ model was applied to each of $M = 300$ datasets generated from Model 2. The $\text{SGPR}_{t_{10}}$ model can be obtained from the SGPR model (6) by setting $\delta(\omega) = 1/\omega$ and $\omega \sim \mathcal{G}(10/2, 10/2)$. If we set $\rho = 0$, the methodology could also be used to obtain posterior samples to estimate the $\text{GPR}_{t_{10}}$ model (GPR model with t_{10} errors). The results of the simulation appear in the right panel of Figure 2 and Table 1. Graphs in the right panel of Figure 2 depict the prediction results of $\boldsymbol{\eta}_{n_2}$ based on the $\text{SGPR}_{t_{10}}$ and $\text{GPR}_{t_{10}}$ models. The graphs clearly show that the sample-selection bias in predicting $\eta(\boldsymbol{x}_i)$'s based on the $\text{GPR}_{t_{10}}$ model is too large to allow for a prediction of the true regression function $\boldsymbol{\eta}_{n_2}$ (or future regression function). However, the proposed methodology using the $\text{SGPR}_{t_{10}}$ model correctly predicted the true regression function; see 95% prediction interval and $\hat{\boldsymbol{\eta}}_{n_2}$, i.e., red dashed line. The prediction of $\boldsymbol{\eta}_{n_2}$ based on the $\text{SGPR}_{t_{10}}$ model is far better than that based on the $\text{GPR}_{t_{10}}$ model. Compared to the $\text{GPR}_{t_{10}}$ model, the methodology based on the $\text{SGPR}_{t_{10}}$ model yields smaller MAB and smaller RMSE of the parameter estimates; see Table 1. Table 1 shows that the parameter estimates of the $\text{SGPR}_{t_{10}}$ model with heavy-tailed errors tend to produce larger estimation errors (MAB and RMSE) than those of the SGPR_N model. The results of the above simulation demonstrate the superior performance of the $\text{SGPR}_{t_{10}}$ model and the usefulness of the proposed Bayesian hierarchical methodology to remedy the sample-selection bias in the prediction that occurred in the $\text{GPR}_{t_{10}}$ analysis of the sample-selection data.

4.2.3. Data Generated from Model 3 with Normal Mixture Errors

We generated datasets from Model 3 with size $n = 300$. Model 3 was defined by the model (22) with independent bivariate normal mixture errors: viz.

$$0.4\,\mathcal{N}_2(\mathbf{0}, \Sigma_{(1)}) + 0.2\,\mathcal{N}_2(\mathbf{0}, \Sigma_{(2)}) + 0.2\,\mathcal{N}_2(\mathbf{0}, \Sigma_{(4)}) + 0.1\,\mathcal{N}_2(\mathbf{0}, \Sigma_{(8)}) + 0.1\,\mathcal{N}_2(\mathbf{0}, \Sigma_{(16)}),$$

where 50% of the outcomes were missing in each dataset and $\Sigma_{(k)}$ was equal to Σ whose value of σ^2 was $9k$. The generated dataset was fitted to the SGPR_N, SGPR_{t_5}, and $\text{SGPR}_{t_{10}}$ models in turn. Based on posterior samples, we calculated the Bayes estimates of the three models' parameters together with their deviance information criterion (DIC) values introduced by [30]. The average DIC values obtained from the dataset were found to be 2727.06, 1477.43, 1396.24 for the SGPR_N, $\text{SGPR}_{t_{10}}$, and SGPR_{t_5} models, respectively. This suggested that the SGPR_{t_5} model is the best fitting model among the three models, while the SGPR_N model is the worst.

