한국통계학회논문집
2012, 19권, 6호, 885–898

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.6.885

A Comparative Study on the Performance of Bayesian Partially Linear Models

Yoonsung Woo[a], Taeryon Choi[1,a], Wooseok Kim[a]

[a]Department of Statistics, Korea University

Abstract

In this paper, we consider Bayesian approaches to partially linear models, in which a regression function is represented by a semiparametric additive form of a parametric linear regression function and a nonparametric regression function. We make a comparative study on the performance of widely used Bayesian partially linear models in terms of empirical analysis. Specifically, we deal with three Bayesian methods to estimate the nonparametric regression function, one method using Fourier series representation, the other method based on Gaussian process regression approach, and the third method based on the smoothness of the function and differencing. We compare the numerical performance of three methods by the root mean squared error(RMSE). For empirical analysis, we consider synthetic data with simulation studies and real data application by fitting each of them with three Bayesian methods and comparing the RMSEs.

Keywords: Partially linear models, Fourier series, Gaussian process priors, smoothness, root mean squared error.

1. 서론

부분선형모형(partially linear model; Engle 등, 1986)은 반응변수를 설명하기 위한 회귀모형으로서 모수적 회귀 모형과 비모수적 회귀 모형이 결합된 준모수적 가법 회귀 모형(semiparametric additive regression model)의 하나이다. 부분선형모형은 반응변수를 설명하는 두 가지 형태의 공변량(covariate)에 대해서 하나의 공변량은 반응변수와 선형 회귀 함수를 통해 선형적으로 연결되고, 다른 하나의 공변량은 반응변수와 비모수적 회귀 함수를 통해 비선형적으로 연결되는 모형으로서 식 (1.1)과 같은 형태로 나타난다.

$$y_i = \mathbf{d}_i^T \boldsymbol{\beta} + \eta(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{1.1}$$

식 (1.1)의 모형에서 볼 수 있듯이 두 가지 형태의 공변량$(\mathbf{d}_i, \mathbf{x}_i)$이 반응변수 y_i를 설명하고 있으며, p-차원 공변량 $\mathbf{d}_i = (d_{i1}, \ldots, d_{ip})^T$에 대한 선형회귀함수와 k-차원 공변량 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$에 대한 비모수 회귀함수가 가법형태로 결합된 회귀모형을 가정한다. 특정한 도시에서의 전기 사용량에 대한 계량경제학적 모형을 설명하기 위해 Engle 등 (1986)에 의해 처음으로 제안된 부분선형모형은 해석이 용이하며 모형적합이 쉽다는 선형회귀모형의 장점과 다양한 형태의 회귀모형을 유연하게 설명할 수 있다는 비모수적 회귀모형의 장점을 동시에 갖게 되며, 또한 특정한 선형구조를 가정해야 하는 선형회귀모

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2012-0002541).

[1] Corresponding author: Associate Professor, Department of Statistics, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-701, Korea. E-mail: trchoi@korea.ac.kr

형의 단점과 공변량의 차원이 커짐에 따라 발생하는 비모수적 회귀모형의 단점(예, 차원의 저주(curse of dimensionality))을 극복할 수 있게 된다. 따라서 이러한 장점들 때문에 통계학 뿐 아니라 계량 경제학 같은 사회과학 또는 여러 가지 응용과학 분야에서 부분선형모형은 많이 사용되고 있다 (Härdle 등, 2000; Li와 Racine, 2007, 등). 부분선형모형에서의 통계적 추론은 식 (1.1)의 모형에서의 선형회귀계수 $\boldsymbol{\beta}$의 추론문제와 비모수적 회귀함수 $\eta(\cdot)$에 대한 추론문제와 아울러 회귀모형의 선택에 관한 문제에 집중되어 왔으며 전통적인 접근방식(classical approach)과 베이지안 접근방식 모두에서 다양한 준모수적(semiparametric) 회귀모형으로 확대 연구 되어 왔다 (Na와 Kim, 2002; Yu와 Ruppert, 2002; Aerts 등, 2004; Ruppert 등, 2009; Choi 등, 2009, 등 참조).

본 논문에서는 이러한 부분선형모형에 대한 베이지안 접근방식을 고려하고, 특히 베이지안 부분선형모형 가운데 많이 활용되는 세 가지 베이지안 부분선형회귀모형에 대해서 고찰하고 각 모형의 성능을 실증적 분석을 통해 비교해보도록 한다. 구체적으로는 부분선형모형의 비모수적 회귀 함수부분을 적합하기 위한 세 가지 베이지안 적합 방식 - 푸리에 급수(Fourier series)를 이용한 방식, 가우지안 확률과정 회귀모형(Gaussian process regression)을 이용한 방식, 차분(differencing)과 평활도(smoothness)를 이용한 방식 -에 대하여 각각의 적합방법에 대해서 고찰하고 실제 자료적합을 통한 실증적 분석을 실시하고 세 가지 방식의 성능을 비교하도록 한다. 이를 위하여 모의실험 자료를 바탕으로 각 모형을 적합하고 모수추정 결과를 제곱근평균제곱오차(root mean squared error; RMSE)를 기준으로 모형 간의 성능을 비교해본다. 아울러 기존의 부분선형모형에서 사용되었던 실제 사례 데이터를 이용하여 세 가지 방식을 적합해보고 결과를 비교하도록 한다. 이를 위하여 먼저 제 2절에서는 세 가지 베이지안 부분선형모형에 대해서 설명하고 각 모형에서의 모수에 대한 사후분포를 설명하고, 마르코프 체인 몬테 카를로(Markov chain Monte Carlo; MCMC)방법을 통한 사후추론에 대해서 고찰한다. 제 3절에서는 모의실험자료와 사례연구를 통해 2절에서 설명한 세 가지 부분선형모형에 대한 실증적 성능 비교분석을 실시한다. 마지막으로 제 4절에서는 본 논문의 결론과 향후 과제를 논의한다.

2. 베이지안 부분선형 모형

베이지안 부분선형회귀모형에서의 연구는 식 (1.1)의 비모수적 회귀 항 $\eta(\cdot)$에 대한 다양한 적합방식에 대한 방법론 연구에 집중되어 왔으며, 본 논문에서는 푸리에 급수를 이용한 부분선형회귀모형 (Lenk, 1999), 가우지안 확률과정 회귀모형 (Shi와 Choi, 2011)을 통한 부분선형회귀모형, 차분(differencing)과 평활도(smoothness)를 이용한 부분선형회귀모형 (Koop과 Poirier, 2004)의 세 가지 방식을 고찰하도록 한다. 이 경우, 식 (1.1)에서 설명된 바와 같이 비모수적 함수 $\eta(\cdot)$는 일반적으로 k-차원 회귀함수를 고려할 수 있으나 2절부터의 논의에서는 일차원($k = 1$) 회귀함수의 적합문제에 국한하도록 하며, 이러한 결과들은 다차원($k > 1$) 회귀 함수의 문제로 일반화될 수 있다.

