The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.4.365
33(4), 365–377

Bayesian analysis of latent factor regression model

Minjung Kyung[a,1]

[a]Department of Statistics, Duksung Women's University

(Received February 4, 2020; Revised June 11, 2020; Accepted June 17, 2020)

Abstract

We discuss latent factor regression when constructing a common structure inherent among explanatory variables to solve multicollinearity and use them as regressors to construct a linear model of a response variable. Bayesian estimation with LASSO prior of a large penalty parameter to construct a significant factor loading matrix of intrinsic interests among infinite latent structures. The estimated factor loading matrix with estimated other parameters can be inversely transformed into linear parameters of each explanatory variable and used as prediction models for new observations. We apply the proposed method to Product Service Management data of HBAT and observe that the proposed method constructs the same factors of general common factor analysis for the fixed number of factors. The calculated MSE of predicted values of Bayesian latent factor regression model is also smaller than the common factor regression model.

Keywords: Bayesian latent factor model, LASSO prior, Gibbs sampling

1. 서론

일반적인 선형회귀모형 $\mathbf{y} = \mathbf{X}\beta + \epsilon$에서 $\mathbf{y}$는 길이 n의 반응값 벡터이고, $\mathbf{X}$는 $n \times p$ 예측변수의 행렬이고, β는 길이 p의 선형모수 벡터이며, $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$는 서로 독립인 길이 n의 오차항 벡터이다. 예측변수의 수가 관측개수 보다 큰 경우, $p >> n$, 차원축소를 목적으로 하는 많은 방법들이 발단하였다. 예를 들면, 주성분회귀모형(principal component regression) (Kendall, 1957; Hotelling, 1957; Jeffers, 1967), 능형회귀모형(ridge regression) (Hoerl과 Kennard, 1950), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) 등이 있다. 그 중 주성분회귀모형은 선형모형에서 모수의 추론 및 모형의 적합성에 문제가 되는 설명변수의 개수가 관측개수보다 많은 자료와 두개 이상의 설명변수들 사이에 존재하는 다중공선성(multicollinearity) 문제를 서로 직교가 되는 분산이 큰 부분집합 주성분을 회귀변수로 사용하여 추정변수의 수를 줄이는 변수축소의 방법이다. 이러한 주성분회귀모형은 인자회귀모형(factor regression)의 특수한 경우로, 인자모형은 다차원 자료에 대하여 공분산 구조를 몇 개의 관측 불가능한 인자로 설명한다. 인자모형을 사용하는 목적은 변수들 간에 내재되어 있는 공통의 구조를 파악하고, 데이터의 특성을 몇 개의 인자로 축약하여 설명함으로써 분석에 필요한 변수의 차원을 줄이는 것이다.

내재된 인자회귀모형(latent factor regression model)은 인자모형의 확장된 형태로, 예측변수들에 대한 공통의 구조를 관측되지 않은(내재된) 인자들(factors)로 설명하고, 각 인자의 적재값을 반응변수를

This research was supported by a Duksung Women's University research grants 3000003265.
[1]Department of Statistics, Duksung Women's University, 33 Samyangro 144-Gil, Dobong-Gu, Seoul 01369, Korea. E-mail: mkyung@duksung.ac.kr

설명하기 위한 예측변수로 사용하는 모형이다. 이러한 내재된 인자회귀모형은 행동과학 및 사회과학에 많이 사용되는 모형으로, 인자는 관찰되지 않는 특정 심리적 특성의 잠재요인으로 자연스럽게 해석된다. 이러한 내재된 인자들의 적절한 개수는 다양한 방법으로 정할 수 있다. 특히, 베이지안 방법을 적용한 내재된 인자모형(latent factor model)에서 Bhattacharya와 Dunson (2011)은 내재된 인자의 수를 무한대로 가정한 후 인덱스가 큰 인자의 적재값을 0으로 수렴하게 하는 감마분포의 곱을 활용한 축소 사전분포를 적재값에 적용하였다. West (2003)는 $p >> n$인 선형모형에서 각각의 인자적재행렬에 t-분포를 가정하고 일반적인 축소 사전분포를 적용하였다. 이는 t-사전분포를 계층적 구조(hierarchical structure)로 표현하기 위하여 인자 적재에 대하여 정규성을 가정하고 가정한 정규분포의 분산에 감마분포를 가정하여 혼합하는 방법을 사용하였고, 정규성 가정에 일반적인 축소 사전분포를 사용하여 구현하였다.

우리는 이 논문에서 내재된 인자회귀모형의 베이지안 분석법을 사용하고, 무한개로 가정 가능한 내재된 인자 중 유의미한 인자적재행렬(factor loading matrix)을 구성하기 위하여 LASSO 사전분포를 인자적재행렬에 적용한다. Kyung 등 (2010)은 선형모형에서 다중공선성 문제를 고려하기 위하여 LASSO 사전분포를 사용한 벌점회귀모형을 실제자료에 적용하였다. LASSO 사전분포를 적용한 베이지안 분석법에서 벌점모수(penalty parameter)는 LASSO 사전분포의 분산의 함수로 표현할 수 있으며, 사전분포를 적용하여 추정할 수 있는 장점이 있다. 그러나 모수에 대한 사후 추정량의 표준오차는 구할 수 있으나, Gibbs 표집과정에서 유동적으로 값이 변하기 때문에 예측변수를 선택하거나 차원을 축소하는 방법으로는 한계가 있다. LASSO 사전분포에서 벌점모수의 값이 작은 경우 무한개로 가정 가능한 인자 중 유의미한 인자적재행렬을 구성하는 것이 아닌, 임의적인 모든 인자적재행렬을 모형에 사용할 수 있다. 그러므로 벌점모수의 값이 작은 LASSO 사전분포를 사용하면 다중공선성이 있는 설명변수들을 일반 선형회귀모형에 적합시키는 결과로 얻어질 수 있는 모든 회귀모수가 유의하지 않거나 추정된 회귀모수가 유일하지 않다는 문제가 발생할 수 있다. 이 논문에서는 벌점모수(λ)의 사전분포의 평균값을 큰 값으로 설정하고 내재된 인자의 인자적재행렬에 LASSO 사전분포를 적용하여 예측변수에 내재된 잠재요인 중 적절한 인자적재행렬을 결정한다. 이를 통해 잠재요인에 대한 해석과 결정된 인자적재행렬을 예측행렬로 사용하여 반응변수와의 관계를 설명하는 선형회귀모형을 베이지안 방법으로 분석한다.