The graphs in the left panel of Figure 3 show the true regression function (black line) and estimated regression functions ($\hat{\boldsymbol{\eta}}_{n_1}$) under the best fitting SGPR_{t_5} model (red line) and the SGPR_N model (blue line). The graphs in the right panel of Figure 3 depict predicted regression function ($\hat{\boldsymbol{\eta}}_{n_2}$) based on the best fitted model (in red), the SGPR_N model (in blue), and the true regression function (in black). Even though the best fitted model based on bivariate t_5 error distributions was misspecified, 95% posterior intervals (or prediction intervals) of $\eta(x_i)$ obtained from the SGPR_{t_5} model did include the true regression function values (see gray bands in Figure 3). The prediction result of the SGPR_{t_5} and SGPR_N models are very wild due to outliers generated by the normal mixture errors, while the graphs of the SGPR_{t_5} are more robust for the model misspecification.

5. Conclusions

This study considered a MaxEnt approach to develop a Bayesian nonparametric regression analysis of non-normal data with the sample selection. For this purpose, by using Boltzmann's maximum entropy theorem, we introduced a MaxEnt process regression model that reflects partial prior information for an uncertain regression function. We found that a special case of the MaxEnt regression model reduced to the well-known GPR model. Second, we generalized the GPR model to propose the RSGPR model and explored its theoretical properties. These properties showed that the new model was well-designed to correct the sample-selection bias and implement a robust GPR analysis. Third, we developed a hierarchical RSGPR model based on a stochastic representation of the RSGPR model and proposed a Bayesian hierarchical methodology for the RSGPR analysis of a non-normal data with sample selection. A simulation study showed that the finite sample performance of the proposed methodology eliminated the sample selection bias and estimated the population model parameters with robustness and high accuracy. In a comparative numerical study on the analysis of nonparametric regression models with sample selection data, we found that the estimation results using the RSGPR model outperformed those using the GPR model for both in-sample estimation and out-of-sample forecasts.

The theoretical results of the RSGPR model and the methodology for the RSGPR analysis proposed in this study have several interesting issues that are worth considering further. First, the RSGPR framework using the MaxEnt process prior can be generalized to the so called *stochastically constrained RSGPR regression* that uses the constrained MaxEnt process as the prior distribution of the regression function with uncertain constraints. Second, an empirical study with real data as well as an asymptotic evaluation, such as consistency, would be particularly noteworthy to explore. For example, estimating monotone regression function with or without uncertainty and testing the monotonicity of the regression function can be considered in the context of a constrained RSGPR analysis with sample-selection data. Finally, the Bayesian hierarchical methodology can be broadened in various regression models with the general class of skew-$\mathcal{SMN}$ error distributions considered by [11]. For example, this methodology can be applied to a von Bertalanffy growth curve analysis

of heavy-tailed fishery data with sample selection (see, e.g., [28]). We hope to address all of these in the future.

Acknowledgments: Research of Hea-Jung Kim was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01057106).

Author Contributions: Hea-Jung Kim: Initiated project plan, led research effort, worked theoretical part, wrote software code, wrote paper. Daehwa Jin: Collected majority of data, wrote software code, analyzed data, contributed to writing of paper. Mihyang Bae: Wrote software code, analyzed data, contributed to writing of paper.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

Appendix A

Appendix A.1. Proof of Lemma 1

Proof. The proof proceeds along the lines of Corollary 1 of [18] by changing the partial prior information on $\boldsymbol{\theta}$ to that on $\boldsymbol{\eta}_n(\mathbf{x})$. □

Appendix A.2. Proof of Lemma 2

Proof. Equation (8) shows that the distribution of $[y_i|\eta(\boldsymbol{x}_i), s_i = 1]$ is skew-$\mathcal{SMN}_p(C_i;\boldsymbol{\theta}_i,\Sigma,\kappa,G)$ with the density (8). Thus, the result by [13] yields a stochastic representation of a conditional skew-$\mathcal{SMN}_p(C_i;\boldsymbol{\theta},\Sigma,\kappa,G)$ distribution given on ω, which is

$$[y_i|\omega,\eta(\boldsymbol{x}_i),s_i=1] \quad \stackrel{d}{=} \quad \eta(\boldsymbol{x}_i)+\rho\sigma Z_{C_i}+\sigma(1-\rho^2)^{1/2}U_i,$$

where U_i is independent of Z_{C_i}. Introducing $\omega \sim G(\omega)$ to the conditional stochastic representation, we have the two-stages of distributional hierarchy for the distribution of $[y_i|\eta(\boldsymbol{x}_i)s_i = 1]$. □

