Journal of the Korean Data & Information Science Society
2017, **28**(3), 709–720

http://dx.doi.org/10.7465/jkdi.2017.28.3.709
한국데이터정보과학회지

Bayesian smoothing under structural measurement error model with multiple covariates

Jinseub Hwang[1] · Dal Ho Kim[2]

[1] Department of Computer Science and Statistics, Daegu University
[2] Department of Statistics, Kyungpook National University

Received 17 April 2017, revised 23 May 2017, accepted 23 May 2017

Abstract

In healthcare and medical research, many important variables have a measurement error such as body mass index and laboratory data. It is also not easy to collect samples of large size because of high cost and long time required to collect the target patient satisfied with inclusion and exclusion criteria. Beside, the demand for solving a complex scientific problem has highly increased so that a semiparametric regression approach could be of substantial value solving this problem. To address the issues of measurement error, small domain and a scientific complexity, we conduct a multivariable Bayesian smoothing under structural measurement error covariate in this article. Specifically we enhance our previous model by incorporating other useful auxiliary covariates free of measurement error. For the regression spline, we use a radial basis functions with fixed knots for the measurement error covariate. We organize a fully Bayesian approach to fit the model and estimate parameters using Markov chain Monte Carlo. Simulation results represent that the method performs well. We illustrate the results using a national survey data for application.

Keywords: Bayes, multivariable, radial basis functions, small area, structural measurement error.

1. Introduction

In healthcare and medical research, there are frequently appearing variables such that heigh, weigh, blood pressure, cholesterol levels and amount of hemoglobin. These are very important but observed with measurement errors because of unknown effects. This measurement error makes complex the statistical analysis and this problem is generally called measurement error problem and the statistical models considering this error are called measurement error models (Fuller, 1987). Goo and Kim (2013) and Hwang (2015) refered about measurement error modeling. There are two versions, the first is called structural measurement error model where measurement error covariate is considered as a random variable.

[1] Assistant professor, Department of Computer Science and Statistics, Daegu University, Gyeongsangbuk-do 38453, Korea.
[2] Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. E-mail: dalkim@knu.ac.kr

And the other version is called functional measurement error model where measurement error covariate is considered as a non-random variable.

To collect many sample is not easy in healthcare and medical research because the high cost and long time are required in order to collect the target patient satisfied with inclusion and exclusion criteria. So, the research is seldom possible to collect a large enough sample to support precision of estimates for small patient group by ages and sex. Small patients group refers to the term "small area" in this context. To deal with this problem there are two of the more common small area models, the first is "Fay-Herriot model" and the other is "nested area unit level regression model" (Rao and Molina, 2015).

The real data is often too complicated to understand for the human mind and semiparametric regression models that combined the parametric and non-parametric models can reduce complex data sets. For non-parametric component there are many "smooth" functions such as a penalized spline, B-splines, natural cubic splines and a radial basis function.

For dealing with measurement error, small area and semiparametric regression, we developed Bayesian curve-fitting using penalized splines with functinoal measurement error model based on "nested area unit level regression model" (Hwang and Kim, 2010). Also, Hwang and Kim (2016) developed the multivariable version under functional measurement error model.

The purpose of this paper is to develop the semiparametric small area model under structural measurement error with multiple covariates. Especially, we use a radial basis functions for the smoothing because the truncated polynomial basis functions are often numerically non-stable when the smoothing parameter close to zero and the number of knots is large. Radial basis functions are available for the dealing with this problem (Ruppert *et al.*, 2003). For radial basis functions, we use fixed knots using a equally spaced sample quantiles of a measurement error covariate. To apply the model and estimate parameters, we conduct a hierarchical Bayesian (HB) method based on Markov chain Monte Carlo (MCMC), specifically Metropolis-Hastings (M-H) and Gibbs sampling.

We start with a overview of the model specification and notations in Section 2. In Section 3, we explain the MCMC implementation for the proposed hierarchical Bayes procedure and prove the propriety of the posterior because we use non-informative priors for some hyperparameters. Also we show full conditional distributions of all parameters in Section 3. We conduct simulation studies for checking the performance of our model based on the root mean square error (RMSE) in Section 4. Section 5 include the result of a real data analysis and we compare models based on the posterior predictive p-value (Meng, 1994), the mean logarithmic conditional predictive ordinate (Carlin and Louis, 2009) and deviance information criterion (Spiegelhalter *et al.*, 2002). Finally, we discuss results and some possible extensions of our model in Section 6.

2. Model specification and Notations

In this paper, we consider only one dimension radial basis functions (RBF) and then RBF is defined as follows

$$|x_1 - \tau_1|, |x_1 - \tau_2|, \cdots, |x_1 - \tau_k|,$$

where x_1 is a measurement error covariate, $|\ \ |$ is the function of absolute value, $\tau = (\tau_1, \tau_2, \cdots, \tau_k)^T$ are knots based on sample quantiles of measurement error covariate ($\tau_1 <$

$\tau_2 < \cdots < \tau_k$) and k is the number of knots.

We presume that there are m strata and each strata size is N_i. And let $(y_{ij}, X_{1ij}, x_{2i}, \cdots, x_{pi})$ denote the observed response and covariates of the j^{th} unit ($j = 1, 2, \cdots, N_i$) in the i^{th} stratum ($i = 1, 2, \cdots, m$), respectively. Here, we assume that covariates $x_2, x_3, \cdots, x_p$ have not a measurement error. Then we can express our superpopulation model based on the unit level nested error regression with RBF as follows

$$y_{ij} = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_p x_{pi} + \sum_{k=1}^{K} |x_{1i} - \tau_k| + u_i + e_{ij}. \tag{2.1}$$

Here a random effect u_i is the area-specific effects.

