The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.3.281
33(3), 281–296

Comparison study of modeling covariance matrix for multivariate longitudinal data

Na Young Kwak[a] · Keunbaik Lee[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received March 5, 2020; Revised April 2, 2020; Accepted April 9, 2020)

Abstract

Repeated outcomes from the same subjects are referred to as longitudinal data. Analysis of the data requires different methods unlike cross-sectional data analysis. It is important to model the covariance matrix because the correlation between the repeated outcomes must be considered when estimating the effects of covariates on the mean response. However, the modeling of the covariance matrix is tricky because there are many parameters to be estimated, and the estimated covariance matrix should be positive definite. In this paper, we consider analysis of multivariate longitudinal data via two modeling methodologies for the covariance matrix for multivariate longitudinal data. Both methods describe serial correlations of multivariate longitudinal outcomes using a modified Cholesky decomposition. However, the two methods consider different decompositions to explain the correlation between simultaneous responses. The first method uses enhanced linear covariance models so that the covariance matrix satisfies a positive definiteness condition; in addition, and principal component analysis and maximization-minimization algorithm (MM algorithm) were used to estimate model parameters. The second method considers variance-correlation decomposition and hypersphere decomposition to model covariance matrix. Simulations are used to compare the performance of the two methodologies.

Keywords: correlation matrix, hypersphere decomposition, modified Cholesky decomposition, positive definite, variance-correlation decomposition

1. 서론

한 개체를 여러 번 반복 측정한 자료를 경시적 자료(longitudinal data)라 한다. 같은 개체에서 측정값이 여러 번 측정되므로 경시적 자료에서는 모든 관측치가 서로 독립이라는 통계적 가정이 성립하지 않는다. 이러한 통계적 가정이 성립하지 않은 채로 공변량의 효과를 추정한다면 그 결과는 편향이 일어날 수 있다. 따라서 경시적 자료를 분석하기 위해서는 흔히 사용되는 횡단자료(cross-sectional data) 분석과는 달리, 관측치 간의 상관성을 설명하는 공분산행렬의 모형화에 초점을 맞추어야 한다. 이미 일변량 경시적 자료의 공분산행렬(covariance matrix) 모형화를 제안하는 방법들은 활발히 연구되었다. 그

This project was supported by Basic Science Research Program through the National Research Foundation of Korea (KRF) funded by the Ministry of Education, Science and Technology (NRF-2019R1F1A1058553).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: keunbaik@skku.edu

러나 다변량 경시적 자료(multivariate longitudinal data)에 대한 공분산행렬 모형화의 경우, 추정해야 할 모수가 많고, 추정된 공분산행렬이 양정치성(positive-definiteness)을 만족해야 하므로 일변량인 경우에 비해 많은 연구가 이루어지지 않았다. 따라서 본 연구에서는 다변량 경시적 자료에 대한 공분산행렬 모형화에 초점을 맞추도록 한다.

다변량 경시적 자료분석에서 공분산행렬의 모형화에서 다음의 3가지 상관관계를 고려해야 한다 (Kim과 Zimmerman, 2012). 1) 같은 속성의 응답변수들이 반복측정될 때에 발생하는 상관관계; 2) 서로 다른 속성들 간의 시간에 따라 발생하는 상관관계; 3) 같은 시간에서 변수들 간의 상관관계. 여기서 1)과 2)의 상관관계는 시간에 의한 상관관계(serial correlations)이며, 3)의 상관관계는 시간과는 무관한 상관관계이다. 이러한 3가지의 상관관계는 공분산행렬에서 설명되며, 이 행렬을 모형화 하기 위한 방법들이 몇몇 논문에서 제시되었다. 이 논문들은 모두 Pourahmadi (1999)의 수정된 콜레스키 분해(modified Cholesky decomposition; MCD)를 이용하여 모형화 하였다. Pourahmadi (1999)는 단변량 경시적 자료의 공분산행렬을 모형화하기 위해 MCD를 이용하였고, 이 때 MCD는 공분산행렬을 직접 분해하는 것이 아니라, 공분산행렬의 역행렬을 분해할 때에 적용하였다. 즉, 공분산행렬의 역행렬을 일반화 자기회귀모수(generalized autoregressive parameteres; GARPs)와 혁신분산(innovation variances; IVs)으로 분해함으로써, 제약은 없으면서도 통계적으로 의미 있는 모수로 분해하였고, 이렇게 만들어진 공분산행렬은 양정치성을 만족하게 된다. 또한, GARPs와 IVs을 선형/로그선형모형 식으로 표현하여 추정함으로써 모수의 수를 줄일 수 있다.

다변량 경시적 자료분석에서 MCD를 이용한 방법으로 Xu와 Mackenzie (2012)와 Kim과 Zimmerman (2012)이 있다. 이 두 논문 모두 공분산행렬을 일반화 자기회귀행렬(generalized autoregressive matrices; GARMs)과 혁신공분산행렬(innovation covariance matrices; ICMs)로 분해하였다. 공분산행렬을 위의 두 행렬로 분해함으로써 공분산행렬을 간단한 모형으로 나타낼 수 있으며 통계적인 해석을 하기에도 유용하다. GARMs는 블록 하삼각행렬(block lower triangular matrix) 형태이며, 각 요소들은 다변량 경시적 자료에서 발생하는 시간에 의한 상관관계인 1)과 2)의 상관관계를 설명할 수 있다. ICMs는 블록 대각행렬(block diagonal matrix)으로 나타낼 수 있고, 이 행렬은 다변량 경시적 자료에서 발생하는 마지막 상관관계인 3)의 상관관계를 설명할 수 있다. 이러한 두 행렬로 공분산행렬이 분해되므로 이 두 행렬은 유일하게 존재하며, IVMs가 양정치성을 만족하면 자연스럽게 공분산행렬도 양정치성을 만족한다 (Kim과 Zimmerman, 2012). 따라서 IVMs가 양정치성을 만족하도록 모형화 하는 것이 중요하다. 이 ICMs를 모형화 하는 방법에 따라서 여러 가지 방법들이 제안되었다 (Xu와 Mackenzie, 2012; Kim과 Zimmerman, 2012; Kohli 등, 2016; Lee 등, 2019).

Xu와 Mackenzie (2012)는 혁신공분산행렬을 스펙트럴 분해(spectral decomposition)를 이용하여 고유벡터(eigenvector)로 구성된 직교행렬과 고유값(eigenvalues)을 가지는 대각행렬로 분해한다. 그리고 ICMs의 로그변환은 결국 고유값의 로그변환으로 연결되면 이는 고유값의 양수를 보장받을 수 있다. 그리고 이를 통하여 IVMs의 양정치성을 만족시키는 모형이다. GARMs의 요소와 IVMs에서 만들어진 고유치는 선형회귀 및 로그 선형회귀 모형으로 모형화 할 수 있다. Xu와 Mackenzie (2012)의 공분산행렬 모형화 방법은 자료가 단조적인 중도탈락(monotone dropout)에 의해 불균형 형태일 때에도 개체별로 공분산행렬을 모형화 하는 것이 가능하다는 장점이 있다. 그러나 각 시점의 값 간의 상관관계를 설명하기 위한 공분산 행렬인 IVMs의 모수들을 추정하기 어렵고 해석하기 어렵다는 단점이 있다. Kim과 Zimmerman (2012)는 Xu와 Mackenzie (2012)와 같이 공분산행렬을 MCD를 이용하여 GARMs과 IVMs로 분해하였고, IVMs를 MCD를 다시 이용하여 모형화 하였다. IVMs는 MCD를 이용하여 분해함에 따라 모형화에 대한 제약을 제거하여 간단화 하였으나, 동일 시점에서의 반응값들에 순서를 부여하여 실제 데이터에 적용하기에는 한계가 있다는 단점이 있다. Kohli 등 (2016)은 Kim과 Zimmerman

(2012)과 Xu와 Mackenzie (2012)와 같이 MCD를 이용하고, IVMs를 알려진 양정치행렬과 미지의 양수인 모수의 선형 결합으로 나타냄으로써 IVMs의 양정치성을 만족시켰다. 따라서 알려진 양정치 행렬을 선택하는 것이 핵심이며, 총 세가지 방법으로 양정치행렬을 선택하였다. IVMs를 직접적으로 모형화한 기존 방법들과는 달리, Lee 등 (2019)은 IVMs를 분산-상관 분해(variance-correlation decomposition)를 하여, 혁신표준편차(innovation standard deviations; ISDs)와 상관계수(correlations)로 분해하여 모형화 하였다. 상관계수행렬을 모형화 할 때에는 양정치성과 모든 원소들이 -1과 1 사이의 값을 가져야하며 대각원소는 반드시 1이여야 한다는 제약조건이 있다. Lee 등 (2019)은 상관행렬을 모형화하기 위해 초구분해(hypersphere decomposition; HD) 방법을 이용하였다. 분산-상관 분해를 이용하여 IVMs 모형화 했기 때문에 Kohli 등 (2016)과는 달리 해석이 가능하다는 장점이 있다. 또한, Kim과 Zimmerman (2012)과는 달리 같은 시점에서의 반응값에 순서를 부여하지 않고, MCD를 이용하여 공분산 행렬을 모형화 하였다. 우리는 위에서 제시한 다변량 경시적 자료에 대한 공분산행렬의 모형화 방법 중에 Kohli 등 (2016)의 방법과 Lee 등 (2019)의 방법을 중점적으로 살펴보고자 한다. 그 이유는 Kohli 등 (2016) 방법은 IVMs의 모형화에 다양한 방법을 제안하였고, Lee 등 (2019)은 자연스러운 해석이 가능한 모형이라서 선택하였다. 그리고 Lee 등 (2019)의 모형이 Kim과 Zimmerman (2012)의 모형 보다 편향과 효율성에서 더 나은 것으로 확인되었기 때문이다 (Lee 등, 2019). 이 논문에서 우리는 이 방법들을 모의실험을 통해 여러 경우에 대해 비교해 보고자 한다.