Appendix A.3. Proof of Lemma 3

Proof. For the RSGPR model, let $\mathbf{z}_1 = (z_1,\ldots,z_{n_1})^\top$ be the latent variables vector which corresponds to the observed vector $\mathbf{y}_{n_1}$. Then, for fixed $\boldsymbol{\eta}_{n_1}$ and ω, the joint distribution of $\mathbf{y}_{n_1}$ and $\mathbf{z}_1$ is $(\mathbf{y}_{n_1}^\top,\mathbf{z}_1^\top)^\top \sim \mathcal{N}_{2n_1}(\boldsymbol{\mu}^*,\delta(\omega)\Sigma\otimes I_{n_1})$, where $\boldsymbol{\mu}^* = (\boldsymbol{\eta}_{n_1}^\top,\boldsymbol{\mu}_1^\top)^\top$ and $\boldsymbol{\mu}_1 = (v_1^\top\gamma,\ldots,v_{n_1}^\top\gamma)^\top$. This yields the density of selected observations (i.e., $[\mathbf{y}_{n_1}|(\delta(\omega))^{-1/2}(\mathbf{z}_1-\boldsymbol{\mu}_1)\in\mathbf{C}_{\boldsymbol{\alpha}}]$ is given by

$$f(\mathbf{y}_{n_1}|\boldsymbol{\eta}_{n_1},\Psi)=\left[\phi_{n_1}\left(\mathbf{y}_1;\boldsymbol{\eta}_{n_1},\delta(\omega)\sigma^2 I_{n_1}\right)\bar{\Phi}_{n_1}\left(\mathbf{C}_{\boldsymbol{\alpha}};\frac{\rho}{\sigma}(\mathbf{y}_{n_1}-\boldsymbol{\eta}_{n_1}),(1-\rho^2)I_{n_1}\right)\right]/\bar{\Phi}_{n_1}\left(\mathbf{C}_{\boldsymbol{\alpha}};\mathbf{0},I_{n_1}\right),$$

where $\phi_p(\cdot;\boldsymbol{\mu},\Sigma)$ is the p-variate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, and $\bar{\Phi}_p(\mathbf{C};\boldsymbol{\mu},\Sigma)=\int_{\mathbf{C}}\phi_p(\mathbf{v};\boldsymbol{\mu},\Sigma)\mathrm{d}\mathbf{v}$. Applying the Gaussian process prior $p_0(\boldsymbol{\eta}_{n_1})\propto\exp\left\{-\frac{1}{2}((\boldsymbol{\eta}_{n_1}-m_1(\mathbf{x}))^\top K_{11}(\mathbf{x})^{-1}(\boldsymbol{\eta}_{n_1}-m_1(\mathbf{x}))\right\}$ to the likelihood yields a conditional posterior density:

$$p(\boldsymbol{\eta}_{n_1}|\mathbf{y}_{n_1},\Psi)=\left[\phi_{n_1}\left(\boldsymbol{\eta}_{n_1};\boldsymbol{\theta}_1,\Omega_1\right)\bar{\Phi}_{n_1}\left(\mathbf{C}_{\boldsymbol{\alpha}};\boldsymbol{\theta}_2+\Gamma\Omega^{-1}(\boldsymbol{\eta}_{n_1}-\boldsymbol{\theta}_1),\Omega_2-\Gamma\Omega_1^{-1}\Gamma^\top\right)\right]/\bar{\Phi}_{n_1}\left(\mathbf{C}_{\boldsymbol{\alpha}};\boldsymbol{\theta}_2,\Omega_2\right),$$

which is the skew-normal distribution whose properties were well developed by [13,23]. According to this literature, we can easily obtain the stochastic representation (10). □

