The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.4.627
29(4), 627–641

A nonparametric Bayesian seemingly unrelated regression model

Seongil Jo[a] · Inhae Seok[a] · Taeryon Choi[a,1]

[a]Department of Statistics, Korea University

(Received February 24, 2016; Revised April 5, 2016; Accepted April 25, 2016)

Abstract

In this paper, we consider a seemingly unrelated regression (SUR) model and propose a nonparametric Bayesian approach to SUR with a Dirichlet process mixture of normals for modeling an unknown error distribution. Posterior distributions are derived based on the proposed model, and the posterior inference is performed via Markov chain Monte Carlo methods based on the collapsed Gibbs sampler of a Dirichlet process mixture model. We present a simulation study to assess the performance of the model. We also apply the model to precipitation data over South Korea.

Keywords: seemingly unrelated regression model, Dirichlet process mixture model, collapsed Gibbs sampling, precipitation prediction

1. 서론

겉보기 무관 회귀(seemingly unrelated regression; SUR)모형은 서로 관련이 없는 것처럼 보이는 여러 개의 연립 방정식으로 표현되는 다중 회귀 방정식들을 동시에 적합하는 모형으로서, 통계학 뿐 아니라 계량 경제학, 금융, 심리학, 사회과학 등의 다양한 분야에서 활용되며, 많은 연구가 이루어져 왔다 (Aliprantis 등 2007; Ando와 Zellner, 2010; Wang, 2010; Zellner와 Chen, 2002; 등). Zellner (1962, 1963)에 의해 처음 제안된 겉보기 무관 회귀 모형은 주어진 여러 개의 회귀식들의 설명변수간에는 상관관계는 적고 오차(error)항들 간에는 상관관계가 높을 때, 이러한 종속적 구조를 공분산 행렬(covariance matrix) 반영함으로써, 오차항들간에 독립성을 가정하는 선형 회귀 모형에서의 추정 방법보다 효율적인 방법으로 알려져 있다.

SUR 모형의 모수 추정을 위한 대표적인 베이지안 접근 방법으로는 마코프 체인 몬테 칼로(Markov chain Monte Carlo; MCMC) 알고리듬을 이용한 방법 (Zellner, 1971), 베이지안 적률법(Bayesian method of moments)을 이용한 방법 (Zellner와 Tobias, 2001), 그리고 최근 Zellner와 Ando (2010a)에 의해 제안된 디렉트 몬테 칼로(direct Monte Carlo; DMC) 알고리듬을 이용한 방법이 있다. 한편, 빈도론적 방법(frequentist method)으로는 Aitken의 일반화 최소 제곱방법(generalized least square

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2013R1A1A2074463).
[1]Corresponding author: Department of Statistics, Korea University, 145 Anam-ro 145, Seongbuk-Gu, Seoul 02841, Korea. E-mail: trchoi@korea.ac.kr

method; Aitken, 1935; Zellner, 1962)과 우도함수를 이용한 접근 방법(likelihood approach; Fraser 등, 2005)이 있다.

대표적인 SUR 모형의 장점은 여러 개의 다중 회귀 방정식의 계수들을 동시에 추정하기 때문에 주어진 다중 회귀 모형의 구조를 동시에 파악할 수 있다는 것이다. 하지만, 지금까지 사용해온 모형은 오차항의 분포를 다변량 정규분포(multivariate normal distribution)로 가정하기 때문에, 이상치(outlier)가 존재하는 경우나, 정규분포를 따르지 않는 관측값에 대한 분석 등에 있어서는 이를 적용하는데 한계가 있어왔다 (Kowalski 등, 1999; Ng, 2002; Zellner와 Ando, 2010b). 이러한 단점들을 해결하기 위하여 Kowalski 등 (1999)과 Ng (2002)는 오차의 분포를 다변량 스튜던트-t 분포(multivariate student-t distribution)로 가정한 SUR 모형을 개발하였고, 최근, Zellner와 Ando (2010b)는 스튜던트-t 오차 분포에 대해 디렉트 몬테 칼로 방법을 이용한 SUR 모형을 개발하였다. 그러나 이러한 t-분포를 바탕으로 한 SUR 모형 역시 오차항의 분포가 치우침(skewness)이 있거나, 다봉성(multi-modality)을 갖거나, 특정한 분포로 가정하는 것이 어려운 관측값들을 설명하기 위한 SUR 모형이 필요한 경우에는 적용하기 어렵다는 것을 알 수 있다. 따라서, 이러한 오차항 분포에 대하여 보다 영향을 받지 않고, 다양한 분포에서 적용 가능한 새로운 베이지안 SUR 모형의 개발이 필요하며, 이를 위하여 본 논문에서는, 비모수 베이지안(nonparametric Bayesian)에서 대표적으로 사용되는 모형인 디리슈레 프로세스 혼합모형(Dirichlet process mixture model)을 오차항의 분포로 가정한 비모수 베이지안 SUR 모형을 제안하고자 한다.

디리슈레 프로세스 혼합모형은 Ferguson (1973)과 Antoniak (1974)에 의하여 제안된 디리슈레 프로세스 사전분포(Dirichlet process prior)를 혼합분포(mixing distribution)로 사용한 것으로, 현재 가장 널리 사용되는 비모수 베이지안 모형 중 하나이다 (Hjort 등, 2010; Müller와 Rodríguez, 2013; 등). 디리슈레 프로세스 혼합모형의 가장 큰 장점 중의 하나는 무한차원(infinite dimension)의 혼합모형이기 때문에, 정규분포(normal distribution) 또는 균등분포(uniform distribution)와 같은 적절한 분포를 바탕으로 한 커널함수(kernel function)의 디리슈레 프로세스 혼합모형을 사용하면, 주어진 확률 분포 함수를 특정한 모수적 가정없이 추정할 수 있다는 것이다 (Ferguson, 1983; Escobar와 West, 1995; Rodríguez와 Walker, 2014). 이와 같은 디리슈레 프로세스 혼합모형을 바탕으로하여 오차항 분포를 모형화 하는 비모수 베이지안 SUR 모형을 제안하도록 한다. 이러한 비모수 베이지안 SUR 모형은 오차항의 분포를 특정한 모수적인 분포(parametric distribution)로 가정하지 않기 때문에, 기존의 모수적 SUR 모형에 비해 오차항의 가정에 있어서 제약을 덜 받기 때문에 더욱 다양한 실제 자료에 적용이 가능하다.

논문의 나머지 구성은 다음과 같다. 먼저 2장에서는 SUR 모형과 이에 따른 베이지안 추정 방법을 소개한다. 이 후 3장에서는 디리슈레 혼합모형에 기반한 비모수 베이지안 SUR 모형을 제안하고, 이러한 모형 하에서의 각 모수들의 사후분포를 계산하는 알고리듬을 설명한다. 이를 바탕으로, 마코프 체인 몬테 칼로 방법에 기반한 사후추론에 대해서 설명하고, 모형 비교를 위한 주변 가능도를 계산하는 방법을 설명한다. 4장에서는 모의실험과 실제 자료분석을 통해 제안된 비모수 베이지안 SUR 모형과 기존의 모수적 SUR 모형간의 성능을 비교하고 5장에서 결론과 향후 연구방향에 대해서 고찰하도록 한다.