논문의 구성은 다음과 같다. 2장에서 MCD를 자세하게 설명한 후, MCD를 기반으로 한 Kohli 등 (2016)의 방법과 Lee 등 (2019)의 방법을 제시한다. 3장에서는 이 두 가지 방법을 이용하여 모의실험을 진행하고 그 결과를 논의한다. 4장에서는 3장에서의 결과를 바탕으로 결론을 내린다.

2. 공분산 행렬의 모형화

2.1. 단변량 경시적 자료에서의 MCD를 이용한 모형화

Pourahmadi (1999)는 공분산행렬을 모형화 하기 위해서 다음의 모형을 제안하였다. $Y_i = (Y_{i1}, \ldots, Y_{in})^T$는 i번째 대상자의 응답변수 벡터이다 $(i = 1, \ldots, N)$. 여기서 Y_{it}는 i번째 대상자의 t번째 응답변수이다 $(t = 1, \ldots, n)$. $\hat{Y}_{it}$를 $\hat{Y}_{it} = \mu_{it} + \sum_{j=1}^{t-1} \phi_{it,j}(Y_{ij} - \mu_{ij})$라고 정의한다. 그리고 계산의 편의를 위해 $\mu_{ij} = E(Y_{ij}) = 0$으로 가정한다. 따라서 $\hat{Y}_{it} = \sum_{j=1}^{t-1} \phi_{it,j} Y_{ij}$이고, 이것은 마치 회귀식에서의 추정식과 유사하다. 이 때의 예측오차는 ε_{it}이며 $\sigma_{it}^2 = \mathrm{var}(\varepsilon_{it})$이다. 즉,

$$\varepsilon_{it} = Y_{it} - \hat{Y}_{it} = Y_{it} - \sum_{j=1}^{t-1} \phi_{it,j} Y_{ij}, \quad t = 1, \ldots, n \tag{2.1}$$

이며, 연속된 예측오차끼리는 서로 독립이다. 대각원소가 1이고 하삼각원소가 $\hat{Y}_{it} = \sum_{j=1}^{t-1} \phi_{it,j} Y_{ij}$의 회귀계수에 대한 음수 값인 행렬을 T_i, 대각원소가 $\sigma_{it}^2 = \mathrm{var}(Y_{it} - \hat{Y}_{it})$인 행렬을 D_i라고 정의한다. 즉, $\varepsilon_i = T_i Y_i$으로 표현할 수 있고, 여기에 분산을 취하면 다음과 같다.

$$\mathrm{cov}(\varepsilon_i) = T_i \mathrm{cov}(Y_i) {T_i}^T = T_i \Sigma {T_i}^T = D_i \tag{2.2}$$

라는 식이 도출되는 것을 확인할 수 있다. 식 (2.2)에서 행렬 T_i의 요소를 일반화 자기회귀모수, 행렬 D_i를 혁신분산이라 지칭한다. T_i가 하삼각 행렬이고 D_i가 대각행렬이므로 Σ는 대칭행렬이 된다. T_i와 D_i는 통계적으로 해석이 용이하며, Σ를 추정하기 위해서는 $(1/2)n(n-1)$개의 $\phi_{it,j}$와 $\log(\sigma_{it}^2)$만 추정하면 된다. Σ_i를 추정하기 위해 T_i와 D_i의 원소만 추정하면 되며, 모수의 수를 줄이기 위한 선형 및 로

그 선형모형이 제안되었다. T_i의 원소인 $\phi_{it,j}$와 D_i는 주대각 요소인 σ_{it}의 모형화는 다음과 같다.

$$\phi_{it,j} = z_{it,j}^T \gamma, \tag{2.3}$$

$$\log(\sigma_{it}^2) = h_{it}^T \lambda. \tag{2.4}$$

이때 $z_{it,j}$와 h_{it}는 개체-특징적 공변량이며, γ와 λ는 알려지지 않은 모수이다.

2.2. 다변량 경시적 자료에서의 MCD를 이용한 모형화

Pourahmadi (1999)가 제안한 MCD는 단변량 경시적 자료에 한정된 방법이기 때문에 다변량 경시적 자료에서 MCD를 사용하기 위해서는 수정된 블록 콜레스키 분해(modified Cholesky block decomposition; MCBD) 방법을 사용해야 한다. 따라서 MCD에서는 하나의 값이었던 T_i와 D_i의 원소들이 MCBD에서는 행렬로 표현되어야 한다.

2.2.1. 수정된 블록 콜레스키 분해 K개의 특성이 있는 다변량 경시적 자료에서 Y_{itk}는 i번째 대상자의 t번째 반복수의 k번째 속성을 나타낸다($i = 1, \ldots, N$; $t = 1, \ldots, n$; $k = 1, \ldots, K$). 이 경우 다음과 같은 선형모형으로 표현할 수 있다.

$$\begin{aligned} Y_{i1k} &= x_{i1}^T \beta_k + \epsilon_{i1k}, \\ Y_{itk} &= x_{it}^T \beta_k + \sum_{j=1}^{t-1} \sum_{g=1}^{K} \phi_{itk,kg} \left(Y_{ijg} - x_{ij}^T \beta_g \right) + \epsilon_{itk}, \end{aligned} \tag{2.5}$$

여기서 x_{it}는 $p \times 1$ 공변량 벡터, β_k는 미지의 회귀계수 벡터이다. 식 (2.5)를 행렬 형태로 나타내면

$$T_i(Y_i - X_i\beta) = \epsilon_i \tag{2.6}$$

이다. 여기서

$$T_i = \begin{pmatrix} I & 0 & \cdots & 0 \\ -\Phi_{i21} & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\Phi_{in1} & -\Phi_{in2} & \cdots & I \end{pmatrix}, \quad Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix}, \quad X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{in} \end{pmatrix}, \quad \epsilon_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in} \end{pmatrix},$$

여기서 Φ_{itj}를 $K \times K$의 행렬로 그 요소는 $\phi_{itj,lm}$ $(l, l = 1, \ldots, K)$이며, $Y_{it} = (Y_{it1}, \ldots, Y_{itK})^T$, $X_{it} = I_K \bigotimes x_{it}^T$는 $K \times p$ 행렬이며, 그리고 $\epsilon_{it} = (\epsilon_{it1}, \ldots, \epsilon_{itn})^T$이다. $\bigotimes$는 크로레크곱을 의미한다. 그리고 Φ_{itj}를 자기회귀모수행렬(generalized autoregressive parameter matrix; GARPM)이라고 지칭한다.

단변량 경시적 자료에서의 MCD와 마찬가지로 예측오차벡터인 ϵ_i는 다음을 만족한다고 가정한다.

$$\epsilon_i = \left(\epsilon_{i1}^T, \ldots, \epsilon_{in}^T \right)^T \sim N(0, D_i),$$

여기서 $D_i = \text{diag}\{D_{i1}, \ldots, D_{in}\}$를 혁신분산행렬(innovation covariance matrix; ICM)이라 지칭하고, $D_{it} = \text{var}(\epsilon_{it})$이다.

MCD와 마찬가지로 MCBD에서의 T_i도 회귀계수 행렬과 같이 제약을 받지 않고, 서론에서 제시한 시

간에 의한 응답변수들의 상관관계(상관관계 1)과 2))를 설명이 가능하다. GARPM의 원소인 $\phi_{itj,lm}$은 시간이나 개체특징적 공변량 벡터인 w_{itj}을 이용하여 추정할 수 있다.

$$\phi_{itj,lm} = w_{itj}^T \alpha_{lm}, \tag{2.7}$$

이때 α_{lm}은 미지의 모수로 이루어져 있는 $a \times 1$벡터이며, w_{itj}은 개체특징적 공변량 벡터이다. 그러나 D_i는 K개의 결과에서의 횡단적인 종속을 의미하며, 양정치성을 만족해야 하는 제약이 있다. D_i는 통계적으로 해석이 가능한 동시에 최대한 제약받지 않도록 추정하는 것이 중요하다. 이러한 조건들에 부합하는 D_i를 추정하기 위하여 Kohli 등 (2016)과 Lee 등 (2019)은 각자 다른 방법론을 제시하였다. D_i의 모형화는 여러 결과값들의 순서를 필요로 하지 않으며, 추정된 Σ의 양정치성을 만족시킨다.