We are able to rewrite our model with structural measurement error covariate x_1 and other auxiliary covariates $x_2, x_3, \cdots, x_p$, free of measurement error, based on (2.1)

$$\begin{aligned} y_{ij} &= \boldsymbol{b}^T \boldsymbol{x}_i + \boldsymbol{\gamma}^T \boldsymbol{z}_i + u_i + e_{ij} \\ &= \theta_i + e_{ij}, \\ X_{1ij} &= x_{1i} + \eta_{ij}, \end{aligned}$$

where $\boldsymbol{b} = (b_0, b_1, \cdots, b_p)^T$, $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, \cdots, x_{pi})^T$, $\boldsymbol{z}_i = (|x_{1i} - \tau_1|, |x_{1i} - \tau_2|, \cdots, |x_{1i} - \tau_k|)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_k)^T$, e_{ij} and u_i are sampling errors and random effects with identically distributed and independent normal random variables, respectively. Also η_{ij} is the measurement error with normal distribution. We regard a structural measurement error model, so we let x_{1i} is a normal random variable. Here $\boldsymbol{z}_i$ presents a radial basis associated with the measurement error covariate x_{1i} with k-knots, $\boldsymbol{b} = (b_0, b_1, \cdots, b_p)^T$ is the vector of regression coefficients and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_K)^T$ is the vector of spline coefficients. We suppose that x_{1i}, u_i, e_{ij} and η_{ij} are mutually independent with $x_{1i} \sim N(\mu, \sigma_x^2), u_i \sim N(0, \sigma_u^2), e_{ij} \sim N(0, \sigma_e^2)$ and $\eta_{ij} \sim N(0, \sigma_\eta^2)$, respectively. Finally, we want to estimate small area means $\theta_i (= \boldsymbol{b}^T \boldsymbol{x}_i + \boldsymbol{\gamma}^T \boldsymbol{z}_i + u_i)$.

3. Hierarchical Bayesian approach to adaptive model

3.1. Hierarchical Bayesian framework

For fitting the model and estimating parameters based on sample n_i from the i^{th} stratum, we use a hierarchical Bayesian framework based on the following stages:

Stage 1. $y_{ij} = \theta_i + e_{ij}$ ($j = 1, \cdots, n_i; i = 1, \cdots, m$).

Stage 2. $\theta_i = \boldsymbol{b}^T \boldsymbol{x}_i + \boldsymbol{\gamma}^T \boldsymbol{z}_i + u_i$ ($i = 1, \cdots, m$).

Stage 3. $X_{1ij} = x_{1i} + \eta_{ij}$ ($j = 1, \cdots, n_i; i = 1, \cdots, m$).

Stage 4. $x_{1i} \sim N(\mu_x, \sigma_x^2)$.

Stage 5. $\boldsymbol{\gamma} \sim N(0, \sigma_\gamma^2 I)$.

Stage 6. $\boldsymbol{b} = (b_0, b_1, \cdots, b_p)^T$, μ_x, σ_e^2, σ_γ^2, σ_u^2, σ_η^2 and σ_x^2 are mutually independent with $\boldsymbol{b}$ and $\mu_x \sim$ uniform$(-\infty, \infty)$, respectively; $(\sigma_e^2)^{-1} \sim G(a_e, b_e)$,$(\sigma_\gamma^2)^{-1} \sim G(a_\gamma, b_\gamma)$, $(\sigma_u^2)^{-1} \sim G(a_u, b_u)$, $(\sigma_x^2)^{-1} \sim G(a_x, b_x)$ and $(\sigma_\eta^2)^{-1} \sim G(a_\eta, b_\eta)$ where $G(\alpha, \beta)$ denotes gamma distribution with rate parameter β and shape parameter α having the expression $f(x) \propto exp(-\beta x)x^{\alpha-1}$.

First, we confirm the propriety of the joint posterior because we define uniform distribution for the prior of regression coefficients $\boldsymbol{b}$ and mean parameter μ_x of x_{1i} that is a non-informative improper priors. In order to prove the propriety of the joint posterior more convenient, we factorize the full posterior by the conditional independence properties as follow

$$\begin{aligned}&\left[\boldsymbol{\theta}, \boldsymbol{x}_1, \boldsymbol{b}, \boldsymbol{\gamma}, \mu_x, \sigma_x^2, \sigma_e^2, \sigma_u^2, \sigma_\eta^2, \sigma_\gamma^2 | \boldsymbol{X}, \mathbf{y}\right] \\ &\propto \left[\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2\right]\left[\boldsymbol{\theta}|\boldsymbol{x}_1, \boldsymbol{b}, \boldsymbol{\gamma}, \sigma_u^2, \boldsymbol{X}\right]\left[\boldsymbol{X}|\boldsymbol{x}_1, \sigma_\eta^2\right]\left[\boldsymbol{x}_1|\mu_x, \sigma_x^2\right]\left[\boldsymbol{\gamma}|\sigma_\gamma^2\right][\boldsymbol{b}][\mu_x]\left[\sigma_x^2\right]\left[\sigma_e^2\right]\left[\sigma_u^2\right]\left[\sigma_\eta^2\right]\left[\sigma_\gamma^2\right].\end{aligned}$$

Theorem 3.1 Assume that a_e, a_γ, $(a_u + m/2 - p/2)$, $(a_x + m/2 - 1)$ and $(a_\eta + n_t/2 - m/2)$ are all positive where $p = rank(\boldsymbol{X}_*)$ and $n_t = \sum_{i=1}^m n_i$. Then the joint posterior is proper.