2. 겉보기 무관 회귀 모형: seemingly unrelated regression (SUR) model

본 절에서는 비모수 베이지안 SUR 모형의 기본이 되는 모수적 SUR 모형을 소개한다. 이를 위해 $\boldsymbol{y}_j$는 j번째 회귀 모형의 반응변수를 포함한 $n \times 1$ 벡터라 하고, $\boldsymbol{X}_j$를 j번째 회귀모형의 설명변수(explanatory variables)에 대응하는 $n \times p_j$차원의 계획행렬(design matrix)이라 하자. 이 때, m개의

회귀 모형으로 구성된 모수적 SUR 모형은 다음과 같다.

$$\begin{aligned} \boldsymbol{y}_1 &= X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1, \\ \boldsymbol{y}_2 &= X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2, \\ &\vdots \\ \boldsymbol{y}_m &= X_m\boldsymbol{\beta}_m + \boldsymbol{\epsilon}_m, \end{aligned} \tag{2.1}$$

여기서 $\boldsymbol{\epsilon}_j$는 $n \times 1$ 오차항 벡터이고, $\boldsymbol{\beta}_j$는 j번째 회귀 모형의 p_j차원의 계수 벡터(vector of coefficients)를 나타낸다. 모형 (2.1)에서 j번째 회귀 모형은 l번째 회귀 모형과 서로 다른 설명변수를 갖지만, 회귀 모형 사이에는 오차항 벡터를 통해 서로 종속관계가 있다고 가정한다. 즉, $E(\boldsymbol{\epsilon}_j\boldsymbol{\epsilon}_l^T) = \omega_{jl}I_n$, $j,l = 1,\ldots,m$이라고 가정하며, I_n는 n차원의 단위행렬을 나타낸다. 또한, 모형 (2.1)은 다음과 같이 행렬의 형태로 표현할 수 있다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \Omega \otimes I), \tag{2.2}$$

여기서 $\boldsymbol{y} = (\boldsymbol{y}_1^T,\ldots,\boldsymbol{y}_m^T)^T$, $\boldsymbol{X} = \mathrm{diag}\{X_1,\ldots,X_m\}$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_m^T)^T$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T,\ldots,\boldsymbol{\epsilon}_m^T)^T$이고, $\otimes$는 크로네커 곱(kronecker product)을 나타낸다. 그리고 분산-공분산(variance-covariance) 행렬 Ω는 $m \times m$차원의 양정치 행렬(positive definite matrix)로써 대각원소는 $\{w_1^2,\ldots,w_m^2\}$이고 (j,l)번째 비대각 원소 $\Omega_{j,l}$는 w_{jl}이다 ($j \neq l$, $j,l = 1,\ldots,m$).

행렬을 이용한 식 (2.2)의 표현방법은 SUR 모형을 일반적인 선형회귀모형과 같은 방식으로 나타낼 수 있기 때문에 사후분포의 계산을 쉽게 만드는 장점이 있다. 구체적으로, 식 (2.2)를 바탕으로 모수 $\boldsymbol{\beta}$와 Ω의 완전 조건부 분포(full conditional distribution)는 다음과 같이 각각 정규분포와 역-위샤트(Inverse-Wishart; IW) 분포로 계산되며, 깁스 표집(Gibbs sampling)을 사용하는 마코프 체인 몬테 칼로 방법을 통해 베이지안 추론을 실시한다.

- $\boldsymbol{\beta}$의 완전 조건부 분포: $\boldsymbol{\beta} \mid \Omega, \boldsymbol{y}, \boldsymbol{X} \sim N(\boldsymbol{\beta}_n, \Sigma_n)$

$$\Sigma_n^{-1} = \Sigma_0^{-1} + \boldsymbol{X}^T\left(\Omega^{-1} \otimes I\right)^{-1}\boldsymbol{X} \quad \text{and} \quad \boldsymbol{\beta}_n = \Sigma_n\left\{\Sigma_0\boldsymbol{\beta}_0 + \boldsymbol{X}^T\left(\Omega^{-1} \otimes I\right)^{-1}\boldsymbol{y}\right\},$$

- Ω의 완전 조건부 분포: $\mathrm{IW}(\Omega; s+n, S_n)$

$$S_n = S_0 + R, \quad R = (r_{ij})_{ij}, \quad r_{ij} = (\boldsymbol{y}_i - X_i\boldsymbol{\beta}_i)^T(\boldsymbol{y}_j - X_j\boldsymbol{\beta}_j).$$

이러한 두가지 완전 조건부 분포 계산에 있어서, 식 (2.2) 모형에 대한 준 공액사전분포(semi-conjugate prior)인 정규분포와 역-위샤트 분포를 $\boldsymbol{\beta}$와 Ω에 대하여 다음과 같이 가정하였다.

$$p(\boldsymbol{\beta}) = N(\boldsymbol{\beta}_0, \Sigma_0), \quad p(\Sigma) = \mathrm{IW}(s, S_0),$$

여기서 $s > m-1$이고, S_0는 $m \times m$ 양정치 행렬이다.

오차항의 분포를 정규분포로 가정하는 모형 (2.2)는 두꺼운 꼬리를 갖는 오차항을 다루기 위하여, 오차항의 분포가 자유도(degree of freedom)를 $\nu > 0$로 하는 다변량 스튜던트-t 분포를 따르는 경우, 즉, $\boldsymbol{\epsilon} \sim t_\nu(0, \Omega \otimes I)$로 확장될 수 있다. 이 경우, 스튜던드-$t$ 분포는 식 (2.3)과 같이 정규분포에 대한 척도 모수 혼합 모형(scale mixture model)으로 설명될 수 있으며, 이를 바탕으로 앞서 설명한 정규분포 오차항을 바탕으로 한 SUR 모형에서의 마코프 체인 몬테 칼로 방법을 쉽게 확장하여 적용할 수 있다

(Gelman 등, 2014).

$$\mathrm{t}_\nu(0, V) = \int N(0, \tau V) d\tau, \quad \tau \sim \mathrm{IG}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \tag{2.3}$$

여기서 IG(·) 역-감마 분포(inverse-gamma distribution)를 나타낸다.

표기상의 편의를 위하여, 본 논문의 나머지 부분에서는 다변량 정규분포를 가정한 SUR 모형에 대한 마코프 체인 몬테 칼로 적합 방식을 MCMC-SUR로 다변량 스튜던트-t 분포를 가정한 SUR 모형에 대한 적합 방식은 MCMC-SURT로 표기하도록 한다.

3. 비모수 베이지안 겉보기 무관 회귀모형

본 절에서는 앞 절에서 설명한 다변량 정규분포 또는 다변량 스튜던트-t 분포의 모수적 오차항 분포를 갖는 SUR 모형을 확장하여, 오차항의 분포에 무관한 비모수적 베이지안 SUR 모형을 제안하도록 한다. 구체적으로 오차항이 특정한 모수적 분포를 따르지 않는 비모수적 분포를 가정하도록 하며, 이러한 비모수적 오차항의 분포가 디리슈레 프로세스 혼합모형(Dirichlet process mixtures)에 의해 설명되는 비모수 베이지안 SUR 모형을 제안하고, 제안된 모형의 사후분포를 계산하는 마코프 체인 몬테 칼로 알고리듬을 설명하고자 한다.