2.2.2. 향상된 혁신공분산행렬을 이용한 모형화 Kohli 등 (2016)은 D_i를 향상된 선형 공분산 모형(enhanced linear covariance models; LCM)을 이용하여 추정하는 것이다. 이를 식으로 표현하면 다음과 같다.

$$D_{it}^{(q)} = \tau_{t1} M_{t1} + \cdots + \tau_{tq} M_{tq}, \tag{2.8}$$

이때의 M_{tl} $(l = 1, \ldots, q)$는 알려진 양정치 행렬이며 τ_{tl}는 미지의 음수가 아닌 모수를 의미한다. 또한 τ_{tl}은 $D_{it}^{(q)}$의 변동을 설명하는 모수로 해석되기도 한다. 이러한 제약은 추정된 Σ의 양정치성을 보장한다. 따라서 LCM 모형을 사용하는 데 있어 가장 중요한 사안은 적절한 M_{it}를 선택하는 것이다. Kohli 등 (2016)은 적절한 양정치행렬 M_{it}를 선택하기 위해 세 가지 방법을 제안하였다. 첫 번째 방법은 모수적인 방법으로써, 알려진 공분산 기반 행렬 묶음에서 M_{it}를 선택하는 것이다. 두 번째 방법은 비모수적인 방법으로써, 주성분분석(principal component analysis; PCA)을 이용하여 혁신행렬의 고유벡터를 이용해 M_{it}를 생성하는 것이다. 마지막 방법은 단변량 regressogram 개념을 확장하여 M_{it}의 양정치성을 만족시키는 것이다.

우선 첫 번째 모수적인 방법은 compound symmetry (CS)나 AR(1)과 같은 모수적 상관관계 구조에서 M_{it}를 선택하는 것이다. Qu 등 (2000)과 Zhou와 Qu (2012)는 CS와 AR(1)모형이 경시적 자료의 상관관계나 의존관계를 잘 설명할 수 있다는 것을 증명하였다. 따라서 M_{it}를 AR(1), CS, AR-CS 상관관계 모형에서 선택하는 것을 고려한다. 모수적 방법은 M_{it}를 쉽게 선택할 수 있다는 장점이 있지만, 상관관계에 대한 사전 정보가 없으면 잘못된 추론을 할 가능성이 있다는 단점이 존재한다. 두 번째 방법은 D의 표본을 스펙트럴 분해하는 비모수적 방법이다. D의 표본의 작은 고유값에 대응하는 고유벡터를 이용하여 M_{it}를 추정할 수 있다. 이 방법은 M_{it}를 잘못 구축할 위험은 피할 수 있으나, 불균형한 데이터의 경우 표본 공분산행렬이 존재하지 않을 수도 있다는 단점이 있다. 세 번째 방법은 그래프를 이용하여 D_t^{-1}나 $\log(D_t)$의 원소들을 선택하는 방법이다.

2.2.3. 초구분해를 이용한 모형화 Lee 등 (2019)은 Kohli 등 (2016)과는 달리 이분산성까지 고려하여 공분산행렬을 모형화 하고자 하였다. 이분산성과 양정치성을 만족시키기 위해 D_{it}를 직접 모형화하는 것이 아니라 분산-상관 분해를 이용하여 혁신분산과 상관행렬로 분해하여 모형화 하였다. 이렇게 혁신분산과 상관행렬의 모수를 추정함으로써 모형화 결과가 해석 불가능한 Kohli 등 (2016)의 경우와는 달리, Lee 등 (2019)의 모형화의 결과는 해석이 가능하다는 장점이 있다. 또한, 이 방법론은 반응변수 간의 순서가 없어도 D_{it}의 모형화가 가능하므로 실제 데이터로의 적용이 더 자연스럽다. 상관행렬을 모형화 하는 방법론으로 Lee 등 (2019)은 초구분해 방법을 제안하였다. 초구분해가 상관행렬의 양정치성을 만족시키므로, 자연스럽게 D_i의 양정치성까지 만족시킬 수 있다. Lee 등 (2019)의 따름정리에 의해

D_i가 양정치성을 만족하면 Σ_i도 자동적으로 양정치성을 만족한다. 따라서 D_i의 양정치성을 만족시키기 위해 D_{it}를 다음과 같은 식으로 표현한다.

$$D_{it} = C_{it} R_i C_{it}, \tag{2.9}$$

이때 C_{it}는 혁신표준편차를 대각원소로 가지는 대각행렬이며, R_i는 각 시점에서의 반응변수들 간의 상관행렬이다. 따라서 C_{it}와 R_i를 각각 모형화하면 D_{it}를 추정할 수 있다. 우선 C_{it}의 대각원소인 혁신표준편차를 추정하기 위한 식은 다음과 같다.

$$\log(\sigma_{itk}) = h_{it}^T \lambda_k, \tag{2.10}$$

여기서 λ는 알려지지 않은 모수이며 h_{it}는 시간이나 개체 특정적 공변량을 의미한다. $\log(\sigma_{itk})$가 혁신표준편차를 로그 변환한 형태이므로 $\log(\sigma_{itk})$은 제약이 없다. 따라서 D_{it}의 양정치성은 R_i의 양정치성 여부에 영향을 받는다.

상관계수행렬 R_i을 모형화 할 때 고려해야 할 사항으로 총 세 가지가 있다. 첫 번째 고려사항은 R_i가 양정치성을 만족해야 하는 것이다. 두 번째는 R_i의 주대각 원소를 제외한 모든 원소는 -1과 1 사이의 값을 가져야 하며, 마지막 고려사항은 R_i의 대각원소는 반드시 1이어야 한다는 것이다. 이 세 가지 고려사항을 모두 만족시키는 R_i을 모형화하기 위하여 Lee 등 (2019)은 초구 분해 방법을 사용하였다. 초구 분해 방법을 이용하면 R_i는 다음과 같이 분해된다.

$$R_i = F_i F_i^T, \tag{2.11}$$

여기서

$$F_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ f_{i21} & f_{i22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{iK1} & f_{iK2} & f_{iK3} & \cdots & f_{iKK} \end{pmatrix}, \tag{2.12}$$

각각의 F_i의 원소인 f_{ilm}은 다음과 같은 식으로 구할 수 있다.

$$f_{ilm} = \begin{cases} \cos(\omega_{ilm}), & \text{for } m = 1,\ l = 2, \ldots, K; \\ \cos(\omega_{ilm}) \prod_{r=1}^{m-1} \sin(\omega_{ilr}), & \text{for } 2 \le m < l \le K; \\ \prod_{r=1}^{m-1} \sin(\omega_{ilr}), & \text{for } l = m;\ m = 2, \ldots, K; \end{cases}$$

따라서 제약이 없는 $\omega_{ilm} \in (0, \pi)$를 추정하기 위해 알려지지 않은 모수인 υ와 공변량 벡터인 g_{ilm}을 이용하여 다음과 같은 식으로 표현한다.

$$\log\left(\frac{\omega_{ilm}}{\pi - \omega_{ilm}}\right) = g_{ilm}^T \upsilon. \tag{2.13}$$

모형 (2.13)과 같은 식으로 ω_{ilm}를 재모수화함에 따라 ω_{ilm}는 $(0, \pi)$의 범위를 만족한다. ω_{ilm}를 추정함으로써 R_i를 추정할 수 있고, 그 결과로 D_{it}를 추정할 수 있다.

3. 모의실험

이 절에서 모의실험을 통하여 제 2절에서 언급한 두 가지 방법의 성능을 비교하고자 한다. 모의실험마다 500개의 자료집합을 생성하였다. 각각의 자료집합에는 응답변수를 3종류의 속성($K = 3$)으로 하고, 표본의 크기(N)와 반복수(n)는 경우에 따라서 다르게 하였다. 구체적인 수는 각각의 소절에서 제시하였다.