2.1. 푸리에 급수를 이용한 부분선형모형

푸리에 급수(Fourier series)를 이용한 비모수적 회귀 함수의 적합에서는 주어진 함수를 푸리에 급수를 통해 표현하는 방식으로 보다 일반적으로는 직교급수확장(orthogonal series expansion)을 통한 비모수적 함수 추정방법을 고려할 수 있다. Lenk (1999)는 이러한 푸리에 급수를 통한 비모수적 회귀함수 적합방법을 베이지안 부분선형모형에 적용하였으며 구체적으로는 식 (2.1)과 같이 비모수적 회귀 항 $\eta(\cdot)$에 푸리에 급수를 사용하는 다음과 같은 모형을 제시하였다.

$$y_i = \mathbf{d}_i^T \boldsymbol{\beta} + \eta(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \ \eta(x) = \sum_{k=1}^{\infty} \theta_k \phi_k(x), \ a \le x \le b. \tag{2.1}$$

식 (2.1)에서 볼 수 있는 것처럼 $\eta(x)$는 푸리에 기저(Fourier basis)의 무한 합으로 표현된다고 가정하며 구체적으로는 다음과 같은 코사인 함수(cosine function)들을 고려한다.

$$\phi_k(x) = \left(\frac{2}{b-a}\right)^{\frac{1}{2}} \cos\left\{\pi k\left(\frac{x-a}{b-a}\right)\right\}, \quad k = 1, 2, \ldots, \quad \theta_k = \int_a^b \eta(x)\phi(x)dx.$$

Lenk (1999)에서는 베이지안 추론을 위하여 코사인 함수 $\phi_k(x)$의 계수 θ_k에 대하여 평균이 0이며 분산이 $\tau^2\exp(-\gamma c_k)$인 다음과 같은 정규사전분포 $\theta_k \sim N\{0, \tau^2\exp(-\gamma c_k)\}$를 사용하고 c_k와 γ에 대해서 기하평활기(geometric smoother)와 대수평활기(algebraic smoother)로 명명되는 두 가지 모수화(parametrization)를 고려하였으며, 본 논문의 실증적 분석에서는 $c_k = k$로 설정하고 $\gamma > 0$으로 설정하는 기하 평활기의 방식을 적용하였다. 이러한 서로 다른 모수화를 통한 분석에서는 실증적으로 큰 차이는 없었으며 아울러 이론적으로도 두 가지 모수화 하에서 베이즈 인자 일치성(Bayes factor consistency)과 같은 점근적 성질이 성립함이 증명되었다 (Choi 등, 2009). 이 경우 τ^2과 γ에 대해서는 추가적인 역감마(Inverse Gamma; IG) 사전분포 $\tau^2 \sim \text{IG}(u_0/2, v_0/2)$와 지수분포 $\gamma \sim \text{Exp}(w_0)$를 각각 사용하였다. 또한 비모수 회귀 항 $\eta(\cdot)$ 뿐만 아니라 그 이외의 주요 관심 모수 $\boldsymbol{\beta}, \sigma^2$에 대해서는 다음과 같은 정규-역감마 사전분포(normal-inverse gamma prior)를 고려하고 모수들 간에 서로 독립이라고 가정하였고,

$$\boldsymbol{\beta} \sim N(b_0, B_0), \quad \sigma^2 \sim \text{IG}\left(\frac{r_0}{2}, \frac{s_0}{2}\right). \tag{2.2}$$

추가적인 초모수(hyperparameter)들은 알려져 있다고 가정하였다. 아울러 푸리에 급수를 이용한 회귀함수의 적합을 위한 실제 구현에 있어서는 무한 합을 고려할 수 없기 때문에, η 대신 사전에 적절히 선택된 $K(< n)$개의 푸리에 기저를 사용하는 절단된(truncated) $\eta_K = \sum_{k=1}^K \theta_k\phi_k(x)$를 바탕으로 식 (2.3)과 같은 벡터형식의 표현을 사용하도록 한다.

$$\mathbf{y} = \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\Theta} + \epsilon. \tag{2.3}$$

식 (2.3)에서 $\mathbf{y}$는 n개의 관측값의 벡터, $\mathbf{D}$는 $n \times p$ 설계행렬(design matrix), Φ는 $n \times K$행렬로서 (i,k)번째 원소는 $\Phi(i,k) = \phi_k(x_i)$로 구성되며, Θ는 푸리에 계수 θ_k들로 이루어진 벡터이다. 이 경우, θ_k에 대한 정규사전분포를 $\theta_k \sim N\{0, \tau^2\exp(-\gamma c_k)\}$ 바탕으로 Θ는 평균이 0이고 공분산 행렬이 $\tau^2\Psi$인 다변량 정규분포를 따름을 알 수 있으며, Ψ는 $\exp(-\gamma c_k)$를 원소로 갖는 대각행렬(diagonal matrix)이다. 따라서 η_K에 대한 사전분포는 평균함수가 0이며 다음과 같은 공분산함수(covariance function)를 갖는 가우지안 확률과정(Gaussian process), $\eta_K \sim \text{GP}(0, C(\cdot,\cdot))$으로 유도됨을 쉽게 알 수 있으며 공분산 함수는 $\mathbf{C}(u,x) = \text{Cov}(\eta_K(u), \eta_K(x)) = \tau^2\sum_{k=1}^K \exp(-rc_k)\phi_k(u)\phi_k(x)$와 같이 주어진다. 따라서 푸리에 급수를 이용한 베이지안 비모수 회귀모형은 2.2절에서 설명할 가우지안 확률과정회귀모형(Gaussian process regression)의 특수한 경우로도 이해할 수 있다.

2.1.1. 조건부 사후분포 (Conditional Posterior Distribution)

식 (2.3)의 모형과 앞서 언급된 사전분포를 바탕으로 모수에 대한 사후분포를 유도하고 이를 바탕으로 마르코프 체인 몬테 카를로(Markov chain Monte Carlo; MCMC)방법을 통한 사후추론을 실시하도록 한다. 이 경우, 모수에 대한 결합사후분포(joint posterior distribution) 또는 주변사후분포(marginal posterior distribution)를 명시적인 형태로 계산하는 것이 어려우므로, 다른 모수들이 주어졌을 때의 조건부 사전분포(conditional posterior distribution)을 유도하고 MCMC방법 중에서 조건부 사전분포를 이용하는 깁스표집(Gibbs sampling)을 사용하도록 한다. 먼저 $\boldsymbol{\beta}, \sigma^2$와 추가적인 초모수 τ^2, γ들이 주어졌

을 때의 Θ의 조건부 사전분포는 베이지안 선형모형 또는 다변량 정규분포의 일반적인 결과 (Lindley와 Smith, 1972)의 결과들을 바탕으로 식 (2.4)와 같은 정규분포를 따름을 알 수 있다. 보다 자세한 결과들은 다음 절의 가우지안 확률과정 회귀모형에서 설명하도록 한다.

$$\begin{aligned} \Theta &\sim N(\nu_n, \Psi_n), \\ \nu_n &= \sigma^{-2}\Psi_n\Phi^T(\mathbf{y} - \mathbf{D}\boldsymbol{\beta}), \\ \Psi_n &= \left(\sigma^{-2}\Phi^T\Phi + \tau^{-2}\Psi^{-1}\right)^{-1}. \end{aligned} \tag{2.4}$$

추가적으로 관심 모수 $\boldsymbol{\beta}$와 σ^2의 경우는 정규-역 감마 사전분포를 사용했기 때문에 조건부 사전분포 역시 정규-역 감마 사전분포로 쉽게 유도 될 수 있으며 아울러 τ^2의 사후분포 역시 이와 비슷한 방식으로 유도될 수 있다. θ_k의 초모수인 γ의 경우에는 양의 값을 갖는 평활모수(smoothing parameter)의 역할을 하기 때문에 지수분포와 같은 사전분포를 고려하며, 이 경우 조건부 사후분포가 명시적인 형태로 유도되지 않기 때문에 깁스표집대신 분할 표집(slice sampling; Damien 등, 1999)을 고려하는 Lenk (1999)의 접근방법을 고려하였다. 아울러 푸리에 급수 모형을 적합함에 있어서는 푸리에 급수의 항의 갯수 K값을 결정해야한다. K값은 일종의 평활모수(smoothing parameter)의 역할을 하며, K값이 작은 경우에는 추정된 곡선이 부드럽게 (또는 평평하게, smooth) 적합되나 부정확하며 K값이 큰 경우에는 적합자체에 있어서는 좋은 결과를 보이나 추정된 곡선이 평평하지 못한 과적합(overfitting)양상을 보인다. 이러한 K값의 추정을 위해서는 reversible jump MCMC를 이용하는 방법, BIC와 같은 모형선택 기준이나 RMSE를 최소로 하는 값을 K값으로 선택하는 방법을 고려해볼 수 있으며, 본 논문에서는 RMSE를 최소로 하는 방법을 고려하였다. 예를 들어 3절에서 설명되는 traffic data의 경우에서는 RMSE를 최소로 하는 방법로 추정된 값으로 $K = 50$을 사용하였다. 이를 바탕으로 한 실증적 분석결과는 3절에 구체적으로 설명되어 있다.