Proof: Let the basic full parameter space is $\boldsymbol{\Omega} = \left\{\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{x}_1, \mu_x, \sigma_x^2, \sigma_e^2, \sigma_u^2, \sigma_\eta^2, \sigma_\gamma^2\right\}$. And let

$$\begin{aligned} I &= \int \cdots \int p(\boldsymbol{\Omega}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{\Omega} \\ &= \int \cdots \int [\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2][\boldsymbol{\theta}|\boldsymbol{x}_1, \boldsymbol{b}, \boldsymbol{\gamma}, \sigma_u^2, \boldsymbol{X}][\boldsymbol{X}|\boldsymbol{x}_1, \sigma_\eta^2][\boldsymbol{x}_1|\mu_x, \sigma_x^2] \\ &\qquad \times [\boldsymbol{\gamma}|\sigma_\gamma^2][\boldsymbol{b}][\mu_x][\sigma_e^2][\sigma_u^2][\sigma_\eta^2][\sigma_\gamma^2][\sigma_x^2] d\boldsymbol{\Omega}. \end{aligned}$$

We prove the propriety of the joint posterior by showing $I \leq$M (any finite positive constant).

First, we integrate with respect to μ_x based on $exp\left[-(2\sigma_x^2)^{-1}\sum(x - \bar{x})^2\right] \leq 1$,

$$\begin{aligned} I_{\mu_x} &= \int [\boldsymbol{x}_1|\mu_x, \sigma_x^2][\mu_x] d\mu_x \qquad (3.1) \\ &= (\sigma_x^2)^{-\frac{m}{2}} \int exp\left[-\frac{1}{2\sigma_x^2}\sum_{i=1}^m (x_{1i} - \mu_x)^2\right] d\mu_x \\ &= (\sigma_x^2)^{-\frac{m}{2}} exp\left[-\frac{1}{2\sigma_x^2}\sum_{i=1}^m (x_{1i} - \bar{x}_1)^2\right] \int exp\left[-\frac{1}{2\sigma_x^2} m(\mu_x - \bar{x}_1)^2\right] d\mu_x \\ &\leq K_1 \cdot (\sigma_x^2)^{-\frac{m-1}{2}}. \end{aligned}$$

Here K_1 is a constant.

Second, let $\boldsymbol{X}^* = (\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_m^T)^T$, $p = rank(\boldsymbol{X}_*)$. Based on $\mathbf{w}^T(I - P_{\mathbf{X}_*})\mathbf{w} \geq 0$, we

integrate for $\boldsymbol{b}$,

$$\begin{aligned} I_{\boldsymbol{b}} &= \int [\boldsymbol{\theta}|\boldsymbol{b},\boldsymbol{\gamma},\boldsymbol{x}_1,\sigma_u^2,\mathbf{X}][\boldsymbol{b}]d\boldsymbol{b} \qquad (3.2)\\ &= (\sigma_u^2)^{-\frac{m}{2}} \int exp\left[-\frac{1}{2\sigma_u^2}\sum_{i=1}^{m}\left(\theta_i - \boldsymbol{b}^T\boldsymbol{x}_i + \boldsymbol{\gamma}^T\boldsymbol{z}_i\right)^2\right] d\boldsymbol{b}\\ &= (\sigma_u^2)^{-\frac{m}{2}} \int exp\left[-\frac{1}{2\sigma_u^2}\sum_{i=1}^{m}\left(w_i - \boldsymbol{b}^T\boldsymbol{x}_i\right)^2\right] d\boldsymbol{b}\\ &= (2\pi)^{\frac{m}{2}}(\sigma_u^2)^{-\frac{m}{2}}(\sigma_u^2)^{\frac{2}{p}}|\mathbf{X}_*^T\mathbf{X}_*|^{-\frac{1}{2}} exp\left[-\frac{1}{2\sigma_u^2}\mathbf{w}^T\left(I - P_{\mathbf{X}_*}\right)\mathbf{w}\right]\\ &\leq K_2\cdot(\sigma_u^2)^{-\frac{(m-p)}{2}}\cdot|\mathbf{X}_*^T\mathbf{X}_*|^{-\frac{1}{2}}. \end{aligned}$$

Here K_2 is a constant, $w_i = \theta_i - \boldsymbol{z}_i^T\boldsymbol{\gamma}$ and $P_{\mathbf{X}_*} = \mathbf{X}_*(\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_*^T$.

Next, we integrate with respect to $\boldsymbol{x}_1$ based on referred method by Ghosh, Sinha and Kim (2006).

$$\begin{aligned} I_{\boldsymbol{x}_1} &= \int [\mathbf{X}|\boldsymbol{x}_1,\sigma_\eta^2]|\mathbf{X}_*^T\mathbf{X}_*|^{-\frac{1}{2}}d\boldsymbol{x}_1 \qquad (3.3)\\ &\propto (\sigma_\eta^2)^{-\frac{n_t}{2}} exp\left[-\frac{1}{2\sigma_\eta^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{1ij} - \overline{X}_{1i}\right)^2\right]\int |\mathbf{X}_*^T\mathbf{X}_*|^{-\frac{1}{2}}\\ &\quad \times exp\left[-\frac{1}{2\sigma_\eta^2}\sum_{i=1}^{m} n_i(\overline{X}_{1i} - x_{1i})^2\right] d\boldsymbol{x}\\ &\leq K_3'\cdot(\sigma_\eta^2)^{-\frac{n_t-m}{2}} exp\left[-\frac{1}{2\sigma_\eta^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{1ij} - \overline{X}_{1i}\right)^2\right]\\ &\leq K_3\cdot(\sigma_\eta^2)^{-\frac{n_t-m}{2}}. \end{aligned}$$

Here K_3' and K_3 are constants and $n_t = \sum_{i=1}^{m} n_i$.

