The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.211
28(2), 211–219

Survey of Models for Random Effects Covariance Matrix in Generalized Linear Mixed Model

Jiyeong Kim[a] · Keunbaik Lee[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received March 13, 2015; Revised March 30, 2015; Accepted March 30, 2015)

Abstract

Generalized linear mixed models are used to analyze longitudinal categorical data. Random effects specify the serial dependence of repeated outcomes in these models; however, the estimation of a random effects covariance matrix is challenging because of many parameters in the matrix and the estimated covariance matrix should satisfy positive definiteness. Several approaches to model the random effects covariance matrix are proposed to overcome these restrictions: modified Cholesky decomposition, moving average Cholesky decomposition, and partial autocorrelation approaches. We review several approaches and present potential future work.

Keywords: Longitudinal data, categorical data, modified Cholesky decomposition, moving average Cholesky decomposition, partial autocorrelation matrix.

1. 서론

경시적 자료(longitudinal data)는 같은 개체(subject)에서 반복 측정된 자료를 말한다. 이 경우 같은 개체에서 나온 측정치들은 서로 상관관계(correlation)를 가질 수 있고, 이러한 특성은 경시적 자료 분석 시에 고려되어야 한다 (Diggle 등, 2002). 따라서 경시적 자료분석을 위한 모형들은 이러한 반복 측정치들의 관련성을 설명하기 위한 모형화에 집중하고 있다. 이 논문에서는 특히 경시적 범주형 자료에 초점을 맞추도록 한다. 경시적 범주형자료(longitudinal categorical data) 분석을 위한 선행연구들을 보면 크게 두 가지로 분류된다. 모집단-평균효과(population-average effects)와 개체-특정적 효과(subject-specific effects)에 관심을 두는 분석으로 분류된다 (Agresti, 2013). 모집단-평균효과에 관심이 있을 경우 일반적으로 일반화추정방정식(generalized estimation equation; GEE, Liang와 Zeger, 1986)을 이용한 모형과 주변화모형 (Heagerty, 1999, 2002)이 많이 사용되고 있다. 일반화 추정방정식의 경우 반복 측정치들의 상관관계를 설명하기 위한 가상 상관계수행렬(working correlation matrix)을 이용하고, 주변화모형의 경우 앞의 시점의 결과치를 이용하는 마코프(Markov)구조와 임의효과(random effects)를 이용한다. 개체-특정적 효과에 관심이 있는 경우 일반화 선형혼합모형(generalized linear mixed models; GLMM)을 주로 사용한다 (Breslow와 Clayton, 1993). 이 모형의 경우 임의효과를

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-2014R1A1A2054997).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-Gu, Seoul 110-745, Korea. E-mail: keunbaik@skku.edu

이용하여 측정치들의 상관관계를 설명하고, 그것의 공분산행렬은 개체 내 변동을 또한 설명한다. 이 논문에서 우리는 일반화 선형혼합모형에 초점을 맞추고자 한다.