2.2. 가우지안 확률과정 부분선형모형

가우지안 확률과정 회귀모형(Gaussian process regression; GPR)이란 베이지안 회귀모형에서 알려지지 않은 회귀함수의 사전분포를 가우지안 확률과정으로 사용하고 이를 바탕으로 회귀함수를 추론하는 베이지안 분석방법을 지칭한다 (Shi와 Choi, 2011). O'Hagan (1978)에서 처음으로 가우지안 사전분포를 함수공간에서의 사전분포로 사용하는 방법, 즉 가우지안 확률과정을 비모수 함수에 대한 사전분포로 사용하는 방법이 제안된 이래로, 가우지안 확률과정은 회귀모형, 분류모형 등에서 활용되어 왔고, 다양한 응용분야, 특히 베이지안 기계 학습 분야에서 많이 사용되어 왔다 (Rasmussen과 Williams, 2006). 최근에 와서는 이러한 GPR의 방법론이 더욱 확대되어서 혼합모형(mixture modeling), 다차원 또는 고차원(multidimensional or high dimensional) 자료 분석, 함수적 자료(functional data analysis) 분석등에서도 응용되고 있다 (Shi 등, 2007; Yi 등, 2011; Choi 등, 2011, 등).

GPR을 이용한 부분선형의 기본적인 모형구조는 식 (1.1)의 모형과 동일하며 이 경우, 회귀함수 $\eta(\cdot)$ 자체를 확률변수로 간주하여 임의함수(random function)에 대한 사전분포를 확률과정(stochastic process), 특히 평균함수 $\mu(\cdot)$와 적절하게 선택된 공분산 함수 $\mathbf{C}(\cdot)$를 갖는 가우지안 확률과정(Gaussian process; GP), GP$(\mu(\cdot), \mathbf{C}(\cdot, \cdot))$으로 사용하는 것이며, 이러한 사전분포를 가우지안 확률과정 사전분포라고 부르며 식 (2.5)와 같이 표현된다.

$$\eta(\cdot) \sim \mathrm{GP}(0, \mathbf{C}(\cdot, \cdot)), \quad \mathbf{C}(\cdot, \cdot) = \nu_0 \exp\left(-\frac{1}{2}\omega(x_1 - x_2)^2\right) \tag{2.5}$$

식 (2.5)의 가우지안 확률과정 사전분포 설정에서는 실제 응용문제에서 많이 사용되는 (Kennedy와 O'Hagan, 2001; Oakley와 O'Hagan, 2004; Rasmussen과 Williams, 2006, 등) 제곱지수 공분산함수

(squared exponential covariance function)을 사용하였으며 이 경우 ν_0와 ω는 공분산 함수의 초모수로서 사전에 알려지지 않는 경우가 일반적이다. 그 이외의 관심 모수 $\sigma^2, \boldsymbol{\beta}$는 $\eta(\cdot)$와 서로 독립이라고 가정하며 2.1절과 유사한 일반적인 정규-역 감마 사전분포를 사용한다.

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{B}_0), \qquad \sigma^2 \sim \text{IG}(A, B).$$

이 경우 초모수 $\mathbf{B}_0$와 A, B는 모두 알려져 있다고 가정하며, 3절의 실제 모형 적합에서는 무정보적 사전분포와 유사한 효과를 주기 위하여 매우 큰 분산을 사용하도록 하였다.

2.1절의 푸리에 급수를 사용하는 비모수 회귀모형이 가우지안 확률과정 회귀모형으로 이해될 수 있다고 언급한 바와 같이 가우지안 확률과정 회귀모형은 다양한 공분산함수의 사용을 통해 여러 가지 베이지안 회귀모형을 포함하게 된다. 예를 들어, 식 (2.5)의 제곱지수 공분산함수는 정상(stationary) 공분산함수의 대표적인 형태이며 이 외에는 Matérn 공분산함수가 많이 활용되고 있으며 2.1절의 푸리에 급수를 이용한 회귀모형의 경우에는 비정상(nonstationary) 공분산함수를 사용하고 있음을 알 수 있다. 아울러 2.3절에서 설명할 차분과 평활도를 활용한 회귀모형의 경우도 특정한 공분산 함수를 사용하는 가우지안 확률과정 회귀모형의 한 형태로 간주될 수 있음을 알 수 있다.

2.2.1. 조건부 사후분포 (Conditional Posterior Distribution)

주 관심 모수와 초모수들이 모두 주어져 있다고 가정하면, 즉, $\boldsymbol{\beta}, \sigma^2$가 주어지고 아울러 공분산 함수에서의 초모수 ν_0, ω가 주어진 경우에는, $\eta(\cdot)$함수에 대한 조건부 사후분포는 다변량 정규분포로 주어진다. 구체적으로는 n개의 공변량 $\mathbf{x} = (x_1, \ldots, x_n)^T$과 이에 대응하는 함수값 $\eta_n = (\eta(x_1), \ldots, \eta(x_n))^T$의 사후분포는 가우지안 확률과정 사전분포를 바탕으로 한 η_n에 대한 다변량 정규분포와 관측값 $\mathbf{y}$를 바탕으로하는 다변량 정규분포, 두 개의 곱에 비례하며 이는 식 (2.6)과 같은 다변량 정규분포로 정리된다.

$$\begin{aligned} p(\boldsymbol{\eta}_\mathbf{n}|\mathbf{y}, \sigma^2, \boldsymbol{\beta}, \nu_0, \omega) &\propto p\left(\mathbf{y}|\boldsymbol{\eta}_\mathbf{n}, \sigma^2, \boldsymbol{\beta}, \nu_0, \omega\right) \cdot p\left(\boldsymbol{\eta}_\mathbf{n}|\sigma^2, \boldsymbol{\beta}, \nu_0, \omega\right) \\ &= N\left(\mathbf{y} - \mathbf{D}\boldsymbol{\beta}, \sigma^2 I_n\right) \cdot N\left(\mathbf{0}, \mathbf{C_n}\right), \\ \boldsymbol{\eta}_\mathbf{n}|\mathbf{y}, \sigma^2, \boldsymbol{\beta}, \nu_0, \omega &\sim N\left(\mu_n, \Sigma_n\right) \\ \mu_n &= \mathbf{C}_n\left(\mathbf{C}_n + \sigma^2\mathbf{I_n}\right)^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\beta}) \\ \Sigma_n &= \sigma^2\mathbf{C}_n\left(\mathbf{C}_n + \sigma^2\mathbf{I}_n\right)^{-1}. \end{aligned} \tag{2.6}$$

식 (2.6)에서 $N(a; \mathbf{A})$는 평균벡터가 a이고 공분산 행렬이 $\mathbf{A}$인 정규분포를 의미하고 $\mathbf{C}_n$은 가우지안 확률과정 사전분포의 공분산함수에 의해서 결정되는 공분산 행렬로서 $\mathbf{C}_n(i, j) = \text{Cov}(\eta(x_i), \eta(x_j)) = \mathbf{C}(x_i, x_j)$을 의미하며 식 (2.5)에서 주어진 제곱지수 공분산 함수를 사용한다. 즉, η_n의 완전 조건부(full conditional) 사후분포는 평균벡터 $\mathbf{C}_n(\mathbf{C}_n + \sigma^2\mathbf{I_n})^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\beta})$와 공분산행렬 $\sigma^2\mathbf{C}_n(\mathbf{C}_n + \sigma^2\mathbf{I}_n)^{-1}$를 가지는 n-차원 정규분포가 된다. $\boldsymbol{\beta}$와 σ^2의 조건부 사후분포 역시 비슷한 방법을 통해 유도가 가능하며 그 결과는 식 (2.7)과 같이 각각 p-차원 정규분포와 역 감마 분포를 따름을 알 수 있다.

$$\begin{aligned} \boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\eta_n}, \sigma^2, \nu_0, \omega &\sim N(\mathbf{B}_1\mathbf{b}, \mathbf{B}_1), \\ \mathbf{B^{-1}} &= \frac{1}{\sigma^2}\mathbf{D}^T\mathbf{D} + \mathbf{B}_0^{-1}, \quad \mathbf{b} = \frac{1}{\sigma^2}\mathbf{D}^T(\mathbf{y} - \boldsymbol{\eta}_n), \\ \sigma^2|\mathbf{y}, \eta_n, \boldsymbol{\beta}, \nu_0, \omega &\sim \text{IG}\left(A + \frac{n}{2}, B^{-1} + \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mathbf{d}_i^T\boldsymbol{\beta} - \eta(x_i)\right)^2\right). \end{aligned} \tag{2.7}$$

