The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.103
30(1), 103–117

Comparison of the covariance matrix for general linear model

Sang Ah Nam[a] · Keunbaik Lee[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received October 18, 2016; Revised December 18, 2016; Accepted December 29, 2016)

Abstract

In longitudinal data analysis, the serial correlation of repeated outcomes must be taken into account using covariance matrix. Modeling of the covariance matrix is important to estimate the effect of covariates properly. However, It is challenging because there are many parameters in the matrix and the estimated covariance matrix should be positive definite. To overcome the restrictions, several Cholesky decomposition approaches for the covariance matrix were proposed: modified autoregressive (AR), moving average (MA), ARMA Cholesky decompositions. In this paper we review them and compare the performance of the approaches using simulation studies.

Keywords: longitudinal data analysis, modified Cholesky decomposition, moving average Cholesky decomposition, general linear model

1. 서론

경시적 연구(longitudinal study)는 일정기간 동안 반복 측정된 자료를 분석하는 연구이다. 이 경우에 같은 개체(subject)에서 결과치들이 반복 측정 되어지고, 이 결과치들은 서로 시간에 따른 상관관계를 가지게 된다. 이러한 상관관계는 경시적 자료분석에서 공변량(covariates)의 효과를 올바르게 추정하려면 반드시 고려되어야 한다 (Diggle 등, 2002). 따라서 경시적 자료를 올바르게 분석하기 위한 모형들은 이러한 상관관계를 설명하기 위한 모형화에 집중하고 있다. 이 논문에서는 특히 경시적 연속형 자료에서 측정치들의 관련성을 설명하기 위하여 공분산행렬의 모형화에 초점을 맞추도록 한다. 공분산 행렬은 다양한 구조를 가질수 있지만 경시적 자료분석에서는 특히 자기회귀(autoregressive; AR) 구조 (Pourahmadi, 1999), 이동평균(moving average; MA) 구조 (Zhang과 Leng, 2012), 그리고 자기회귀-이동평균(autoregressive moving average; ARMA) 구조 (Lee 등, 2017)를 주로 가정하고, 이러한 구조를 가지는 공분산 행렬의 모형화를 위한 방법들이 개발되고 있다 (Kim과 Lee, 2015).

자기회귀 구조의 공분산 행렬을 모형화 하기 위한 방법으로 수정된 콜레스키 분해(modified Cholesky decomposition)가 개발되었다 (Pourahmadi, 1999). 이 방법에서는 공분산 행렬의 역행렬을 일반화

This project was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2014R1A1A2054997, NRF-2016R1D1A1B03930343).

[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: keunbaik@skku.edu

자기회귀 모수(generalized autoregressive parameters; GARPs)와 혁신분산(innovation variances; IVs)으로 분해하여 공분산행렬의 모수들을 직접 추정화 하지 않고 이 새로운 모수들을 모형화한다. 이를 위하여 각각 회귀모형과 로그 선형모형을 이용하여 모수의 숫자를 줄이면서 추정된 공분산 행렬의 양정치성을 만족시키게 된다. Daniels와 Pourahmadi (2002)과 Daniels와 Zhao (2003)는 수정된 콜레스키 분해에서 모수들의 추정을 위한 베이지안 방법을 제안하였다. Pan과 MacKenzie (2003, 2006)는 불균형 경시적 자료를 고려한 수정된 콜레스키 분해를 이용하여 경시적 자료를 분석하였다. Lee (2013)와 Lee 등 (2012)은 수정된 콜레스키 분해를 일반화 선형혼합모형(generalized linear mixed models; GLMM)에 적용하였고, Lee와 Sung (2014)은 이를 주변화 임의효과 모형(marginalized random effects models)으로 확장 하였다.

이동평균 구조의 공분산 행렬은 경시적 자료에서 반복이 많은 경우에 주로 사용한다. 앞에서 살펴 본 수정된 콜레스키 분해를 적용하면 반복 수가 증가함에 따라 그 차수를 높여야 함을 알 수 있다. 따라서 무한차수의 자기회귀모형은 유한차수의 이동평균모형으로 변환할 수 있다는 점을 이용할 수 있다. 따라서 이동평균 구조의 공분산행렬의 모형화를 위하여 이동평균 콜레스키 분해(moving average Cholesky decomposition)가 제안되었다 (Zhang과 Leng, 2012). 이 방법은 수정된 콜레스키 분해와 달리 공분산 행렬을 일반화 이동평균 모수(generalized moving average parameters; GMAPs)와 혁신 분산으로 직접 분해하는 방법을 제시하였다. 그리고 이 모수들을 수정된 콜레스키 분해에서 처럼 회귀모형과 로그 선형 모형을 이용하여 모형화 한다. 수정된 콜레스키 분해와 유사하게 혁신 분산이 양의 값을 가지므로 공분산 행렬은 항상 양정치성을 가진다. Lee와 Yoo (2014)는 베이지안 추정을 이용하여 반복수가 많은 경시적 이항 자료를 분석할 때 이동평균 콜레스키 분해를 사용하였다. Kim 등 (2016)은 경시적 순서 자료 분석을 위해서 Lee와 Yoo (2014)의 모형을 확장하였고, 자기회귀와 이동평균 콜레스키 분해방법을 비교하여 각 방법들이 정도행렬(precision matrix)과 공분산행렬의 모형화에 적합함을 보였다.

자료를 자기회귀모형이나 이동평균모형만으로 표현하려면 추정해야 할 모수의 수가 너무 많아질 수 있으며, 그 결과 추정의 효율성이 떨어지고 결과의 해석에 어려움이 생긴다 (Lee 등, 2017). 따라서 자기회귀-이동평균 모형(autoregressive-moving average models)을 통해 이러한 문제점을 해결하여 적은 수의 모수로 다양한 구조의 공분산 행렬을 추정하며, 그 결과 모형 해석이 쉬워지고 모수 추청을 안정적으로 할 수 있다. 또한 복잡한 형태의 자기회귀 모형 혹은 이동평균 모형보다 더 나은 예측을 제공한다. Lee 등 (2017)은 자기회귀-이동평균 구조를 가지는 공분산 행렬을 자기회귀-이동평균 콜레스키 분해를 제안하였고, 최대 가능도 추청 방법을 이용하여 모수를 추정하였다.

우리는 위에서 제시된 3가지의 공분산행렬의 분해방법들을 고찰하고, 이들 방법들의 장단점을 모의실험을 통하여 제시하고자 한다. 이를 위하여 본 논문의 구성은 다음과 같다. 2장에서는 우선 일반 선형모형(general linear model)과 앞에서 제시된 공분산 행렬의 분해를 좀 더 자세히 제시한다. 3장에서는 앞서 설명한 모형들을 여러 조건 하에서 다양한 모의실험을 시행하여 그 방법들의 장단점들을 비교하여 그 결과를 논의하도록 한다. 마지막으로 4장에서는 앞서 논의한 내용들을 토대로 결론을 다루도록 한다.

2. 공분산 행렬의 모형화

이 절에서 우리는 연속형 자료를 분석하기 위한 일반 선형모형을 고려하고, 그 모형에서 공분산 행렬의 모형화에 대한 여러 방법들을 고찰한다. 우선 수정된 콜레스키 분해는 경시적 연속형 자료분석에서 공분산행렬의 모형화를 위하여 처음 제안되었다 (Pourahmadi, 1999, 2000). 공분산 행렬은 항상 양정치성을 만족해야 하고, 고차원이므로 추정하는 데 어려움이 있다. 이 절에서는 이러한 어려움을 해소하고 시간에 따른 변동과 개체들 간의 변동을 잘 설명해주는 모형들 중에 수정콜레스키 분해 (Pourahmadi,

1999, 2000), 이동평균 콜레스키분해 (Zhang과 Leng, 2012), 그리고 자기회귀 이동평균 콜레스키 분해 (Lee 등, 2017)를 사용하는 모형들을 살펴본다.