3.1. 모형 설정

비모수 베이지안 SUR 모형은 식 (2.1)과 (2.2)의 SUR 모형에서 오차항의 분포를 정규분포를 가정하지 않고, 알려지지 않은 분포를 따른다고 가정하며 다음과 같이 디리슈레 프로세스 혼합모형(Dirichlet process mixture model)을 이용하여 정의한다.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im}) \overset{iid}{\sim} F(\boldsymbol{\epsilon}_i) = \int N(\boldsymbol{\epsilon} \mid \mu, V) G(d\mu, dV), \quad G \sim \mathrm{DP}(\alpha G_0), \tag{3.1}$$

여기서 α는 정밀도 모수(precision parameter) 또는 질량 모수(mass parameter)라 불려지는 양의 값을 가지는 모수이고 G_0는 G의 중심 위치를 나타내는 기저 분포(base distribution)로서, $E(G) = G_0$이다. 그리고 DP(·)는 디리슈레 프로세스 사전분포(Dirichlet process prior)로써 Sethurman (1994)의 막대 자르기(stick-breaking) 표현을 이용하여 다음과 같이 정의된다.

$$\begin{aligned} G(\cdot) &= \sum_{h=1}^{\infty} \left[v_h \prod_{l<h} (1 - v_l) \right] \delta_{\theta_h}(\cdot), \\ v_h &\overset{iid}{\sim} \mathrm{Beta}(1, \alpha), \quad h = 1, 2, \ldots, \\ \theta_h &\overset{iid}{\sim} G_0, \quad h = 1, 2, \ldots, \end{aligned} \tag{3.2}$$

여기서 $\delta_\theta(\cdot)$는 θ에서 확률 1을 가지는 퇴화분포(degenerate distribution)이다. 디리슈레 프로세스 사전분포의 대표적인 장점 중의 하나는, 식 (3.1)과 같이 정규분포에 대한 디리슈레 프로세스 위치-척도 모수 혼합모형(location-scale mixture model)이 임의의 절대 연속 분포(absolutely continuous distribution)를 설명할 수 있다는 것으로서 (Lo, 1984), 본 논문에서 제안하는 비모수 SUR 모형은 오차항 분포에 대한 제약을 받지 않음을 알 수 있다. 디리슈레 프로세스에 대한 더욱 자세한 내용과 응용은 Hjort 등 (2010)과 Müller와 Rodríguez (2013) 그리고 Müller 등 (2015) 등에서 참조할 수 있다.

제안된 모형 (3.1)에 대한 적합을 위하여 다음과 같이 사전분포를 정하도록 한다. 먼저 정밀 모수 α는

형태모수(shape parameter) a_α와 척도모수(scale parameter) $1/b_\alpha$를 갖는 감마 분포(gamma distribution)를 할당하고, 기저 분포 G_0는 공액 분포인 정규-역 위샤트(normal-inverse Wishart) 분포를 가정한다.

$$G_0 = N(\mu \mid \mu_0, \tau V)\text{Wishart}\left\{V^{-1} \mid s, (sS_0)^{-1}\right\}. \tag{3.3}$$

본 논문에서는 기저분포의 평균 μ_0가 0이라고 가정하지 않기 때문에 모형의 식별성(identifiability)을 만족하기 위해 모형 (3.1)의 계획행렬에 절편(intercept)를 포함하지 않도록 한다. 계획행렬에 절편을 포함하기 위해서는 기저분포의 평균을 0으로 가정할 수 있다 (Chib과 Greenberg, 2010).

마지막으로 기저분포의 모수 (μ_0, τ, S_0)와 회귀계수 $\boldsymbol{\beta}$에 대하여 다음의 사전분포를 가정한다.

$$\boldsymbol{\beta} \sim N(\boldsymbol{b}_0, \boldsymbol{B}_0), \quad \mu_0 \sim N(a, A_0), \quad \tau^{-1} \sim \text{Gamma}(a_\tau, b_\tau), \quad S_0 \sim \text{Wishart}\left(q, q^{-1}Q_0\right), \tag{3.4}$$

여기서 $\boldsymbol{b}_0$와 a는 각각 $p = \sum_{j=1}^{m} p_j$차원과 m차원의 실수 값을 갖는 벡터이고, $\boldsymbol{B}_0$는 $p \times p$차원의 양정치 행렬, A_0와 Q_0는 $m \times m$ 차원의 양정치 행렬이다. 그리고 a_τ, b_τ, s, q는 양의 값을 가지는 고정된 상수이다. 각각의 사전분포의 초모수 $(\boldsymbol{b}_0, \boldsymbol{B}_0, a, A_0, a_\tau, b_\tau, s, q, Q_0)$의 값에 대해서는, 각 사전분포들이 무정보적 사전분포(noninformative prior)를 나타낼 수 있도록 설정한다.

식 (3.1)의 디리슈레 혼합 모형은 계층적 모형(hierarchical model)을 이용하여 다음과 같이 표현할 수 있다.

$$\begin{gathered}(\epsilon_{i1}, \ldots, \epsilon_{im}) \mid \mu_i, \quad V_i \overset{ind.}{\sim} N(\mu_i, V_i), \quad i = 1, \ldots, n, \\ \theta_i = (\mu_i, V_i) \mid G \overset{iid}{\sim} G, \quad i = 1, \ldots, n, \\ G \sim \text{DP}(\alpha G_0).\end{gathered} \tag{3.5}$$

본 논문의 나머지 부분에서는 식 (3.1)의 디리슈레 혼합모형을 이용한 비모수 SUR 모형을 DPM-SUR로 나타내도록 한다.

3.2. 사후 분포의 계산

3.1절에서 정의된 DPM-SUR로부터 사후표본(posterior sample)을 추출하기 위하여 Neal (2000)에 의해 제안된 붕괴깁스표집방법(collapsed Gibbs sampling method)을 이용한다. Neal (2000)의 붕괴깁스표집방법은 공액사전분포가 사용되지 않을 때에도 적용 가능한 알고리듬으로, DPM-SUR 모형에 다음과 같은 방식으로 쉽게 구현될 수 있다. 즉, c_i를 i번째 관측치에 해당되는 오차 $\boldsymbol{\epsilon}_i$를 디리슈레 혼합모형의 구성성분에 할당하는 잠재변수(latent variable), K를 $\boldsymbol{c} = (c_1, \ldots, c_n)$의 최대값의 갯수, h를 새로운 구성성분을 나타내는 보조변수(auxiliary variable)의 갯수, 그리고 μ_k^*, V_k^*를 디리슈레 혼합모형의 k번째 구성성분에 해당하는 모수라고 하면, 사후표본은 다음의 단계를 통해 추출된다.

- 단계 1. $i = 1, \ldots, n$에 대하여 K_{-i}를 $\boldsymbol{c}_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$에서 유일한 최대값의 갯수로 정의하고, H와 n_i를 다음과 같이 $H = K_{-i} + h$와 $n_i = \sum_{l:l\neq i} I(c_i = c_l)$라고 각각 정의한 후, $\boldsymbol{c}_{-i}$를 $\{1, \ldots, K_{-i}\}$의 값을 바탕으로 재할당한다. 이 경우, n_i가 0이라면 $c_i = K_{-i} + 1$이라 하고 (μ_k^*, V_k^*), $k = K_{-i} + 2, \ldots, H$를 G_0로 부터 추출한다. 만약, n_i가 0이 아니라면 (μ_k^*, V_k^*), $k = K_{-i} + 1, \ldots, H$를 G_0로 부터 추출한다. 그 후 c_i를 다음의 확률로부터 다시 추출한다.

$$\Pr\left(c_i = k \mid c_{-i}, \boldsymbol{r}_i, \mu_1^*, \ldots, \mu_H^*, V_1^*, \ldots, V_H^*\right) \propto \begin{cases} n_{-i,k} N\left(\boldsymbol{r}_i; \mu_k^*, V_k^*\right), & \text{for } 1 \leq k \leq K_{-i}, \\ \left(\dfrac{\alpha}{h}\right) N\left(\boldsymbol{r}_i; \mu_k^*, V_k^*\right), & \text{for } K_{-i} < k \leq H, \end{cases} \tag{3.6}$$

여기서 $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{im})^T$, $r_{ij} = y_{ij} - x_{ij}^T \beta_j$이고 $n_{-i,k} = \sum_{l:l\neq i} I(c_{-i,l} = k)$이다.

- 단계 2. 디리슈레 프로세스 혼합모형의 정밀도 모수 α는 Escobar와 West (1995)가 제안한 자료확대(data augmentation)방법을 이용하여 두 단계로 추출한다.