Table 3.1. $N = 100$, $n = 5$

β	Lee		Kohli(AR)		Kohli(CS)		Kohli(AR-CS)	
	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)
β_{10}	0.312	3.962	0.282	5.756	0.292	2.396	0.289	3.630
(0.3)	$(0.865_{0.870})$	88.000	$(0.288_{0.288})$	96.000	$(0.309_{0.276})$	96.600	$(0.315_{0.279})$	98.400
β_{11}	−0.110	8.309	−0.075	24.490	−0.088	11.740	−0.082	17.401
(−0.1)	$(0.703_{0.672})$	85.000	$(0.337_{0.311})$	96.800	$(0.360_{0.320})$	97.000	$(0.368_{0.326})$	97.800
β_{12}	0.262	15.309	0.190	4.885	0.197	1.167	0.192	3.861
(0.2)	$(0.829_{0.702})$	89.200	$(0.130_{0.121})$	96.000	$(0.139_{0.123})$	96.800	$(0.142_{0.125})$	97.200
β_{13}	0.281	8.934	0.301	0.353	0.300	0.075	0.300	0.259
(0.3)	$(0.730_{0.698})$	86.800	$(0.014_{0.013})$	95.200	$(0.015_{0.014})$	96.600	$(0.016_{0.014})$	97.400
β_{20}	0.223	13.050	0.219	9.900	0.204	20.157	0.202	10.179
(0.2)	$(0.594_{0.509})$	89.200	$(0.307_{0.281})$	96.200	$(0.325_{0.295})$	97.800	$(0.333_{0.279})$	98.000
β_{21}	−0.129	30.493	−0.125	25.745	−0.109	9.902	−0.107	7.526
(−0.1)	$(0.532_{0.483})$	91.000	$(0.366_{0.339})$	96.600	$(0.143_{0.134})$	96.800	$(0.147_{0.129})$	96.400
β_{22}	0.217	12.942	0.210	5.244	0.204	2.162	0.199	0.171
(0.2)	$(0.692_{0.629})$	82.600	$(0.137_{0.125})$	96.600	$(0.143_{0.134})$	96.800	$(0.147_{0.129})$	96.400
β_{23}	0.339	24.583	0.298	0.412	0.299	0.169	0.300	0.083
(0.3)	$(0.822_{0.821})$	89.200	$(0.015_{0.014})$	96.600	$(0.016_{0.015})$	96.800	$(0.016_{0.014})$	96.600
β_{30}	0.237	15.834	0.190	4.960	0.221	10.844	0.200	0.136
(0.2)	$(0.998_{0.984})$	90.000	$(0.315_{0.300})$	95.400	$(0.342_{0.300})$	98.000	$(0.348_{0.280})$	98.800
β_{31}	−0.228	13.200	−0.184	7.561	−0.213	6.786	−0.206	3.359
(−0.2)	$(0.702_{0.698})$	85.000	$(0.370_{0.345})$	95.800	$(0.404_{0.353})$	97.600	$(0.412_{0.338})$	98.800
β_{32}	0.253	24.196	0.193	3.425	0.204	2.023	0.203	1.725
(0.2)	$(0.824_{0.794})$	84.200	$(0.133_{0.125})$	95.400	$(0.145_{0.127})$	97.800	$(0.148_{0.121})$	98.400
β_{33}	0.434	6.292	0.400	0.203	0.399	0.132	0.399	0.076
(0.4)	$(0.883_{0.829})$	89.200	$(0.014_{0.013})$	95.600	$(0.015_{0.014})$	97.600	$(0.016_{0.013})$	98.200
F.norm	4.838		4.922		5.373		5.502	

3.1. Kohli 등 (2016)에서 생성한 자료집합

우선 Kohli 등 (2016)이 제안한 방식을 이용하여 평균과 AR(1)구조의 공분산행렬을 아래에 제시된 방법으로 다변량 정규분포에서 난수를 발생시켰다.

$$Y_{itk} = x_{it}^T \beta_k + \sum_{j=1}^{t-1} \sum_{g=1}^{K} \phi_{itj,kg} \left(Y_{ijg} - x_{ij}^T \beta_g \right) + e_{it}, \tag{3.1}$$

여기서 공변량은 $x_{it}^T = (1, \text{Group}_i, \text{Time}_{it}, \text{Time}_{it} \times \text{Group}_i)$이며, $\text{Time}_{it} = (t-1)/10$이며 Group_i는 0 또는 1의 값을 가진다. 이 경우 회귀계수의 차원은 $p = 4$이며 그 값은 아래와 같다.

$$\beta_1 = (0.3, -0.1, 0.2, 0.3)^T, \quad \beta_2 = (0.2, -0.1, 0.2, 0.3)^T, \quad \beta_3 = (0.2, -0.2, 0.2, 0.4)^T.$$

일반화자기회귀모수의 회귀식 (2.7)에 있는 모수인 α_{lm}과 향상된 혁신공분산행렬 (2.8)의 모수인 τ_{tl}의 값은 아래와 같다.

$$(\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33}) = (0.3, 0.4, 0.1, 0.1, 0.3, 0.1, 0.1, 0.3, 0.2),$$
$$(\tau_1, \tau_2, \tau_3) = (1.0, 0.5, 0.7).$$

Table 3.2. $N = 100$, $n = 10$

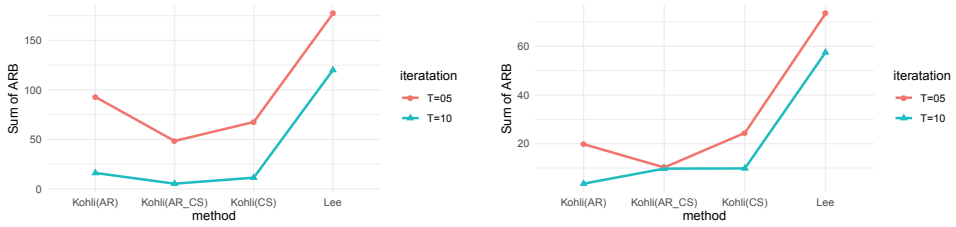
β	Lee		Kohli(AR)		Kohli(CS)		Kohli(AR-CS)	
	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)
β_{10}	0.318	6.309	0.291	2.712	0.292	2.460	0.298	0.429
(0.3)	$(0.794_{0.759})$	88.000	$(0.157_{0.137})$	97.600	$(1.862_{0.142})$	98.800	$(0.195_{0.144})$	99.600
β_{11}	−0.115	14.392	−0.093	6.112	−0.095	4.860	−0.099	0.213
(−0.1)	$(0.643_{0.534})$	90.200	$(0.133_{0.117})$	97.000	$(0.157_{0.120})$	98.800	$(0.164_{0.124})$	99.200
β_{12}	0.263	8.394	0.198	0.990	0.198	0.677	0.199	0.041
(0.2)	$(0.584_{0.539})$	85.000	$(0.042_{0.037})$	97.000	$(0.050_{0.038})$	98.800	$(0.053_{0.039})$	99.000
β_{13}	0.285	4.742	0.300	0.063	0.300	0.004	0.300	0.013
(0.3)	$(0.498_{0.438})$	88.400	$(0.004_{0.003})$	96.800	$(0.004_{0.003})$	98.800	$(0.005_{0.003})$	99.000
β_{20}	0.220	6.306	0.205	2.561	0.201	0.795	0.201	0.834
(0.2)	$(0.434_{0.423})$	89.600	$(0.164_{0.150})$	96.400	$(0.200_{0.153})$	98.400	$(0.211_{0.160})$	99.200
β_{21}	−0.113	9.329	−0.102	2.298	−0.098	1.085	−0.098	1.940
(−0.1)	$(0.403_{0.307})$	90.200	$(0.138_{0.128})$	97.200	$(0.168_{0.130})$	98.400	$(0.176_{0.133})$	99.200
β_{22}	0.220	10.391	0.201	0.639	0.200	0.047	0.199	0.177
(0.2)	$(0.594_{0.528})$	83.600	$(0.048_{0.045})$	96.000	$(0.057_{0.045})$	98.400	$(0.061_{0.046})$	99.200
β_{23}	0.363	7.807	0.299	0.064	0.299	0.027	0.299	0.009
(0.3)	$(0.693_{0.684})$	88.200	$(0.005_{0.004})$	95.800	$(0.006_{0.005})$	96.400	$(0.006_{0.004})$	99.400
β_{30}	0.284	14.236	0.200	0.135	0.201	0.652	0.197	1.295
(0.2)	$(0.959_{0.892})$	90.000	$(0.124_{0.114})$	95.200	$(0.143_{0.118})$	97.000	$(0.150_{0.107})$	99.200
β_{31}	−0.218	6.493	−0.199	0.093	−0.201	0.552	−0.200	0.297
(−0.2)	$(0.632_{0.611})$	89.200	$(0.108_{0.100})$	95.600	$(0.125_{0.101})$	98.600	$(0.131_{0.096})$	99.900
β_{32}	0.264	27.394	0.199	0.402	0.199	0.151	0.200	0.010
(0.2)	$(0.813_{0.743})$	84.600	$(0.033_{0.030})$	94.800	$(0.038_{0.030})$	98.400	$(0.040_{0.029})$	99.600
β_{33}	0.433	4.302	0.400	0.030	0.400	0.018	0.400	0.013
(0.4)	$(0.843_{0.892})$	92.000	$(0.003_{0.002})$	95.600	$(0.003_{0.002})$	98.600	$(0.003_{0.002})$	99.600
F.norm	4.493		4.530		5.248		5.395	

표본의 크기는 100, 300, 500으로 하였고, 각각 500개의 자료집합을 만들었다.

이러한 500개의 자료집합을 제 2절에서 제시한 Kohli 등 (2016)의 모형과 Lee 등 (2019)이 제안한 모형을 적합시켰다. Kohli 등 (2016)의 모형 적합 시에는 공분산행렬을 모수적인 방법인 AR(1), CS, 그리고 AR-CS구조로 가정한 모형을 각각 적합시켰다. 그 적합한 결과들은 각 모수의 추정치 평균(mean), 상대편향의 절댓값(absolute relative bias; ARB), 추정량의 표준오차의 평균(standard error; SE), 추정치들의 표준편차(standard deviation; SD), 그리고 포함확률(coverage probability; CP)으로 요약되었다. 또한 Kohli 등 (2016)이 제안한 모형과 Lee 등 (2019)이 제안한 모형 간의 비교를 위해 프로베니우스 노름(Frobenius norm)을 이용하여 추정된 공분산행렬을 비교하고자 한다.

$$\text{F.norm}\left(\hat{\Sigma}\right) = \text{tr}\left\{\left(\hat{\Sigma}\Sigma^{-1} - I\right)^T\left(\hat{\Sigma}\Sigma^{-1} - I\right)\right\},$$

Tables 3.1과 3.2는 표본의 크기가 100이며, 각각 반복수가 5, 10일 때에 공분산행렬을 앞서 제시한 4가지 모형을 가정하고 적합한 결과를 나타낸 것이다. 그리고 Figure 3.1의 (a)는 모든 평균모수들의 추정 정도를 한 번에 파악하기 위해 ARB값의 합인 $\sum_{i=1}^{3}\sum_{j=0}^{3}(|\hat{\beta_{ij}} - \beta_{ij}|/\beta_{ij}) \times 100$를 계산하여 그 결과를 그림으로 나타내었다. 반복수가 증가하면 편향이 줄어드는 것을 알 수 있다. 그리고 Kohli 등 (2016)의 모든 방법들이 Lee 등 (2019)의 방법보다는 편향이 적음을 알 수 있다. 이는 공분산행렬의 구조를 난



(a) $N = 100$ (b) $N = 500$

Figure 3.1. Figure 3.1: Sum of ARBs of mean parameters in Lee *et al.* (2020)'s and Kohli *et al.* (2016)'s models using the data generated from Kohli *et al.* (2016)'s model.