식 (2.6)과 (2.7)의 조건부 사후분포 결과를 바탕으로 깁스 표집을 통한 마르코프 체인 몬테칼로 방법을 바탕으로 모수에 대한 사후추론을 실시하도록 한다. 이 경우 공분산 함수의 초모수 (ν_0, ω)가 주어졌다고 가정하였으나 실제 적합에서는 여러 가지 방식을 통한 초모수 추정을 고려하도록 한다. 본 논문에서는 초모수에 대한 주변 가능도(marginal likelihood)를 최대화하는 값으로 초모수를 추정하는 경험적 베이즈 방법(empirical Bayes method)을 사용하도록 한다. 이를 위하여, 전체 모수의 가능도(likelihood)에 주 관심모수 $(\eta_n, \boldsymbol{\beta}, \sigma^2)$의 사전분포가 곱해진 값을 주 관심모수에 대하여 적분하고 나면 식 (2.8)과 같은 주변 가능도를 얻게 되는데, 이 경우 오차항의 정규분포 가정과 주 관심모수에 대한 가우지안 확률과정 사전분포 및 정규-역감마 사전분포를 사용함으로써 주변 가능도 역시 식 (2.8)과 n-차원 정규분포를 도출할 수 있다.

$$\mathbf{y}|\sigma^2, \nu_0, \omega \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}_n + \mathbf{C}_n + \mathbf{D}\mathbf{B_0}\mathbf{D}^{\mathbf{T}}\right) \tag{2.8}$$

따라서 주변 가능도 (2.8)을 최대화하는 값 $(\hat{\nu}_0, \hat{\omega})$을 (ν_0, ω)의 추정 값으로 사용하고 $(\hat{\nu}_0, \hat{\omega})$을 식 (2.6)과 (2.7)에 대입한 조건부 사전분포를 이용하여 주 관심모수에 대한 깁스표집을 실시하도록 한다. 이러한 적합방식은 가우지안 확률과정 회귀모형을 통한 응용문제에서 많이 이용되어 왔으며 (Shi와 Wang, 2008; Choi 등, 2011; Choi와 Woo, 2012, 등) 특히 Choi와 Woo (2012)에서는 이러한 방법론을 바탕으로 가우지안 확률과정 부분선형 회귀모형에서의 모형 선택 문제에 대해서도 연구하였다. 이를 바탕으로 한 실증적 분석결과는 3절에서 논의하도록 한다.

2.3. 차분과 평활도를 이용한 부분선형모형

2.1절과 2.2절에서 각각 고찰한 푸리에 급수를 이용한 방식과 가우지안 확률과정 회귀 모형을 이용한 비모수적 회귀 함수의 베이지안 적합방식은 주로 통계적 모형에서 활용되었으며 계량 경제학적 관점에서도 대안적인 비모수적 또는 준모수적 회귀모형이 활용되어 왔다 (Yatchew, 1998; Li와 Racine, 2007, 등). Koop와 Poirier (2004)에서는 계량경제학적 관점에서의 베이지안 부분선형 모형을 제안하였으며 차분과 평활도를 이용한 비모수 함수추정을 고려하였다. Yatchew (1998)에서는 전통적 관점에서의 부분선형 회귀모형에서의 접근방식을 고찰하고 특히 비모수적 효과(nonparametric effect)를 차분 추정량(differencing estimatior)을 통해 제거하는 방식에 집중하였으며, Koop와 Poirier (2004)에서는 회귀함수의 차분과 평활도(smoothness)를 이용한 베이지안 적합방식을 고려하였다. 구체적으로는 식 (1.1)의 비모수 함수 $\eta(\cdot)$가 Lipshitz 연속조건과 같은 평활도를 만족한다고 가정하고 이를 바탕으로 사전분포를 설정하였다. 즉 $\eta(\cdot)$의 1차 도함수의 절대값이 특정한 상수 $C > 0$에 의해 제한되어 있다고 가정한다면, $|\eta(x_i) - \eta(x_{i-1})| \leq C|x_i - x_{i-1}|$, 유한구간 $a \leq x \leq b$에서의 공변량 x_i들의 갯수가 증가함에 따라 $|x_i - x_{i-1}|$은 점점 작아지고 이에 대응하는 함수 값의 차이 $|\eta(x_i) - \eta(x_{i-1})|$는 Lipschitz 연속조건으로부터 0으로 수렴할 것이라는 생각을 바탕으로 다음과 같은 차분과 평활도에 기반한 사전분포를 고려하였다. 이를 위하여 먼저 $\eta(x_i) = \eta_i$로 표기하고 $\triangle_j = x_j - x_{j-1}$, $j = 2, \ldots, n$라고 정의한다. 이를 바탕으로 식 (2.9)와 같은 차분행렬 $\mathbf{H}$를 정의하고 식 (2.10)과 같이 n-차원 정규사전분포를 정의한다.

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \triangle_2^{-1} & -\left(\triangle_2^{-1} + \triangle_3^{-1}\right) & \triangle_3^{-1} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \triangle_3^{-1} & -\left(\triangle_3^{-1} + \triangle_4^{-1}\right) & \triangle_4^{-1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \triangle_{n-1}^{-1} & -\left(\triangle_{n-1}^{-1} + \triangle_n^{-1}\right) & \triangle_n^{-1} \end{bmatrix}, \tag{2.9}$$

$$\mathbf{H}\boldsymbol{\eta}_n|\nu \sim N\left(\begin{bmatrix}\mu_1\\ 0\end{bmatrix}\begin{bmatrix}V_1 & 0\\ 0 & \nu I_{n-2}\end{bmatrix}\right) \equiv N(\boldsymbol{\mu}, \mathbf{V}), \quad \boldsymbol{\eta}_n = (\eta_1, \ldots, \eta_n)^T. \tag{2.10}$$

Koop과 Poirier (2004)에서는 식 (2.10)과 같은 분포를 X-사전분포(X-prior)라고 지칭하였으며 이 경우 $\gamma_n = \mathbf{H}\eta_n$로 재모수화(reparametrization)를 하면 식 (2.11)과 같은 행렬이 생성되며 실제 모형적합에서는 $\boldsymbol{\gamma}_n$를 이용하였다.

$$\boldsymbol{\gamma}_n = \mathbf{H}\boldsymbol{\eta}_n = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \dfrac{\eta_3 - \eta_2}{\Delta_3} - \dfrac{\eta_2 - \eta_1}{\Delta_2} \\ \dfrac{\eta_4 - \eta_3}{\Delta_4} - \dfrac{\eta_3 - \eta_2}{\Delta_3} \\ \vdots \\ \dfrac{\eta_n - \eta_{n-1}}{\Delta_n} - \dfrac{\eta_{n-1} - \eta_{n-2}}{\Delta_{n-1}} \end{bmatrix}. \tag{2.11}$$

Koop과 Poirier (2004)에서는 이러한 비모수적 함수에 대한 사전분포와 그 외의 주관심모수$(\boldsymbol{\beta}, \sigma^2)$에 대한 무정보적 사전분포(noninformative)를 바탕으로 사후분포를 유도하고 모형적합과 모형비교를 통한 사후추론을 실시하였다. 본 논문에서는 식 (2.9)–(2.11)에서 설명된 사전분포를 바탕으로 하는 부분선형모형 역시 2.2절의 가우지안 확률과정 회귀모형의 한 형태로 간주하여 2.2절에서 설명된 방식과 같은 방식을 바탕으로 사후추론을 실시한다. 구체적으로는 다음 절에서 설명되는 조건부 사후분포를 통한 깁스추출을 이용하고 식 (2.10)의 추가적인 초모수 ν에 대해서는 2.2절에서 고려되었던 경험적 베이즈 추정량을 사용하도록 한다.