Now, we assume that a_u, a_x and a_η are all positive and integrate with respect to σ_u^2, σ_x^2 and σ_η^2 based on gamma distribution

$$I_{\sigma_u^2} = \int [\sigma_u^2](\sigma_u^2)^{-(m-p)/2}d\sigma_u^2 = \int (\sigma_u^2)^{-(m/2+a_u-p/2)-1} exp\left(-b_u/\sigma_u^2\right) d\sigma_u^2 = K_4, \qquad (3.4)$$

$$I_{\sigma_x^2} = \int [\sigma_x^2](\sigma_x^2)^{-(m-1)/2}d\sigma_x^2 = \int (\sigma_x^2)^{-(m/2+a_x-1)-1} exp\left(-b_x/\sigma_x^2\right) d\sigma_x^2 = K_5, \qquad (3.5)$$

$$I_{\sigma_\eta^2} = \int [\sigma_\eta^2](\sigma_\eta^2)^{-(n_t-m)/2}d\sigma_\eta^2 = \int (\sigma_\eta^2)^{-(n_t/2+a_\eta-m/2)-1} exp\left(-b_\eta/\sigma_\eta^2\right) d\sigma_\eta^2 = K_6. \qquad (3.6)$$

Here K_4, K_5 and K_6 are constants.

Through combining (3.1)∼(3.6), we have

$$I \leq K_1K_2K_3K_4K_5K_6 \int \cdots \int \left[\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2\right] \left[\boldsymbol{\gamma}|\sigma_\gamma^2\right] \left[\sigma_e^2\right] \left[\sigma_\gamma^2\right] d\boldsymbol{\Omega}^*, \tag{3.7}$$

where $\boldsymbol{\Omega}^* = (\boldsymbol{\Omega} - \mu_x - \boldsymbol{b} - \boldsymbol{x}_1 - \sigma_u^2 - \sigma_x^2 - \sigma_\eta^2)$.

The right term of (3.7) would be finite since the remaining components of the integrand have proper distributions when a_e and a_γ are all positive. Thus, the joint posterior I is finite and the propriety of the posterior is proved. □

3.2. Inference and diagnostic

In this model, we implement the Bayesian procedure based on MCMC numerical integration technique, in particular the M-H algorithm and Gibbs sampler because the full conditional distribution of x_{1i} is not a closed form. For all parameters $\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{x}_1, \mu_x, \sigma_x^2$, $\sigma_e^2, \sigma_u^2, \sigma_\eta^2$, we generate samples from full conditional distribution of each parameter using M-H algorithm and Gibbs sampling. The full conditional distribution for each parameter is as follow:

(1) $\left[\theta_i|\boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{x}_1, \mu_x, \sigma_x^2, \sigma_e^2, \sigma_\gamma^2, \sigma_u^2, \sigma_\eta^2, \boldsymbol{y}, \boldsymbol{X}\right]$
$\overset{iid}{\sim} N\left[(1 - C_i)\,\bar{y}_i + \left(\boldsymbol{b}^T\boldsymbol{x}_i + \boldsymbol{\gamma}^T\boldsymbol{z}_i\right) C_i, \sigma_e^2/\left(1 - C_i\right) n_i\right]$
where $C_i = \sigma_e^2/\left(n_i\sigma_u^2 + \sigma_e^2\right)$.

(2) $\left[\boldsymbol{b}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{x}_1, \mu_x, \sigma_x^2, \sigma_e^2, \sigma_\gamma^2, \sigma_u^2, \sigma_\eta^2, \boldsymbol{y}, \boldsymbol{X}\right] \sim N\left[\left(\boldsymbol{X}_*^T\boldsymbol{X}_*\right)^{-1}\boldsymbol{X}_*^T\boldsymbol{w}, \left(\boldsymbol{X}_*^T\boldsymbol{X}_*\right)^{-1}\sigma_u^2\right]$
where $\boldsymbol{X}_* = \left(\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_m^T\right)^T$, $w_i = \theta_i - \boldsymbol{\gamma}^T\boldsymbol{z}_i$, $\mathbf{w} = (w_1, \cdots, w_m)^T$.

(3) $\left[\boldsymbol{\gamma}|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{x}_1, \mu_x, \sigma_x^2, \sigma_e^2, \sigma_\gamma^2, \sigma_u^2, \sigma_\eta^2, \boldsymbol{y}, \boldsymbol{X}\right] \sim N\left[\left(\frac{I}{\sigma_\gamma^2} + \frac{\boldsymbol{Z}^T\boldsymbol{Z}}{\sigma_u^2}\right)^{-1}\frac{\boldsymbol{Z}^T}{\sigma_u^2}\mathbf{t}, \left(\frac{I}{\sigma_\gamma^2} + \frac{\boldsymbol{Z}^T\boldsymbol{Z}}{\sigma_u^2}\right)^{-1}\right]$
where $\mathbf{Z} = \begin{pmatrix} |x_{11} - \tau_1| & \cdots & |x_{11} - \tau_k| \\ \vdots & \vdots & \vdots \\ |x_{1m} - \tau_1| & \cdots & |x_{1m} - \tau_k| \end{pmatrix}$, $t_i = \theta_i - \boldsymbol{b}^T\boldsymbol{x}_i$, $\mathbf{t} = (t_1, \cdots, t_m)^T$.