우선 응답변수 Y_{it}는 시간 t $(t = 1, \ldots, n_i)$에서의 개체 i $(i = 1, \ldots, N)$의 반응변수(response variable)라고 하고, X_{it}는 Y_{it}에 상응하는 $p \times 1$ 공변량 벡터(covariate vector)이다. 그리고 다른 개체의 응답과는 독립이라고 가정한다. 만약 X_i가 확률적이라면 $(Y_i, X_i), \ldots, (Y_i, X_i)$은 독립이다. 또한 X_i가 주어졌을 때, $E(Y_i \mid X_i) = X_i\beta$, $\mathrm{var}(Y_i \mid X_i) = \sum_i$이다. 공분산행렬 Σ_i는 같은 그룹 내에서의 측정치들의 상관성을 나타낸다. 공분산은 공변량에 의해 변하며 시간에 관한 함수이다.

2.1. 공분산 행렬의 자기회귀를 포함한 수정 콜레스키 분해

일반 선형모형에서 자기회귀 구조를 가지는 공분산 행렬은 수정 콜레스키 분해(modified Cholesky decomposition)를 이용하여 모형화 할 수 있다 (Pourahmadi, 1999). 이를 설명하기 위하여 응답변수 Y_{it}를 다음과 같이 가정한다.

$$\begin{aligned} y_{i1} - \mu_{i1}(\beta) &= e_{i1}, \\ y_{it} - \mu_{it}(\beta) &= \sum_{j=1}^{t-1} \phi_{itj} \left(y_{ij} - \mu_{ij}(\beta)\right) + e_{it}, \quad \text{for } t = 2, \ldots, n_i, \end{aligned}$$

여기서 $\mu_{it}(\beta) = x_{it}^T\beta$, $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T$는 평균이 0, 분산이 $\boldsymbol{D}_i$인 정규분포를 가정하고, $\boldsymbol{D}_i = \mathrm{diag}(\sigma_{i1}^2, \ldots, \sigma_{in_i}^2)$, ϕ_{itj}는 일반화 자기회귀모수라고 부르며, ϕ_{itj}의 값은 앞의 결과가 현재의 결과에 영향을 미치는 자기회귀 구조를 가지게 된다. σ_{it}^2는 혁신분산이다. 위의 식을 행렬로 표현하면 다음과 같다.

$$T_i(y_i - \mu_i(\beta)) = e_i, \tag{2.1}$$

여기서 $y_i = (y_{i1}, \ldots, y_{in_i})^T$, $\mu_i(\beta) = (\mu_{i1}(\beta), \ldots, \mu_{in_i}(\beta))^T$, 그리고 T_i는 주대각 요소(diagonal elements)가 1이고, 하대각 요소(lower off-diagonal elements)에서 (t, j)번째 요소가 $-\phi_{itj}$인 $n_i \times n_i$ 차원의 행렬이다.

식 (2.1)에 분산을 취하면 다음과 같은 식을 얻는다.

$$\begin{aligned} T_i\Sigma_iT_i^T &= \mathrm{Var}(e_i) = D_i, \\ \Sigma_i &= T_i^{-1}D_iT_i^{-T} \Leftrightarrow \Sigma_i^{-1} = T_i^TD_i^{-1}T_i. \end{aligned}$$

이 결과로부터 수정된 콜레스키 분해는 공분산 행렬의 정도 행렬(precision matrix)을 직접적으로 모형화할 수 있음을 알 수 있다. 그리고 Σ_i의 모수들인 ϕ_i와 σ_i는 개체의 반복수 n_i가 증가하면 기하급수 적으로 증가하게 된다. 따라서 이 모수들을 시간 그리고/또는 개체-특정적 공변량 벡터인 w_{itj}와 h_{it}를 이용하여 모수의 개수를 줄일 수 있다. 이를 위하여 회귀식과 로그 선형모형(log linear model)을 이용한다.

$$\phi_{itj} = w_{itj}^T\gamma, \qquad \log\sigma_{it}^2 = h_{it}^T\lambda, \tag{2.2}$$

여기서 γ와 λ는 $a \times 1$과 $c \times 1$의 모르는 모수벡터이다 (Daniels와 Zhao, 2003; Lee, 2013; Lee 등, 2012; Pourahmadi, 1999, 2000; Daniels와 Pourahmadi, 2002). 그리고 w_{itj}의 선택을 통하여 공분산행렬을 AR(p)의 구조를 만들 수 있다. 그리고 h_{it}의 선택을 통하여 이공분산성(heteroscedasticity)을 만족하게 할 수 있다. 예를 들면 $w_{itj} = I_{|t-j|=1}$이라고 가정하자. 이는 시차가 1인 경우에만 1이고 나머지는 모두 0인 함수이다. 그러면 ϕ_{itj}는 t와 j의 차가 1인 경우에만 γ가 되고, 나머지는 모두 0이 된다. 따

라서 AR(1)의 구조를 가지게 된다. 그리고 $h_{it} = (1, \text{SEX}_i)$이라고 가정하자. 여기서 SEX_i는 i번째 개체의 성별을 나타내고 남자일 때 1이고 여자일때 0으로 가정한다. 그러면 혁신분산이 성별에 따라 다른 값을 가지게 된다. 따라서 전체 공분산행렬은 성별에 따라 달라지게 된다. 식 (2.2)에서 모수들이 어떠한 제약이 없기 때문에 공변량(w_{itj}, h_{it})의 선택으로 공분산행렬의 다양한 차수의 자기회귀와 이분산성을 가질 수 있는 장점이 있다. 그리고 로그선형모형을 통하여 모든 혁신분산이 항상 양수이므로 이 결과 Σ_i는 양정치성을 만족하게 된다 (Pourahmadi, 1999).

2.2. 공분산 행렬의 이동평균 콜레스키 분해

앞의 2.1절에서 제시된 방법과 유사하게 이동평균 콜레스키 분해는 아래와 같이 모형을 가정한다 (Zhang과 Leng, 2012).

$$y_{i1} - \mu_{i1}(\beta) = e_{i1}, \tag{2.3}$$

$$y_{it} - \mu_{it}(\beta) = \sum_{j=1}^{t-1} l_{itj} e_{ik} + e_{it}, \quad \text{for } t = 2, \ldots, n_i, \tag{2.4}$$

여기서 $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T$는 평균이 0, 공분산행렬이 $\boldsymbol{D}_i$인 정규분포를 가정한다. 그리고 l_{itj}는 일반화 이동평균모수(generalized moving average parameters; GMAPs)이다.

이 모형을 행렬 형식으로 표현하면 다음과 같다.

$$(y_i - \mu_i) = L_i e_i, \tag{2.5}$$

여기서 L_i는 수정된 콜레스키 분해와 비슷하게 주대각 요소가 1이고, 하대각 요소(lower off-diagonal elements)에서 (t, j) 번째 요소가 l_{itj}인 $n_i \times n_i$ 차원의 행렬이다. 수정된 콜레스키 분해와 같이 혁신분산이 양수이면 Σ_i는 양정치성을 만족한다.

식 (2.5)에 분산을 취하여 y_i의 공분산행렬 Σ_i는 아래와 같다.

$$\Sigma_i = L_i D_i L_i^T. \tag{2.6}$$

이 결과로 이동평균 콜레스키 분해는 공분산 행렬을 직접적으로 모형화 할 수 있음을 알 수 있다. 수정된 콜레스키 분해와 같이 일반화 이동평균모수와 혁신분산은 공변량 z_{itj}와 h_{it}을 이용하여 다음과 같이 모형화 할 수 있다.

$$l_{itj} = z_{itj}^T \eta, \qquad \log \sigma_{it}^2 = h_{it}^T \lambda,$$

여기서 η와 λ는 $b \times 1$와 $c \times 1$의 각각의 모르는 모수벡터이다. 여기서 수정 콜레스키분해 방법에서 처럼 벡터 z_{itj}와 h_{it}는 개체-특정적 공변량이고, 이를 이용하여 공분산행렬은 개체의 특성에 따라 다른 차수의 이동평균 구조를 가지게 할 수 있고, 추정된 공분산행렬이 이분산성을 만족하게 된다 (Zhang과 Leng, 2012). 그리고 수정 콜레스키 분해에서처럼 이렇게 만들어진 공분산행렬은 항상 양정치성을 만족한다.