일반적인 인자분석에서 인자의 개수 k를 결정할 때 다음과 같은 몇 가지 기준들을 사용한다. 첫 번째로 각각의 인자가 기여하는 비율의 합을 사용하여 공통인자들에 의해 설명되는 분산의 비율이 전체 변이의 70–90%가 되도록 인자의 개수를 결정하는 방법이 있다. 두 번째로는 공분산 행렬의 고유값들의 평균을 구한 후 고유값이 평균값 이상이 되는 주성분을 사용하는 방법이 있고, 세 번째는 2차원 좌표에 고유값 순서와 고유값 크기로 점을 찍고 점간을 선분으로 연결하는 스크리 그림(scree plot)을 활용하는 방법이 있다. 값이 큰 고유값부터 크기순으로 점을 찍어 가파른 정도를 보고 가파른 부분에 해당하는 고유값까지로 인자의 개수를 결정하는 방법이다. 또 다른 방법으로는 적절한 인자의 개수를 정하고 우도비 검정법을 통해 인자의 개수를 결정하는 Bartlett (1951)의 검정법이 있다. 그러나 우리는 이 연구에서 인자의 개수를 설명변수들의 공분산행렬의 구조를 통해서 구하는 것이 아닌 반응변수와의 선형적 관계를 통하여 선택하며, 베이축지안 분석에서 많이 사용되는 Bayesian information criterion (BIC)을 기준으로 사용한다. BIC의 장점은 (1) 우도함수 값과 모형자유도, 그리고 표본의 크기만으로 쉽게 계산할 수 있다는 것과 (2) 내포모형뿐 아니라 내포되지 않은 모형의 적합도 비교에도 사용할 수 있다는 것 (Raftery, 1995)이다. 우리가 제시한 모형은 고차원의 자료를 다루는 것도 아니며 변수선택의 모형이 아니므로, BIC는 주어진 자료에 적합 가능한 다양한 모형 중 가장 적합한 모형을 선택할 때 사용할 수 있는 통계량이라 할 수 있다.

본 논문의 구성은 다음과 같다. 2절에서는 내재된 인자회귀모형의 구성에 대해서 설명하고, 3절에서는

모수의 사전분포에 대한 가정과 사후분포에 대해 설명한다. 그리고 4절에서는 실제자료에 적용한 결과에 대해 살펴보고, 5절에서는 요약과 결론으로 끝을 맺는다.

2. 내재된 인자회귀모형

내재된 인자회귀모형은 예측변수 행렬에 대한 인자를 찾은 후 그 인자들의 행렬을 반응변수와의 선형회귀모형에서 예측변수 행렬로 사용하는 구조이다. 즉, 다음과 같은 계층적 구조로 설명할 수 있다.

우선, 예측변수의 행렬 $\mathbf{X}$의 i번째 열벡터를 $\mathbf{x}_i'$이라 할 때, 예측변수들에 대한 내재된 인자모형은

$$\mathbf{x}_i = \mathbf{B}\boldsymbol{\lambda}_i + \boldsymbol{\nu}_i \tag{2.1}$$

이고, $i = 1, \ldots, n$에 대하여

$$\boldsymbol{\lambda}_i \sim N(\mathbf{0}, \boldsymbol{\Delta}), \quad \boldsymbol{\nu}_i \sim N(\mathbf{0}, \boldsymbol{\Phi})$$

을 가정한다. 여기에서 $\boldsymbol{\lambda}_i$는 길이 k의 i번째 예측변수들의 관측값에 대한 내재된 인자들의 벡터이고, $\mathbf{B}$는 $p \times k$ 인자적재행렬이고, $\boldsymbol{\nu}_i$는 독립인 오차 벡터이다. 이 가정에서 $\boldsymbol{\Delta}$와 $\boldsymbol{\Phi}$는 대각행렬이다. 일반적으로 인자의 개수인 k는 예측변수의 수 p보다 작고 고정되어 있다. 즉, 예측변수들의 관계를 내재된 k개의 인자와 오차로 설명할 수 있는 모형이다. 이러한 내재된 인자모형에 대한 베이지안 분석법은 Aguilar와 West (2000)에 설명되어있으며, 베이지안 인자모형들에 대한 참고 문헌도 Aguilar와 West (2000)에 자세히 설명되어 있다.

반응변수는 k개의 인자들과 다음과 같은 관계로 설명할 수 있다.

$$y_i = \boldsymbol{\lambda}_i'\boldsymbol{\theta} + \epsilon_i \tag{2.2}$$

이고 $\epsilon_i \sim N(0, \sigma^2)$로 가정한다. 다시 표현하면 $y_i|\boldsymbol{\lambda}_i \sim N(\boldsymbol{\lambda}_i'\boldsymbol{\theta}, \sigma^2)$로 내재된 인자들의 벡터가 설명변수가 되고 $\boldsymbol{\theta}$가 선형회귀모수인 선형회귀모형을 나타낸다. $\boldsymbol{\theta}$는 인자회귀모수이고, $\boldsymbol{\lambda}_i$는 식 (2.1)의 길이 k의 i번째 예측변수들의 관측값에 대한 내재된 인자들의 벡터이다. 즉, 원래의 예측변수들은 식 (2.1)을 통해서 내재된 관계에 대한 정보를 제공하지만, 선형회귀모형에 직접적으로 사용하지는 않다. 그러므로 반응값 y_i는 $\boldsymbol{\lambda}_i$가 주어졌을 때 $\mathbf{x}_i$와 조건부 독립관계가 된다.

위의 모형에서 y_i, $\mathbf{x}_i$와 $\boldsymbol{\lambda}_i$의 결합 분포는

$$f(y_i, \mathbf{x}_i, \boldsymbol{\lambda}_i) = N\left(y_i|\boldsymbol{\lambda}_i'\boldsymbol{\theta}, \sigma^2\right) N(\mathbf{x}_i|\mathbf{B}\boldsymbol{\lambda}_i, \boldsymbol{\Phi}) N(\boldsymbol{\lambda}_i|\mathbf{0}, \boldsymbol{\Delta})$$

이다. y_i와 $\mathbf{x}_i$의 결합 분포는

$$\begin{aligned} f(y_i, \mathbf{x}_i) &= \int f(y_i, \mathbf{x}_i, \boldsymbol{\lambda}_i)\, d\boldsymbol{\lambda}_i \\ &\propto \left(\sigma^2\right)^{-\frac{1}{2}} |\boldsymbol{\Delta}|^{-\frac{1}{2}} |\boldsymbol{\Phi}|^{-\frac{1}{2}} \left|\frac{1}{\sigma^2}\boldsymbol{\theta}\boldsymbol{\theta}' + \boldsymbol{\Delta}^{-1} + \mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{B}\right|^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}y_i^2 - \frac{1}{2}\mathbf{x}_i'\boldsymbol{\Phi}^{-1}\mathbf{x}_i\right) \\ &\quad \times \exp\left\{\frac{1}{2}\left(\frac{1}{\sigma^2}\boldsymbol{\theta}y_i' + \mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{x}_i\right)'\left(\frac{1}{\sigma^2}\boldsymbol{\theta}\boldsymbol{\theta}' + \boldsymbol{\Delta}^{-1} + \mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{B}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{\theta}y_i' + \mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{x}_i\right)\right\} \end{aligned}$$