(4) $\left[x_{1i}|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}, \mu_x, \sigma_x^2, \sigma_e^2, \sigma_\gamma^2, \sigma_u^2, \sigma_\eta^2, \boldsymbol{y}, \boldsymbol{X}\right]$
$\overset{iid}{\sim} exp\left\{-\frac{1}{2\sigma_u^2}(\theta_i - \boldsymbol{x}_i^T\boldsymbol{b} - \boldsymbol{z}_i^T\boldsymbol{\gamma})^2\right\}$
$\times N\left[\left(\sigma_\eta^{-2}n_i + \sigma_x^{-2}\right)^{-1}\left(\sigma_\eta^{-2}n_i\overline{X}_{1i} + \sigma_x^{-2}\mu_x\right), \left(\sigma_\eta^{-2}n_i + \sigma_x^{-2}\right)^{-1}\right]$.

(5) $\left[\mu_x|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{x}_1, \sigma_x^2, \sigma_e^2, \sigma_\gamma^2, \sigma_u^2, \sigma_\eta^2, \boldsymbol{y}, \boldsymbol{X}\right] \sim N\left(\overline{x}_1, \sigma_x^2/m\right)$.

(6) $\left[\sigma_e^{-2}|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{x}_1, \mu_x, \sigma_x^2, \sigma_\gamma^2, \sigma_u^2, \sigma_\eta^2, \boldsymbol{y}, \boldsymbol{X}\right]$
$\sim G\left[\frac{n_t}{2} + a_e, \frac{1}{2}\sum_{i=1}^m\sum_{j=1}^{n_i}\left(y_{ij} - \theta_i\right)^2 + b_e\right]$.

(7) $\left[\sigma_u^{-2}|\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\gamma},\boldsymbol{x}_1,\mu_x,\sigma_x^2,\sigma_e^2,\sigma_\gamma^2,\sigma_\eta^2,\boldsymbol{y},\boldsymbol{X}\right]$
$\sim G\left[\frac{m}{2}+a_u,\frac{1}{2}\sum_{i=1}^{m}\left(\theta_i-\boldsymbol{x}_i^T\boldsymbol{b}-\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)^2+b_u\right].$

(8) $\left[\sigma_\eta^{-2}|\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\gamma},\boldsymbol{x}_1,\mu_x,\sigma_x^2,\sigma_e^2,\sigma_\gamma^2,\sigma_u^2,\boldsymbol{y},\boldsymbol{X}\right]$
$\sim G\left[\frac{n_t}{2}+a_\eta,\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{1ij}-x_{1i}\right)^2+b_\eta\right].$

(9) $\left[\sigma_\gamma^{-2}|\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\gamma},\boldsymbol{x}_1,\mu_x,\sigma_x^2,\sigma_e^2,\sigma_u^2,\sigma_\eta^2,\boldsymbol{y},\boldsymbol{X}\right]\sim G\left[\frac{k}{2}+a_\gamma,\frac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{\gamma}+b_\gamma\right].$

(10) $\left[\sigma_x^{-2}|\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\gamma},\boldsymbol{x}_1,\mu_x,\sigma_e^2,\sigma_\gamma^2,\sigma_u^2,\sigma_\eta^2,\boldsymbol{y},\boldsymbol{X}\right]\sim G\left[\frac{m}{2}+a_x,\frac{1}{2}\sum_{i=1}^{m}(x_{1i}-\mu_x)^2+b_x\right].$

To get samples from full conditional distributions, we conduct L chains and $2d$ iterations for each chain. To diminish the effect of starting samples, from the first to d iterations of each chain are eliminated for all parameters and posterior summaries are calculated based on d remaining samples. The hierarchical Bayes estimators for small area means $\theta_1,\cdots,\theta_m$ are approximated as:

$$\begin{aligned} E\left(\theta_i|\boldsymbol{y},\boldsymbol{X}\right) &= E\left[E\left(\theta_i|\boldsymbol{x}_1,\boldsymbol{b},\boldsymbol{\gamma},\mu_x,\sigma_x^2,\sigma_e^2,\sigma_u^2,\sigma_\gamma^2,\sigma_\eta^2,\boldsymbol{y},\boldsymbol{X}\right)\right] \\ &\simeq \frac{1}{Ld}\sum_{l=1}^{L}\sum_{s=d+1}^{2d}\left[\left(1-C_i^{(lr)}\right)\overline{y}_i+\left(\boldsymbol{b}^{T(lr)}\boldsymbol{x}_i^{(lr)}+\boldsymbol{\gamma}^{T(lr)}\boldsymbol{z}_i^{(lr)}\right)C_i^{(lr)}\right] \end{aligned} \tag{3.8}$$

and the posterior variance is also estimated as:

$$\begin{aligned} V\left(\theta_i|\boldsymbol{y},\boldsymbol{X}\right) &= E\left[V\left(\theta_i|\boldsymbol{x}_1,\boldsymbol{b},\boldsymbol{\gamma},\mu_x,\sigma_x^2,\sigma_e^2,\sigma_u^2,\sigma_\gamma^2,\sigma_\eta^2,\boldsymbol{y},\boldsymbol{X}\right)\right] \\ &\quad +V\left[E\left(\theta_i|\boldsymbol{x}_1,\boldsymbol{b},\boldsymbol{\gamma},\mu_x,\sigma_x^2,\sigma_e^2,\sigma_u^2,\sigma_\gamma^2,\sigma_\eta^2,\boldsymbol{y},\boldsymbol{X}\right)\right] \\ &\simeq \frac{1}{Ld}\sum_{l=1}^{L}\sum_{s=d+1}^{2d}\left(\frac{\sigma_e^{2(lr)}}{n_i}(1-C_i^{(lr)})\right) \\ &\quad +\frac{1}{Ld}\sum_{l=1}^{L}\sum_{s=d+1}^{2d}\left[\left(1-C_i^{(lr)}\right)\overline{y}_i+\left(\boldsymbol{b}^{T(lr)}\boldsymbol{x}_i^{(lr)}+\boldsymbol{\gamma}^{T(lr)}\boldsymbol{z}_i^{(lr)}\right)C_i^{(lr)}\right]^2 \\ &\quad -\left[E(\theta_i|\boldsymbol{X},\boldsymbol{y})\right]^2. \end{aligned}$$

To check the convergence of MCMC, we calculate $\sqrt{\hat{R}_i}$ that is the estimator of a potential scale reduction factors (Gelman and Rubin, 1992). When $\sqrt{\hat{R}_i}$ is close to 1 for all θ_i, it means that the sampling is satisfied with the convergence.

For simulation studies, we calculate the root mean squared errors (RMSE) for each θ_i from independent simulations to confirm the model adequacy. Also we check the posterior predictive p-value (p), deviance information criterion (DIC) and the mean logarithmic conditional predictive ordinate ($\overline{LCPO}_1$) to confirm the performance for application. If p is close to 0.5 and $\overline{LCPO}_1$ and DIC is more small, it means that the model is better supported based on the data.

4. Simulation Studies

To compare the performance, we consider one covariate with measurement error and one covariate free of measurement error. The first true function is

$$y = 5 + 3x_1 + 4x_1^2 + 5x_2, \quad (x_1, x_2) \in [-2, 2]$$

and the second true function is

$$y = 5 + cos(2x_1) + 2exp(-16x_1^2) + 5x_2, \quad (x_1, x_2) \in [-2, 2],$$

where x_1 has a measurement error and x_2 has not a measurement error.

To get simulated data, we sequentially generate x_{1i} and x_{2i} $(i = 1, \cdots, 12)$ on [-2,2] and then we generate X_{1ij} from x_{1i} with error $\eta_{ij} \sim N(0, 0.3^2)$. And θ_i are generated from x_{1i} with random error $u_i \sim N(0, 0.1^2)$ and x_{2i} for the true function. Finally, y_{ij} are generated from θ_i with errors $e_{ij} \sim N(0, 2^2)$. We conduct independently five chains ($L = 5$) and 10,000 samples ($d = 10,000$). We set 1.0 for all hyperparameters $a_u, b_u, a_e, b_e, a_\eta, b_\eta, a_\eta, b_\eta, a_x$ and b_x and we consider one and three knots ($k = 1$ and $k = 3$). We conduct the sensitivity analysis for those hyperparameters and k, the results are not sensitive. For checking the model adequacy, we conduct 100 independent simulations to calculate RMSE as:

$$RMSE_i = \sqrt{\sum_{s=1}^{100} (\theta_i^{(s)} - \hat{\theta}_i^{(s)})^2 / 100}.$$

We compare two models based on two simulated data. For estimating small area means, we only use the measurement error covariate x_1 in Model 1. And we consider x_2 free of measurement error covariate with x_1 in Model 2. In our simulation studies, $\hat{R}_i \simeq 1$ for all θ_i and 2 models. We present the detailed results in Table 4.1 $\sim$ Table 4.4. In Table 4.1 and Table 4.2, we report the sample size, true mean (TM) and small area means for Model 1 and Model 2 based on two simulated data, respectively. And we report RMSE for each strata and overall RMSE in Table 4.3 and Table 4.4.

For the first simulated data, Model 1 with $k = 1$ and $k = 3$ have 9.167 and 8.303 overall RMSE, respectively. And Model 2 with $k = 1$ and $k = 3$ have 8.864 and 8.120 overall RMSE, respectively. Therefore, the performance of Model 2 is better than Model 1 for all k based on overall RMSE. For the second simulated data, Model 2 with $k = 1$ and $k = 3$ have 7.439 and 7.593 overall RMSE and Model 2 with $k = 1$ and $k = 3$ have 7.280 and 7.438 overall RMSE, respectively. Also, Model 2 is better than Model 1 for all k in the second simulation study.

5. Application

For application, we consider the sixth wave (2014) of the Korean National Health and Nutrition Examination Survey (KNHANES), the nationally representative sample. The KNHANES has been annually performed since 1988 by the Korea Centre for Disease Control and Prevention (KCDC). The data of KNHANES consists of demographic such as age and sex, laboratory data such as blood pressure, weight and height, life-style, family history (KCDC, 2013).