2.3. 공분산 행렬의 자기회귀-이동평균 콜레스키 분해

수정된 콜레스키와 이동평균 콜레스키 분해를 결합하여 Lee 등 (2016)은 다음과 같은 모형을 제안하였다.

$$y_{i1} - \mu_{i1}(\beta) = e_{i1},$$
$$y_{it} - \mu_{it}(\beta) = \sum_{j=1}^{t-1} \phi_{itj}(y_{ij} - \mu_{ij}(\beta)) + \sum_{j=1}^{t-1} l_{itj} e_{ik} + e_{it},$$

여기서 $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T$는 평균이 0, 공분산행렬이 $\boldsymbol{D}_i$인 정규분포를 가정한다. 식을 행렬 형식으로 표현하면 다음과 같다.

$$T_i(y_i - \mu_i) = L_i e_i. \tag{2.7}$$

식 (2.7)에 분산을 취하면 다음과 같은 식을 얻는다.

$$T_i \Sigma_i T_i^T = L_i D_i L_i^T \tag{2.8}$$
$$\Leftrightarrow \Sigma_i = T_i^{-1} L_i D_i L_i^T T_i^{-T}. \tag{2.9}$$

따라서 공분산행렬 Σ_i는 일반화 자기회귀모수, 일반화 이동평균모수와 혁신분산을 모수로 가지게 되며, 이들은 공변량 w_{it}, z_{itj}와 h_{it}을 이용하여 다음과 같이 모형화 할 수 있다.

$$\phi_{itj} = w_{itj}^T \alpha, \qquad l_{itj} = z_{itj}^T \gamma, \qquad \log \sigma_{it}^2 = h_{it}^T \lambda, \tag{2.10}$$

여기서 α, γ와 λ는 $a \times 1$, $b \times 1$와 $c \times 1$의 각각의 모르는 모수벡터이다. 여기서 앞의 두 모형에서처럼 벡터 w_{itj}, z_{itj}와 h_{it}는 개체-특성적 공변량이고, 이를 이용하여 공분산행렬은 개체의 특성에 따라 다른 차수의 자기회귀-이동평균 구조를 가지게 할 수 있고, 추정된 공분산행렬이 이분산성을 만족하게 된다 (Lee 등, 2017). 그리고 이렇게 만들어진 공분산행렬은 항상 양정치성을 만족한다.

3. 모의 실험

이 절에서 모의실험을 통하여 공분산행렬의 모형화를 위한 3가지 방법들의 장단점을 알아 보고자 한다. 그리고 일반 선형모형에서 공분산행렬의 잘못된 가정이 회귀계수 추정에 어떻게 영향을 미치는지 알아보고자 한다. 그리고 이를 통하여 소개한 공분산 행렬의 모형화 방법들의 성능을 비교하고자 한다.

우선 자기회귀-이동평균 구조의 공분산행렬을 가지는 일반 선형모형에서 난수를 발생시킨다. 여기에서 공분산행렬이 같은 ARMA$(1,1)$ 구조를 가지면서 등분산성을 가지는 경우와 이분산성을 가지는 경우로 구분하였다. 각각의 경우에 표본수(sample size), 반복 수(replication number)를 변화시키며 난수를 발생시켰다. 표본의 수는 100, 300, 500으로 각각 하였고, 각 경우에 반복 횟수를 5, 10, 20인 경우로 해서 각각의 모의 표본을 완성하였다. 반복되는 응답변수에서 결측이 발생시켰고, 이 결측치는 임의 결측(missing at random; MAR)으로 하였고, 결측의 확률은 아래와 같이 계산되었다.

$$\text{logit}P(\text{dropout} = t \mid \text{dropout} \geq t) = -0.5 + 0.3Y_{it-1}. \tag{3.1}$$

이 결과 결측치의 비율은 40% 정도이다. 이러한 방법으로 500개의 자료집합(data set)을 만들었다.

위에서 제시된 500개의 자료집합을 각각의 자료집합을 가지고 3가지 모형에 적합한다. ARMA모형은 ARMA$(1,1)$ 구조의 공분산행렬을 가지는 모형으로 위에서 난수를 발생시킨 모형과 동일하다. AR모형은 공분산행렬이 AR(1) 구조를 가지는 일반 선형모형이며, MA모형은 공분산 행렬이 MA(1) 구조를 가지는 모형이다.

3.1. 공분산 행렬이 등분산성(homoscedastic)을 가지는 경우

우선 ARMA$(1,1)$ 구조의 등분산성을 만족하는 공분산행렬을 가지는 일반 선형모형에서 난수를 발생시켰다. 그 모형은 아래와 같다.

$$y_{it} - \mu_{it}(\beta) = \sum_{j=1}^{t-1} \phi_{itj}(y_{ij} - \mu_{ij}(\beta)) + \sum_{j=1}^{t-1} l_{itj} e_{ik} + e_{it}, \tag{3.2}$$

여기에서 $t = 1, \ldots, 10$에서,

$$\mu_{ij}(\beta) = \beta_0 + \beta_1 \text{Group}_i + \beta_2 \text{Time}_{it} + \beta_3 \text{Group}_i \times \text{Time}_{it} \tag{3.3}$$

이다. 그리고 회귀계수의 값은 $(\beta_0, \beta_1, \beta_2, \beta_3) = (0.1, -0.1, 0.1, 0.1)$이며, 공변량의 경우는 $\text{Time}_{it} = (t-1)/10$, Group_i은 0 또는 1의 값을 가지면서 각 그룹은 동일한 개체수를 가지게 할당한다. 그리고 GARP와 GMAP들의 값은 각각 $\phi_{itj} = \alpha \times I_{|t-j|=1}$, $l_{itj} = \gamma \times I_{|t-j|=1}$이고, IV는 $\log \sigma_{it}^2 = \lambda_0$이며, 그 모수들의 값들은 $(\alpha, \gamma, \lambda) = (0.7, 0.7, 0.1)$이다.

Tables 3.1–3.3는 표본의 수가 100, 300, 500개인 500개의 자료를 이용하여 모수들의 추정치들의 평균(mean), 참값과의 절대편향(absolute bias; AB), 추정치들의 표준편차(SD), 추정치의 표준오차들의 평균(SE), 그리고 포함확률(coverage probability)을 나타내고 있다. ARMA모형이 표본수나 반복 횟수에 상관 없이 AR모형, MA모형과 비교했을 때 편향이 가장 작은 값을 가진다. 그리고 포함확률도 전체적 으로 ARMA와 AR모형이 MA모형보다 크다는 것을 알 수 있다. 반복 횟수가 5, 10, 20으로 커짐에 따라 이러한 포함확률의 경향은 크게 변화지 않음을 알 수 있다. 그리고 표본수는 100, 300, 500으로 커져도 포함확률의 경향은 크게 차이가 나지 않는다. 이는 일정 표본수 이상이 되면 공분산 행렬의 추정에 영향을 미치지 않는다고 볼 수 있다. 또한 모든 경우에 있어 β_2의 편향과 포함확률을 보면 다른 두 모델과 비교했을 때 ARMA모형에서 가장 잘 추정된 것을 볼 수 있다. 이는 β_2는 시간과 관련된 모수이기 때문이다.