가 된다. 그리고 $\mathbf{x}_i$와 $\boldsymbol{\lambda}_i$의 결합 분포 $N(\mathbf{x}_i|\mathbf{B}\boldsymbol{\lambda}_i,\ \boldsymbol{\Delta})\, N(\boldsymbol{\lambda}_i|\mathbf{0},\ \boldsymbol{\Delta})$로 부터 $\mathbf{x}_i$에 대한 주변 확률 분포

를 구하면

$$f(\mathbf{x}_i) = \int f(\mathbf{x}_i, \boldsymbol{\lambda}_i)\, d\boldsymbol{\lambda}_i$$
$$\propto (\sigma^2)^{-\frac{1}{2}} \left| \frac{1}{\sigma^2}\boldsymbol{\theta}\boldsymbol{\theta}' + \boldsymbol{\Delta}^{-1} + \mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{B} \right|^{\frac{1}{2}} \mid \boldsymbol{\Delta}^{-1} + \mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{B}\mid^{\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{x}_i'\boldsymbol{\Phi}^{-1}\mathbf{x}_i\right)$$
$$\times \exp\left\{ \frac{1}{2} \left(\mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{x}_i\right)' \left(\boldsymbol{\Delta}^{-1} + \mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{B}\right)^{-1} \left(\mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{x}_i\right) \right\}$$

가 되어 위의 식들로 부터, $\boldsymbol{\lambda}_i$가 주어졌을 때 y_i의 조건부 분포함수는 $f(y_i, \mathbf{x}_i)/f(\mathbf{x}_i)$를 통해 구할 수 있다. 즉, $\boldsymbol{\lambda}_i$가 주어졌을 때 y_i의 조건부 분포는

$$y_i|\mathbf{x}_i \sim N\left(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2 + \boldsymbol{\theta}'\boldsymbol{\Sigma}\boldsymbol{\theta}\right) \tag{2.3}$$

이 되며, 여기에서

$$\boldsymbol{\beta} = \boldsymbol{\Phi}^{-1}\mathbf{B}\boldsymbol{\Sigma}\boldsymbol{\theta}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Delta}^{-1} + \mathbf{B}'\boldsymbol{\Phi}^{-1}\mathbf{B}. \tag{2.4}$$

그러므로 y_i에 대한 $\mathbf{x}_i$의 관계는 선형회귀모형으로 설명가능하며, 저차원(low-dimension) 인자회귀모수 $\boldsymbol{\theta}$의 고차원(high-dimension) $\boldsymbol{\beta}$로의 확장으로 모형화할 수 있다.

이러한 내재된 인자회귀모형에서 인자의 개수 k를 예측변수의 수 p보다 작은 수로 고정하지 않고, 무한개의 가능한 인자로 가정 한 후 유의한 인자를 선택하는 방법으로 확장시킨 내재된 인자회귀모형에 대한 베이지안 분석법을 West (2003)가 제안하였다. West (2003)가 제안한 방법은 무한대로 확장 가능한 인자적재행렬에 스파이크와 슬래브 사전분포(spike-and-slab prior)를 사용하여 마르코프 체인 몬테 카를로(Markov chain Monte Carlo; MCMC) 샘플링에서 유의한 인자만 선택하는 분석법이다. 이러한 스파이크와 슬래브 사전분포를 이용한 방법은 인자적재행렬의 사후분포에서 각 관측값에 대한 각각의 인자 적재값 $\mathbf{B}_{ij}$ 에 대해 각각의 샘플링을 해야 하는 단점이 있다. 이에 이 연구에서는 벌점모수의 값을 매우 큰 값으로 고정한 LASSO 사전분포를 적용하여, 깁스표본 안에서 각 인자별 벡터의 형태로 샘플링하는 방법을 설명한다. 사전분포와 사후 분포에 대한 자세한 내용은 다음 장에 서술한다.

3. 베이지안 추정법

내재된 인자회귀모형에 대한 베이지안 추정법은 크게 두 단계로 나누어 생각한다. 첫 번째 단계는 예측변수들에 대한 내재된 인자모형의 모수들에 대한 베이지안 추정법이고, 두 번째 단계는 적절한 인자적재행렬을 구성한 후 k개의 인자들과 반응변수의 선형회귀모형에 대한 베이지안 추정법 적용이다. 이 연구에서 우리는 인자의 개수 k의 값을 K에서 1까지 변화시키며 각 k에서 첫 번째 단계를 M_1번 반복하여 인자적재행렬을 구성하고 두 번째 단계를 M_2번 반복하여 회귀모수에 대한 추정값을 얻고 다음과 같은 BIC를 얻는다.

$$\text{BIC} = -2 \times \ln(\hat{\pi}) + p\ln(n)$$

으로 $\hat{\pi}$는 베이지안 사후 최빈값(Bayesian posterior mode)이고 p는 모수의 개수이다. 이 연구에서 BIC를 고려하는 이유는 인자의 개수인 k를 결정하는데 있어서 적절한 기준을 제시하기 위함이다. 인자회귀모형에서 비교하는 모형들은 인자의 개수를 선택하는 문제이므로 단순히 설명변수들의 관계식으로 만들어지는 인자의 개수를 선택하는 문제가 아닌 반응변수와의 설명력이 높은 내재된 부분집합 인자를 선택하기 위하여 BIC를 사용한다.

3.1. LASSO 사전분포

LASSO의 L_1 norm 벌점함수는 이중 지수분포의 핵함수(kernel function)로 표현가능하다. 이에 일반적인 선형회귀모형 $\mathbf{y} = \mathbf{X}\beta + \epsilon$에서 Park과 Casella (2008)는 라플라스 분포를 정규분포의 척도모수에 지수분포를 혼합하여 표현한 베이지안 계층모형으로 깁스 표본(Gibbs sampling) 방법을 제안하였다. 그들은 다음과 같은 조건부 라플라스 사전분포를 사용하여 완전 베이즈 사후분포를 서술하였다.