  - 단계 2-1. 잠재변수 η를 Beta$(\alpha+1, n)$에서 추출한다.
  - 단계 2-2. 추출된 잠재변수를 이용하여 α를 다음의 혼합 감마분포로 부터 추출한다.

$$p(\alpha|\eta, K) = \frac{a_\alpha + K - 1}{a_\alpha + K - 1 + n\{b_\alpha - \log(\eta)\}} \text{Gamma}(a_\alpha + K, b_\alpha - \log(\eta)) + \frac{n(b_\alpha - \log(\eta))}{a_\alpha + K - 1 + n\{b_\alpha - \log(\eta)\}} \text{Gamma}(a_\alpha + K - 1, b_\alpha - \log(\eta)). \tag{3.7}$$

- 단계 3. $k = 1, \ldots, K = \max\{c_1, \ldots, c_n\}$에 대하여 (μ_k^*, V_k^*)를 다음의 조건부 사후분포로 부터 추출한다.

$$p\left(\mu_k^* \mid \boldsymbol{r}_i, V_k^*, \mu_0, \tau, S_0, \boldsymbol{c}, K\right) = N\left(\mu_k^*; \hat{\mu}_k, \hat{\tau} V_k^*\right), \tag{3.8}$$

$$p\left(V_k^* \mid \boldsymbol{r}_i, \mu_k^*, s, q, Q_0, \boldsymbol{c}, K\right) = \text{Wishart}\left\{V_k^*; s + n_k, \hat{S}_k\right\}, \tag{3.9}$$

여기서 $\hat{\tau} = 1/(n_k + 1/\tau)$, $\hat{\mu}_k = (n_k + 1/\tau)^{-1}(\mu_0/\tau + \sum_{i:c_i=k} \boldsymbol{r_i})$, $\hat{S}_k^{-1} = sS_0 + \sum_{i:c_i=k}(\boldsymbol{r}_i - \mu_k^*)(\boldsymbol{r}_i - \mu_k^*)^T$이고 $n_k = \sum_{i=1}^n I(c_i = k)$이다.

- 단계 4. 기저분포의 초모수에 대한 사후표본은 다음의 분포로 부터 추출한다.

$$p\left(\mu_0 \mid a, A_0, \tau, \mu_1^*, \ldots, \mu_K^*, V_1^*, \ldots, V_K^*, K\right) = N\left(\mu_0; \hat{a}, \hat{A}_0\right), \tag{3.10}$$

$$p\left(\tau^{-1} \mid a_\tau, b_\tau, \mu_0, \mu_1^*, \ldots, \mu_K^*, V_1^*, \ldots, V_K^*, K\right) = \text{Gamma}\left(\tau^{-1}; \hat{a}_\tau, \hat{b}_\tau\right), \tag{3.11}$$

$$p\left(S_0 \mid s, q, Q_0, V_1^*, \ldots, V_K^*, K\right) = \text{Wishart}\left\{S_0; q + sK, \hat{Q}_0\right\}, \tag{3.12}$$

여기서 $\hat{A}_0^{-1} = A_0^{-1} + \sum_{k=1}^K V_k^{*-1}/\tau$, $\hat{a} = \hat{A}_0(A_0^{-1}a + \sum_{k=1}^K V_k^{*-1}\mu_k^*)$, $\hat{a}_\tau = a_\tau + mK/2$, $\hat{b}_\tau = b_\tau + \sum_{k=1}^K (\mu_k^* - \mu_0)^T V_k^{*-1}(\mu_k^* - \mu)$이고 $\hat{Q}_0 = (qQ_0^{-1} + s\sum_{k=1}^K V_k^{*-1})^{-1}$이다.

- 단계 5. 마지막으로 설명변수의 계수는 p차원의 다변량 정규분포로 부터 추출한다.

$$p(\boldsymbol{\beta} \mid \boldsymbol{b}_0, \boldsymbol{B}_0, \boldsymbol{r}_1^*, \ldots, \boldsymbol{r}_n^*, \boldsymbol{X}, \mu_1^*, \ldots, \mu_K^*, V_1^*, \ldots, V_K^*, \boldsymbol{c}) = N\left(\boldsymbol{\beta}; \hat{\boldsymbol{b}}, \hat{\boldsymbol{B}}\right), \tag{3.13}$$

여기서 $\hat{\boldsymbol{B}}^{-1} = \boldsymbol{B}_0^{-1} + \sum_{i=1}^n X_i^T V_{c_i}^{*-1} X_i$, $\hat{\boldsymbol{b}} = \hat{\boldsymbol{B}}(\boldsymbol{B}_0^{-1}\boldsymbol{b}_0 + \sum_{i=1}^n X_i^T V_{c_i}^{*-1}\boldsymbol{r}_i^*)$, 그리고 $\boldsymbol{r}_i^* = \boldsymbol{y}_i - \mu_{c_i}^*$를 각각 의미하고, X_i는 $X_i = \text{diag}(x_{i1}, \ldots, x_{im})$로 정의되는 계획행렬을 나타낸다.

4. 모의실험과 자료 분석을 통한 실증분석

이 절에서는 본 논문에서 제안한 DPM-SUR 모형에 대한 실증적 분석을 위하여, 모의실험을 통해 기존의 SUR 모형에 대한 적합 방법과의 성능비교를 실시하고, 아울러 실제 자료 분석을 고려하도록 한다. 4.1절에서는 다양한 형태의 오차분포하에서 생성된 SUR 모형의 적합에 대한 성능비교를 수행하였고, 4.2절에서는 우리나라의 강수량 자료에 대하여 실증적으로 분석하였다. 성능 비교 대상의 된 기존의 SUR 모형으로, 2절에서 설명된 다변량 정규분포 오차항을 가정하는 MCMC-SUR, 다변량 스튜던트-t 분포를 가정하는 MCMC-SURT, 그리고 최대우도 추정량(maximum likelihood estimator; MLE)을 이용한 빈도론적(frequentist) 방법(MLE-SUR로 지칭)을 고려하도록 한다.

MLE-SUR의 경우 Feasible generalized least square(FGLS) 방법을 반복하여 적용하는 R의 system 패키지 (Henningsen과 Hamann, 2007)를 이용한다. FGLS 방법은 추정된 공분산을 이용하는 일반화 최소제곱(generalized least square) 방법으로써 점근적으로 효율적인 방법이며 (Henningsen과 Hamann, 2007), 각 반복마다 이전 단계의 잔차로부터 계산된 잔차 공분산 행렬(residual covariance matrix)을 이용한 SUR 추정량은 점근적으로 MLE가 됨이 알려져 있다 (Greene, 2003).

4.1. 모의실험

성능 비교를 위한 모의실험자료를 생성하기 위하여, 여러가지 형태의 오차항 분포를 사용하는 SUR 모형을 고려하였다. 특히 본 논문에서 제안하는 DPM-SUR 모형이 오차항 분포 가정에 영향을 받지 않는 강건한 모형임을 보이기 위하여, 구체적으로, 정규성이 만족되는 경우, 두꺼운 꼬리를 가지는 경우, 그리고 분포가 다봉성을 띄는 경우의 세 가지로 나누어 모의실험을 진행하였다. 아울러, 모의실험자료 생성을 위한 SUR 모형은 다음과 같이 회귀 방정식 $m = 2$개로 이루어져 있다고 가정하였고, 모든 실험에서 각각 50번의 반복을 실시하였다.

$$\begin{pmatrix} \mathbf{y}_{n1} \\ \mathbf{y}_{n2} \end{pmatrix} = \begin{pmatrix} X_{n1} & 0 \\ 0 & X_{n2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}, \tag{4.1}$$

여기서 X_{n1}과 X_{n2}는 각각 계획행렬로서, 절편을 고려하기 위한 1벡터와, 0과 1 사이의 균일분포, Unif$(0, 1)$에서 독립적으로 100개씩 추출된 2개의 설명변수 벡터들로 구성되었다. 이때, 회귀계수의 참값은 $\beta_1 = (1.5, 4.5, 3.0)^T$, $\beta_2 = (3.5, 3.0, -2.5)^T$이며, 오차항은 다음과 같은 세 가지 분포로부터 생성하도록 한다.