Appendix A.4. Proof of Corollary 2

Proof. Setting $\eta(\omega)=1$ (i.e., the distribution of ω degenerates at $\omega=1$ and $\eta(\omega)=\omega$, the RSGPR model reduces to the SGPR$_N$ model. Applying Lemma 3 for the SGPR$_N$ model and the mean of a truncated normal distribution given in [31] yield the conditional posterior mean of the regression

function: $\mathrm{E}[\boldsymbol{\eta}_{n_1}|\mathbf{y}_{n_1}, \Psi] = \boldsymbol{\theta}_1 + \Gamma\Omega_2^{-1}\mathrm{E}[\mathbf{W}_1^{\mathbf{C}_\beta}]$. Difference between this posterior mean with that in the first $n_1 \times 1$ sub-vector of the Equation (5) gives the result. □

Appendix A.5. Proof of Corollary 3

Proof. Under the RSGPR model $E[y_i|\eta(\boldsymbol{x}_i), s_i = 1] = \eta(\boldsymbol{x}_i) + \rho\sigma E\Big[\delta_1(\boldsymbol{v}_i^\top\boldsymbol{\gamma}, \omega)\Big]$ by Lemma 2 and the expression (see, e.g., [31]) of $E[Z_{C_i}|\omega]$, where $\eta(\boldsymbol{x}_i)$ is the regression function and $\alpha_i = -\boldsymbol{v}_i^\top\boldsymbol{\gamma}$. A straightforward derivation of $E[y_i|\eta(\boldsymbol{x}_i), s_i = 1]$ with respect to x_{ki} gives

$$\frac{\partial E[y_i|\eta(\boldsymbol{x}_i), s_i = 1]}{\partial x_{ki}} = \frac{\partial\boldsymbol{\eta}(\boldsymbol{x}_i)}{\partial x_{ki}} + \gamma_k\rho\sigma E_\omega\Big[\frac{1}{\delta(\omega)}\Big(\delta_2(\boldsymbol{v}_i^\top\boldsymbol{\gamma}, \omega) - \delta_1(\boldsymbol{v}_i^\top\boldsymbol{\gamma}, \omega)^2\Big)\Big].$$

Comparing with $\frac{\partial E[y_i|\boldsymbol{\eta}(\boldsymbol{x}_i)]}{\partial x_{ki}} = \frac{\partial\boldsymbol{\eta}(\boldsymbol{x}_i)}{\partial x_{ki}}$ for the GPR model, we see that the expression (11) is the magnitude of the sample-selection bias in estimating the marginal effect of the k-th independent variable. □

Appendix A.6. Proof of Theorem 1

Proof. The first four stages of the distributional hierarchy reduce to the stochastic representation in Lemma 2 where marginal density of $[y_i|s_i = 1]$ is $h(y_i)$ which is given by Equation (8). We also see that the 2nd to 4th stages of the distributional hierarchy yield the probability mass function $p(s_i)$, which is

$$\int_0^\infty\int_{-\infty}^\infty p(s_i|z_i, \omega_i)\phi(z_i; \boldsymbol{v}_i^\top\boldsymbol{\gamma}, \delta(\omega_i))\mathrm{d}z_i\mathrm{d}G(\omega_i) = \bar{F}(C_i; 0, 1)^{s_i}\big(1 - \bar{F}(C_i; 0, 1)\big)^{1-s_i}.$$

This is equivalent to Equation (7). As a result, the logarithm of the joint density of the n pairs of independent observations, (y_i, s_i) under the hierarchy is equal to the right hand side of Equation (12). Thus, the first four stages of the hierarchy defines a hierarchical RSGPR model. Introducing the $\mathcal{GP}$ prior of $\boldsymbol{\eta}_{n_1}$ and priors of Ψ to elicit our prior information about them, we have the Bayesian hierarchical model. □