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\frac{\lambda|\beta_j|}{\sigma}}, \tag{3.1}$$

λ는 벌점모수이고 β_j는 일반적인 선형회귀분석에서의 j번째 설명변수의 회귀모수이며 σ^2는 선형모형의 오차항의 분산이다. 여기에서 $\pi(\sigma^2) = 1/\sigma^2$인 무정보적 척도불변성 주변 사전분포를 사용하였고, 단봉형인(unimodal) 사후분포를 확인하기 위해 σ^2 주어졌다는 조건부 분포의 중요성을 설명하였다. 왜냐하면 단봉형의 사후분포가 아니면 깁스 표본의 수렴이 늦어지며, 점추정값이 의미 없어지기 때문이다. 이러한 사전분포는 설명변수 행렬 $\mathbf{X}$이 주어졌을 때 유의한 변수에 더 많은 설명값을 허용하는 방법으로 사용된다. 벌점모수에 사전분포를 적용하여 사후 추정량과 사후 추정량의 표준오차는 구할 수 있으나, Gibbs 표집과정에서 유동적으로 값이 변하기 때문에 예측변수를 선택하거나 차원을 축소하는 방법으로는 한계가 있다. 그러므로 이 논문에서는 무한개의 가능한 인자로 부터 유의미한 인자적재행렬을 얻기 위하여 벌점모수(λ)의 사전분포의 평균값을 큰 값으로 설정한 LASSO 사전분포를 내재된 인자의 인자적재행렬에 적용하여 예측변수에 내재된 잠재요인 중 유의미한 인자적재행렬을 구성한다. 인자의 개수 $k\,(k = 1, \ldots, K)$와 벌점모수의 값이 주어졌을 때, 인자적재행렬에 가정하는 사전분포는 다음과 같다. $j = 1, \ldots, p$에 대하여

$$\begin{aligned} \boldsymbol{b}_j' \,|\, \tau_{1j}^2, \ldots, \tau_{kj}^2 &\sim N_k\left(\mathbf{0}_k, \boldsymbol{D}_{\tau_j}\right), \quad \boldsymbol{D}_{\tau_j} = \operatorname{diag}\left(\tau_{1j}^2, \ldots, \tau_{kj}^2\right), \\ \tau_{1j}^2, \ldots, \tau_{kj}^2 &\sim \prod_{h=1}^{k} \frac{\lambda^2}{2} e^{-\frac{\lambda^2 \tau_{hj}^2}{2}}\, d\tau_j^2, \quad \tau_1^2, \ldots, \tau_p^2 > 0, \end{aligned} \tag{3.2}$$

여기에서 $\boldsymbol{b}_j = (b_{1j}, \ldots, b_{kj})$은 인자적재행렬 $\mathbf{B}$의 j번째 행벡터이다.

3.2. 인자모형에 대한 사후분포

첫 번째 단계인 예측변수들에 대한 내재된 인자모형의 모수들에 대한 베이지안 추정법이다.

3.2.1. 조건부 사후분포 내재된 인자모형 (2.1)에서 인자와 오차에 대한 분포 가정으로 $i = 1, \ldots, n$에 대하여

$$\boldsymbol{\lambda}_i \sim N\left(\mathbf{0}, \boldsymbol{\Delta}\right), \quad \boldsymbol{\nu}_i \sim N\left(\mathbf{0}, \boldsymbol{\Phi}\right)$$

을 사용하였다. 여기에서 $\boldsymbol{\Delta}$와 $\boldsymbol{\Phi}$는 대각행렬로 일반성을 잃지 않고 $\boldsymbol{\Delta} = \mathbf{I}_k$를 가정하며 $\boldsymbol{\Phi} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2)$에서 $j = 1, \ldots, p$에 대하여

$$\sigma_j^{-2} | a, b \sim \mathrm{G}a(a, b)$$

감마(Ga) 사전분포로 사용한다. 그리고 $\boldsymbol{\Lambda} \equiv (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n)'$로 정의한다.

인자의 개수 k $(k = 1, \ldots, K)$와 벌점모수의 값이 주어졌을 때 위에서 제시한 사전분포와 오차항의 정규성 가정으로 부터 얻은 조건부 사후분포는 다음과 같다.

P1. $\boldsymbol{B}$의 조건부 사후분포: $j = 1, \ldots, p$에 대하여

$$\boldsymbol{b}_j' \,|\, \boldsymbol{D}_{\tau_j}, \boldsymbol{\Phi}, \boldsymbol{\Lambda}, \boldsymbol{X} \sim N_k \left(\left(\sigma_j^{-2} \sum_{i=1}^{n} \boldsymbol{\lambda}_i \boldsymbol{\lambda}_i^T + \boldsymbol{D}_{\tau_j}^{-1} \right)^{-1} \left(\sigma_j^{-2} \sum_{i=1}^{n} x_{ij} \boldsymbol{\lambda}_i \right), \left(\sigma_j^{-2} \sum_{i=1}^{n} \boldsymbol{\lambda}_i \boldsymbol{\lambda}_i^T + \boldsymbol{D}_{\tau_j}^{-1} \right)^{-1} \right).$$

P2. $\boldsymbol{\Lambda}$의 조건부 사후분포: $i = 1, \ldots, n$에 대하여

$$\boldsymbol{\lambda}_i \,|\, \boldsymbol{D}_\tau, \boldsymbol{\Phi}, \boldsymbol{B}, \boldsymbol{X} \sim N_k \left(\left(\boldsymbol{B}^T \boldsymbol{\Phi}^{-1} \boldsymbol{B} + \mathbf{I} \right)^{-1} \left(\boldsymbol{B}^T \boldsymbol{\Phi}^{-1} \boldsymbol{x}_i \right), \left(\boldsymbol{B}^T \boldsymbol{\Phi}^{-1} \boldsymbol{B} + \mathbf{I} \right)^{-1} \right).$$

P3. $\boldsymbol{\Phi}$의 조건부 사후분포: $j = 1, \ldots, p$에 대하여

$$\sigma_j^{-2} \,|\, \boldsymbol{D}_\tau, \boldsymbol{\Lambda}, \boldsymbol{B}, \boldsymbol{X} \sim \mathrm{Ga} \left(\frac{n}{2} + a, \, \frac{1}{2} \sum_{i=1}^{n} (x_{ij} - \boldsymbol{b}_j \boldsymbol{\lambda}_i)^2 + b \right).$$

P4. $\boldsymbol{D}_\tau$의 조건부 사후분포: $j = 1, \ldots, p$와 $h = 1, \ldots, k$에 대하여

$$\tau_{jh}^{-2} \,|\, \boldsymbol{\Phi}, \boldsymbol{\Lambda}, \boldsymbol{B}, \boldsymbol{X} \sim \text{Inverse Gaussian} \left(\frac{\lambda}{|b_{jh}|}, \lambda^2 \right).$$

3.3. 회귀모형에 대한 사후분포

3.2.1절로 부터 구한 k개의 인자를 설명변수로 사용한 다음의 회귀모형에서

$$y_i = \boldsymbol{\lambda}_i' \boldsymbol{\theta} + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right).$$

모수 $\boldsymbol{\theta}$와 σ^2에 대한 사전분포는 공액사전분포(conjugate prior distirbution)인 다음 분포를 사용한다.

$$\sigma^2 \sim \mathrm{IG}\left(a_0, b_0\right), \quad \boldsymbol{\theta} | \sigma^2 \sim N_{k+1}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right),$$

여기에서 IG는 역감마 분포이고 $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_k)$이다.
$\boldsymbol{\theta}$와 σ^2에 대한 조건부 사후분포는 다음과 같다.