- 다변량 정규분포:

$$\epsilon_i = (\epsilon_{i1}, \epsilon_{i2})^T \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 2 & 1.6 \\ 1.6 & 2 \end{pmatrix}, \ i = 1, \ldots, n.$$

- 다변량 정규혼합분포:

$$\epsilon_i = (\epsilon_{i1}, \epsilon_{i2})^T \sim 0.5N(\mu_1, \Sigma_1) + 0.2N(\mu_2, \Sigma_2) + 0.3N(\mu_3, \Sigma_3), \quad i = 1, \ldots, n,$$

  여기서 $\mu_1 = (0.0, 0.0)^T$, $\mu_2 = (-3.0, 3.0)^T$, $\mu_3 = (2.0, -2.0)^T$이고, 분산은 다음과 같다.

$$\Sigma_1 = \begin{pmatrix} 0.2 & 0.16 \\ 0.16 & 0.2 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 0.8 & 0.64 \\ 0.64 & 0.8 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 0.4 & 0.32 \\ 0.32 & 0.4 \end{pmatrix}.$$

- 다변량 t-분포:

$$\epsilon_i = (\epsilon_{i1}, \epsilon_{i2})' \sim t_2(\mu, \Sigma), \quad \mu = (0, 0)', \ \Sigma = \begin{pmatrix} 2 & 1.6 \\ 1.6 & 2 \end{pmatrix}, \ i = 1, \ldots, n.$$

세 가지 베이지안 모형(DPM-SUR, MCMC-SUR, MCMC-SURT)에서 모수를 추정하기 위하여 각각 15,000번의 소각과정(burn-in)을 거친 후 10,000개의 사후표본(posterior sample)을 생성하였다. Figure 4.1은 각 모형의 마르코프 연쇄에서 나온 회귀계수에 대한 사후표본의 trace plot을 나타내는 것으로서, 전체적으로 주기성이나 추세가 없이 특정한 범위 안에서 흩어져 있음을 알 수 있으며, 이를 통해 사후 표본들이 적절하게 잘 수렴하는 것을 확인할 수 있다.

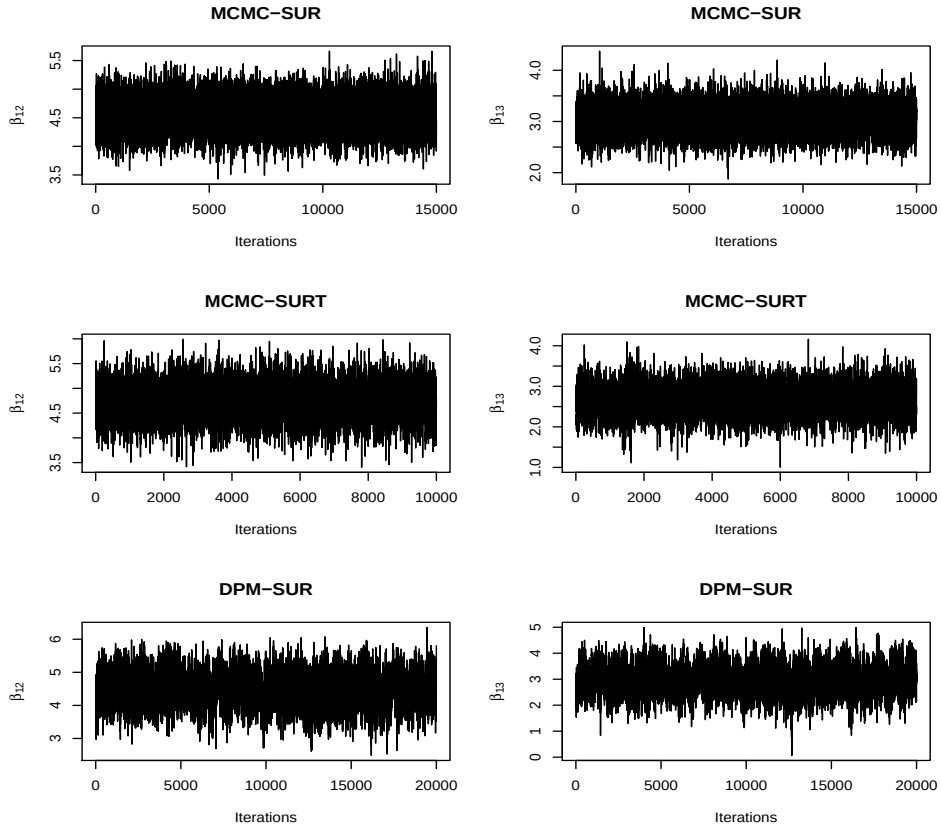

Figure 4.1. Trace plots of posterior samples for SUR coefficients under multivariate normal error.

모형 간의 성능비교를 위하여, 계수의 참 값과 추정된 값(사후평균 또는 MLE)의 차이를 나타내는 편향성(bias)과 실제 관측값과 적합값의 차이를 나타내는 제곱근 평균 제곱 오차(root mean squared error; RMSE), 그리고 로그 유사 주변가능도(log pseudo marginal likelihood; LPML)를 사용하였다. 로그 유사 주변가능도는 베이지안 모형의 적합도(goodness of fit)검정 또는 모형 선택을 위하여 주로 사용되는 측도로서 Geisser과 Eddy (1979)에 의해 제안된 conditional predictive ordinate(CPO) 통계량의 로그 합에 의해 정의되며, LPML 값이 클수록 모형의 적합도가 좋은 것을 의미한다.

편향성과 제곱근 평균 제곱 오차, 그리고 i번째 관측치에 대한 CPO 통계량의 구체적인 계산 방법은 다음과 같다.

$$\text{BIAS}_{jlk} = \left|\hat{\beta}_{jlk} - \beta_{jlk}\right|, \quad j = 1, \ldots, m,\ l = 1, \ldots, p_j,\ k = 1, \ldots, 50, \tag{4.2}$$

$$\text{RMSE}_k = \sqrt{\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{y}_{ijk} - y_{ij})^2}, \quad k = 1, \ldots, 50, \tag{4.3}$$

$$\text{CPO}_i = \left\{\frac{1}{B}\sum_{b=1}^{B}\frac{1}{f\left(y_i|\theta^{(b)}\right)}\right\}^{-1}, \quad i = 1, \ldots, n, \tag{4.4}$$