Table 4.1 Small area means for the first simulated data

i	n_i	TM	Model 1		Model 2	
			$k=1$	$k=3$	$k=1$	$k=3$
1	5	4.996	4.199	4.714	4.566	4.830
2	10	2.645	2.705	2.716	2.663	2.668
3	5	1.281	1.860	1.390	1.591	1.330
4	7	1.024	1.536	1.084	1.317	1.027
5	8	1.819	1.992	1.988	1.944	1.968
6	8	3.711	3.397	3.720	3.481	3.730
7	7	6.559	6.503	6.624	6.374	6.616
8	9	10.612	10.695	10.717	10.754	10.756
9	8	15.577	15.535	15.446	15.723	15.461
10	7	21.625	21.557	21.504	21.736	21.548
11	7	28.880	28.889	28.887	28.912	28.932
12	5	36.996	36.708	36.771	36.434	36.700

Table 4.2 Small area means for the second simulated data

i	n_i	TM	Model 1		Model 2	
			$k=1$	$k=3$	$k=1$	$k=3$
1	5	-5.657	-5.609	-5.582	-5.742	-5.709
2	10	-4.183	-4.062	-4.062	-4.058	-4.054
3	5	-2.185	-1.951	-1.971	-1.911	-1.924
4	7	0.195	0.321	0.316	0.365	0.352
5	8	2.728	2.771	2.771	2.796	2.791
6	8	6.248	5.926	5.923	5.844	5.853
7	7	8.017	7.649	7.648	7.638	7.643
8	9	8.221	8.294	8.285	8.306	8.297
9	8	9.288	9.344	9.332	9.353	9.331
10	7	10.539	10.641	10.645	10.659	10.642
11	7	12.211	12.34	12.352	12.358	12.365
12	5	14.342	14.082	14.101	14.125	14.168

Table 4.3 RMSEs for the first simulated data

i	n_i	Model 1		Model 2	
		$k=1$	$k=3$	$k=1$	$k=3$
1	5	1.037	0.741	0.846	0.730
2	10	0.484	0.477	0.518	0.516
3	5	0.859	0.661	0.832	0.776
4	7	0.758	0.559	0.691	0.605
5	8	0.634	0.603	0.662	0.616
6	8	0.695	0.573	0.664	0.563
7	7	0.700	0.675	0.701	0.596
8	9	0.709	0.691	0.675	0.623
9	8	0.709	0.725	0.675	0.651
10	7	0.838	0.858	0.795	0.772
11	7	0.830	0.837	0.767	0.772
12	5	0.914	0.903	1.038	0.900
overall		9.167	8.303	8.864	8.120

The factors affecting blood pressure are known as sex, age, smoking, obesity and consumption of sodium potassium, vitamin D and so on. In this application, we estimate blood pressure of groups stratified by smoking, gender and ages. We use systolic and diastolic blood pressure (SBP & DPB) as outcome variables and body mass index (BMI) as a measurement error covariate, respectively. And we consider amount of vitamin D (ng/mL) as other covariate that have not measurement error.

Table 4.4 RMSEs for the second simulated data

i	n_i	Model 1		Model 2	
		$k=1$	$k=3$	$k=1$	$k=3$
1	5	0.603	0.625	0.573	0.602
2	10	0.508	0.509	0.483	0.486
3	5	0.724	0.721	0.670	0.680
4	7	0.565	0.591	0.533	0.561
5	8	0.570	0.587	0.522	0.540
6	8	0.643	0.643	0.661	0.658
7	7	0.646	0.652	0.666	0.670
8	9	0.542	0.550	0.540	0.553
9	8	0.537	0.559	0.535	0.561
10	7	0.632	0.652	0.650	0.658
11	7	0.686	0.697	0.688	0.694
12	5	0.783	0.807	0.759	0.775
overall		7.439	7.593	7.280	7.438

We exclude subjects from 7,550 in 2014 based on exclusion criteria (1) over aged 40 years who should watch for hypertension (2) who had hypertension history. So, we use 956 subjects for application. We conduct independently five chains ($L = 5$) and 10,000 samples ($d = 5,000$) and we set 1.0 for all hyperparameters $a_u, b_u, a_e, b_e, a_\eta, b_\eta, a_\eta, b_\eta, a_x$ and b_x. And we consider one and three knots ($k = 1$ and $k = 3$). We conduct the sensitivity analysis for those hyperparameters and k, the results are not sensitive. We check the convergence by $\sqrt{\hat{R}_i}$ and we use $\overline{LCPO}_1$, DIC and p to confirm the performance.

In this application, $\sqrt{\hat{R}} \simeq 1$ for all θ_i and two models (Model 1: BMI only, Model 2: BMI and vitamin D. We report the sample size, small area means and standard error for each strata and we present $\overline{LCPO}_1$, DIC and p in Table 5.1 and Table 5.2.

For SBP, we can see that Model 2 with $k = 3$ has the smallest $\overline{LCPO}_1$ and DIC as 4.711 and 7651.243, respectively. But Model 1 with $k = 1$ is better than other models based on p-value as 0.460 in Table 5.1. Also, for DBP, Model 2 with $k = 1$ has the smallest $\overline{LCPO}_1$ and DIC as 4.328 and 6712.643, respectively. And Model 1 with $k = 1$ is better than other models based on p-value as 0.458 in Table 5.2. Here, the best models are different by model selection criteria and we use $\overline{LCPO}_1$ and DIC in this paper.

Based on Model 2 with three knots, smoker of 50's has the highest SBP as 119.401 in male and non-smoker of 40's has the highest SBP as 120.117 in female. Also, smoker of 40's in male and smoker 50's in female have the highest DBP as 77.829 and 77.240, respectively, based on Model 2 with one knot.

6. Concluding Remarks

We develop multivariable Bayesian smoothing based on radial basis functions under structural measurement error model with fixed knots. Based on simulation studies, we show the availability with additional auxiliary covariate and we apply the real data using our model. So we expect that our model is useful to solve for small sample sizes, a complex scientific and measurement error problems with multiple covariates in many fields as well as healthcare and medical research.