Table 3.3. $N = 500$, $n = 5$

β	Lee		Kohli(AR)		Kohli(CS)		Kohli(AR-CS)	
	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)
β_{10}	0.307	2.394	0.294	1.813	0.305	1.891	0.299	0.090
(0.3)	$(0.382_{0.311})$	91.400	$0.129_{0.125})$	94.800	$(0.138_{0.116})$	97.800	$(0.140_{0.109})$	98.800
β_{11}	−0.114	13.391	−0.093	6.668	−0.110	10.319	−0.102	2.381
(−0.1)	$(0.399_{0.328})$	89.200	$(0.151_{0.146})$	95.200	$(0.161_{0.134})$	97.800	$(0.164_{0.132})$	98.400
β_{12}	0.223	5.730	0.197	1.098	0.204	2.198	0.200	0.403
(0.2)	$(0.528_{0.493})$	90.000	$(0.058_{0.056})$	94.600	$(0.062_{0.053})$	97.800	$(0.063_{0.050})$	99.000
β_{13}	0.292	3.088	0.300	0.065	0.299	0.179	0.299	0.050
(0.3)	$(0.211_{0.209})$	92.600	$(0.006_{0.006})$	95.200	$(0.007_{0.006})$	97.200	$(0.007_{0.005})$	98.800
β_{20}	0.210	4.399	0.195	2.288	0.201	0.531	0.202	1.100
(0.2)	$(0.293_{0.234})$	92.600	$(0.138_{0.122})$	96.400	$(0.145_{0.131})$	97.600	$(0.148_{0.125})$	98.000
β_{21}	−0.109	7.203	−0.094	5.521	−0.098	1.842	−0.101	1.389
(−0.1)	$(0.278_{0.300})$	90.000	$(0.164_{0.147})$	95.800	$(0.172_{0.159})$	97.200	$(0.176_{0.151})$	97.800
β_{22}	0.212	7.392	0.197	1.489	0.198	0.677	0.199	0.142
(0.2)	$(0.494_{0.440})$	89.200	$(0.061_{0.055})$	96.400	$(0.064_{0.059})$	97.200	$(0.065_{0.056})$	97.600
β_{23}	0.311	8.133	0.300	0.138	0.300	0.056	0.300	0.022
(0.3)	$(0.413_{0.401})$	91.600	$(0.006_{0.006})$	96.800	$(0.007_{0.006})$	97.400	$(0.007_{0.006})$	96.400
β_{30}	0.213	7.742	0.199	0.257	0.202	1.300	0.194	2.940
(0.2)	$(0.293_{0.246})$	88.200	$(0.141_{0.128})$	96.000	$(0.153_{0.132})$	97.200	$(0.155_{0.130})$	98.000
β_{31}	−0.209	4.300	−0.199	0.433	−0.207	3.711	−0.197	1.227
(−0.2)	$(0.533_{0.510})$	91.400	$(0.166_{0.152})$	96.200	$(0.181_{0.155})$	96.600	$(0.184_{0.153})$	98.400
β_{32}	0.214	6.483	0.199	0.433	0.203	0.632	0.199	0.446
(0.2)	$(0.429_{0.412})$	92.000	$(0.059_{0.055})$	96.800	$(0.064_{0.056})$	96.200	$(0.066_{0.055})$	97.600
β_{33}	0.417	3.205	0.400	0.003	0.399	0.088	0.400	0.037
(0.4)	$(0.131_{0.117})$	94.600	$(0.006_{0.006})$	96.000	$(0.007_{0.006})$	96.200	$(0.007_{0.006})$	97.300
F.norm	7.193		8.745		11.224		11.914	

수를 발생한 모형과 다르게 가정하였기에 발생한 현상이다. 회귀계수에 대한 정확성을 모형별로 비교하면, 반복수가 증가함에 따라 모든 모형의 SE와 SD는 감소한다. 그러나 포함확률의 변화는 반복수가 증가하는 것과 큰 상관이 없는 것을 알 수 있다.

F.norm 기준으로 살펴보았을 때, 반복수 관계없이 Lee 등 (2019) 모형의 F.norm 값이 가장 작다. 이는 Kohli 등 (2016)의 방법으로 난수를 발생시켜도 Lee 등 (2019) 모형이 공분산 행렬을 더 잘 추정한다고 볼 수 있다. 또한, 4개의 모형 모두 반복수가 커짐에 따라 F.norm 값이 작아졌다.

Table 3.4. $N = 500$, $n = 10$

β	Lee		Kohli(AR)		Kohli(CS)		Kohli(AR-CS)	
	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)
β_{10}	0.314	5.842	0.299	0.249	0.302	0.941	0.297	0.813
(0.3)	$(0.193_{0.188})$	92.000	$(0.007_{0.006})$	95.800	$(0.083_{0.065})$	99.200	$(0.086_{0.064})$	99.800
β_{11}	−0.104	3.180	−0.098	1.169	−0.101	1.814	−0.098	1.998
(−0.1)	$(0.392_{0.228})$	92.600	$(0.060_{0.050})$	95.400	$(0.070_{0.055})$	99.200	$(0.073_{0.055})$	99.600
β_{12}	0.232	4.843	0.199	0.131	0.200	0.401	0.199	0.324
(0.2)	$(0.321_{0.319})$	88.800	$(0.019_{0.017})$	94.800	$(0.022_{0.017})$	99.000	$(0.023_{0.017})$	99.800
β_{13}	0.293	2.385	0.300	0.007	0.299	0.029	0.300	0.029
(0.3)	$(0.132_{0.131})$	92.400	$(0.001_{0.001})$	95.000	$(0.002_{0.001})$	98.600	$(0.002_{0.001})$	99.400
β_{20}	0.208	3.492	0.199	0.230	0.195	2.150	0.203	1.674
(0.2)	$(0.243_{0.203})$	92.600	$(0.075_{0.069})$	95.800	$(0.089_{0.070})$	98.600	$(0.093_{0.070})$	98.800
β_{21}	−0.109	6.923	−0.100	0.402	−0.097	2.733	−0.102	2.920
(−0.1)	$(0.293_{0.249})$	95.000	$(0.063_{0.059})$	95.800	$(0.075_{0.060})$	97.800	$(0.078_{0.059})$	99.200
β_{22}	0.206	6.423	0.199	0.097	0.198	0.643	0.200	0.475
(0.2)	$(0.243_{0.221})$	91.600	$(0.022_{0.020})$	96.800	$(0.025_{0.020})$	98.600	$(0.027_{0.020})$	99.600
β_{23}	0.321	4.312	0.300	0.017	0.300	0.055	0.299	0.030
(0.3)	$(0.219_{0.234})$	92.000	$(0.002_{0.002})$	95.600	$(0.002_{0.002})$	98.600	$(0.002_{0.002})$	99.200
β_{30}	0.232	8.433	0.198	0.941	0.198	0.588	0.198	0.540
(0.2)	$(0.143_{0.148})$	94.600	$(0.054_{0.043})$	96.800	$(0.064_{0.048})$	99.200	$(0.066_{0.049})$	99.200
β_{31}	−0.209	3.543	−0.199	0.187	−0.199	0.082	−0.198	0.671
(−0.2)	$(0.290_{0.242})$	93.200	$(0.049_{0.042})$	96.600	$(0.056_{0.043})$	99.200	$(0.058_{0.044})$	99.200
β_{32}	0.239	5.694	0.200	0.127	0.200	0.223	0.199	0.270
(0.2)	$(0.092_{0.083})$	95.400	$(0.015_{0.013})$	96.400	$(0.017_{0.013})$	99.000	$(0.018_{0.014})$	99.400
β_{33}	0.418	2.438	0.399	0.001	0.399	0.171	0.400	0.013
(0.4)	$(0.149_{0.132})$	95.000	$(0.001_{0.001})$	96.200	$(0.001_{0.001})$	98.200	$(0.001_{0.001})$	99.000
F.norm	3.819		4.527		4.842		4.976	

Tables 3.3과 3.4은 표본의 크기가 500일 때 4가지 모형의 결과를 나타낸 것이다. 그리고 Figure 3.1의 (b)는 모든 평균모수들의 추정정도를 한 번에 파악하기 위해 ARB값의 합을 나타내었다. F.norm 기준으로 살펴보았을 때, 반복수 관계없이 Lee 등 (2019) 모형의 F.norm 값이 가장 작다. 이는 표본의 크기가 100일 때뿐만 아니라 500일 때에도 Lee 등 (2019) 모형이 공분산 행렬을 더 잘 추정한다고 볼 수 있다. 또한, 4개의 모형 모두 반복수가 커짐에 따라 F.norm 값이 작아졌다. 따라서 반복수가 커짐에 따라 모든 모형이 공분산 행렬을 잘 추정한다고 해석할 수 있다. 회귀계수에 대한 정보 기준으로 모형을 비교하면, 반복수가 증가함에 따라 모든 모형의 SE와 SD는 감소한다. 그러나 포함확률의 변화는 반복수가 증가하는 것과 큰 상관이 없는 것을 알 수 있다.