Table 4.1. Summary results of simulation studies

Error distribution	Parameter	TRUE	MCMC-SUR	MCMC-SURT	DPM-SUR	MLE-SUR
			BIAS (SE)	BIAS (SE)	BIAS (SE)	BIAS (SE)
Normal	β_{11}	1.5	**0.045** (0.315)	0.067 (0.237)	-	0.070 (0.242)
	β_{12}	4.5	0.163 (0.488)	0.014 (0.345)	**0.009** (0.486)	0.070 (0.297)
	β_{13}	3.0	0.106 (0.388)	0.054 (0.282)	**0.005** (0.518)	0.057 (0.284)
	β_{21}	3.5	0.059 (0.374)	0.004 (0.278)	-	0.048 (0.213)
	β_{22}	3.0	0.061 (0.449)	0.072 (0.295)	0.065 (0.357)	**0.032** (0.318)
	β_{23}	−2.5	0.092 (0.575)	**0.005** (0.382)	0.045 (0.291)	0.061 (0.305)
	RMSE		1.390 (0.082)	1.394 (0.104)	**1.267** (0.181)	1.386 (0.098)
	LPML		−307.216 (9.822)	−310.185 (8.994)	**−306.466** (10.233)	-
Normal mixture	β_{11}	1.5	0.097 (0.423)	0.214 (0.321	-	**0.032** (0.325)
	β_{12}	4.5	0.280 (0.575)	0.003 (0.382)	**0.002** (0.196)	0.034 (0.464)
	β_{13}	3.0	0.100 (0.619)	0.030 (0.342)	**0.006** (0.210)	0.094 (0.369)
	β_{21}	3.5	0.049 (0.542)	0.166 (0.286)	-	0.018 (0.376)
	β_{22}	3.0	0.103 (0.549)	0.009 (0.366)	0.015 (0.117)	**0.007** (0.508)
	β_{23}	−2.5	0.014 (0.686)	0.058 (0.334)	**0.005** (0.130)	0.059 (0.391)
	RMSE		1.815 (0.122)	1.817 (0.110)	**0.560** (0.056)	1.836 (0.099)
	LPML		−358.838 (10.046)	−357.506 (11.701)	**−230.766** (12.429)	-
t_2	β_{11}	1.5	0.097 (0.586)	**0.022** (0.299)	-	0.054 (0.602)
	β_{12}	4.5	0.093 (0.559)	**0.027** (0.400)	0.198 (0.877)	0.024 (0.764)
	β_{13}	3.0	0.115 (0.681)	0.071 (0.389)	**0.011** (0.713)	0.082 (0.691)
	β_{21}	3.5	0.099 (0.632)	**0.003** (0.269)	-	0.040 (0.629)
	β_{22}	3.0	0.064 (0.887)	0.035 (0.439)	0.084 (0.380)	**0.021** (0.707)
	β_{23}	−2.5	0.120 (0.715)	0.006 (0.327)	**0.001** (0.451)	0.131 (0.591)
	RMSE		3.616 (1.175)	3.774 (1.575)	**2.413** (0.981)	3.751 (2.137)
	LPML		−481.021 (46.930)	−419.765 (22.845)	**−410.650** (73.596)	-

RMSE = root mean squared error, LPML = log pseudo marginal likelihood.

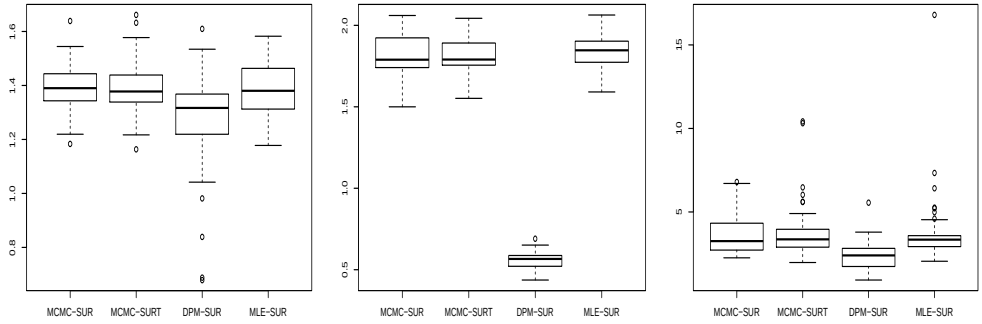

Figure 4.2. Boxplots of RMSE values (first panel shows the boxplot under the multivariate normal error, second under the normal mixture error, and third panel shows the boxplot under the multivariate t error).

여기서 p_j는 j번째 회귀 방정식 계수의 차원이고 $f(y_i|\theta^{(b)})$는 가능도 함수, $\theta^{(b)}$는 사후분포로부터 추출된 b번째 사후표본을 나타낸다.

모의실험의 결과는 Table 4.1과 Figure 4.2에 나타나 있다. Table 4.1의 결과들은 50번 반복을 통해 얻은 세 가지 측도값(BIAS, RMSE, LPML)들의 표본평균과 괄호로 표기된 표본표준편차(SE)를 나타내며, Figure 4.2의 결과들은 50번 반복 측정된 값들의 상자그림을 나타낸다. DPM-SUR의 경우는, 본 논문에서 제안하는 디리슈레 혼합모형의 특성상, 절편에 해당하는 회귀계수들은 추정하지 않기 때문에 Table 4.1에 표시되지 않음을 알 수 있다. Table 4.1에서 진한 글씨로 표기된 값들은, 각각의 모의실험

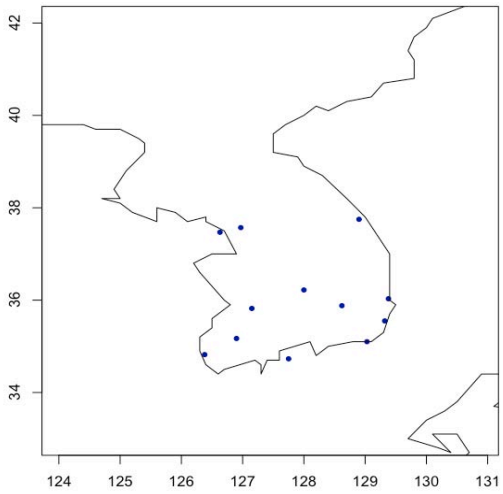

Figure 4.3. The geographical area used for this study (Closed circles indicate the locations of 12 weather stations for the observed precipitation in Korea).

에서 얻어진 측도값들 중 가장 우수한 값들을 나타내며, 이를 통해 알 수 있듯이, 대부분의 경우 계수에 대한 편향성이 DPM-SUR 모형이 가장 작은 값을 나타내고 있고, 평균제곱오차의 관점에서는 상자그림에도 나타나듯이 DPM-SUR이 모든 경우에 가장 작은 값을 나타내고 있다. 특히, 오차항의 분포가 정규혼합분포인 경우, 특정한 분포를 가정하는 다른 세 가지 모형과 본 논문에서 제안하는 DPM-SUR 모형 간의 차이는 두드러지는 것으로 나타났으며, 이를 통해 DPM-SUR 모형이 기존의 모형에 비해 오차항의 분포에 대해 더 강건하며, 아울러 LPML 값들을 통해서 살펴볼 수 있는 모형 선택의 관점에서도 가장 우수한 성능을 보임을 알 수 있었다.

4.2. 강수량 예측을 통한 실증분석

본 절에서는 본 논문에서 제안한 DPM-SUR 모형을 우리나라의 강수량 자료에 대한 실증적 분석을 통해 DPM-SUR 모형의 우수성을 보이고 기존의 모형들과 성능을 비교 분석 하고자 한다. 강수량의 계절 예측에 있어서 가장 널리 사용되는 방법은, 기후의 실제 관측값과 편미분 방정식으로 구현된 동적모형(dynamic model)으로 부터 얻어지는 예측값과의 통계적 함수관계를 이용하는 Model Output Statistics(MOS) 방법 (Glahn과 Lowry, 1972)으로, 동적모형으로 부터 얻어지는 예측값과 실제 관측값의 선형 회귀모형을 통해 편향수정(bias correction)을 실시하는 방식이다. 비록, MOS 방법으로 한 많은 연구가 이루어지고 있으나 (Baran와 Lerch, 2015; Jo 등 2012; Kang 등 2011; Lim 등 2012; 등), 각 지역별로 편향수정을 하기 때문에 기상자료가 가지고 있는 지역별 상관관계를 고려하지 않으며, 강수량의 경우 오차의 정규성 가정이 적절하지 않음이 알려져 왔다 (Raftery 등, 2005). 따라서 본 절에서는 비모수적 SUR모형을 이용하여 기존의 MOS방법이 가진 단점을 보완하고, 동시에 오차의 정규성 가정에 영향을 받지 않는 강수량 예측 실증분석을 하고자 하였다.