SAD에 관한 해석: $\beta_0, \beta_1, \beta_2, \beta_3$는 세 모형에 대해서 포함 확률을 기준으로 직접 비교가 가능했지만 AR모형에서는 오차항과 관련된 γ값이 없고, MA모형에서는 α값이 없으므로 세 모형의 α, γ, λ의 값이 얼마나 잘 추정되었는지 비교할 기준이 없다. 따라서 Σ를 구하여 그의 차를 비교한 값인 sum of absolute difference(SAD)를 이용한다. 표본수나 반복 횟수를 변화시킨 각각의 경우의 SAD값을 살펴보면 포함 확률을 비교했을 때와는 다르게 반복 횟수는 크게 영향을 미치는 것을 볼 수 있다. 반복 횟수가 커짐에 따라 ARMA, AR, MA모형에서 SAD값이 각각 커지는 것으로 나타났다. 그러나 ARMA모형에서는 값이 커지기는 하지 만 그 폭이 작아 어느 정도 잘 추정한다고 볼 수 있지만 AR과 MA 모형에서는 SAD값이 급격하게 증가해 잘 추정하지 못하였다. 결론적으로 표본수와 반복 횟수의 9가지 조합에 있어 포함 확률의 값과 마찬가지로 SAD값 역시 ARMA모형이 월등히 작게 나타났다.

이상의 결과에서 공분산행렬이 잘못 가정되면, 회귀계수의 추정치에 편향이 생김을 알 수 있다. 따라서 공분산행렬에 일반적으로 많이 쓰는 자기회귀구조의 가정을 이동평균 또는 자기회귀-이동평균 구조로 확장하여 가정함이 올바를 것으로 사려된다.

3.2. 공분산 행렬이 이분성(heteroscedastic)을 가지는 경우

여기에서는 ARMA$(1,1)$ 구조를 가지는 이분산성의 공분산 행렬을 가지는 일반 선형모형에서 난수를 발생시켰다. 나머지 형태의 모형은 앞에서 제시된 등분산성을 만족하는 경우의 모형과 동일하며 이를

Table 3.1. Comparison of three models with ARMA, AR, and MA structured covariance matrix (Homogeneous, $n = 100$)

		ARMA		AR		MA	
		Mean SE (SD)	AB Coverage	Mean SE (SD)	AB Coverage	Mean SE (SD)	AB Coverage
$T = 5$	β_0 (0.1)	0.11 0.11 (0.11)	0.01 0.96	0.10 0.16 (0.13)	0.00 0.97	0.09 0.53 (0.68)	0.01 0.87
	β_1 (−0.1)	−0.11 0.16 (0.15)	0.01 0.96	−0.12 0.23 (0.21)	0.02 0.97	−0.08 0.75 (0.97)	0.02 0.86
	β_2 (0.1)	0.09 0.74 (0.76)	0.01 0.94	0.39 0.95 (1.05)	0.29 0.92	0.08 0.64 (0.77)	0.02 0.89
	β_3 (0.1)	0.11 1.06 (1.09)	0.01 0.94	0.06 1.36 (1.58)	0.04 0.91	0.12 0.91 (1.09)	0.02 0.90
	α (0.7)	0.69 0.05 (0.05)	0.01 094	0.97 0.04 (0.04)	0.27 0.00		
	γ (0.7)	0.70 0.05 (0.05)	0.00 0.93			1.03 0.03 (0.03)	0.33 0.00
	λ (0.1)	0.09 0.06 (0.06)	0.01 0.95	0.28 0.07 (0.08)	0.18 0.31	0.39 0.06 (0.08)	0.29 0.03
SAD		1.09		7.23		21.01	
$T = 10$	β_0 (0.1)	0.10 0.11 (0.12)	0.00 0.93	0.09 0.17 (0.15)	0.01 0.96	0.11 0.17 (0.32)	0.01 0.72
	β_1 (−0.1)	−0.10 0.16 (0.16)	0.00 0.95	−0.08 0.24 (0.21)	0.02 0.98	−0.09 0.25 (0.46)	0.01 0.72
	β_2 (0.1)	0.10 0.35 (0.37)	0.00 0.94	0.67 0.49 (0.49)	0.57 0.79	0.12 0.24 (0.36)	0.02 0.81
	β_3 (0.1)	0.11 0.49 (0.50)	0.01 0.93	0.06 0.70 (0.69)	0.04 0.94	0.10 0.34 (0.53)	0.00 0.80
	α (0.7)	0.70 0.03 (0.03)	0.00 0.95	0.88 0.02 (0.02)	0.18 0.00		
	γ (0.7)	0.70 0.03 (0.03)	0.00 0.94			0.95 0.02 (0.02)	0.25 0.00
	λ (0.1)	0.09 0.04 (0.05)	0.01 0.95	0.36 0.05 (0.06)	0.26 0.01	0.51 0.04 (0.06)	0.41 0.00
SAD		3.44		34.39		85.45	
$T = 20$	β_0 (0.1)	0.09 0.11 (0.11)	0.00 0.95	0.08 0.17 (0.15)	0.02 0.96	0.09 0.07 (0.15)	0.01 0.65
	β_1 (−0.1)	−0.09 0.16 (0.16)	0.00 0.96	−0.10 0.24 (0.22)	0.00 0.97	−0.08 0.10 (0.21)	0.02 0.64
	β_2 (0.1)	0.11 0.15 (0.15)	0.00 0.96	0.32 0.21 (0.20)	0.22 0.84	0.09 0.10 (0.16)	0.01 0.77
	β_3 (0.1)	0.08 0.22 (0.21)	0.00 0.97	0.10 0.31 (0.28)	0.00 0.98	0.12 0.15 (0.23)	0.02 0.78
	α (0.7)	0.70 0.02 (0.02)	0.00 0.95	0.86 0.01 (0.01)	0.16 0.00		
	γ (0.7)	0.70 0.02 (0.02)	0.00 0.96			0.92 0.01 (0.01)	0.22 0.00
	λ (0.1)	0.10 0.03 (0.03)	0.00 0.95	0.39 0.03 (0.04)	0.29 0.00	0.58 0.03 (0.05)	0.48 0.00
SAD		5.19		141.55		236.69	

ARMA = autoregressive moving average; AR = autoregressive; MA = moving average; SE = standard error; SD = standard deviation; AB = absolute bias; Coverage = coverage probability; SAD = sum of absolute difference.

Table 3.2. Comparison of three models with ARMA, AR, and MA structured covariance matrix (Homogeneous, $n = 300$)