Kohli 등 (2016)이 제안한 AR, CS, AR-CS 모형을 비교하면, 반복수가 적을 때에는 세 모형의 SE, SD 값의 차이가 크다. AR 모형의 SE와 SD가 가장 작을 때가 많으며, AR-CS 모형의 SE와 SD 값이 가장 큰 경우가 많다. 이러한 사실 때문에 AR-CS 모형에서의 포함확률이 가장 크게 나타난다. 그러나 반복수가 커짐에 따라 각 모형의 SE와 SD 차이가 작아지는 것을 알 수 있다.

3.2. Lee 등 (2019)에서 생성한 자료집합

두 번째로 Lee 등 (2019)이 제안한 모형에서 난수를 발생시켰다. 응답변수의 수, 대상자의 수 및 반

Table 3.5. $N = 100$, $n = 5$

β	Lee		Kohli(AR)		Kohli(CS)		Kohli(AR-CS)	
	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)
β_{10}	0.301	0.350	0.228	23.808	0.156	10.188	0.163	9.9780
(0.3)	$(1.155_{1.152})$	95.400	$(0.124_{0.145})$	85.800	$(0.480_{0.391})$	94.200	$(0.482_{0.383})$	94.200
β_{11}	−0.100	0.170	−0.025	47.429	−0.014	60.813	−0.016	56.410
(−0.1)	$(0.222_{0.217})$	94.800	$(0.730_{0.492})$	82.000	$(0.781_{0.625})$	97.000	$(0.718_{0.650})$	96.600
β_{12}	0.206	3.086	0.170	14.591	0.210	15.318	0.225	25.186
(0.2)	$(0.296_{0.287})$	95.800	$(0.985_{0.829})$	84.600	$(0.184_{0.181})$	78.600	$(1.842_{1.743})$	74.600
β_{13}	0.305	1.866	0.189	36.769	0.153	41.082	0.149	45.461
(0.3)	$(0.423_{0.421})$	95.200	$(0.821_{0.814})$	86.200	$(0.152_{0.137})$	89.800	$(0.152_{0.142})$	89.000
β_{20}	0.202	1.225	0.128	35.887	0.061	46.910	0.125	26.260
(0.2)	$(0.129_{0.124})$	97.000	$(0.107_{0.163})$	87.000	$(0.587_{0.608})$	91.200	$(0.585_{0.634})$	92.000
β_{21}	−0.102	2.919	−0.011	58.841	−0.026	57.399	−0.095	13.440
(−0.1)	$(0.184_{0.183})$	95.200	$(0.671_{0.610})$	82.000	$(0.527_{0.408})$	98.800	$(0.527_{0.419})$	98.800
β_{22}	0.205	2.824	0.176	11.565	0.181	8.907	0.113	26.727
(0.2)	$(0.245_{0.243})$	94.200	$(0.090_{0.128})$	83.600	$(0.170_{0.162})$	89.200	$(1.711_{1.655})$	89.200
β_{23}	0.310	3.347	0.193	35.441	0.247	45.332	0.257	37.365
(0.3)	$(0.350_{0.357})$	94.400	$(0.088_{0.118})$	71.000	$(0.130_{0.132})$	62.400	$(1.296_{1.290})$	86.000
β_{30}	0.203	1.820	0.081	59.089	0.176	10.384	0.159	14.035
(0.2)	$(0.137_{0.140})$	95.400	$(0.123_{0.128})$	82.800	$(0.499_{0.471})$	94.200	$(0.501_{0.482})$	94.000
β_{31}	−0.203	1.866	−0.053	45.382	−0.015	77.547	−0.039	58.033
(−0.2)	$(0.197_{0.200})$	94.800	$(0.712_{0.468})$	92.400	$(0.455_{0.351})$	96.400	$(0.455_{0.369})$	96.800
β_{32}	0.202	1.324	0.199	0.282	0.521	87.758	0.685	99.796
(0.2)	$(0.261_{0.262})$	94.000	$(0.099_{0.107})$	92.400	$(0.209_{0.195})$	95.600	$(0.209_{0.198})$	94.600
β_{33}	0.409	2.395	0.207	48.095	0.863	69.842	1.067	86.998
(0.4)	$(0.373_{0.377})$	94.200	$(0.089_{0.097})$	82.200	$(0.173_{0.167})$	95.600	$(1.725_{1.739})$	93.000
F.norm	4.796		4.802		4.811		4.918	

복수는 소절 3.1에서 제시된 것과 동일하며, 회귀계수 값도 동일하다. 공분산행렬의 구조는 동질한 AR(1)이며 이 공분산행렬을 분해하여 만들어진 모수들인 일반화자기회귀모수, 혁신분산모수 그리고 초구모수의 선형모형인 식 (2.8), (2.10), 그리고 (2.13)에서 만들어지는 모수들의 값은 아래와 같다.

$$(\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33}) = (0.3, 0.4, 0.1, 0.1, 0.3, 0.1, 0.1, 0.3, 0.2),$$
$$(\lambda_1, \lambda_2, \lambda_3) = (0.2, 0.2, 0.2),$$
$$(\nu_1, \nu_2, \nu_3) = (-0.5, -0.4, -0.3).$$

표본의 크기(N)가 100, 500으로 하고 500개의 자료집합을 생성하였다. 이렇게 만들어진 자료집합을 소절 3.1에서 적합시킨 4개의 모형이 다시 적합시켰다. 그 결과로 각 모수의 추정치 평균(mean), 상대편향의 절댓값(ARB), 추정량의 표준오차의 평균(SE), 추정치들의 표준편차(SD), 그리고 포함확률(CP)을 계산하였다. 또한 Kohli 등 (2016)이 제안한 모형과 Lee 등 (2019)이 제안한 모형 간의 비교를 위해 프로베니우스 노름을 다시 사용하여 비교하였다.

Tables 3.5과 3.6은 표본의 크기가 100일 때 4가지 모형의 결과를 표로 나타낸 것이다. 또한, 모든 평균모수들의 추정정도를 한 번에 파악하기 위해 ARB값의 합인 $\sum_{i=1}^{3}\sum_{j=0}^{3}(|\hat{\beta}_{ij} - \beta_{ij}|/\beta_{ij}) \times 100$를 계