Furthermore, we are able to extend our model. First, we do not consider possibility of measurement error for auxiliary covariates (vitamin D and sodium), so we could extend

Table 5.1 The result for SBP

Gender	Smoker	Age	n_i	Model 1				Model 2			
				$k=1$		$k=3$		$k=1$		$k=3$	
				Mean	s.e.	Mean	s.e.	Mean	s.e.	Mean	s.e.
Male	No	40-49	65	116.640	1.335	116.593	1.337	117.206	1.300	117.473	1.314
		50-59	75	114.317	1.236	114.337	1.234	114.164	1.181	114.254	1.150
		60-	106	109.626	1.099	109.618	1.097	109.406	1.077	109.369	1.065
	Yes	40-49	77	115.661	1.214	115.698	1.214	116.500	1.281	116.240	1.275
		50-59	63	119.707	1.385	119.650	1.388	119.517	1.365	119.401	1.344
		60-	54	114.295	1.435	114.275	1.427	114.305	1.353	114.486	1.324
Female	No	40-49	179	120.378	0.850	120.389	0.848	120.210	0.864	120.117	0.874
		50-59	176	113.377	0.844	113.404	0.846	113.209	0.829	113.227	0.821
		60-	139	117.631	0.936	117.648	0.934	117.845	0.918	117.909	0.908
	Yes	40-49	14	115.339	2.297	115.308	2.275	112.718	2.277	112.154	2.237
		50-59	3	109.276	5.233	109.037	5.209	112.981	3.369	113.022	3.106
		60-	5	112.288	4.515	111.423	4.689	112.416	4.124	113.274	4.094
	$\overline{LCPO}_1$			4.728		4.715		4.720		4.711	
	DIC			7662.730		7663.361		7656.962		7651.243	
	p			0.460		0.458		0.447		0.449	

Table 5.2 The result for DBP

Gender	Smoker	Age	n_i	Model 1				Model 2			
				$k=1$		$k=3$		$k=1$		$k=3$	
				Mean	s.e.	Mean	s.e.	Mean	s.e.	Mean	s.e.
Male	No	40-49	65	75.580	0.910	75.564	0.912	75.751	0.918	75.649	0.934
		50-59	75	76.374	0.835	76.376	0.835	76.342	0.845	76.290	0.847
		60-	106	72.687	0.716	72.677	0.717	72.600	0.728	72.648	0.731
	Yes	40-49	77	77.650	0.854	77.663	0.852	77.829	0.862	77.946	0.867
		50-59	63	71.976	0.919	71.958	0.917	71.959	0.925	71.947	0.918
		60-	54	73.654	0.943	73.635	0.942	73.593	0.945	73.570	0.937
Female	No	40-49	179	73.730	0.564	73.739	0.565	73.723	0.565	73.706	0.563
		50-59	176	76.521	0.568	76.530	0.567	76.483	0.575	76.465	0.575
		60-	139	73.198	0.638	73.207	0.639	73.228	0.646	73.233	0.644
	Yes	40-49	14	73.500	1.486	73.474	1.483	72.871	1.605	73.017	1.599
		50-59	3	69.760	3.081	69.74	3.053	70.937	2.522	71.088	2.478
		60-	5	77.402	2.812	77.047	2.924	77.240	2.819	77.053	2.765
	$\overline{LCPO}_1$			4.341		4.331		4.328		4.331	
	DIC			6715.817		6715.951		6712.643		6713.402	
	p			0.458		0.456		4.448		4.446	

multivariable model with multi-dimensional measurement error covariates. Also, we don't consider a measurement error of outcome variable like as blood pressure and we could develop model with measurement error outcome and covariate. Next, we assume the normal distribution for outcome variable and measurement error covariate. So, we may consider other distributions and this can be extended in our model. Finally, we use fixed knots in penalized spline and we can consider free knots based on reversible jump MCMC method (Green, 1995). Additionally, in application, the best models for SBP and DBP are different by model selection criteria. So, we need to confirm the best model selection criteria for our models based on simulation study.

References

Carlin, B. O. and Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed), Chapman & Hall/CRC,

Boca Raton.
Fuller, W. A. (1987). *Measurement error models*, Dekker, New York.
Gelman, A. E. and Rubin, D. (1992). Inference from iterative simulation (with discussion). *Statistical Science*, **7**, 457–511.
Ghosh, M., Sinha, K. and Kim, D. (2006). Empirical and hierachical Bayesian estimation in finite population sampling under structural measurement error models. *Scandinavian Journal of Statistics*, **33**, 591-608.
Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **57**, 87–109.
Goo, Y. M. and Kim, D. H. (2013). Bayesian small area estimations with measurement error. *Journal of the Korean Data & Information Science Society*, **24**, 885-893.
Hwang, J. (2015). Statistical analysis of KNHANES data with measurement error models. *Journal of the Korean Data & Information Science Society*, **26**, 773-779.
Hwang, J. and Kim, D. (2010). Semiparametric Bayesian estimation under functional measurement error model. *Journal of the Korean Data & Information Science Society*, **21**, 379-385.
Hwang, J. and Kim, D. H. (2016). Multivariable Bayesian curve-fitting under functional measurement error model. *Journal of the Korean Data & Information Science Society*, **27**, 1645-1651.
Korea Centers for Disease Control & Prevention (2013). *Statistics on the fourth Korea National Health and Nutrition Examination Survey (KNHANES IV)*, Osong, Korea.
Meng, X. L. (1994). Posterior predictive p-values. *Annals of Statistics*, **22**, 1142-1160.
Rao, J. N. K. and Molina, I. (2015). *Small area estimation* (2nd ed), John Wiley & Sons Inc, New York.
Ruppert, D., Wand, M. and Carroll, R. (2003). *Semiparametric regression*, Cambridge University Press, New York.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583-616.