		ARMA		AR		MA	
		Mean SE (SD)	AB Coverage	Mean SE (SD)	AB Coverage	Mean SE (SD)	AB Coverage
$T = 5$	β_0 (0.1)	0.10 0.07 (0.07)	0.00 0.94	0.10 0.09 (0.08)	0.00 0.98	0.11 0.31 (0.40)	0.01 0.86
	β_1 (−0.1)	−0.10 0.09 (0.10)	0.00 0.94	−0.10 0.13 (0.12)	0.00 0.97	−0.08 0.44 (0.58)	0.02 0.85
	β_2 (0.1)	0.14 0.43 (0.45)	0.04 0.94	0.34 0.56 (0.59)	0.24 0.92	0.11 0.37 (0.45)	0.01 0.88
	β_3 (0.1)	0.08 0.61 (0.63)	0.02 0.94	0.12 0.79 (0.85)	0.02 0.93	0.12 0.53 (0.64)	0.02 0.88
	α (0.7)	0.70 0.03 (0.03)	0.00 0.95	0.98 0.02 (0.02)	0.28 0.00		
	γ (0.7)	0.70 0.03 (0.03)	0.00 0.94			1.03 0.02 (0.02)	0.33 0.00
	λ (0.1)	0.09 0.04 (0.04)	0.01 0.92	0.30 0.04 (0.05)	0.20 0.01	0.40 0.04 (0.05)	0.30 0.00
SAD		0.20		7.76		20.88	
$T = 10$	β_0 (0.1)	0.10 0.07 (0.06)	0.00 0.95	0.09 0.10 (0.08)	0.01 0.98	0.09 0.10 (0.19)	0.01 0.71
	β_1 (−0.1)	−0.10 0.09 (0.10)	0.00 0.95	−0.10 0.14 (0.12)	0.00 0.97	−0.08 0.14 (0.27)	0.02 0.73
	β_2 (0.1)	0.12 0.20 (0.21)	0.02 0.94	0.64 0.29 (0.27)	0.54 0.51	0.09 0.14 (0.22)	0.01 0.80
	β_3 (0.1)	0.09 0.29 (0.29)	0.01 0.95	0.10 0.41 (0.41)	0.00 0.94	0.13 0.20 (0.31)	0.03 0.80
	α (0.7)	0.70 0.02 (0.02)	0.00 0.95	0.88 0.01 (0.01)	0.18 0.00		
	γ (0.7)	0.70 0.02 (0.02)	0.00 0.96			0.95 0.01 (0.01)	0.25 0.00
	λ (0.1)	0.10 0.03 (0.03)	0.00 0.95	0.36 0.03 (0.04)	0.26 0.00	0.52 0.03 (0.04)	0.42 0.00
SAD		1.02		36.88		85.16	
$T = 20$	β_0 (0.1)	0.10 0.07 (0.06)	0.00 0.95	0.08 0.10 (0.09)	0.02 0.96	0.10 0.04 (0.09)	0.00 0.67
	β_1 (−0.1)	−0.10 0.09 (0.10)	0.00 0.93	−0.10 0.14 (0.12)	0.00 0.96	−0.10 0.06 (0.12)	0.00 0.66
	β_2 (0.1)	0.10 0.09 (0.09)	0.00 0.96	0.32 0.12 (0.12)	0.22 0.59	0.10 0.06 (0.10)	0.00 0.75
	β_3 (0.1)	0.10 0.13 (0.13)	0.00 0.96	0.10 0.18 (0.17)	0.00 0.97	0.10 0.09 (0.15)	0.00 0.77
	α (0.7)	0.70 0.01 (0.01)	0.00 0.98	0.86 0.01 (0.01)	0.16 0.00		
	γ (0.7)	0.70 0.01 (0.01)	0.00 0.94			0.92 0.01 (0.01)	0.22 0.00
	λ (0.1)	0.10 0.02 (0.02)	0.00 0.94	0.40 0.02 (0.02)	0.30 0.00	0.58 0.03 (0.03)	0.48 0.00
SAD		1.54		145.49		236.45	

ARMA = autoregressive moving average; AR = autoregressive; MA = moving average; SE = standard error; SD = standard deviation; AB = absolute bias; Coverage = coverage probability; SAD = sum of absolute difference.

Table 3.3. Comparison of three models with ARMA, AR, and MA structured covariance matrix (Homogeneous, $n = 500$)

		ARMA		AR		MA	
		Mean SE (SD)	AB Coverage	Mean SE (SD)	AB Coverage	Mean SE (SD)	AB Coverage
$T = 5$	β_0 (0.1)	0.10 0.05 (0.06)	0.00 0.94	0.10 0.07 (0.06)	0.00 0.96	0.11 0.24 (0.31)	0.01 0.87
	β_1 (−0.1)	−0.10 0.07 (0.08)	0.00 0.93	−0.10 0.10 (0.09)	0.00 0.96	−0.11 0.34 (0.45)	0.01 0.86
	β_2 (0.1)	0.10 0.34 (0.35)	0.00 0.94	0.35 0.43 (0.46)	0.25 0.90	0.12 0.29 (0.34)	0.02 0.89
	β_3 (0.1)	0.10 0.48 (0.47)	0.00 0.95	0.13 0.61 (0.63)	0.00 0.94	0.08 0.41 (0.50)	0.02 0.90
	α (0.7)	0.70 0.02 (0.02)	0.00 0.96	0.98 0.02 (0.02)	0.28 0.00		
	γ (0.7)	0.70 0.02 (0.02)	0.00 0.96			1.03 0.01 (0.01)	0.33 0.00
	λ (0.1)	0.10 $0.03_{0.03}$	0.00 0.94	0.30 $0.03_{0.04}$	0.20 0.00	0.40 $0.03_{0.04}$	0.30 0.00
SAD		0.30		7.76		20.81	
$T = 10$	β_0 (0.1)	0.10 0.05 (0.05)	0.00 0.94	0.09 0.08 (0.07)	0.01 0.97	0.11 0.08 (0.14)	0.01 0.71
	β_1 (−0.1)	−0.10 0.07 (0.07)	0.00 0.94	−0.10 0.11 (0.10)	0.00 0.96	−0.12 0.11 (0.21)	0.02 0.69
	β_2 (0.1)	0.10 0.16 (0.16)	0.00 0.94	0.62 0.22 (0.21)	0.52 0.34	0.10 0.11 (0.16)	0.00 0.80
	β_3 (0.1)	0.12 0.22 (0.22)	0.02 0.95	0.12 0.32 (0.32)	0.02 0.97	0.09 0.15 (0.24)	0.01 0.78
	α (0.7)	0.70 0.01 (0.01)	0.00 0.95	0.88 0.01 (0.01)	0.18 0.00		
	γ (0.7)	0.70 0.01 (0.01)	0.00 0.95			0.95 0.01 (0.01)	0.25 0.00
	λ (0.1)	0.10 0.02 (0.02)	0.00 0.94	0.36 0.02 (0.03)	0.26 0.00	0.52 0.02 (0.03)	0.42 0.00
SAD		0.16		37.28		85.11	
$T = 20$	β_0 (0.1)	0.10 0.05 (0.05)	0.00 0.96	0.08 0.08 (0.07)	0.02 0.97	0.10 0.03 (0.06)	0.00 0.69
	β_1 (−0.1)	−0.10 0.07 (0.07)	0.00 0.96	−0.11 0.11 (0.09)	0.01 0.98	−0.11 0.05 (0.09)	0.01 0.67
	β_2 (0.1)	0.10 0.07 (0.07)	0.00 0.96	0.31 0.10 (0.09)	0.21 0.41	0.11 0.05 (0.07)	0.01 0.79
	β_3 (0.1)	0.10 0.10 (0.09)	0.00 0.96	0.11 0.14 (0.12)	0.01 0.97	0.09 0.07 (0.11)	0.01 0.74
	α (0.7)	0.70 0.01 (0.01)	0.00 0.93	0.86 0.01 (0.00)	0.16 0.00		
	γ (0.7)	0.70 0.01 (0.01)	0.00 0.95			0.92 0.00!(0.00)	0.22 0.00
	λ (0.1)	0.10 0.01 (0.01)	0.00 0.96	0.40 0.02 (0.02)	0.30 0.00	0.58 0.01 (0.02)	0.48 0.00
SAD		1.68		147.01		236.15	

ARMA = autoregressive moving average; AR = autoregressive; MA = moving average; SE = standard error; SD = standard deviation; AB = absolute bias; Coverage = coverage probability; SAD = sum of absolute difference.

위한 실제 모형은 다음과 같다.

$$\phi_{itj} = \alpha \times I_{|t-j|=1}, \qquad l_{itj} = \gamma \times I_{|t-j|=1}, \qquad \log \sigma_{it}^2 = \lambda_0 + \lambda_1 \text{Group}_i,$$

여기서 $(\alpha, \gamma, \lambda_0, \lambda_1) = (0.7, 0.7, 0.1, 0.1)$이다.