2.3.1. 조건부 사후분포 (Conditional Posterior Distribution)

식 (1.1)의 부분선형모형으로부터 관측된 n개의 관측 값 $\mathbf{y}$와 식 (2.9)–(2.11)에서 설명된 $\boldsymbol{\eta}_n$에 대한 사전분포와 2.2절에서 고려한 $(\boldsymbol{\beta}, \sigma^2)$에 대한 정규-역 감마 사전분포를 사용하면, $\boldsymbol{\eta}_n$에 대한 가능도, 사전분포 및 조건부 사후분포를 식 (2.12)와 같은 벡터형식으로 표현할 수 있다.

$$\begin{aligned} \mathbf{y} &= \mathbf{D}\boldsymbol{\beta} + \mathbf{H}^{-1}\boldsymbol{\gamma}_n + \boldsymbol{\epsilon} \\ \mathbf{y}|\boldsymbol{\gamma}_n &\sim N\left(\mathbf{D}\boldsymbol{\beta} + \boldsymbol{\eta}_n, \sigma^2\mathbf{I}_n\right), \quad \boldsymbol{\eta}_n = \mathbf{H}^{-1}\boldsymbol{\gamma}_n \\ \boldsymbol{\gamma}_n &\sim N(0, \mathbf{C}_n), \quad \mathbf{C}_n = \mathbf{H}^{-1}\mathbf{V}\left(\mathbf{H}^{-1}\right)^T \\ \boldsymbol{\eta}_n|\mathbf{y}, \sigma^2, \boldsymbol{\beta}, \nu &\sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \\ \boldsymbol{\mu}_n &= \mathbf{C}_n\left(\mathbf{C}_n + \sigma^2\mathbf{I}_n\right)^{-1}(\mathbf{y} - \mathbf{D}\boldsymbol{\beta}) \\ \boldsymbol{\Sigma}_n &= \sigma^2\mathbf{C}_n\left(\mathbf{C}_n + \sigma^2\mathbf{I}_n\right)^{-1}. \end{aligned} \tag{2.12}$$

식 (2.12)의 결과에서 알 수 있듯이 $\boldsymbol{\eta}_n$의 조건부 사후분포는 2.2절의 가우지안 확률과정 회귀모형을 바탕으로 한 $\boldsymbol{\eta}_n$의 조건부 사후분포인 식 (2.6)의 결과와 같은 형태를 가짐을 알 수 있다. 즉, Koop과 Poirier (2004)의 X-사전분포 역시 특정한 비정상 공분산 함수(nonstationary covariance function)를 가우지안 확률과정 사전분포로부터 도출되었다고 이해할 수 있을 것이다. 마찬가지 방식으로 $\boldsymbol{\beta}$와 σ^2의 조건부 사후분포는 쉽게 유도가 될 수 있으며 각각 정규분포와 역 감마 분포가 됨을 알 수 있다.

Koop과 Poirier (2004)의 결과에서는 $(\boldsymbol{\beta}, \sigma^2)$에 대하여 무정보적 사전분포를 사용하여 사후추론을 실시하였으나 본 논문에서는 앞서 언급한 바와 같이 $(\boldsymbol{\beta}, \sigma^2)$에 대하여 정규-역 감마 분포를 고려하고 이 경우, 무정보적 사전분포와 유사한 효과를 주기 위하여 사전 분산을 매우 크게 설정하도록 한다.

주 관심모수의 사전분포의 초모수설정은 실제 구현에 있어서 고려해야할 사항이며 특히 $\boldsymbol{\eta}_n$의 설정에 있어서 비모수 함수의 평활도와 관련이 있는 식 (2.10)의 공분산 초모수 ν에 대한 추정은 여러 가지 다양한 방법이 고려될 수 있다. Koop과 Poirier (2004)에서는 간단한 형태의 교차타당성 입증(cross-validation)을 통하여 초모수에 대한 추정치(estimate)를 결정하였으나, 본 논문에서는 2.2절에서 설명된 바와 같은 경험적 베이즈 방법을 통한 추정치를 사용하도록 한다. 이와 관련한 보다 자세한 사항들은 다음 절의 모의 자료와 사례 연구를 통한 실증적 분석에서 다루도록 한다.

3. 실증적 분석을 통한 성능비교

3절에서는 2절에서 고찰한 세 가지 베이지안 부분선형모형의 모형적합에 대한 실증적 분석을 실시하고 이를 통하여 각 모형 간의 성능을 비교한다. 이를 위하여 모의실험 자료를 바탕으로 각 모형을 적합하고 사후추론을 통한 모형 간의 모수 추정 결과를 비교한다. 아울러 사후평균을 바탕으로 하는 부분선형 회귀함수의 추정치와 실제 회귀함수 참 값간의 비교를 위해 식 (3.1)과 같이 정의되는 제곱근 평균제곱오차(root mean squared error; RMSE)를 기준으로 사용하고 이를 통한 세 가지 모형 간의 성능을 비교해본다.

$$\begin{aligned} \text{RMSE}(\eta) &= \sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\{\eta_0(x_i) - \hat{\eta}_0(x_i)\}^2\right]}, \\ \text{RMSE}(y) &= \sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\{y - \mathbf{d}_i^T\hat{\boldsymbol{\beta}}_0 - \hat{\eta}_0(x_i)\}^2\right]}. \end{aligned} \tag{3.1}$$

식 (3.1)에서 $\boldsymbol{\beta}_0$와 $\eta_0(x)$는 각각 참인 선형회귀계수와 비모수 회귀함수를, $\hat{\boldsymbol{\beta}}_0$와 $\hat{\eta}_0(x)$는 이에 대응하는 사후평균을 나타내며 RMSE가 작을수록 더 정확한 모형적합을 의미한다. 아울러, 모의실험 자료 뿐만 아니라 기존의 부분선형모형 관련 연구에서 사용되었던 실제 사례 데이터를 이용하여 세 가지 방식을 적합해보고 모형 간의 결과를 비교하도록 한다.

3.1. 모의실험 자료를 통한 베이지안 부분선형모형 비교

모의실험자료 적합을 통한 베이지안 부분선형모형의 성능비교를 하기 위하여 식 (3.2)와 같은 부분선형모형으로부터 자료를 생성하도록 한다.

$$\begin{aligned} &y_i = d_i \times 2 + \eta(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0,1),\ i = 1,\ldots,n, \\ &x_i = \frac{2i-1}{2n}, \quad i = 1,\ldots,n, \\ &d_i = 2(x_i - 0.5) + \delta_i, \quad \delta_i \sim N\left(0, 0.5^2\right),\ i = 1,\ldots,n. \end{aligned} \tag{3.2}$$

식 (3.2)의 비모수적 회귀함수 $\eta(x_i)$에 대해서는 기존의 부분선형 모형 (Yatchew, 1998; Lenk, 1999; Koop과 Poirier, 2004)에서 사용되었던 비선형 회귀함수를 고려한다. 구체적으로, 식 (3.3)과 같이 각각 (1) double normal, (2) linexp, (3) lincos, (4) yatchew라고 지칭되는 네 가지 종류의 함수를 이용하도

Table 1: Summary results for simulation study 1

	double normal			linexp			lincos			yatchew		
	GP	Lenk	Koop	GP	Lenk	Koop	GP	Lenk	Koop	GP	Lenk	Koop
$\alpha(\alpha_0 = 0)$	0.030	0.018	5.065	1.704	0.967	2.130	0.490	0.305	1.015	0.015	0.006	0.203
$\beta(\beta_0 = 2)$	1.962	2.082	1.989	2.015	1.970	1.998	1.939	1.954	2.006	2.047	2.007	2.009
$\sigma^2(\sigma_0^2 = 1)$	0.993	1.028	0.972	1.034	0.996	1.019	0.973	0.969	0.992	1.156	1.018	1.045
RMSE (β)	0.303	0.328	0.306	0.302	0.310	0.307	0.307	0.297	0.307	0.237	0.287	0.302
RMSE (η)	2.041	0.572	5.608	2.317	1.073	3.182	0.846	0.459	2.680	0.547	0.387	2.492
RMSE (y)	0.996	1.009	0.976	1.016	0.993	0.999	0.986	0.979	0.986	1.075	1.004	1.012

록 한다.