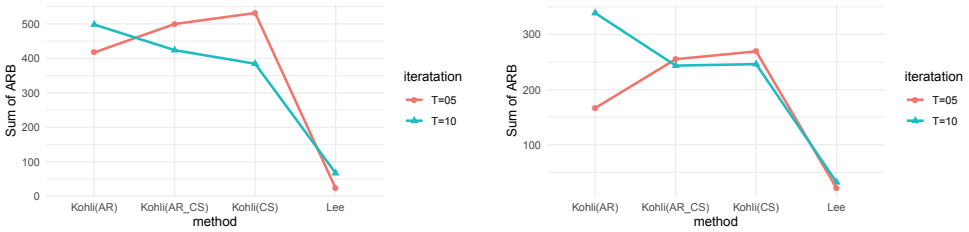
Table 3.6. $N = 100$, $n = 10$

β	Lee		Kohli(AR)		Kohli(CS)		Kohli(AR-CS)	
	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)
β_{10}	0.304	1.424	0.122	32.536	0.231	22.950	0.233	22.048
(0.3)	$(0.165_{0.170})$	95.000	$(0.483_{0.438})$	93.400	$(0.129_{0.161})$	85.200	$(0.126_{0.154})$	87.200
β_{11}	−0.102	2.483	−0.033	36.643	−0.037	33.234	−0.025	40.025
(−0.1)	$(0.235_{0.240})$	95.200	$(0.719_{0.698})$	95.000	$(0.071_{0.048})$	81.200	$(0.722_{0.502})$	78.200
β_{12}	0.231	0.667	0.126	23.063	0.164	17.978	0.156	21.600
(0.2)	$(0.659_{0.667})$	94.800	$(0.808_{0.819})$	72.400	$(0.102_{0.142})$	83.400	$(0.100_{0.141})$	79.800
β_{13}	0.285	4.729	0.553	47.774	0.184	38.416	0.184	38.371
(0.3)	$(0.942_{0.938})$	95.400	$(0.511_{0.426})$	87.000	$(0.082_{0.123})$	65.200	$(0.082_{0.119})$	64.800
β_{20}	0.206	3.446	0.190	20.467	0.124	37.633	0.124	37.778
(0.2)	$(0.147_{0.149})$	95.000	$(0.583_{0.716})$	89.400	$(0.108_{0.182})$	72.400	$(0.108_{0.174})$	72.600
β_{21}	−0.109	9.537	−0.126	37.305	−0.055	20.055	−0.077	17.922
(−0.1)	$(0.209_{0.213})$	95.000	$(0.526_{0.438})$	98.000	$(0.674_{0.601})$	71.100	$(0.673_{0.613})$	70.000
β_{22}	0.216	8.388	1.195	67.659	0.172	13.861	0.172	13.779
(0.2)	$(0.577_{0.578})$	94.600	$(1.689_{1.624})$	88.000	$(0.090_{0.140})$	78.800	$(0.090_{0.134})$	81.200
β_{23}	0.323	7.807	0.257	57.099	0.187	37.614	0.180	39.938
(0.3)	$(0.824_{0.821})$	95.600	$(0.977_{1.005})$	56.600	$(0.092_{0.129})$	66.600	$(0.089_{0.128})$	64.000
β_{30}	0.201	0.752	0.396	20.269	0.072	63.551	0.086	56.767
(0.2)	$(0.154_{0.163})$	93.800	$(0.502_{0.543})$	91.800	$(0.127_{0.141})$	80.600	$(0.125_{0.138})$	81.000
β_{31}	−0.206	3.306	−0.093	39.534	0.104	50.521	−0.048	79.975
(−0.2)	$(0.220_{0.231})$	93.600	$(0.453_{0.362})$	98.000	$(0.699_{0.459})$	84.000	$(0.705_{0.477})$	91.100
β_{32}	0.248	24.196	0.337	68.571	0.199	0.267	0.189	5.238
(0.2)	$(0.610_{0.622})$	93.800	$(1.058_{1.083})$	92.400	$(0.103_{0.121})$	89.400	$(0.101_{0.119})$	90.000
β_{33}	0.401	0.424	0.686	47.170	0.206	48.311	0.198	50.457
(0.4)	$(0.872_{0.878})$	95.200	$(1.005_{1.052})$	90.200	$(0.092_{0.110})$	75.400	$(0.090_{0.108})$	79.600
F.norm	4.121		5.343		6.522		7.131	

산하였고 그 결과를 Figure 3.2의 (a)에 나타내었다. 회귀계수는 우리가 예상하는 바와 같이 Lee 등 (2019)의 모형으로 적합한 경우에 편향이 적었다. 하지만 Kohli 등 (2016)의 모형을 적합시에는 편향이 많이 있음을 알 수 있다. 이는 3.1소절에서 Lee 등 (2019) 모형적합에서 나타나는 편향보다 훨씬 더 큼을 알 수 있다. 이를 통하여 Lee 등 (2019) 모형이 Kohli 등 (2016)의 모형적합 보다 좀 더 강건함(robust)을 알 수 있다.

F.norm 기준으로 살펴보았을 때, 반복수와 관계없이 Lee 등 (2019) 모형의 F.norm 값이 가장 작다. 이는 Kohli 등 (2016)의 방법들보다 Lee 등 (2019) 모형이 공분산 행렬을 더 잘 추정한다고 볼 수 있다. 반복수가 증가함에 따라 Lee 등 (2019) 모형의 F.norm 값은 작아졌지만, Kohli 등 (2016)의 세 모형의 F.norm 값은 커졌다. 회귀계수에 대한 정보 기준으로 모형을 비교하면, 반복수와는 상관없이 Lee 등 (2019) 모형의 포함확률이 가장 큰 것을 알 수 있다. 상대편향의 절댓값 측면에서도 대부분 Lee 등 (2019) 모형에서의 값이 가장 작으므로, Lee 등 (2019) 모형이 회귀계수 또한 잘 추정한다고 해석할 수 있다.

Tables 3.7과 3.8은 표본의 크기가 500일 때 4가지 모형의 결과를 표로 나타낸 것이다. 그리고 모든 평균모수들의 추정정도를 한 번에 파악하기 위해 ARB값의 합인 $\sum_{i=1}^{3}\sum_{j=0}^{3}(|\hat{\beta_{ij}} - \beta_{ij}|/\beta_{ij}) \times 100$를 계산하였고 그 결과를 Figure 3.2의 (b)에 나타내었다. F.norm 기준으로 살펴보았을 때, 반복수와 관계



(a) $N = 100$ (b) $N = 500$

Figure 3.2. Sum of ARBs of mean parameters in Lee *et al.* (2020)'s and Kohli *et al.* (2016)'s models using the data generated from Lee *et al.* (2020)'s model.

Table 3.7. $N = 500$, $n = 5$

β	Lee		Kohli(AR)		Kohli(CS)		Kohli(AR-CS)	
	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)
β_{10}	0.296	1.114	0.185	21.165	0.163	28.308	0.172	24.333
(0.3)	$(0.074_{0.073})$	95.800	$(0.212_{0.182})$	92.000	$(0.209_{0.223})$	92.400	$(0.183_{0.139})$	91.000
β_{11}	−0.098	1.357	−0.060	33.910	−0.075	20.845	−0.081	17.379
(−0.1)	$(0.106_{0.108})$	94.400	$(0.317_{0.287})$	95.400	$(0.359_{0.373})$	91.200	$(0.366_{0.364})$	91.200
β_{12}	0.198	0.532	0.128	24.082	0.224	14.396	0.208	3.385
(0.2)	$(0.299_{0.287})$	96.000	$(0.452_{0.426})$	87.200	$(0.093_{0.100})$	90.000	$(0.283_{0.244})$	92.600
β_{13}	0.298	0.448	0.408	23.850	0.160	39.848	0.156	40.097
(0.3)	$(0.423_{0.411})$	97.000	$(0.673_{0.600})$	89.000	$(0.158_{0.139})$	87.600	$(0.228_{0.283})$	90.000
β_{20}	0.198	0.789	0.194	4.015	0.173	15.288	0.181	12.633
(0.2)	$(0.066_{0.063})$	96.200	$(0.254_{0.290})$	92.400	$(0.232_{0.298})$	88.400	$(0.265_{0.233})$	88.200
β_{21}	−0.099	0.049	−0.104	3.594	−0.088	17.391	−0.090	20.112
(−0.1)	$(0.945_{0.954})$	95.000	$(0.232_{0.187})$	97.000	$(0.247_{0.235})$	91.400	$(0.173_{0.169})$	92.400
β_{22}	0.193	3.350	0.134	10.011	0.179	9.395	0.125	14.838
(0.2)	$(0.261_{0.259})$	94.800	$(0.752_{0.693})$	86.200	$(0.399_{0.323})$	87.000	$(0.098_{0.093})$	88.400
β_{23}	0.303	1.313	0.273	9.494	0.267	20.934	0.261	23.776
(0.3)	$(0.370_{0.380})$	94.400	$(0.567_{0.562})$	93.400	$(0.632_{0.655})$	88.200	$(0.533_{0.518})$	89.800
β_{30}	0.196	1.892	0.194	3.022	0.187	7.394	0.188	8.332
(0.2)	$(0.070_{0.067})$	97.200	$(0.220_{0.215})$	92.400	$(0.238_{0.277})$	90.400	$(0.163_{0.132})$	92.200
β_{31}	−0.192	3.771	−0.179	7.001	−0.025	46.384	−0.083	35.365
(−0.2)	$(0.099_{0.098})$	96.400	$(0.201_{0.165})$	88.200	$(0.211_{0.236})$	89.000	$(0.234_{0.284})$	90.800
β_{32}	0.201	0.888	0.247	8.382	0.372	32.686	0.402	36.443
(0.2)	$(0.276_{0.272})$	95.000	$(0.406_{0.388})$	93.600	$(0.118_{0.106})$	87.200	$(0.100_{0.094})$	89.200
β_{33}	0.377	5.687	0.663	18.109	0.636	16.325	0.688	18.355
(0.4)	$(0.392_{0.380})$	95.600	$(0.751_{0.779})$	88.600	$(0.073_{0.078})$	89.400	$(0.097_{0.093})$	90.000
F.norm	4.139		4.852		6.389		8.378	

없이 Lee 등 (2019) 모형의 F.norm 값이 가장 작다. 이는 표본의 크기가 100일 때뿐만 아니라 500일 때에도 Lee 등 (2019) 모형이 공분산 행렬을 더 잘 추정한다고 볼 수 있다. 또한, 4개의 모형 모두 반복수가 커짐에 따라 F.norm 값이 작아졌다. 따라서 반수가 커짐에 따라 모든 모형이 공분산 행렬을 잘 추정한다고 해석할 수 있다. 회귀계수에 대한 정보 기준으로 모형을 비교하면, 반복수가 증가함에 따라 모든 모형의 SE와 SD는 감소한다. 그러나 포함확률의 변화는 반복수가 증가하는 것과 큰 상관이 없는 것