Tables 3.4–3.6은 이 경우의 추정치들의 평균, 절대편향, 추정치들의 표준편차, 추정치의 표준오차들의 평균, 그리고 포함확률를 나타내고 있다. 공분산 행렬이 등분산성을 가지는 경우와 같이 ARMA모형이 표본수나 반복 횟수에 상관 없이 AR모형, MA모형과 비교할 때 편향이 가장 작은 값을 가지고, 포함확률은 전체적으로 ARMA모형이 다른 두 모형에 비해 더 좋음을 알 수 있다. 반복 횟수가 5, 10, 20으로 커짐에 따라 같은 경향을 보임을 알 수 있다.

그러나 표본수는 100, 300, 500으로 커져도 포함확률의 경향은 크게 차이가 나지 않음을 알 수 있다. 이는 일정 표본수 이상이 되면 공분산 행렬의 추정에 영향을 미치지 않는다고 볼 수 있다. 그러나 앞의 경우와 다르게 공분산 행렬이 이분산성을 가지는 경우에는 모든 경우에 있어 λ_0의 편향과 포함확률을 보면 ARMA모형과 비교했을 때 MA(1) 구조를 가진 MA모형뿐만 아니라 AR(1) 구조를 가진 AR모형에서도 잘 추정하지 못하는 것을 볼 수 있다. 이는 난수를 발생시킨 모형은 이분산성을 공분산을 가지는 모형이지만 우리가 적합을 위해서 가정한 모형들은 모두 등분산성을 가정한 모 형이기에 발생하는 현상이다.

SAD해석: 공분산 행렬이 이분산성을 가진 경우에도 역시 AR모형에서는 오차항과 관련된 γ값이 없고, MA모형에서는 α값이 없으므로 SAD를 이용하여 $\alpha, \gamma, \lambda_0, \lambda_1$을 비교한다. 표본수나 반복 횟수를 변화시킨 각각의 경우를 살펴보면 포함 확률을 비교했을 때와는 다르게 반복 횟수가 SAD값에 영향을 미치는 것을 볼 수 있다. 앞의 경우에서처럼 반복 횟수가 증가 함에 따라 모든 모형에서 SAD값이 커짐을 볼 수 있고, 표본수는 많은 영향을 미치지 않는 것으로 나타났다. 역시 일정 표본수 이상이면 행렬의 추정에 영향을 미치지 않는다는 결과이다. 공분산 행렬이 등분산성을 가지 경우처럼 표본수와 반복 횟수의 9가지 조합에 있어 SAD값이 ARMA모형에서 월등히 작게 나타났고, ARMA, AR, MA모형 순서로 SAD값이 작았다.

4. 결론

이 논문에서 경시적 연속형 자료분석을 위한 일반 선형모형을 고려하였다. 그리고 공분산행렬의 모수추정을 위한 여러가지 방법들을 고찰하였다. 자기회귀/이동평균/자기회귀-이동평균을 가지는 공분산행렬의 모형화를 위한 수정된 콜레스키분해 방법을 고찰하였고, 이 분해를 통하여 나오는 일반화 자기회귀/이동평균/자기회귀-이동평균 모수들을 선형회귀모형을 통하여 모형화하였다. 그리고 원하는 차수의 자기회귀/이동평균/자기회귀-이동평균의 구조를 가지는 공분산행렬을 추정하였다. 혁신분산에 로그선형모형을 적용하여 공분산행렬이 이공분산성을 가질수 있도록 추정할 수 있었다. 그리고 이를 통하여 추정된 공분산행렬이 항상 양정치성을 만족하게 하였고, 행렬의 모수의 개수도 줄일 수 있었다.

자기회귀구조를 가지는 공분산행렬의 추정을 위한 수정 콜레스키분해 방법은 많은 장점에도 불구하고 반복의 수가 많을 때에는 자기회귀의 차수를 높여야 한다. 이 경우 모수의 수가 같이 증가함을 알 수 있다. 이동평균 콜레스키분해 방법의 경우 이동평균의 특성상 지정된 차수 이상으로 시차가 벌어지면 자기공분산이 0이 됨을 알 수 있다. 따라서 이들 자기회귀와 이동평균의 특성을 결합한 자기회귀-이동평균 콜레스키분해를 고려하였다. 이 경우 수정 콜레스키 방법에서 차수가 큰 경우 이를 이용한 통계적 모형의 비효율성을 극복할 수 있었다.

모의실험을 통하여 세 가지 모형에서 공변량의 효과의 추정치들의 편향은 잘못된 공분산 행렬의 가정으

Table 3.4. Comparison of three models with ARMA, AR, and MA structured covariance matrix (Heterogeneous, $n = 100$)

		ARMA		AR		MA	
		Mean	AB	Mean	AB	Mean	AB
		SE (SD)	Coverage	SE (SD)	Coverage	SE (SD)	Coverage
$T = 5$	β_0	0.10	0.00	0.11	0.01	0.11	0.01
	(0.1)	0.11 (0.12)	0.94	0.16 (0.14)	0.96	0.53 (0.70)	0.85
	β_1	−0.10	0.00	−0.10	0.00	−0.14	0.04
	(−0.1)	0.17 (0.18)	0.93	0.24 (0.21)	0.97	0.77 (1.00)	0.88
	β_2	0.09	0.01	0.30	0.20	0.12	0.02
	(0.1)	0.74 (0.75)	0.95	0.95 (1.02)	0.93	0.64 (0.78)	0.89
	β_3	0.04	0.06	0.25	0.15	0.04	0.06
	(0.1)	1.09 (1.06)	0.96	1.39 (1.51)	0.93	0.94 (1.12)	0.91
	α	0.70	0.00	0.98	0.28		
	(0.7)	0.05 (0.05)	0.95	0.04 (0.04)	0.00		
	γ	0.70	0.00			1.03	0.33
	(0.7)	0.05 (0.04)	0.95			0.03 (0.03)	0.00
	λ_0	0.09	0.01	0.28	0.18	0.38	0.28
	(0.1)	0.09 (0.09)	0.94	0.10 (0.11)	0.57	0.09 (0.12)	0.18
	λ_1	0.10	0.00	0.10	0.00	0.10	0.00
	(0.1)	0.13 (0.12)	0.95	0.15 (0.16)	0.92	0.13 (0.16)	0.87
SAD		0.79		7.52		21.07	
$T = 10$	β_0	0.10	0.00	0.09	0.01	0.10	0.00
	(0.1)	0.11 (0.11)	0.95	0.17 (0.14)	0.97	0.17 (0.33)	0.72
	β_1	−0.09	0.01	−0.10	0.00	−0.08	0.02
	(−0.1)	0.17 (0.17)	0.94	0.25 (0.21)	0.97	0.26 (0.47)	0.73
	β_2	0.08	0.02	0.64	0.54	0.09	0.01
	(0.1)	0.35 (0.35)	0.94	0.49 (0.47)	0.82	0.24 (0.37)	0.80
	β_3	0.12	0.02	0.13	0.03	0.13	0.03
	(0.1)	0.51 (0.52)	0.94	0.72 (0.72)	0.95	0.35 (0.53)	0.81
	α	0.69	0.01	0.88	0.18		
	(0.7)	0.03 (0.03)	0.93	0.02 (0.02)	0.00		
	γ	0.70	0.00			0.95	0.25
	(0.7)	0.03 (0.03)	0.95			0.02 (0.02)	0.00
	λ_0	0.09	0.01	0.35	0.25	0.51	0.41
	(0.1)	0.06 (0.06)	0.95	0.07 (0.08)	0.08	0.06 (0.09)	0.00
	λ_1	0.10	0.00	0.10	0.00	0.10	0.00
	(0.1)	0.09 (0.09)	0.95	0.10 (0.12)	0.93	0.09 (0.13)	0.84
SAD		3.93		34.89		85.54	
$T = 20$	β_0	0.11	0.01	0.07	0.03	0.11	0.01
	(0.1)	0.11 (0.11)	0.95	0.17 (0.14)	0.97	0.07 (0.14)	0.68
	β_1	−0.10	0.00	−0.08	0.02	−0.10	0.00
	(−0.1)	0.16 (0.17)	0.95	0.25 (0.21)	0.99	0.11 (0.19)	0.71
	β_2	0.10	0.00	0.33	0.23	0.11	0.01
	(0.1)	0.15 (0.15)	0.95	0.21 (0.19)	0.84	0.10 (0.17)	0.76
	β_3	0.10	0.00	0.11	0.01	0.10	0.00
	(0.1)	0.22 (0.22)	0.96	0.31 (0.28)	0.98	0.15 (0.26)	0.76
	α	0.70	0.00	0.86	0.16		
	(0.7)	0.02 (0.02)	0.94	0.01 (0.01)	0.00		
	γ	0.70	0.00			0.92	0.22
	(0.7)	0.02 (0.02)	0.94			0.01 (0.01)	0.00
	λ_0	0.10	0.00	0.39	0.29	0.58	0.48
	(0.1)	0.04 (0.04)	0.96	0.05 (0.06)	0.00	0.04 (0.07)	0.00
	λ_1	0.10	0.00	0.10	0.00	0.10	0.00
	(0.1)	0.06 (0.06)	0.95	0.07 (0.08)	0.91	0.06 (0.10)	0.80
SAD		6.19		140.19		236.75	