본 절에서, 강수량 예측을 위한 실증분석에서 사용한 실제 관측자료는 1979년부터 2007년까지 여름철(6, 7, 8월 자료의 평균)의 기간 동안 기상청(Korea Meteorological Administration)에서 측정한 자료로, Figure 4.3에 나타나 있는 바와 같이 총 12개 지역에 대한 자료이다. 동적모형자료는 기상청의 Global Data Assimilation and Prediction System(GDAPS; Park과 Hong, 2007) 모형으로부터 예측

Table 4.2. Summary results for precipitation data over Korea

Performance Measure	MCMC-SUR	MCMC-SURT	DPM-SUR	MLE-SUR
RMSE	0.5121	0.5134	**0.3294**	0.5127
LEPS	0.2477	0.2481	**0.1221**	0.2481

RMSE = root mean squared error, LPML = log pseudo marginal likelihood.

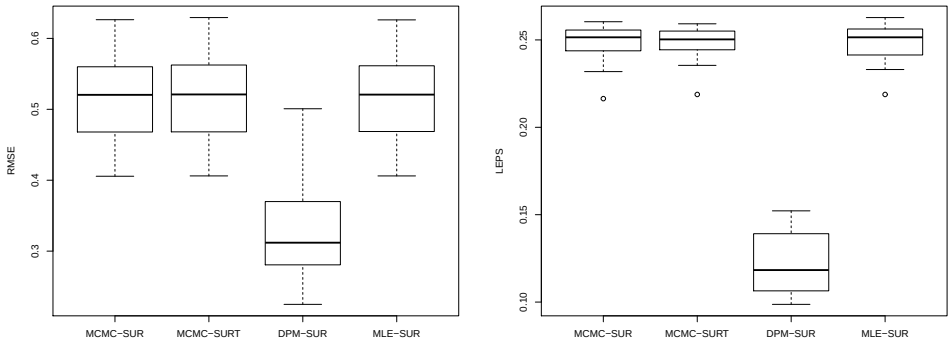

Figure 4.4. Root mean squared error (RMSE) and log pseudo marginal likelihood (LEPS) values for precipitation data over Korea.

된 값으로 총 29개의 앙상블(ensemble)로 이루어져 있다. 자료분석에서는 각 지역별 앙상블들의 평균을 설명변수(x)로 이용하였다. 실제 관측값과 앙상블 평균 간의 선형관계를 설명하기 위해서는 12개 지역 간의 종속적인 관계를 반영하는 것이, 보다 합리적인 분석이라고 할 수 있으며, 이에 다음과 같이, 12개 지역을 설명하는 SUR 모형을 가정하도록 한다.

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, 12, \ i = 1, \ldots, 29, \\ \boldsymbol{\epsilon}_i &= (\epsilon_{i,1}, \ldots, \epsilon_{i,12})^T \sim F(\boldsymbol{\epsilon}_i), \quad i = 1, \ldots, 29, \end{aligned} \tag{4.5}$$

여기서 i는 1979년부터 2007년까지의 년도를 나타내고, j는 12개의 지역을 나타낸다.

4.1절의 모의실험과 마찬가지로, 사후표본을 추출하기 위해 15,000번의 소각과정을 거친 후 10,000개의 표본을 생성하였다. 성능비교 측도로는 평균제곱오차(RMSE) 외에 기상학에서 널리 사용되는 linear error in probability space score(LEPS; Deque, 2003; Jo 등, 2012; Potts 등, 1996)을 추가로 사용하였다. LEPS는 확률공간에서 오차를 측정하는 측도로서 아래와 같이 계산된다.

$$\text{LEPS}_j = \frac{1}{29} \sum_{i=1}^{29} |F_o(\hat{y}_{ij}) - F_o(y_{ij})|, \quad j = 1, \ldots, 12, \tag{4.6}$$

여기서 F_o는 실제 관측치의 경험적 분포함수(empirical cumulative distribution function)를 나타내고, LEPS는 0과 1사이의 값을 가짐을 알 수 있다.

Table 4.2는 각각의 베이지안 SUR 모형(DPM-SUR, MCMC-SUR, MCMC-SURT)과 MLE-SUR을 적용한 결과를 나타내며, 평균제곱오차 RMSE와 LEPS의 관점에서 본 논문에서 제안하는 DPM-SUR이 자료를 가장 잘 적합하는 것으로 나타났다. 또한 Figure 4.4에서 알 수 있듯이 각 지역별 예측값에 있어서도 DPM-SUR이 가장 잘 적합하는 결과를 나타내고 있다. 따라서 오차항의 가정에 의존하지 않

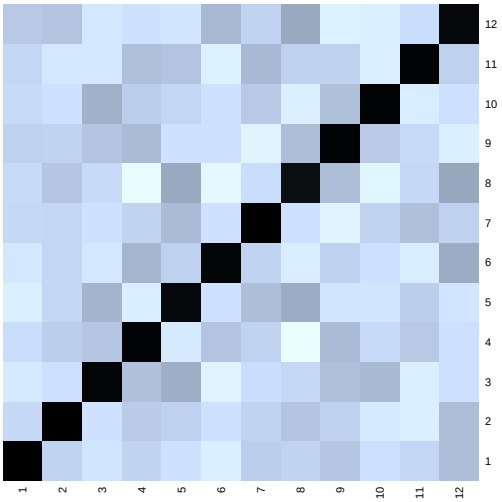

Figure 4.5. Estimated correlation matrix for precipitation data over Korea.

는 DPM-SUR이 강수량 예측에 있어서 적절한 모형이라 할 수 있다. Figure 4.5는 DPM-SUR모형으로 부터 추정된 상관행렬을 나타낸다. 특별한 오차항의 공분산 행렬에 대한 가정없이, DPM-SUR 모형을 적용하였으며, 이를 통하여 상관계수 행렬을 추정할 수 있었다. 이 경우 추정된 상관계수의 범위는 약 -0.1756에서 0.2507사이를 나타내고 있으며, 이러한 상관계수 행렬로 부터 지역간 강수량의 상관관계를 파악할 수 있을 것이다. 예를 들어, 울산(위치 8)은 여수(위치 12, $r = 0.2507$), 포항(위치 5, $r = 0.2324$), 광주(위치 9, $r = 0.1375$)와 양의 상관관계를 갖는 것으로 나타났으며, 추풍령(위치 4, $r = -0.1756$), 대구(위치 6, $r = -0.1601$), 부산(위치 10, $r = -0.1461$)과는 음의 상관관계를 갖는 것으로 나타났다. 그 외에, 강릉(위치 1, $r = -0.005$), 서울(위치 2, $r = 0.096$), 인천(위치 3, $r = -0.004$), 전주(위치 7, $r = -0.0199$), 목포(위치 11, $r = 0.006$)와는 거의 독립으로 나타났다. 즉, 울산은 1979년 부터 2007년까지 여름철 강수량의 패턴이 여수, 포항, 광주와는 유사하지만, 추풍령, 대구, 부산과는 다소 상이한 양태를 나타냄을 알 수 있었다.

5. 결론

본 논문에서는 비모수 베이지안에서 주로 사용되는 디리슈레 프로세스 혼합모형을 바탕으로 한 DPM-SUR 모형을 제안하였다. DPM-SUR 모형은, 오차항의 분포에 자유로우며, 특정한 모수적 가정에 영향을 받지 않는 강건한 모형으로서, 기존에 연구된 베이지안 SUR 모형에 비해 여러가지 측면에서 우수한 성능을 보임을 확인할 수 있었다. DPM-SUR 모형에 대한 사후추론을 위하여, SUR 모형과 디리슈레 프로세스 혼합 모형의 사후 추론 기법을 결합한 마코프 체인 몬테 칼로 알고리듬을 제시하였고, 모의실험 뿐 만 아니라, 강수량 예측을 위한 실제 자료 에 대한 실증적 성능비교를 통하여 제안된 DPM-SUR 모형의 우수성을 입증하였다.