P1. $\boldsymbol{\theta}^*$의 조건부 사후분포:

$$\boldsymbol{\theta} \,|\, \sigma^2, \boldsymbol{\Lambda}, \boldsymbol{y} \sim N_{k+1} \left(\left(\boldsymbol{\Lambda}' \boldsymbol{\Lambda} + \mathbf{I} \right)^{-1} \boldsymbol{\Lambda}' \boldsymbol{y}, \; \sigma^2 \left(\boldsymbol{\Lambda}' \boldsymbol{\Lambda} + \mathbf{I} \right)^{-1} \right),$$

여기에서 $\boldsymbol{\Lambda} \equiv (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n)'$이다.

P2. σ^2의 조건부 사후분포:

$$\sigma^2 \,|\, \boldsymbol{\theta}, \boldsymbol{\Lambda}, \boldsymbol{y} \sim \mathrm{IG} \left(\frac{n}{2} + \frac{k+1}{2} + a_0, \; \frac{1}{2} (\boldsymbol{y} - \boldsymbol{\Lambda}\boldsymbol{\theta})' (\boldsymbol{y} - \boldsymbol{\Lambda}\boldsymbol{\theta}) + \frac{1}{2} \boldsymbol{\theta}' \boldsymbol{\theta} + b_0 \right).$$

4. 제품 서비스 관리 자료 분석

제품 서비스 관리 자료는 다변량 분석기법을 설명하기 위하여 Hair 등 (2014)이 사용한 데이터로, 종이 제품 업체인 HBAT 산업 고객의 설문조사를 한 것이다. 이 자료는 HBAT 고객의 시장 세분화를 목적으로 수집된 자료로 $p = 18$개의 변수에 대해 $n = 100$개의 관측값으로 구성되어 있다. 이 중 제안한 방법론, 내재된 회귀 인자선형모형의 베이지안 분석법, 적용을 위해 사용하는 반응변

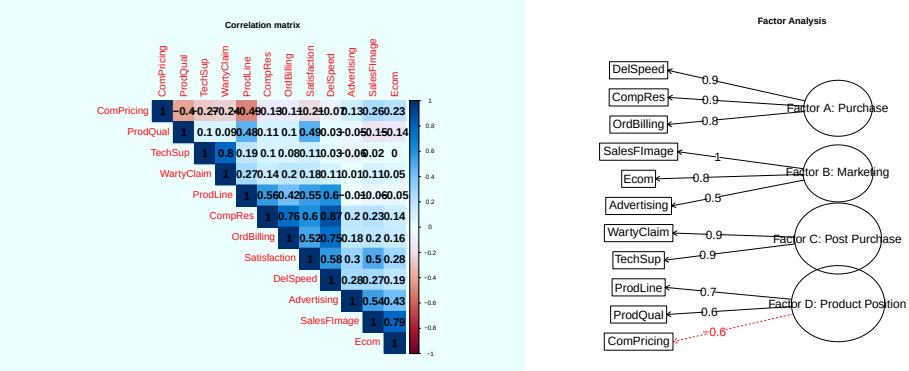

Figure 4.1. Correlation coefficients and common 4 factor plot of market segment data of HBAT customers.

수는 고객만족도(Satisfaction)이고 설명변수들은 제품의 품질 (ProdQual), 사용자 편의의 웹 접근성(Ecom), 기술적 지원(TechSup), 불만 해결(CompRes), 광고(Advertising), 제품 라인(Prodline), 영업(SalesFImage), 가격 경쟁력(ComPricing), 보증 및 요구(WartyClaim), 주문 및 청구(OrdBilling), 배송 속도(DelSpeed)이다. Figure 4.1의 상관계수행렬 그림에서 설명변수들 사이에 다중공선성이 의심되는 상관계수 값들을 발견할 수 있고, 모든 설명변수들을 사용하여 고객만족도에 대해 선형모형을 사용하면 다중공선성 문제로 인하여 유일하지 않는 선형모수의 추정값과 안정되지 않는 분산 추정값으로 모수의 추론의 정확성을 의심할 수 있다.

제품 서비스 관리 자료의 설명변수행렬에 주성분 분석을 통하여 공통인자분석을 수행하였을 때 최적의 인자의 개수는 $k = 4$로 베리맥스 회전(varimax rotation)을 적용한 인자적재행렬 추정값은 Figure 4.1에서 확인할 수 있다 (최적의 인자의 개수를 선택하기 위한 스크리 그림은 생략한다). 첫 번째 인자는 배송 속도, 불만 해결, 주문 및 청구 변수의 선형조합으로 설명되는 영업 전략에 대한 인자라 할 수 있고, 두 번째 인자는 전체적 영업에 대한 느낌을 나타내는 영업 변수, 사용자 편의의 웹 접근성, 광고 변수의 선형조합인 마케팅 점수, 세 번째 인자는 보증 및 요구와 기술적 지원에 해당하는 판매 후 서비스 점수, 그리고 제품라인, 제품 품질, 가격 경쟁력의 선형조합인 가격 결정 인자로 구분할 수 있다. 그러나 이러한 설명변수행렬에 대한 인자분석은 설명변수들만의 선형조합으로 내재된 관계를 찾아내고, 최적의 인자의 개수를 결정하는 한계점이 있다. 고객만족도라는 반응변수와의 관계는 최적의 인자의 개수와 인자적재행렬이 결정되고 난 후 인자들과 반응변수와의 선형모형으로만 설명되어지며, 새로운 자료들이 관측되었을 때 고객만족도를 예측할 수 있는 모형적인 장점은 기대할 수 없다. 이는 기존의 인자선형모형의 문제점으로 이러한 문제를 고려하기 위하여 3장에서 제안한 내재된 인자선형모형의 베이지안 분석법을 적용한다.