ARMA = autoregressive moving average; AR = autoregressive; MA = moving average; SE = standard error; SD = standard deviation; AB = absolute bias; Coverage = coverage probability; SAD = sum of absolute difference.

Table 3.5. Comparison of three models with ARMA, AR, and MA structured covariance matrix (Heterogeneous, $n = 300$)

		ARMA		AR		MA	
		Mean	AB	Mean	AB	Mean	AB
		SE (SD)	Coverage	SE (SD)	Coverage	SE (SD)	Coverage
$T = 5$	β_0	0.10	0.00	0.10	0.00	0.08	0.02
	(0.1)	0.07 (0.07)	0.94	0.09 (0.08)	0.97	0.31 (0.41)	0.88
	β_1	−0.10	0.00	−0.10	0.00	−0.09	0.01
	(−0.1)	0.10 (0.10)	0.96	0.14 (0.13)	0.97	0.45 (0.61)	0.85
	β_2	0.09	0.01	0.38	0.28	0.08	0.02
	(0.1)	0.43 (0.44)	0.93	0.56 (0.57)	0.91	0.37 (0.46)	0.90
	β_3	0.15	0.05	0.12	0.02	0.12	0.02
	(0.1)	0.63 (0.64)	0.95	0.81 (0.82)	0.94	0.54(0.67)	0.89
	α	0.70	0.00	0.98	0.28		
	(0.7)	0.03 (0.03)	0.95	0.02 (0.02)	0.00		
	γ	0.70	0.00			1.03	0.33
	(0.7)	0.03 (0.03)	0.95			0.02 (0.02)	0.00
	λ_0	0.10	0.00	0.29	0.19	0.40	0.30
	(0.1)	0.05 (0.05)	0.95	0.06 (0.07)	0.13	0.05 (0.07)	0.00
	λ_1	0.10	0.00	0.10	0.00	0.09	0.01
	(0.1)	0.07 (0.07)	0.94	0.08 (0.10)	0.91	0.07 (0.09)	0.89
SAD		0.21		7.52		21.07	
$T = 10$	β_0	0.10	0.00	0.09	0.01	0.11	0.01
	(0.1)	0.07 (0.07)	0.94	0.10 (0.09)	0.96	0.10 (0.18)	0.72
	β_1	−0.10	0.00	−0.10	0.00	−0.11	0.01
	(−0.1)	0.10 (0.10)	0.94	0.14 (0.13)	0.96	0.15 (0.27)	0.73
	β_2	0.09	0.01	0.63	0.53	0.11	0.01
	(0.1)	0.20 (0.20)	0.95	0.29 (0.28)	0.55	0.14 (0.20)	0.82
	β_3	0.11	0.01	0.17	0.07	0.10	0.00
	(0.1)	0.29 (0.29)	0.96	0.42 (0.41)	0.95	0.20 (0.30)	0.82
	α	0.70	0.00	0.88	0.18		
	(0.7)	0.02 (0.02)	0.96	0.01 (0.01)	0.00		
	γ	0.70	0.00			0.95	0.25
	(0.7)	0.02 (0.02)	0.94			0.01 (0.01)	0.00
	λ_0	0.10	0.00	0.36	0.26	0.52	0.42
	(0.1)	0.04 (0.04)	0.95	0.04 (0.05)	0.00	0.04 (0.05)	0.00
	λ_1	0.10	0.00	0.10	0.00	0.10	0.00
	(0.1)	0.05 (0.05)	0.94	0.06 (0.07)	0.91	0.05 (0.07)	0.83
SAD		0.86		36.18		85.27	
$T = 20$	β_0	0.10	0.00	0.07	0.03	0.10	0.00
	(0.1)	0.07 (0.06)	0.95	0.10 (0.09)	0.96	0.04 (0.09)	0.68
	β_1	−0.10	0.00	−0.10	0.00	−0.10	0.00
	(−0.1)	0.10 (0.09)	0.95	0.14 (0.12)	0.98	0.06 (0.13)	0.63
	β_2	0.10	0.00	0.32	0.22	0.10	0.00
	(0.1)	0.09 (0.09)	0.96	0.12 (0.11)	0.60	0.06 (0.10)	0.77
	β_3	0.10	0.00	0.11	0.01	0.10	0.00
	(0.1)	0.13 (0.13)	0.95	0.18 (0.17)	0.97	0.09 (0.14)	0.77
	α	0.70	0.00	0.86	0.16		
	(0.7)	0.01 (0.01)	0.95	0.01 (0.01)	0.00		
	γ	0.70	0.00			0.92	0.22
	(0.7)	0.01 (0.01)	0.94			0.01 (0.01)	0.00
	λ_0	0.10	0.00	0.40	0.30	0.58	0.48
	(0.1)	0.03 (0.03)	0.95	0.03 (0.03)	0.00	0.03 (0.04)	0.00
	λ_1	0.10	0.00	0.10	0.00	0.10	0.00
	(0.1)	0.04 (0.04)	0.95	0.04 (0.05)	0.90	0.04 (0.05)	0.80
SAD		0.86		143.99		236.36	

ARMA = autoregressive moving average; AR = autoregressive; MA = moving average; SE = standard error; SD = standard deviation; AB = absolute bias; Coverage = coverage probability; SAD = sum of absolute difference.

Table 3.6. Comparison of three models with ARMA, AR, and MA structured covariance matrix (Heterogeneous, $n = 500$)