일반화 선형혼합모형에서 임의효과의 분포는 주로 다변량 정규분포(multivariate normal distribution)를 가정한다. 이 분포의 공분산행렬(covariance matrix)은 반복 측정된 결과치들의 상관관계를 설명하므로 시간에 따른 변동과 개체들 간의 변동을 동시에 설명하게 된다. 이 공분산행렬을 추정할 때에 그 추정치의 양정치성(positive definite)을 만족해야 하며, 그리고 그 추정 공분산행렬이 보통 고차원(high dimension)이므로 모수의 개수도 많다. 이러한 제약들은 또한 상관계수행렬(correlation matrix)의 추정에도 마찬가지로 존재하는 문제이고, 부가적으로 상관계수행렬의 경우 주대각요소(diagonal element)들이 모두 일이어야 하는 제약조건이 더 필요하다. 이러한 제약조건을 만족하는 공분산/상관계수행렬의 추정은 위에서 제시된 제약조건들 때문에 쉽지 않은 문제이다 (Lee 등, 2012). 일반화 선형 혼합모형에서는 일반적으로 이러한 제약조건을 피하기 위해서 공분산/상관계수행렬을 단순한 AR(1)과 같은 구조를 가정하였고, 그리고 동질적 공분산(homogeneous covariance)을 가정하였다. 하지만 이러한 가정은 너무 강하고 이로 인해 편의(bias)가 발생할 수 있다 (Heagerty와 Kurland, 2001). 따라서 위의 공분산/상관계수행렬의 추정시의 제약조건을 만족하면서 일반적인 구조인 AR 또는 MA구조를 가지는 이질적(heterogeneous) 공분산/상관계수행렬의 추정방법들이 제안되었다 (Pourahmadi, 1999, 2000; Daniels와 Pourahmadi, 2002; Daniels와 Zhao, 2003; Pan와 Mackenzie, 2003, 2006; Lee 등, 2012; Zhang와 Leng, 2012; Lee와 Yoo, 2014). 이 논문에서 우리는 일반화 선형혼합모형의 임의효과 공분산/상관행렬의 추정을 위한 모형화를 살펴본다.

본 논문의 구성은 다음과 같다. 2장에서는 일반화 선형혼합모형에 대해 소개하고, 3장에서는 임의효과 공분산행렬의 모형화를 제시한다. 여기서 자기회귀와 이동평균을 각각을 포함한 수정한 콜레스키 분해와 부분자기상관계수 행렬를 이용한 상관행렬에 대한 다양한 모형화에 대해 소개한다. 마지막으로 4장에서는 결론을 제시한다.

2. 일반화 선형혼합모형

이 절에서는 일반화 선형혼합모형을 우선 설명한다. 여기서 $Y_{i,t}$는 시간 $t\,(t = 1, \ldots, n_i)$에서의 개체 $i\,(i = 1, \ldots, N)$의 응답변수(response variables)라고 하고, $x_{i,t}$는 $Y_{i,t}$에 상응하는 $p \times 1$ 공변량 벡터(covariate vector)이다. $Y_{i,t}$는 임의효과 $b_{i,t}$가 주어졌을 때, 조건부 독립(conditional independence)을 가정하고, 응답변수의 조건부 분포는 지수산포족(exponential dispersion family) 형태를 따른다고 가정한다.

$$f(y_i; b_{i,t}, \theta_i, \phi) = \exp\left[\frac{\{y_i\theta_i(b_{i,t}) - \psi(\theta_i(b_{i,t}))\}}{a(\phi)} + c(y_i; \phi)\right],$$

여기서 $a(\cdot)$, $\psi(\cdot)$, 그리고 $c(\cdot)$는 알려진 함수이고, $\theta_i(b_{it})$와 ϕ는 임의효과가 주어진 자연모수(natural parameter)와 산포모수(dispersion parameter)이다. 일반화 선형모형(generalized linear model)과 같이 체계적 성분(systematic component)을 가지므로 아래와 같은 관계를 가진다.

$$\eta_{i,t}(b_{i,t}) = x_{i,t,j}^T\beta + b_{i,t}, \tag{2.1}$$

여기서 $\mu_i(b_{i,t}) = E(Y_i; b_{i,t})$, β는 $p \times 1$ 미지 회귀계수(unknown regression coefficient)이다. 연결함수(link function), $g(\cdot)$는 선형 가역성을 만족하므로 반응변수의 조건부 평균과 공변량벡터는 다음과 같은 관계를 만족한다.

$$\eta_{i,t}(b_{i,t}) = g(\mu_{i,t}(b_{i,t})), \tag{2.2}$$

따라서 식 (2.1)과 (2.2)로부터 다음의 모형을 가진다.

$$g(\mu_{i,t}(b_{i,t})) = x_{i,t,j}^T\beta + b_{i,t},$$

여기서 임의효과는 시간에 따른 변동과 개체들 간의 변동을 동시에 설명하게 되는데 그 분포는 다변량 정규분포를 가정한다.

$$b_i = (b_{i,1}, \ldots, b_{i,n_i})^T \sim N(0, \Sigma_i). \tag{2.3}$$

다음 절에서는 임의효과 공변량행렬, Σ_i의 이공분산성과 다양한 구조에 대한 모형들에 대해 알아보도록 한다.