제 3장에서 설명한 사전분포를 고려한 MCMC 과정을 $k = 1, \ldots, 10$ $(< p = 11)$의 범위에 모두 적용한 후 각각의 k값에 대한 BIC를 구하고 모수를 추정하여, 식 (2.3)을 활용하여 각 설명변수의 모수 $\boldsymbol{\beta}$로 역변환을 한다. 표본과정은 각각의 k값에 대하여 20,000번 반복하여 10,000은 burn-in으로 제거한 후 나머지 10,000개의 표본 중 매 5번째 표본만을 선택하여 최종 2,000개의 표본을 내재된 인자모형의 사후추론에 사용한다. LASSO 사전분포에서 벌점모수 λ의 사전분포에 대하여 인자의 개수 $k = 4$로 고정한 후 평균이 100이고 분산이 1인 감마분포와, 평균이 50이고 분산이 5인 감마분포, 평균이 25이고 분산이 5인 감마분포, 평균이 10이고 분산이 1인 감마분포, 평균이 5이고 분산이 1인 감마분포를 하였다. 각 사전분포에 대한 내재된 인자적재행렬을 비교한 결과 사전분포의 평균이 25보다 큰 감마분포를 적용한 경우의 인자적재행렬은 주성분 분석을 통한 공통인자분석에서의 인자적재행렬 추정값과 부호는 다르

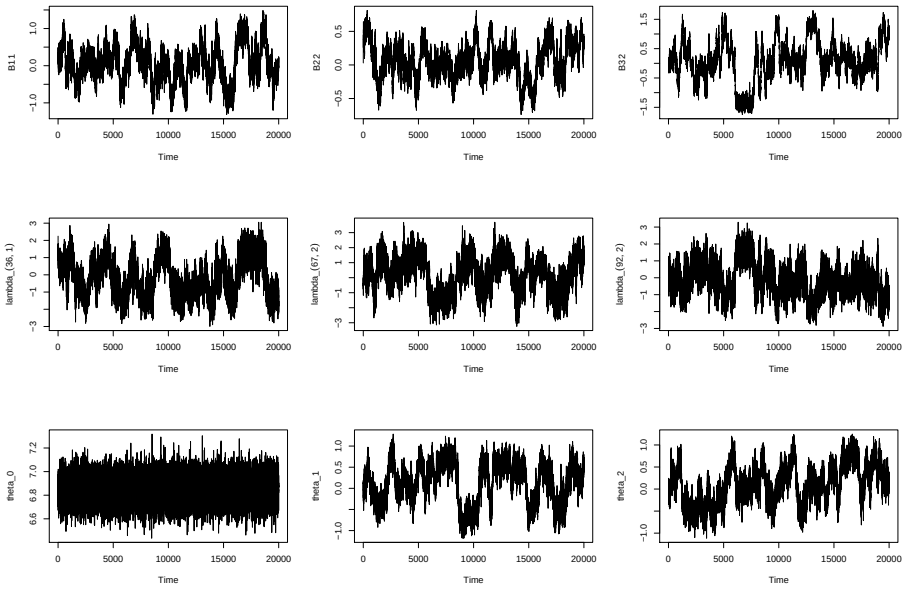

Figure 4.2. Trace plots of posterior sampling of factor loading $\mathbf{B}$, factor $\mathbf{\Lambda}$ and factor regression coefficients θ based on LASSO prior with Gamma(mean = 5, var = 1) prior on penalty parameter λ.

나 관계성 및 수치가 비슷한 결과가 나왔다. 그러나 사전분포의 평균이 10보다 작은 감마분포를 사용한 경우 인자적재행렬의 추정값의 수치가 작으며, 4개의 인자를 확실히 구분할 만한 결과를 제시하지 못하는 것을 확인할 수 있었고, 인자적재행렬($\mathbf{B}$)과 인자벡터(λ_i), 그리고 인자회귀모수(θ)의 마르코프 체인(Markov chain)의 수렴이 만족하지 않는 것을 확인할 수 있었다. 이는 평균이 5이고 분산이 1인 감마분포를 LASSO 사전분포의 벌점모수 λ의 사전분포에 적용한 추정모수들의 사후 표본 그림의 일부인 Figure 4.2에서 확인할 수 있다. 그리고 평균이 25 보다 큰 사전분포를 LASSO의 벌점모수 λ에 적용한 경우, 추정된 인자적재행렬에 대한 사후추론에 차이가 없어, 평균이 100이고 분산이 1인 감마분포를 LASSO의 벌점모수 λ의 사전분포로 사용한다 (각 평균별 추정된 인자적재행렬의 값은 생략한다).

Figure 4.3는 $k = 1, \ldots, 10$개의 인자에 대하여 내재된 선형인자회귀모형을 적용한 BIC값으로 평균적으로 가장 작은 BIC값은 인자의 개수가 $k = 4$인 경우로, 설명변수들만의 인자분석에서 최적의 인자의 개수가 $k = 4$개로 설정된 결과 같은 결과를 보여주는 것을 확인할 수 있었다. 그리고 4개보다 작은 인자의 개수에 대한 BIC값의 평균과 중위수는 비슷한 값을 나타내며 이는 인자의 개수가 작아질수록 다중공선성 문제를 효과적으로 해결하지 못하는 것으로 생각할 수 있다.

공통인자분석에서 선택한 최적의 인자의 개수 $k = 4$에 대한 내재된 선형인자회귀모형의 베이지안 분석법으로 추정된 인자분석 행렬 추정값은 Table 4.1에 있다. 표에 주어진 각 인자는 설명변수들의 인자분석으로 확인된 영업 전략 인자, 마케팅 인자, 판매 후 서비스 인자와 가격 결정 인자인 4개의 인자와 일치하는 것을 확인할 수 있다. 이는 내재된 선형인자회귀모형의 베이지안 분석법이 설명변수들의 인자분석 방법도 구현하며, 반응변수와의 선형관계를 통하여 새로운 관측값에 대한 예측 역시 가능하다는 것을 확인한 결과라 할 수 있다.