본 논문에서 제안한 DPM-SUR 모형은 더욱 확대되어, 실제 자료분석에 활용되고, 다양한 방식의 SUR 모형으로 확대될 예정이다. 본 논문에서 고려한 DPM-SUR 모형은 설명변수와 반응변수간에 선형적인 관계만을 고려하였으나, 실제 응용 연구와 자료분석에서는, 선형적인 관계 외에 비선형적인 관계를 가지는 변수에 대한 모형화, 임의효과를 반영하는 복잡한 구조의 자료들에 대한 모형화, 그리고 차원 증

가에 따른 변수 선택 문제 등이 활발히 연구되고 있다 (Koop 등, 2005; Lang 등, 2003; Wang, 2010; 등). 따라서, 본 논문의 연구결과를 바탕으로 한 향후 과제로써, 선형과 비선형 관계의 설명변수를 모두 포함하는 고차원 부분선형 겉보기 회귀모형(partially linear seemingly unrelated regression)과 임의효과(random effects)를 반영하는 SUR 모형에 대한 연구를 진행 중에 있으며, 기상자료 분석에 있어서 고려해야 할 특성 중의 하나인 공간정보(spatial information)를 반영하는 SUR 모형으로 확장하고, 이 경우 디리슈레 혼합모형을 활용하는 연구 역시 진행 중에 있다. 또한 이상치(outlier)에 강건(robust)한 비대칭 라플라스 분포(asymmetric Laplace distribution)를 오차분포로 활용하는 겉보기 무관 분위수 회귀모형(seemingly unrelated quantile regression)에 관한 연구를 진행할 예정이다.

References

Aitken, A. C. (1935). On least-squares and linear combination of observations. In *Proceedings of the Royal Society of Edinburgh*, **55**, 42–48.

Aliprantis, C. D., Barnett, W. A., Cornet, B., and Durlauf, S. (2007). The interface between econometrics and economic theory, *Journal of Econometrics*, **136**, 325–724.

Ando, T. and Zellner, A. (2010). Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equations models using a combination of direct Monte Carlo and importance sampling techniques, *Bayesian Analysis*, **5**, 65–96.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics*, **2**, 1152–1174.

Baran, S. and Lerch, S. (2015). Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting, *Quarterly Journal of the Royal Meteorological Society*, DOI:10.1002/qj.2521

Chib, S. and Greenberg, E. (2010). Additive cubic spline regression with Dirichlet process mixture errors, *Journal of Econometrics*, **156**, 322–336.

Deque, M. (2003). "Continuous Variable" Chapter 5, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, New York.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, **90**, 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, **1**, 209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In M. Rizvi, J. Rustagi and D. Siegmund (Ed), *Recent Advances in Statistics* (pp. 287–302), Academic Press, New York.

Fraser, D. A. S., Rekkasb, M., and Wong, A. (2005). Highly accurate likelihood analysis for the seemingly unrelated regression problem, *Journal of Econometrics*, **127**, 17–33.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection, *Journal of the American Statistical Association*, **74**, 153–160.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed), Chapman & Hall/CRC, Florida.

Glahn, H. R. and Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting, *Journal of Applied Meteorology*, **11**, 1203–1211.

Greene, W. H. (2003). *Econometric Analysis* (5th ed), Prentice Hall, New Jersey.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Henningsen, A. and Hamann, J. D. (2007). systemfit: A package for estimating systems of simultaneous equations in R, *Journal of Statistical Software*, **23**, 1–40.

Jo, S., Lim, Y., Lee, J., Kang, H., and Oh, H. (2012). Bayesian regression model for seasonal forecast of precipitation over Korea, *Asia-Pacific Journal of Atmospheric Sciences*, **48**, 205–212.

Kang, J., Suh, M., Hong, K., and Kim, C. (2011). Development of updateable Model Output Statistics

(UMOS) System for Air Temperature over South Korea, *Asia-Pacific Journal of Atmospheric Sciences*, **47**, 199–211.

Koop, G., Poirier, D. J., and Tobias, J. (2005). Semiparametric Bayesian inference in multiple equation models, *Journal of Applied Econometrics*, **20**, 723–747.

Kowalski, J., Mendoza-Blanco, J. R., Tu, X. M., and Gleser, L. J. (1999). On the difference in inference and prediction between the joint and independent t-error models for seemingly unrelated regressions, *Communications in Statistics - Theory and Methods*, **28**, 2119–2140.

Lang, S., Adebayo, S. B., Fahrmeir, L., and Steiner, W. J. (2003). Bayesian geoadditive seemingly unrelated regression, *Computational Statistics*, **18**, 263–292.

Lim, Y., Jo, S., Lee, J., Oh, H., and Kang, H. (2012). An improvement of seasonal climate prediction by regularized canonical correlation analysis, *International Journal of Climatology*, **32**, 1503–1512.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates, *The Annals of Statistics*, **12**, 351–357.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*, Springer Series in Statistics.

Müller, P. and Rodríguez, A. (2013). *Nonparametric Bayesian Inference*, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 9, Institute of Mathematical Statistics.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Ng, V. M. (2002). Robust Bayesian inference for seemingly unrelated regressions with elliptical errors, *Journal of Multivariate Analysis*, **83**, 409–414.

Park, H. and Hong, S.-Y. (2007). An evaluation of a mass-flux cumulus parameterization scheme in the kma global forecast system, *Journal of the Meteorological Society of Japan*, **85**, 151–169.

Potts, J. M., Folland, C. K., Jolliffe, I. T., and Secton, D. (1996). Revised LEPS scores for assessing climate model simulations and long-range forecasts, *Journal of Climate*, **9**, 34–54.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles, *Monthly Weather Review*, **133**, 1155–1174.

Rodríguez, C. E. and Walker, S. G. (2014). Univariate Bayesian nonparametric mixture modeling with unimodal kernels, *Statistics and Computing*, **24**, 35–49.

Sethurman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica*, **4**, 639–650.

Wang, H. (2010). Sparse seemingly unrelated regression modelling: applications in finance and econometrics, *Computational Statistics and Data Analysis*, **54**, 2866–2877.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association*, **57**, 348–368.

Zellner, A. (1963). Estimators for seemingly unrelated regression equations: some exact finite sample results, *Journal of the American Statistical Association*, **58**, 977–992.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.

Zellner, A. and Ando, T. (2010a). A direct Monte Carlo approach for Bayesian analysis for the seemingly unrelated regression model, *Journal of Econometrics*, **159**, 33–45.

Zellner, A. and Ando, T. (2010b). Bayesian and non-Bayesian analysis of the seemingly unrelated regression model with Student-t errors, and its application for forecasting, *International Journal of Forecasting*, **26**, 413–434.

Zellner, A. and Chen, B. (2002). Bayesian modeling of economies and data requirements, *Macroeconomic Dynamics*, **5**, 673–700.

Zellner, A. and Tobias, J. (2001). Further results on Bayesian method of moments analysis of the multiple regression model, *International Economic Review*, **42**, 121–139.

비모수 베이지안 겉보기 무관 회귀모형

조성일[a] · 석인혜[a] · 최태련[a,1]

[a]고려대학교 통계학과

(2016년 2월 24일 접수, 2016년 4월 5일 수정, 2016년 4월 25일 채택)

요 약

본 논문에서는 겉보기 무관 회귀모형을 고려하고 디리크레 프로세스 혼합모형을 오차항의 분포로 하는 비모수 베이지안 방법을 제안한다. 제안된 모형을 바탕으로 사후분포를 유도하고 디리크레 프로세스 혼합모형의 붕괴깁스표집 방법을 통해 마코프 체인 몬테 칼로 알고리듬을 구성하고 사후추론을 실시한다. 모형의 성능을 비교하기 위해 모의실험을 실시하고, 더 나아가 한국지역의 강수량 예측에 대한 실제 자료에 적용해 본다.

주요용어: 겉보기 무관 회귀모형, 디리크레 프로세스 혼합모형, 붕괴깁스표집, 강수량 예측

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2013R1A1A2074463).

[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: trchoi@korea.ac.kr