$$
\begin{aligned}
&(1)\ \eta(x) = 5 - 10x + 8\exp\left\{-100(x-0.3)^2\right\} - 8\exp\left\{-100(x-0.7)^2\right\} \qquad (3.3)\\
&(2)\ \eta(x) = 2 - 5x + \exp\{5(x-0.6)\}\\
&(3)\ \eta(x) = x + \cos(4x)\\
&(4)\ \eta(x) = x\cos(4\pi x)
\end{aligned}
$$

구체적으로 식 (3.3)의 (1)부터 (4)까지의 비선형 함수 $\eta(x)$를 바탕으로 식 (3.2)의 부분선형모형으로부터 각각 $100(= n)$개의 자료를 생성하여 푸리에 급수를 이용한 베이지안 부분선형모형(Lenk), 가우지안 확률과정 부분선형 모형(GP), 차분과 평활도를 바탕으로 한 베이지안 부분선형모형(Koop), 세 가지 모형을 적합해보았다. β와 σ^2에 대해서는 세 가지 모형 모두 공통적으로 $\beta \sim N(0, 10)$와 $\sigma^2 \sim \text{IG}(0.01, 0.01)$인 무정보적 사전분포를 사용하였고, 그 이외의 추가적인 초모수에 대해서는 $\tau^2 \sim \text{IG}(1, 1/2)$, $\gamma \sim \text{Exp}(1)$, $V_1 = 10I$, $\mu_1 = (0, 0)^T$와 같이 설정하였다. 이러한 사전분포를 이용한 세 가지 적합 방법 하에서의 각 모수의 사후평균(posterior mean)값으로 요약되는 추정치를 구하여 실제 참 값과 비교해보고 식 (3.1)에 주어진 RMSE를 계산하여 각 모형간의 성능을 비교해보았으며 이러한 절차를 100회 반복하였다. 모형적합을 위한 MCMC 구현을 위해서 총 10,000번의 깁스표집 과정을 통해 사후표본을 추출하였으며, 이중 처음 8,000번의 사후표본을 소각(burn-in)과정으로 제외하고, 남아있는 2,000번의 사후표본을 분석에 사용하였으며 이러한 결과는 Table 1에 요약되어 있다. Table 1에서는 주 관심모수들의 사후평균값을 요약하였으며, 먼저 y절편을 포함하는 부분선형모형 $y = \alpha + \beta d + \eta(x) + \epsilon$을 고려하였다. Table 1에 요약된 결과와 같이 β의 사후평균은 참 값 ($\beta_0 = 2$)과 유사한 결과를 보였으나 y절편 α의 사후평균은 참 값($\alpha = 0$)과는 상이한 결과를 보여주고 있다. 이러한 결과들은 세 가지 베이지안 부분선형모형, GP, Lenk, Koop 모두 다 공통적으로 나타나고 있으며, 이러한 결과들은 부분선형모형적합에서 비모수 항 $\eta(x)$를 포함시킴으로써 발생하는 y절편과 비모수 항 간의 식별가능성(identifiability)문제와 관련이 있다. 다만 이 경우에서는 세 모형 모두 공통적으로 좋은 적합결과를 보여주지는 못하지만 차분과 평활도를 이용한 방법에 비해서 가우지안 확률과정 회귀모형이나 푸리에 급수를 이용한 방식이 다소 나아보임을 알 수 있다. 이러한 면에서 차분과 평활도를 이용한 방식에 비해서 나머지 두 방법이 효과적인 방법이라고 할 수 있으나 식별가능성 이외의 문제점에 있는지의 여부에 대해서는 추후에 더 연구해보아야 할것이다. RMSE(η)의 결과에서는 푸리에 급수를 이용하는 부분선형모형이 가장 좋은 성능을 보여주고 있음을 알 수 있었으며 RMSE(β)와 RMSE(y)의 경우에서는 세 가지 방법에서 큰 차이가 없음을 알 수 있었다. RMSE(β)의 경우는 식 (3.1)에 정의된 RMSE와 달리 모의실험자료의 표본의 크기(n)이 아니라 100회의 반복횟수(m)와 각 반복에서의 사후평균 값과 참 값(β_0)과의 차이로 계산하였다.

베이지안 부분선형 모형을 적합하는데 있어서 고려해야 할 점 중의 하나는 두 가지 형태의 공변량($\mathbf{d}, \mathbf{x}$)이 가능한 서로 관련이 없어야 한다는 것이며, 아울러 이러한 이유로 두 공변량이 서로 연관

Table 2: Summary results for simulation study 2

	double normal $d = 2(x-0.5)+\delta$ $\delta \sim N(0, 0.1^2)$				double normal $d = 2(x-0.5)+\delta$ $\delta \sim N(0, 2^2)$				double normal $d = 2(x-0.5)^2+\delta$ $\delta \sim N(0, 0.5^2)$			
	GP	Lenk	Koop	lin.reg.	GP	Lenk	Koop	lin.reg.	GP	Lenk	Koop	lin.reg.
$\beta(\beta_0 = 2)$	−2.666	8.075	1.270	−5.817 (0.015)	2.002	2.011	2.003	1.338 (−0.026)	1.971	1.999	1.995	1.760 (0.054)
$\sigma^2(\sigma_0^2 = 1)$	8.299	1.565	1.000	14.123	0.990	1.017	0.974	35.024	0.967	1.001	0.955	37.042
RMSE (β)	4.984	6.435	1.480	7.842	0.071	0.073	0.073	0.730	0.274	0.272	0.277	1.364
RMSE (η)	3.748	3.929	1.037		0.488	0.530	0.552		0.504	0.542	0.565	
RMSE (y)	2.529	1.238	0.987	3.756	0.992	1.001	0.974	5.917	0.980	0.993	0.965	6.087

Table 3: Summary results for simulation study 3

	double normal				linexp				lincos				yatchew			
	GP	Lenk	Koop	lin.reg.	GP	Lenk	Koop	lin.reg.	GP	Lenk	Koop	lin.reg.	GP	Lenk	Koop	lin.reg.
$\beta(\beta_0 = 2)$	1.959	2.081	1.945	−2.692 (0.011)	2.100	1.986	1.982	2.126 (0.987)	1.905	1.950	1.996	1.625 (0.311)	2.057	2.010	2.008	2.051 (0.010)
$\sigma^2(\sigma_0^2 = 1)$	0.993	1.024	0.976	23.529	1.808	2.032	1.019	1.949	0.987	1.069	0.992	1.094	1.157	1.025	1.045	1.184
RMSE (β)	0.290	0.310	0.296	4.734	0.289	0.397	0.291	0.266	0.299	0.290	0.293	0.404	0.203	0.276	0.288	0.196
RMSE (η)	0.517	0.565	0.580		0.901	1.099	0.386		0.344	0.460	0.335		0.423	0.384	0.408	
RMSE (y)	0.993	1.004	0.974	4.844	1.335	1.419	0.997	1.393	0.990	1.027	0.984	1.043	1.072	1.004	1.009	1.086

된 경우에도 상관계수가 작을수록 두 가지 주 관심 모수 $(\beta, \eta(\cdot))$가 더 잘 적합될 것으로 예상할 수 있다. 이러한 두 공변량 간의 상관계수는 식 (3.2)의 δ_i의 분산에 의해 결정되며 Table 1의 결과에서는 $\delta_i \sim N(0, 0.5^2)$를 사용하였다. 아래의 Table 2에서는 두 공변량 간의 상관관계에 따른 부분선형모형의 적합성 정도를 알아보기 위하여 공변량 $\mathbf{d}$에 대한 세 가지 서로 다른 생성방식을 고려하고 이를 바탕으로 베이지안 부분선형모형을 적합해보았다. 첫 번째는 두 공변량 간의 상관계수가 큰 경우, 즉 δ_i의 분산이 작은 경우, $\delta_i \sim N(0, 0.1^2)$, 두 번째는 두 공변량 간의 상관계수가 작은 경우, 즉 δ_i의 분산이 큰 경우, $\delta_i \sim N(0, 2^2)$, 그리고 마지막으로는 공변량 $\mathbf{d}$가 공변량 $\mathbf{x}$와 비선형적 관계를 갖는 경우, $d = 2(x-0.5)^2 + \delta_i$, $\delta_i \sim N(0, 0.5^2)$를 바탕으로 세 가지 방법을 적합해 보았다. 이 경우 식 (3.2)의 모형을 참 모형으로 가정하고 공변량 $\mathbf{d}$를 앞서 소개한 세 가지 방식으로 생성하였고, 두 공변량 간의 상관관계에 따른 부분선형모형적합이 주 관심사였기 때문에 비모수 회귀 함수는 식 (3.3)의 (1)에 해당하는 double normal 함수만을 사용하였다.