		ARMA		AR		MA	
		Mean	AB	Mean	AB	Mean	AB
		SE (SD)	Coverage	SE (SD)	Coverage	SE (SD)	Coverage
$T = 5$	β_0	0.10	0.00	0.10	0.00	0.11	0.01
	(0.1)	0.05 (0.05)	0.95	0.07 (0.07)	0.97	0.24 (0.32)	0.85
	β_1	−0.11	0.01	−0.10	0.00	−0.11	0.01
	(−0.1)	0.08 (0.07)	0.95	0.11 (0.10)	0.98	0.35 (0.45)	0.87
	β_2	0.10	0.00	0.38	0.28	0.11	0.01
	(0.1)	0.34 (0.34)	0.94	0.43 (0.48)	0.88	0.29 (0.36)	0.89
	β_3	0.10	0.00	0.12	0.02	0.09	0.01
	(0.1)	0.49 (0.50)	0.95	0.63 (0.71)	0.91	0.42 (0.50)	0.91
	α	0.70	0.00	0.98	0.28		
	(0.7)	0.02 (0.02)	0.93	0.02 (0.02)	0.00		
	γ	0.70	0.00			1.03	0.33
	(0.7)	0.02 (0.02)	0.94			0.01 (0.01)	0.00
	λ_0	0.09	0.01	0.30	0.20	0.40	0.30
	(0.1)	0.04 (0.04)	0.93	0.05 (0.05)	0.01	0.04 (0.05)	0.00
	λ_1	0.10	0.00	0.10	0.00	0.10	0.00
	(0.1)	0.06 (0.06)	0.96	0.07 (0.07)	0.94	0.06 (0.07)	0.86
SAD		0.26		7.91		20.82	
$T = 10$	β_0	0.10	0.00	0.09	0.01	0.10	0.00
	(0.1)	0.05 (0.05)	0.95	0.08 (0.07)	0.96	0.08 (0.14)	0.73
	β_1	−0.11	0.01	−0.10	0.00	−0.11	0.01
	(−0.1)	0.08 (0.07)	0.95	0.11 (0.10)	0.97	0.11 (0.21)	0.72
	β_2	0.10	0.00	0.62	0.52	0.11	0.01
	(0.1)	0.34 (0.34)	0.94	0.22 (0.21)	0.36	0.11 (0.16)	0.82
	β_3	0.10	0.00	0.15	0.05	0.09	0.01
	(0.1)	0.49 (0.50)	0.95	0.32 (0.31)	0.96	0.16 (0.23)	0.82
	α	0.70	0.00	0.88	0.18		
	(0.7)	0.02 (0.02)	0.93	0.01 (0.01)	0.00		
	γ	0.70	0.00			0.95	0.25
	(0.7)	0.02 (0.02)	0.94			0.01 (0.01)	0.00
	λ_0	0.09	0.01	0.36	0.26	0.52	0.42
	(0.1)	0.04 (0.04)	0.93	0.03 (0.04)	0.00	0.03 (0.04)	0.00
	λ_1	0.10	0.00	0.10	0.00	0.10	0.00
	(0.1)	0.06 (.06)	0.96	0.05 (0.05)	0.91	0.04 (0.06)	0.83
SAD		1.1		37.1		85.19	
$T = 20$	β_0	0.10	0.00	0.08	0.02	0.10	0.00
	(0.1)	0.05 (0.05)	0.95	0.08 (0.07)	0.96	0.03 (0.07)	0.66
	β_1	−0.10	0.00	−0.11	0.01	−0.10	0.00
	(−0.1)	0.07 (0.08)	0.94	0.11 (0.09)	0.98	0.05 (0.10)	0.67
	β_2	0.10	0.00	0.32	0.22	0.10	0.00
	(0.1)	0.07 (0.07)	0.95	0.10 (0.09)	0.41	0.05 (0.08)	0.75
	β_3	0.11	0.01	0.12	0.02	0.10	0.00
	(0.1)	0.10 (0.11)	0.93	0.14 (0.13)	0.97	0.07 (0.11)	0.78
	α	0.70	0.00	0.86	0.16		
	(0.7)	0.01 (0.01)	0.95	0.01 (0.00)	0.00		
	γ	0.70	0.00			0.92	0.22
	(0.7)	0.01 (0.01)	0.97			0.00 (0.00)	0.00
	λ_0	0.10	0.00	0.40	0.30	0.58	0.48
	(0.1)	0.02 (0.02)	0.95	0.02 (0.02)	0.00	0.02 (0.03)	0.00
	λ_1	0.10	0.00	0.10	0.00	0.10	0.00
	(0.1)	0.03 (0.03)	0.96	0.03 (0.04)	0.89	0.03 (0.04)	0.80
SAD		1.80		117.05		236.26	

ARMA = autoregressive moving average; AR = autoregressive; MA = moving average; SE = standard error; SD = standard deviation; AB = absolute bias; Coverage = coverage probability; SAD = sum of absolute difference.

로 인하여 영향을 받는다는 것을 알 수 있었다. 그리고 반복 횟수가 커질 때 편향과 포함 확률은 거의 달라지지 않았으나 추정된 공분산행렬의 편향이 잘못된 공분산행렬을 가지는 모형에서 일정하게 증가함을 볼 수 있다. 결론적으로 공분산 행렬이 등분산성/이분산성을 가질 때에 올바른 공분산 행렬의 가정이 필요함을 알 수가 있다.

References

Daniels, J. M. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data, *Biometrika*, **89**, 553–566.

Daniels, J. M. and Zhao, Y. D. (2003). Modelling the random effects covariance matrix in longitudinal data, *Statistics in Medicine*, **22**, 1631–1647.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data* (2nd Ed), Oxford University Press, Oxford.

Kim, J. and Lee, K. (2015). Survey of models for random effects covariance matrix in generalized linear mixed model, *The Korean Journal of Applied Statistics*, **28**, 211–219.

Kim, J., Sohn, I., and Lee, K. (2016). Bayesian modeling of random effects precision/covariance matrix in cumulative logit random effects models, *Communications for Statistical Applications and Methods*, **24**, 81–96.

Lee, K. (2013). Bayesian modeling of random effects covariance matrix for generalized linear mixed models, *Communications for Statistical Applications and Methods*, **20**, 235–240.

Lee, K., Baek, C., and Daniels, M. J. (2017). ARMA Cholesky factor models for the covariance matrix of linear models, *Computational Statistics & Data Analysis*, working paper.

Lee, K. and Sung, S. A. (2014). Autoregressive Cholesky factor modeling for marginalized random effects models, *Communications for Statistical Applications and Methods*, **21**, 169–181.

Lee, K. and Yoo, J. (2014). Bayesian Cholesky factor models in random effects covariance matrix for generalized linear mixed models, *Computational Statistics & Data Analysis*, **80**, 111–116.

Lee, K., Yoo, J. K., Lee, J., and Hagan, J. (2012). Modeling the random effects covariance matrix for the generalized linear mixed models, *Computational Statistics & Data Analysis*, **56**, 1545–1551.

Pan, J. X. and Mackenzie, G. (2003). Model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239–244.

Pan, J. X. and MacKenzie, G. (2006). Regression models for covariance structures in longitudinal studies. *Statistical Modelling*, **6**, 43–57.

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

Pourahmadi, M. (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix, *Biometrika*, **87**, 425–435.

Zhang, W. and Leng, C. (2012). A moving average Cholesky factor model in covariance modeling for longitudinal data, *Biometrika*, **99**, 141–150.

일반 선형 모형에 대한 공분산 행렬의 비교

남상아[a] · 이근백[a,1]

[a]성균관대학교 통계학과

(2016년 10월 18일 접수, 2016년 12월 18일 수정, 2016년 12월 29일 채택)

요 약

경시적 자료분석에서 공변량 효과를 추정할 때 반복 측정된 결과들의 상관성은 고려되어야 한다. 따라서 공분산 행렬을 모형화하는 것은 매우 중요하다. 그러나 공분산 행렬의 추정은 모수들의 수가 많고 추정된 공분산행렬이 양정치성을 만족해야 하므로 쉽지 않은 문제이다. 이러한 제한을 극복하기 위해, 공분산행렬의 모형화를 위한 여러가지 방법을 제안하였다: 자기회귀/이동평균/자기회귀-이동평균 구조를 각각 적용한 수정 콜레스키분해 (Pourahmadi, 1999), 이동평균 콜레스키분해 (Zhang과 Leng, 2012)와 자기회귀-이동평균 콜레스키 분해 (Lee 등, 2017) 이들 구조를 가지는 공분산 행렬의 특징을 비교연구하고자 한다. 이 세 가지 모형의 성능을 비교하기 위한 모의실험을 실시한다.

주요용어: 경시적 자료분석, 수정된 단냠레스키 분해, 이동평균 콜레스키 분해, 일반 선형모형

이 연구는 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2014R1A1A2054997, NRF-2016R1D1A1B03930343).

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