Appendix A.7. Derivation of Conditional Posterior Distributions

(1) Full conditional distribution of $\boldsymbol{\eta}_{n_1}$: Equation (13) states that the full conditional density of $\boldsymbol{\eta}_{n_1}$ is

$$\begin{aligned} p(\boldsymbol{\eta}_{n_1} \mid \Theta_{\setminus\boldsymbol{\eta}_{n_1}}, \mathcal{D}_n) &\propto \phi_{n_1}(\mathbf{y}_{n_1}; \boldsymbol{\eta}_{n_1} + \zeta\mathbf{z}_C, \tau^2 D_1(\delta(\boldsymbol{w})))\phi_{n_1}(\boldsymbol{\eta}_{n_1}; \mathrm{m}_1(\mathbf{x}), K_{11}(\mathbf{x})), \\ &\propto \exp\Big\{-\frac{1}{2}\boldsymbol{\eta}_{n_1}^\top\Sigma_{\boldsymbol{\eta}_{n_1}}^{-1}\boldsymbol{\eta}_{n_1} + \boldsymbol{\theta}_{\boldsymbol{\eta}_{n_1}}^\top\Sigma_{\boldsymbol{\eta}_{n_1}}^{-1}\boldsymbol{\eta}_{n_1}\Big\}, \\ &\propto \exp\Big\{-\frac{1}{2}(\boldsymbol{\eta}_{n_1} - \boldsymbol{\theta}_{\boldsymbol{\eta}_{n_1}})^\top\Sigma_{\boldsymbol{\eta}_{n_1}}^{-1}(\boldsymbol{\eta}_{n_1} - \boldsymbol{\theta}_{\boldsymbol{\eta}_{n_1}})\Big\}, \end{aligned}$$

which is the kernel of $\mathcal{N}_{n_1}(\boldsymbol{\theta}_{\boldsymbol{\eta}_{n_1}}, \Sigma_{\boldsymbol{\eta}_{n_1}})$ distribution.

(2) Full conditional distribution of τ^2: We see from Equation (13) that the full conditional posterior density is

$$\begin{aligned} p(\tau^2 \mid \Theta_{\setminus\tau^2}, \mathcal{D}_n) &\propto \prod_{i=1}^{n_1}\phi(y_i; \eta(\boldsymbol{x}_i) + \zeta z_{C_i}, \delta(\omega_i)\tau^2)IG(\tau^2; c, d)\phi(\zeta; \theta_0, \sigma_0\tau^2), \\ &\propto \tau^{-(n_1+2c+3)}\exp\Big\{-\Big(d + \frac{1}{2}\sum_{i=1}^{n_1}\frac{(y_i - \eta(\boldsymbol{x}_i) - \zeta z_{C_i})^2}{\delta(\omega_i)} + \frac{(\zeta - \theta_0)^2}{2\sigma_0}\Big)/\tau^2\Big\}. \end{aligned}$$

This is the kernel of $\mathcal{IG}\Big(c + \frac{n_1+1}{2},\ d + \frac{1}{2}\sum_{i=1}^{n_1}\frac{(y_i-\eta(\boldsymbol{x}_i)-\zeta z_{C_i})^2}{\delta(\omega_i)} + \frac{(\zeta-\theta_0)^2}{2\sigma_0}\Big)$ distribution.

(3) Full conditional distribution of ζ: Equation (13) gives the full conditional density of ζ given by

$$\begin{aligned} p(\zeta \mid \Theta_{\setminus\zeta}, \mathcal{D}_n) &\propto \prod_{i=1}^{n_1} \phi(y_i; \eta(\boldsymbol{x}_i) + \zeta z_{C_i}, \delta(\omega_i)\tau^2)\phi(\zeta; \theta_0, \sigma_0\tau^2), \\ &\propto \exp\Big\{ -\frac{\zeta^2 - 2\theta_\zeta \zeta}{2\sigma_\zeta^2} \Big\} \propto \exp\Big\{ -\frac{(\zeta - \theta_\zeta)^2}{2\sigma_\zeta^2} \Big\}, \end{aligned}$$

which is the kernel of $\mathcal{N}(\theta_\zeta, \sigma_\zeta^2)$ distribution.