Table 2에 요약된 결과에서 알 수 있는 것처럼, 두 공변량 간의 상관관계가 큰 경우($\delta \sim N(0, 0.1^2)$)에 주 관심모수 $(\beta, \eta(\cdot))$에 대한 적합결과와 실제 참값과의 차이가 매우 큼을 알 수 있었다. 즉, GP, Lenk, Koop 세 가지 모형 모두에서 β의 사후평균은 참값인 2와 매우 상이했으며 이에 대응하는 RMSE의 값들 역시 모두 큰 값을 나타냄을 알 수 있었다. 또한 이 경우, 부분선형모형을 고려하지 않는 선형회귀모형(lin.reg)의 적합결과 역시 만족스럽지 못한 결과를 나타내었으며 RMSE의 값들 역시 큰 값을 나타내었다. 선형회귀모형의 결과에서는 y절편 α를 포함하는 결과이며 괄호 안에 y절편에 관한 적합결과를 명시하였다. 두 공변량 간의 상관관계가 작을 때($\delta \sim N(0, 2^2)$)나 비선형 관계가 있을 때에는 세 가지 부분선형모형의 적합결과 모두 실제 참값과 매우 유사하며 RMSE의 값들 역시 작은 값을 나타내었다. 이에 반해 선형회귀모형의 적합결과는 여전히 참 값과 차이를 보이고 있으며 아울러 RMSE의 값 역시 큰 값을 나타내었다. 즉, 두 공변량이 서로 연관되지 않고 참 모형이 부분선형모형으로부터 생성된 경우에는 선형회귀모형만으로는 모형을 설명하기에는 부족하며, 이 경우 부분선형모형을 고려해야함을 알 수 있었다. 다음의 Table 3은 이러한 결과를 바탕으로 추가 모의실험자료에 대한

Table 4: Summary results for traffic data

	GP	Lenk	Koop
constant	4.5759	4.4267	4.1557
log(unemp)	−0.5177	−0.4555	−0.4613
spring	−0.0558	−0.0572	−0.0559
summer	−0.0833	−0.0953	−0.0955
fall	0.0141	0.0240	0.0207
σ^2	0.0208	0.0350	0.0256
RMSE (y)	0.1438	0.1601	0.1581

Table 5: Gelman-Rubin's diagnostics for traffic data based on Koop

	constant	log(unemp)	spring	summer	fall	σ^2
Point Est.	1.33	1.15	1.05	1.02	1.02	1.00
Upper C.I.	1.81	1.36	1.12	1.05	1.06	1.01

적합결과를 나타낸다. Table 3의 결과를 위하여 Table 2의 세 번째 경우와 같은 두 공변량 사이의 비선형 관계 $d = 2(x - 0.5)^2 + \delta$, $\delta \sim N(0, 0.5^2)$를 고려하였으며, Table 1에서 고려하였던 식 (3.3)의 네 가지 종류의 비선형 함수를 사용하였다.

Table 3의 결과에서는 세 가지 부분선형 모두 실제 참 값인 모수와 매우 유사한 사후평균 추정치를 나타내고 RMSE값들 모두 작은 값을 보여주고 있다. 아울러 세 가지 부분선형모형 간에도 큰 차이는 나타내지 않고 있다. 반면에 선형회귀모형의 적합결과는 실제 참 값과 많이 다르며 이 경우 Table 2의 결과와 마찬가지로 부분선형모형 적합이 필요함을 확인할 수 있었다.

3.2. 사례 데이터 분석을 통한 베이지안 부분선형모형 비교

3.2.1. 교통사고 자료분석 (Traffic data analysis)

사례 데이터 분석을 위하여 첫 번째로 Lenk (1999)에서 사용되었던 교통사고 자료(traffic data)를 세 가지 베이지안 부분선형모형에 적용해보았다. traffic data는 1979년 1월 1일부터 1987년 12월 31일까지 108개월 동안 미시간 주에서 일어난 교통사고에 관한 것으로서 미시간 대 교통 연구소(University of Michigan Transportation Research Institute)에 기록된 모든 자료들 중 0.1%만 무작위로 추출된 자료이다. 반응변수는 교통사고 횟수의 로그값으로, 연속형 변수이며 공변량으로는 연속형 변수인 로그실업률(실업률에 로그변환을 한 변수, log(unemployment rate))과 이산형 변수인 봄, 여름, 가을의 계절효과(seasonal effect)이다.

세 가지 베이지안 부분선형모형 적합에 있어서 3.1절의 모의실험자료 적합과 마찬가지로 깁스표집 반복 횟수는 10,000번이며, 처음 8,000번을 제외한 2,000번의 사후표본을 분석에 사용하였다. 적합결과는 Table 4와 Figure 1에 요약되어 있으며 이러한 결과들을 바탕으로 볼 때 traffic data 적합에서는 세 가지 모형 중에서 GP 모형이 가장 작은 RMSE를 나타내었으며 이러한 관점에서 GP 모형을 통한 traffic data 적합이 가장 좋은 결과를 보여준다고 할 수 있다.

MCMC의 수렴여부를 확인하기 위하여 Gelman-Rubin 통계량 Brooks와 Gelman (1998)값을 확인하였고 **R**의 **CODA** 패키지가 제공하는 **Gelman.diag**함수를 사용하였다. 이를 위하여 두 개의 체인(chain)을 사용하였고 Gelman-Rubin 통계량의 값이 대부분 1에 매우 가까움을 알 수 있었다. 이 경우 가우지안 확률과정 회귀모형과 푸리에 급수 회귀모형에 비해 차분과 평활도를 이용한 방식에서는 수렴정도가 다른 두 가지 방법에 비해서 다소 떨어짐을 확인할 수 있었으며 Table 5에 요약되어 있는 것처럼 Gelman-Rubin 통계량 값이 1보다는 약간씩 크다는 것을 확인할 수 있었다.

Table 6: Summary results for wage data

	GP	Lenk	Koop	np package
constant	−0.2492	−0.2156	−0.2793	−0.2636
married	0.1258	0.1647	0.1666	0.1183
educ	0.1083	0.1267	0.0907	0.0893
tenure	0.0182	0.0212	0.0138	0.0175
σ^2	0.1845	0.1972	0.1768	
RMSE (y)	0.4297	0.4394	0.4193	0.4175

3.2.2. 임금 자료분석 (Wage data analysis)

사례 데이터 분석을 위하여 고려하는 또 다른 자료는 **R**의 **np** package에서 제공하는 임금 자료(wage data)로서 계량경제학 모형에서 활용되는 자료이다 (Wooldridge, 2003). **np** package는 부분선형모형을 포함하는 다양한 준모수 회귀모형에 대한 전통적 접근방법에서의 적합을 위한 **R** package이다 (Hayfield와 Racine, 2008). wage data를 세 가지 베이지안 부분선형모형을 이용하여 모형을 적합하고 아울러 **np** package에서 제공하는 **npplreg** 함수를 사용하여 전통적 접근방식에서의 부분선형모형 적합과 비교해보았다.