Table 3.8. $N = 500$, $n = 10$

β	Lee		Kohli(AR)		Kohli(CS)		Kohli(AR-CS)	
	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)	Mean (SE_{SD})	ARB CP(%)
β_{10}	0.297	0.852	0.234	21.858	0.236	19.134	0.242	16.332
(0.3)	$(0.070_{0.069})$	95.000	$(0.058_{0.065})$	77.000	$(0.073_{0.081})$	82.000	$(0.078_{0.069})$	84.200
β_{11}	−0.091	8.216	−0.083	10.239	−0.073	17.389	−0.071	18.373
(−0.1)	$(0.100_{0.098})$	94.600	$(0.341_{0.242})$	92.000	$(0.032_{0.038})$	89.200	$(0.227_{0.221})$	88.200
β_{12}	0.202	1.330	0.160	19.895	0.163	17.733	0.158	21.229
(0.2)	$(0.134_{0.136})$	95.000	$(0.045_{0.063})$	76.600	$(0.051_{0.068})$	78.000	$(0.068_{0.073})$	81.400
β_{13}	0.290	3.251	0.184	38.534	0.186	37.110	0.187	36.385
(0.3)	$(0.190_{0.183})$	96.600	$(0.037_{0.051})$	84.200	$(0.044_{0.062})$	81.200	$(0.053_{0.066})$	83.800
β_{20}	0.195	2.430	0.126	36.892	0.153	18.331	0.151	19.989
(0.2)	$(0.058_{0.056})$	96.200	$(0.050_{0.074})$	86.800	$(0.056_{0.099})$	82.600	$(0.052_{0.087})$	87.200
β_{21}	−0.094	5.104	−0.124	25.391	−0.077	10.378	−0.081	9.341
(−0.1)	$(0.083_{0.084})$	94.200	$(0.317_{0.312})$	86.400	$(0.277_{0.243})$	87.600	$(0.192_{0.188})$	88.400
β_{22}	0.205	2.677	0.171	14.112	0.168	16.776	0.175	10.378
(0.2)	$(0.111_{0.109})$	95.000	$(0.041_{0.057})$	78.000	$(0.046_{0.076})$	81.400	$(0.065_{0.078})$	85.000
β_{23}	0.294	1.7000	0.183	38.783	0.198	20.311	0.197	21.534
(0.3)	$(0.157_{0.152})$	96.000	$(0.040_{0.058})$	76.400	$(0.048_{0.064})$	81.400	$(0.044_{0.073})$	83.400
β_{30}	0.196	1.840	0.079	60.452	0.093	49.433	0.097	47.988
(0.2)	$(0.062_{0.062})$	95.000	$(0.058_{0.057})$	79.400	$(0.064_{0.069})$	82.600	$(0.076_{0.072})$	86.000
β_{31}	−0.195	2.171	−0.132	20.061	−0.132	19.998	−0.123	23.423
(−0.2)	$(0.088_{0.091})$	94.200	$(0.333_{0.199})$	92.600	$(0.208_{0.200})$	89.600	$(0.174_{0.182})$	90.000
β_{32}	0.201	0.975	0.194	2.947	0.195	1.733	0.192	3.112
(0.2)	$(0.118_{0.120})$	94.600	$(0.046_{0.051})$	91.000	$(0.046_{0.051})$	90.200	$(0.063_{0.076})$	91.400
β_{33}	0.393	1.535	0.202	49.496	0.332	17.777	0.347	15.379
(0.4)	$(0.167_{0.169})$	94.200	$(0.041_{0.046})$	88.600	$(0.057_{0.051})$	82.600	$(0.046_{0.042})$	91.600
F.norm	4.912		9.273		11.634		12.337	

을 알 수 있다.

Kohli 등 (2016)이 제안한 AR, CS, AR-CS 모형을 비교하면, 반복수가 적을 때에는 세 모형의 SE, SD 값의 차이가 크다. AR 모형의 SE와 SD가 가장 작을 때가 많으며, AR-CS 모형의 SE와 SD 값이 가장 큰 경우가 많다. 이러한 사실 때문에 AR-CS 모형에서의 포함확률이 가장 크게 나타난다. 그러나 반복수가 적어짐에 따라 각 모형의 SE와 SD 차이가 작아지는 것을 알 수 있다.

4. 결론

본 논문에서 다변량 경시적 자료분석을 위해 공분산행렬을 모형화 하는 여러 방법을 고찰하였다. Pourahmadi (1999)는 단변량 경시적 자료의 공분산 행렬을 모형화하기 위해 MCD를 이용하였다. MCD에 의하여 만들어진 모수들은 선형/로그선형모형으로 모형화하여 모수의 수를 줄이고 양정치성을 만족시킬수 있다. 이러한 MCD의 장점 때문에, Kohli 등 (2016)과 Lee 등 (2019)은 MCD를 이용하여 다변량 경시적 자료의 공분산 행렬을 모형화하는 방법론을 제안하였다. 두 방법론 모두 MCD를 이용하였다는 공통점이 있지만 혁신공분산행렬을 모형화하는 방법이 다르다는 차이점이 있다. 우선 Kohli 등 (2016)은 혁신공분산행렬을 직접적으로 추정하지 않고, 이를 미지의 음수가 아닌 모수와 알려진 양정치

행렬의 선형결합 식으로 변환함으로써 추정하였다. 따라서 알려진 양정치 행렬을 어떤 구조에서 모수적 모형화 하느냐에 따라 AR / CS / AR-CS 유형으로 나뉜다. Lee 등 (2019) 또한 혁신공분산행렬을 혁신분산을 대각원소로 가지는 대각행렬과 각 시점에서의 반응변수에 대한 상관행렬의 식으로 분해함으로써 해석 가능한 추정을 하였다.

Kohli 등 (2016)과 Lee 등 (2019)의 성능을 비교하기 위해 모의실험을 진행하였다. 회귀계수와 공분산 행렬을 얼마나 잘 추정했는지 확인하기 위해 평균 회귀계수들의 편향, 포함확률과 프로베니우스 노름 등을 계산하였다. 표본의 크기와 반복수와 관계없이 Lee 등 (2019)의 모형이 프로베니우스 노름 값이 가장 작았다. 즉, 주어진 상황 내에서는 Lee 등 (2019)이 공분산 행렬을 가장 정확하게 추정한다는 것을 알 수 있다. 그리고 Lee 등 (2019)은 회귀계수에 추정에서 Kohli 등 (2016)보다 강건함을 보이고 있음을 알 수 있었다.

References

Kim, C. and Zimmerman, D. L. (2012). Unconstrained models for the covariance structure of multivariate longitudinal data, *Journal of Multivariate Analysis*, **107**, 104–118.

Kohli, P., Garcia, T. P., and Pourahmadi, M. (2016). Modeling the Cholesky factors of covariance matrices of multivariate longitudinal data, *Journal of Multivariate Analysis*, **145**, 87–100.

Lee, K., Cho, H., Kwak, M. S. and Jang, E. J. (2019). Estimation of covariance matrix of multivariate longitudinal data using modified Choleksky and hypersphere decompositions, *Biometrics*, **76**, 75–86.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, *Biometrika*, **86**, 677–690.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions, *Biometrika*, **87**, 823–836.

Xu, J. and Mackenzie, G. (2012). Modelling covariance structure in bivariate marginal models for longitudinal data, *Biometrika*, **99**, 649–662.

Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data, *Journal of the American Statistical Association*, **107**, 701–710.

다변량 경시적 자료 분석을 위한 공분산 행렬의 모형화 비교 연구

곽나영[a] · 이근백[a,1]

[a]성균관대학교 통계학과

(2020년 3월 5일 접수, 2020년 4월 2일 수정, 2020년 4월 9일 채택)

요 약

같은 개체로부터 반복 측정한 자료를 경시적 자료(longitudinal data)라고 한다. 이러한 자료를 분석하려면 흔히 사용되는 횡단 자료 분석과는 다른 분석 방법이 필요하다. 즉, 경시적 자료에서 공변량의 효과를 추정할 때에는 반복 측정된 결과 간의 상관성을 고려해야 하며, 따라서 공분산행렬을 모형화 하는 것이 매우 중요하다. 그러나 추정해야 할 모수가 많고, 추정된 공분산행렬이 양정치성을 만족해야 하므로 공분산 행렬의 모형화는 쉽지 않다. 특히 다변량 경시적 자료분석을 위한 공분산행렬의 모형화는 더욱더 심층적인 방법론을 사용해야 한다. 본 논문은 다변량 경시적 자료분석을 위한 공분산행렬을 모형화하기 위해 두 가지 방법론을 고찰한다. 두 방법 모두 수정된 콜레스키 분해(modified Cholesky decomposition)를 이용하여 시간에 따른 응답변수들의 상관관계를 설명하고 있다. 하지만 같은 시간에서 관측된 응답변수들간의 상관관계를 설명하는 방법이 다르다. 첫 번째 방법론에서는 향상된 선형 공분산 모형(enhanced linear covariance models)을 사용하여 공분산행렬이 양정치성을 만족하도록 한다. 두 번째 방법론에서는 분산-공분산 분해(variance-correlation decomposition)와 초구분해(hypersphere decomposition)을 이용하여 공분산 행렬을 모형화 한다. 이 두 방법론의 성능을 비교하고자 모의실험을 진행한다.

주요용어: 다변량 경시적 자료, 공분산 행렬 모형화, 수정된 콜레스키 분해, 양정치성, 초구분해

이 논문은 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (과제번호: NRF-2019R1F1A1058553).

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: keunbaik@skku.edu