제안한 내재된 선형인자회귀모형의 베이지안 분석법을 일반적인 인자선형모형과 비교하기 위하여, $n =$ 100개의 자료의 70%는 모수의 추정에 사용하고 나머지 30%의 설명변수들은 새로운 관측값으로 추정

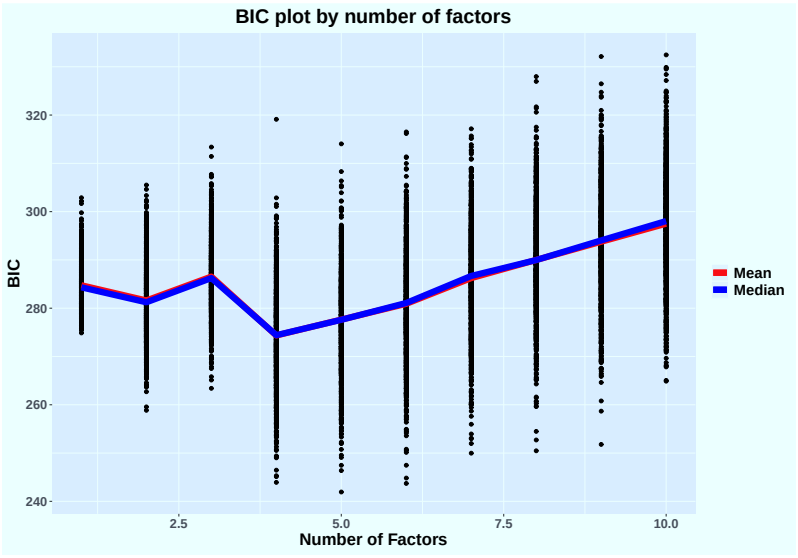

Figure 4.3. BIC of various number of factors k for Bayesian latent factor regression model.

Table 4.1. Estimated factors based on Bayesian latent factor regression model with $k = 4$

	영업 전략 (인자1)	마케팅 (인자2)	가격 결정 (인자3)	판매 후 서비스 (인자4)
ProdQual	0.023	−0.033	0.733	−0.007
Ecom	0.020	0.494	−0.044	0.008
TechSup	0.015	−0.011	0.078	0.989
CompRes	0.950	0.069	0.044	0.021
Advertising	0.095	0.516	−0.033	−0.035
ProdLine	0.602	−0.040	0.694	0.076
SalesFImage	0.086	0.820	-0.108	0.034
ComPricing	−0.057	0.172	−0.748	−0.177
WartyClaim	0.051	0.043	0.060	0.623
OrdBilling	0.633	0.045	0.000	0.028
DelSpeed	0.565	0.079	0.031	−0.021

값으로 부터 역변환된 선형모형의 모수와 설명변수들을 예측하는데 사용하였다. 일반적인 인자선형모형은 $n = 100$개의 관측값에 대하여 설명변수들의 공통 인자분석 시행 후, $100 \times k$ 인자행렬에서 70행을 임의로 선택하여 반응변수와의 선형모형에 적용하여 선형모수의 추정값을 구한 후 나머지 30행의 인자행렬은 반응변수를 예측하는데 사용하였다. 두 모형의 예측값을 비교하기 위하여 예측된 값들의 평균제곱 오차(mean squared error; MSE)값을 구하였고,

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 . \tag{4.1}$$

$k = 1, \ldots, 10$개의 인자에 대하여 구한 두 방법의 MSE값은 Figure 4.4에서 확인할 수 있다. 우선 일반적인 인자분석은 설명변수의 개수 $p = 11$에 대하여 최대 $K = 6$개의 인자만 구할 수 있었고, 식 (2.3)을 활용하여 구한 베이지안 분석법의 내재된 인자선형모형의 MSE값(실선)이 일반적인 인자선형

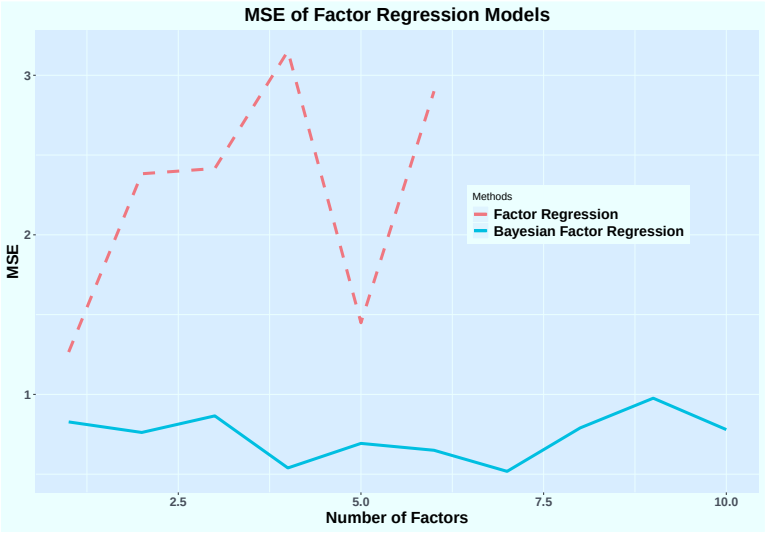

Figure 4.4. The calculated MSE of predicted values for each number of factors for the comparison between Bayesian latent factor regression model (bold line) and the common factor regression model (dashed line).

Table 4.2. Estimated regression parameters based on linear regression model and Bayesian latent factor regression model with $k = 4$

	Linear regression		Bayesian factor regression	
	Estimates	Std.Error	Estimates	Std.Error
Intercept	−0.67	0.81	6.85	1.06
ProdQual	0.37	0.05	0.09	0.03
Ecom	−0.44	0.13	0.19	0.06
TechSup	0.03	0.06	−0.01	0.05
CompRes	0.17	0.10	0.16	0.05
Advertising	−0.03	0.06	0.06	0.02
ProdLine	0.14	0.08	0.23	0.06
SalesFImage	0.81	0.10	0.24	0.07
ComPricing	−0.04	0.05	−0.07	0.03
WartyClaim	−0.10	0.12	0.03	0.04
OrdBilling	0.15	0.10	0.08	0.03
DelSpeed	0.17	0.20	0.19	0.05

모형의 MSE값(점선) 보다 작다는 것을 확인할 수 있다. 일반적인 인자선형모형의 경우 $k = 1$과 $k = 5$일 때 최소 MSE값을 가졌고, 베이지안의 경우 인자의 개수가 4개 ($k = 4$)인 경우 최소 MSE값을 갖는 것을 알 수 있다.