(4) Full conditional distributions of z_i's: Equation (13) indicates that the full conditional posterior densities of z_is are mutually independent, and that for each i,

$$\begin{aligned} p(z_i \mid \Theta_{\setminus_i}, \mathcal{D}_n) &\propto \Big[\phi\big(y_i; \eta(\boldsymbol{x}_i) + \zeta(z_i - \boldsymbol{v}_i^\top \boldsymbol{\gamma}), \delta(\omega_i)\tau^2\big)\phi(z_i; \boldsymbol{v}_i^\top \boldsymbol{\gamma}, \delta(\omega_i))\Big]^{s_i} \Big[\phi(z_i; \boldsymbol{v}_i^\top \boldsymbol{\gamma}, \delta(\omega_i))\Big]^{1-s_i} \\ &\propto \Big[\phi(z_i; \theta_{z_i}, \sigma_{z_i}^2)]^{s_i} \Big[\phi(z_i; \boldsymbol{v}_i^\top \boldsymbol{\gamma}, \delta(\omega_i))\Big]^{1-s_i}. \end{aligned}$$

Since the support of z_i is $\{z_i; z_i \geq 0\}$ for $s_i = 1$, while $\{z_i; z_i < 0\}$ for $s_i = 0$, we have the truncated normal distributions.

(5) Full conditional distribution of $\boldsymbol{\gamma}$: The full conditional posterior density of $\boldsymbol{\gamma}$ is given by

$$\begin{aligned} p(\boldsymbol{\gamma} \mid \Theta_{\setminus\boldsymbol{\gamma}}, \mathcal{D}_n) &\propto \prod_{i=1}^{n_1} \phi\big(y_i; \eta(\boldsymbol{x}_i) + \zeta(z_i - \boldsymbol{v}_i^\top \boldsymbol{\gamma}), \delta(\omega_i)\tau^2\big) \\ &\times \phi_q(\boldsymbol{\gamma}; \boldsymbol{\gamma}_0, \Omega_0) \prod_{i=1}^{n} \phi(z_i; \boldsymbol{v}_i^\top \boldsymbol{\gamma}, \delta(\omega_i)) \\ &\propto \exp\Big\{ -\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\theta}_\gamma)^\top \Sigma_\gamma^{-1} (\boldsymbol{\gamma} - \boldsymbol{\theta}_\gamma) \Big\}, \end{aligned}$$

which is the kernel of $\mathcal{N}_q(\boldsymbol{\theta}_\gamma, \Sigma_\gamma)$ distribution.

References

1. Cox, G.; Kachergis, G.; Shiffrin, R. Gaussian process regression for trajectory analysis. In Proceedings of the Annual Meeting of the Cognitive Science Society, Sapporo, Japan, 1–4 August 2012; Volume 34.
2. Rasmussen, C.E.; Nickisch, H. Gaussian process for machine learning (gpml) toolbox. *J. Mach. Learn. Res.* **2010**, *11*, 3011–3015.
3. Liutkus, A.; Badeau, R.; Richard, G. Gaussian processes for underdetermined source separation. *IEEE Trans. Signal Process.* **2011**, *59*, 3455–3167.
4. Caywood, M.S.; Roberts, D.M.; Colombe, J.B.; Greenward, H.S.; Weiland, M.Z. Gaussian Process Regression for Predictive But Interpretable Machine Learning Models: An Example of Predicting Mental Worklosd across Tasks. *Front. Hum. Neurosci.* **2017**, *10*, 1–19.
5. Contreras-Reyes, J.E.; Arellano-Valle, R.B.; Canales, T.M. Comparing growth curves with asymmetric heavy-tailed errors: Application to the southern blue whiting (Micromesistius australis). *Fish. Res.* **2014**, *159*, 88–94.
6. Heckman, J.J. Sample selection bias as a specification error. *Econometrica* **1979**, *47*, 153–161.
7. Marchenko, Y.V.; Genton, M.G. A Heckman selection-*t* model. *J. Am. Stat. Assoc.* **2012**, *107*, 304–317.
8. Ding, P. Bayesian robust inference of sample selection using selection *t*-models. *J. Multivar. Anal.* **2014**, *124*, 451–464.
9. Hasselt, V.M. Bayesian inference in a sample selection model. *J. Econ.* **2011**, *165*, 221–232.
10. Arellano-Valle, R.B.; Contreras-Reyes, J.E.; Stehlík, M. Generalized skew-normal negentropy and its application to fish condition factor time series. *Entropy* **2017**, *19*, 528, doi:10.3390/e19100528.
11. Kim, H.-J.; Kim, H.-M. Elliptical regression models for multivariate sample-selection bias correction. *J. Korean Stat. Soc.* **2016**, *45*, 422–438.