wage data는 1976년 미국의 현재인구조사(current population survey)에서 얻어진 526명의 사람들에 관한 횡단면 자료(cross-sectional data)이다. 본 사례분석에서는 시간당 임금에 대한 로그변환을 한 값(log(wage))을 반응변수로 사용하고 교육받은 기간으로 표현되는 학력(educ), 관련 업무에서의 경력(exper)과 현 직장에서의 근무기간(tenure)을 연속형 설명변수로, 성별(female)과 결혼 유무(married)를 이산형 설명변수로 고려하였다. 이 경우 반응변수를 설명하기 위한 부분선형모형에서 경력(exper)변수가 비모수 회귀함수로 연결되는 모형을 고려하였다. 이 경우 경력(exper)변수는 관련 업무에서의 잠재적 경험(potential experience)을 연도(year)로 측정한 값이기 때문에 자연수 값을 가지며, 동일한 값들이 관찰되고 있다. 이러한 변수 값의 형태는 부분선형모형 적합 시 문제가 되었는데 특히 차분과 평활도를 바탕으로 한 베이지안 부분선형모형인 Koop방법으로는 적합할 수가 없었다. 이러한 문제는 특히 Koop방법에서는 식 (2.9)–식 (2.11)에 주어진 것과 같이 차분행렬을 사용하기 때문에 비모수 회귀함수에 대응되는 공변량 x는 크기 순서로 정렬이 되어 있어야 하고 동일한 값을 갖지 말아야하는 제약조건에 기인한 것으로 보인다. 따라서 이러한 문제를 해결하기 위하여 약간의 잡음(noise)을 더해서 서로 다른 연속형 자료로 만드는 과정(jittering, $x = \text{exper} + N(0, 0.1^2)$)을 통해 공변량 x를 변형한 후에 자료를 부분선형모형에 적합하였다. 이러한 적합결과는 Table 6에 요약되어 있으며, 전체적으로 세 가지 모형의 적합결과는 유사하지만 Koop 모형이 가장 작은 RMSE를 나타내었으며 세 가지 베이지안 부분선형모형 모두 **np** package를 이용한 전통적인 방식에서의 모형적합결과와 유사하게 나타났다.

4. 결론

본 논문에서는 부분선형모형에 대한 베이지안 접근방식을 고려하고 세 가지 베이지안 부분선형회귀모형에 대해서 설명하고 각 모형의 성능을 비교해보았다. 구체적으로는 베이지안 부분선형모형의 비모수적 회귀 모형부분을 적합하기 위한 세 가지 방식 - 푸리에 급수를 이용한 방식, 가우지안 확률과정 회귀모형을 이용한 방식, 차분과 평활도를 이용한 방식 -을 바탕으로 각 모형의 적합방법에 대해서 고찰하고, 실증적 분석을 실시하였다. 이를 위하여 다양한 비선형 회귀함수로부터 생성되는 모의실험 자료를 바탕으로 세 가지 모형을 적합하고 모수추정 결과 비교해보고 제곱근평균제곱오차를 기준으로 모형 간의 성능을 비교해보았다. 아울러 기존의 부분선형모형에서 사용되었던 두 가지 실제 사례 데이

터를 이용하여 세 가지 방식을 적합해보고 모형 간의 성능을 비교해보았다. 다양한 경우에서 생성되는 모의실험 자료와 사례 데이터 분석 모두에서 제곱근평균제곱오차를 바탕으로 하는 세 모형간의 성능비교에서는 큰 차이를 발견할 수는 없었으나 세 모형 간의 일관된 차이보다는 주어진 자료에 따라 다소간의 차이를 발견할 수 있었다. 아울러 실제 자료 분석에 있어서는 가우지안 확률과정모형 적합 시에는 비모수 회귀모형에 대한 공변량의 형태가 큰 문제가 되지는 않았으나 푸리에 급수를 이용한 방식이나 차분과 평활도를 이용한 방식에서는 공변량이 특정한 형태가 만족되는 경우 외에는 적합시 문제가 발생함을 확인할 수 있었다. 아울러 차분과 평활도를 이용한 방식에서는 다른 두 가지 방법에 비해 수렴여부나 모수 적합에 있어서 성능이 다소 떨어짐을 확인할 수 있었다.

본 논문에서는 비모수 회귀모형에 대한 공변량이 1차원인 경우만을 고려하였는데 다차원 공변량의 경우에서 모형간의 성능비교를 향후과제로 고려해 볼 수 있을 것이다. 푸리에 급수를 이용하는 경우에는 다차원 공변량으로의 확장이 쉽지 않으며, 특히 다차원 문제를 보다 효율적으로 처리할 수 있는 것으로 알려진 (Shi와 Choi, 2011) 가우지안 확률과정회귀모형을 이용한 부분선형모형이나 다차원으로 쉽게 확장될 수 있을 것으로 예상되는 차분과 평활도를 이용한 접근방식이 더 효율적일 것이라고 생각되며 이를 위한 실증적 성능비교 연구를 향후과제로 수행하고자 한다. 아울러 본 논문에서는 가우지안 확률과정회귀모형과 차분과 평활도를 이용한 방식에서의 초모수 추정을 위하여 경험적 베이즈 방법을 사용하였는데, 이에 대안적으로 교차타당성 입증(cross-validation)이나 초사전분포를 통한 완전 베이즈 방법(full Bayes method)를 고려해보는 것도 추가적으로 고려해 볼 사항이라 하겠다.

준모수 회귀모형은 본 연구에서 고려한 부분선형모형 이외에 일반화 가법모형(generalized additive model), 일반화 부분선형 가법모형(generalized partially linear additive model)등을 포함하는 구조적 가법모형(structured additive model)으로 다양하게 확장되어지고 있다 (Kneib 등, 2011). 따라서 본 논문에서 고려했던 세 가지 베이지안 접근방식을 확장하여 다양한 구조적 베이지안 가법모형으로 발전시키는 일은 매우 흥미롭고 유의미한 향후과제라고 할 수 있겠다.

References

Aerts, M., Claeskens, G. and Hart, J. D. (2004). Bayesian-motivated tests of function fit and their asymptotic frequentist properties, *The Annals of Statistics*, **32**, 2580–2615.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations, *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Choi, T., Lee, J. and Roy, A. (2009). A note on the Bayes factor in a semiparametric regression model, *Journal of Multivariate Analysis*, **100**, 1316–1327.

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

Choi, T. and Woo, Y. (2012). A partially linear model using a Gaussian process prior, *submitted.,*

Damien, P., Wakefield, J. and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 331–344.

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

Härdle, W., Liang, H. and Gao, J. (2000). *Partially linear Models*, Physica-Verlag, Heidelberg.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package, *Journal of Statistical Software*, **27**, 1–32.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 425–464.

Kneib, T., Konrath, S. and Fahrmeir, L. (2011). High dimensional structured additive regression models: Bayesian regularization, smoothing and predictive performance, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **60**, 51–70.

Koop, G. and Poirier, D. J. (2004). Bayesian variants of some classical semiparametric regression techniques, *Journal of Econometrics*, **123**, 259–282.

Lenk, P. J. (1999). Bayesian inference for semiparametric regression using a Fourier representation, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 863–879.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics, Theory and Practice*, Princeton University Press, Princeton, New Jersey.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **34**, 1–41.

Na, J. and Kim, J. (2002). Bayesian model selection and diagnostics for nonlinear regression model, *Korean Journal of Applied Statistics*, **15**, 139–151.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 751–769.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **40**, 1–42.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA.

Ruppert, D., Wand, M. P. and Caroll, R. J. (2009). Semiparametric regression during 2003–2007, *Electronic Journal of Statistics*, **3**, 1193–1256.

Shi, J. Q. and Choi, T. (2011). *Gaussian Process Regression Analysis for Functional Data*, Chapman & Hall/CRC Press, New York.

Shi, J. Q., Murray-Smith, R. and Titterington, D. M. (2007). Gaussian process function regression modeling for batch data, *Biometrics*, **63**, 714–723.

Shi, J. Q. and Wang, B. (2008). Curve prediction and clustering with mixtures of Gaussian process functional and regression models, *Statistics and Computing*, **18**, 267–283.

Wooldridge, J. M. (2003). *Introductory Econometrics, A Modern Approach*, MIT Press, Cambridge, MA.

Yatchew, A. (1998). Nonparametric regression technique in Economics, *Journal of Economic Literature*, **36**, 669–721.

Yi, G., Shi, J. Q. and Choi, T. (2011). Penalized Gaussian process regression and classification for high-dimensional nonlinear data, *Biometrics*, **67**, 1285–1294.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models, *Journal of the American Statistical Association*, **97**, 1042–1054.

2012년 8월 20일 접수; 2012년 10월 22일 수정; 2012년 11월 13일 채택