제품 서비스 만족도 자료에 고객만족도를 반응변수로 하고 다른 변수들을 설명변수로 하여 적합한 회귀모형의 모수의 추정값과 인자의 개수를 $k = 4$개로 하여 적용한 베이지안 분석법에서 추정된 Table 4.1의 인자분석 행렬 추정값을 바탕으로 계산한 식 (2.3)의 값은 Table 4.2에 있다. 다중공선성이 존재하는 자료를 고객만족도(Satisfaction)를 반응변수로 하는 일반선형회귀모형에 적합하였을 때 회귀모수의 95% 신뢰구간(credible interval)에 0을 포함하지 않는 유의한 변수는 제품의 품질(ProdQual), 사용자 편의의 웹 접근성(Ecom), 그리고 영업(SalesFImage)변수이다. 이 변수들의

회귀모수 추정값은 다른 변수들의 회귀모수 추정값에 비해 큰 값을 갖는 것을 확인할 수 있다. 반면 베이지안 인자회귀모형을 바탕으로 추정한 각 설명변수들의 회귀모수 추정값을 회귀모수의 95% 신뢰구간을 바탕으로 살펴보면, 고객만족도(Satisfaction)를 유의하게 설명하는 변수는 기술적 지원(TechSup)과 보증 및 요구(WartyClaim)를 제외한 제품의 품질(ProdQual), 사용자 편의의 웹 접근성(Ecom), 불만 해결(CompRes), 광고(Advertising), 제품 라인(Prodline), 영업(SalesFImage), 가격 경쟁력(ComPricing), 주문 및 청구(OrdBilling), 배송 속도(DelSpeed)이다. 설명변수들의 조합에 의해 만들어지는 4개의 직교 공간(인자)을 사용하여 다시 재구성한 회귀모형의 표준 오차값들은 매우 작았으며, 절편의 추정값이 다른 모수의 추정값에 비해 매우 크게 나타나는 것을 알 수 있다. 이는 직교성을 이용한 베이지안 인자분석 회귀모형이 근사적으로 설명변수의 공간을 다시 구성하여 반응변수와 설명변수들의 관계를 설명하는 회귀모형의 정확성을 높이며, 예측 가능한 모형을 생성한다는 것을 설명한다고 할 수 있다.

5. 결론

선형모형에서 모수의 추론 및 모형의 적합성에 문제가 되는 두개 이상의 설명변수들 사이에 존재하는 다중공선성 문제를 설명변수들의 선형결합들로 구성하는 인자들을 회귀변수로 사용하여 해결하는 인자회귀모형에 대하여 논의하였다. 이는 차원축소의 문제와 모수 추정의 문제를 동시에 수행할 수 있는 선형모형으로 의미가 있지만, 반응변수의 예측을 목적으로 하는 분석에서는 적합한 모형이 아닌 것을 확인할 수 있었다.

이 논문에서 제시한 내재된 인자회귀모형의 베이지안 추정방법은 인자적재행렬에 LASSO 사전분포를 사용하여 인자회귀모수에 연관된 모수들을 추정한 후 다시 원 회귀계수로 역변환 하는 사후 표본 추출과정을 사용한 방법이다. LASSO 사전분포의 벌점모수의 값을 큰 값으로 고정하여 내재된 인자 중 유의미한 인자적재행렬을 구성하고, 이를 통해 잠재요인에 해석과 결정된 인자적재행렬을 예측행렬로 사용하여 반응변수와의 관계를 선형모형으로 설명하는 가능하게 하는 방법론으로 사용할 수 있다는 장점을 확인할 수 있었다.

제시한 방법을 적용한 자료분석 결과 내재된 선형인자회귀모형을 적용한 BIC값은 평균적으로 인자의 개수가 감소할 때 BIC값이 감소하는 것을 확인할 수 있었다. 그리고 공통인자분석에서 선택한 최적의 인자의 개수 $k = 4$에 대한 내재된 선형인자회귀모형의 베이지안 분석법으로 추정된 인자분석 결과는 설명변수들의 인자분석의 결과와 일치하는 것을 확인하였고, 이는 내재된 선형인자회귀모형의 베이지안 분석법이 설명변수들의 인자분석 방법도 구현하며, 반응변수와의 선형관계를 통하여 새로운 관측값에 대한 예측 역시 가능하다는 것을 확인한 결과라 할 수 있다. 게다가 베이지안 분석법의 내재된 인자선형모형의 MSE값이 일반적인 인자선형모형의 MSE값 보다 작다는 것을 확인할 수 있어서, 예측모형으로의 활용 역시 가능함을 다시 한 번 확인할 수 있었다.

References

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation, *Journal of Business and Economic Statistics*, **18**, 338–357.

Bartlett, M. S. (1951). The effect of standardization on a χ^2 approximation in factor analysis, *Biometrika*, **38**, 337–344

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models, *Biometrika*, **98**, 291–306.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2014). *Multivariate Data Analysis : Pearson*

New International Edition (7th ed), Pearson.

Hoerl, A. E. and Kennard, R. W. (1950). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis, *British Journal of Statistical Psychology*, **10**, 69–79.

Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis, *Applied Statistics*, **16**, 225–236.

Kendall, M. G. (1957). *A Course in Multivariate Analysis*, Griffin, London.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos, *Bayesian Analysis*, **5**, 369–412.

Park, T. and Casella, G. (2008). The Bayesian lasso, *Journal of the American Statistical Association*, **103**, 681–686.

Raftery, A. E. (1995). Bayesian model selection in social research, *Sociological Methodology*, **25**, 111–163.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B*, **58**, 267–288.

West, M. (2003). Bayesian factor regression models in the "Large p, Small n" paradigm, in: J.M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West (Eds.), Bayesian Statistics, Vol. 7, OxfordUniversity Press, Oxford, 723–732

내재된 인자회귀모형의 베이지안 분석법

경민정[a,1]

[a]덕성여자대학교 정보통계학과

(2020년 2월 4일 접수, 2020년 6월 11일 수정, 2020년 6월 17일 채택)

요 약

선형모형에서 두개 이상의 설명변수들 사이에 존재하는 다중공선성 문제를 변수들 간에 내재되어 있는 공통의 구조인 인자를 구성하고, 인자들을 회귀변수로 사용하여 해결하는 인자회귀모형에 대하여 논의한다. 무한개로 가정 가능한 내재된 인자 중 유의미한 인자적재행렬을 구성하기 위하여 벌점모수의 값이 큰 LASSO 사전분포를 적용하는 베이지안 추정법을 사용한다. 결정된 인자적재행렬과 다른 모수들의 추정값을 각 설명변수의 선형모수로 역변환 하여, 새로운 관측값에 대한 예측 모형으로도 사용한다. 제안한 방법을 제품 서비스 관리 자료에 적용하여 정해진 인자의 개수에 대한 인자가 일반적인 공통인자회귀모형과 동일한 결과를 나타냄을 확인하였고, 일반적인 공통인자회귀모형과 비교를 위해 계산한 평균 제곱 오차값이 더 작다는 것을 알 수 있었다.

주요용어: 베이지안 내재된 인자모형, LASSO 사전분포, 깁스표집법

이 연구는 2019년도 덕성여자대학교 연구비 (3000003265) 지원을 받아 수행되었습니다.

[1](01369) 서울특별시 도봉구 삼양로 144길 33, 덕성여자대학교 정보통계학과. E-mail: mkyung@duksung.ac.kr