12. Kim, H.-J. Bayesian hierarchical robust factor analysis models for partially observed sample-selection data. *J. Multivar. Anal.* **2018**, *164*, 65–82.
13. Kim, H.-J. A class of weighted multivariate normal distributions and its properties. *J. Multivar. Anal.* **2008**, *99*, 1758–1771.
14. Lenk, P.J. Bayesian inference for semiparametric regression using a Fourier representation. *J. R. Stat. Soc. Ser. B.* **1999**, *61*, 863–879.
15. Fahrmeir, L.; Kneib, T. *Bayesian Smoothing and Regression for Longitudial, Spatial and Event History Data*; Oxford Statistical Science Series; Oxford University Press: Oxford, UK, 2011; Volume 36.
16. Chakraborty, S.; Ghosh, M.; Mallick, B.K. Bayesian nonlinear regression for large p and small n problems. *J. Multivar. Anal.* **2012**, *108*, 28–40.
17. Leonard, T.; Hsu, J.S.J. *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*; Cambridge University Press: New York, NY, USA, 1999.
18. Kim, H.-J. A two-stage maximum entropy prior of location parameter with a stochastic multivariate interval constraint and its properties. *Entropy* **2016**, *18*, 188, doi:10.3390/e18050188.
19. Shi, J.; Choi, T. *Monographs on Statistics and Applied Probability, Gaussian Process Regression Analysis for Functional Data*; Chapman & Hall: London, UK, 2011.
20. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Process for Machine Learning*; The MIT Press: Cambridge, MA, USA, 2006.
21. Andrews, D.F.; Mallows, C.L. Scale mixtures of normal distributions. *J. R. Stat. Soc. Ser. B* **1974**, *36*, 99–102.
22. Lachos, V.H.; Labra, F.V.; Bolfarine, H.; Ghosh, P. Multivariate measurement error models based on scale mixtures of the skew-normal distribution. *Statistics* **2010**, *44*, 541–556.
23. Arellano-Valle, R.B.; Branco, M.D.; Genton, M.G. A unified view on skewed distributions arising from selection. *Can. J. Stat.* **2006**, *34*, 581–601.
24. Kim, H.J.; Choi, T.; Lee, S. A hierarchical Bayesian regression model for the uncertain functional constraint using screened scale mixture of Gaussian distributions. *Statistics* **2016**, *50*, 350–376.
25. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592.
26. Ntzoufras, I. *Bayesian Modeling Using WinBUGS*; Wiley: New York, NY, USA, 2009.
27. Chib, S.; Greenberg, E. Understanding the Metropolis-Hastings algorithm. *Am. Stat.* **1995**, *49*, 327–335.
28. Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **2006**, *1*, 515–534.
29. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017; ISBN 3-900051-07-0.
30. Spiegelhalter, D.; Best, N.; Carlin, B.; van der Linde, A. Bayesian measure of model complexity and fit (with discussion). *J. R. Stat. Soc. Ser. B* **2002**, *64*, 583–639.
31. Johnson, N.L.; Kotz, S.; Balakrishnan, N. *Distribution in Statistics: Continuous Univariate Distributions*, 2nd ed.; John Wiley & Son: New York, NY, USA, 1994; Volume 1.