3. 임의효과 공분산 및 상관계수 행렬의 모형화

임의효과 공분산행렬은 항상 양정치성을 만족해야 하고, 고차원이기에 추정하는데 어려움이 있다. 이 절에서는 어려움을 해소하고 시간에 따른 변동과 개체들 간의 변동을 잘 설명해주는 모형들 중에 수정 콜레스키분해, 이동평균 콜레스키분해, 그리고 부분 자기상관 방법을 이용하는 모형들을 살펴본다.

3.1. 공분산 행렬의 자기회귀를 포함한 수정 콜레스키 분해

일반화 선형혼합모형의 임의효과 공분산행렬에 자기회귀를 포함한 수정 콜레스키분해(modified Cholesky decomposition)를 제시한 Lee 등 (2012)의 모형을 살펴본다. 임의효과의 다변량 정규분포 (2.3)을 다음과 같이 분해한다.

$$b_{i,1} = e_{i,1}, \tag{3.1}$$

$$b_{i,t} = \sum_{j=1}^{t-1} \phi_{i,t,j} b_{i,j} + e_{i,t}, \quad \text{for } t = 2, \ldots, n_i, \tag{3.2}$$

여기서 $e_i = (e_{i,1}, \ldots, e_{i,n_i})^T \overset{indep}{\sim} N(0, D_i)$를 가정하고, $D_i = \mathrm{diag}(\sigma_{i,1}^2, \ldots, \sigma_{i,n_i}^2)$. ϕ는 일반화 자기회귀모수(generalized autoregressive parameters; GARP)라고 부르며, ϕ의 값은 앞의 결과가 현재의 결과에 영향을 미치는 자기회귀 구조를 가지게 된다. σ_{it}^2는 혁신분산(innovation variances; IV)라 불린다. 위의 식 (3.1)과 (3.2)을 행렬 형식으로 다시 작성하면 다음과 같다.

$$T_i b_i = e_i, \tag{3.3}$$

여기서

$$T_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_{i,2,1} & 1 & 0 & \cdots & 0 \\ -\phi_{i,3,1} & -\phi_{i,3,2} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\phi_{i,n_i,1} & -\phi_{i,n_i,2} & -\phi_{i,n_i,3} & \cdots & 1 \end{pmatrix}$$

이다. 식 (3.3)에 분산을 취하면 다음과 같은 식을 얻는다.

$$T_i \Sigma_i T_i^T = D_i \Leftrightarrow \Sigma^{-1} = T_i^T D_i^{-1} T_i. \tag{3.4}$$

Σ_i의 모수인 ϕ_i와 σ_i의 개수는 개체의 반복수, n_i가 증가하면 가하급수적으로 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 로그선형모형(log linear model)을 이용한다.

$$\begin{aligned} \phi_{i,t,j} &= w_{i,t,j}^T \gamma, \\ \log \sigma_{i,t}^2 &= h_{i,t}^T \lambda. \end{aligned} \tag{3.5}$$

γ과 λ는 $a \times 1$ 그리고 $b \times 1$의 모르는 모수벡터이다 (Pourahmadi, 1999, 2000; Pourahmadi와 Daniels, 2002; Daniels와 Zhao, 2003; Lee 등, 2012; Lee, 2013). 여기서 $w_{i,t,j}$의 선택을 통하여 공분산행렬을 AR(p)의 구조를 만들 수 있다. 그리고 $h_{i,t}$의 선택을 통하여 이공분산성을 만족하게 할 수 있다. 예를 들면 $w_{i,t,j} = I_{|t-j|=1}$이라고 가정하자. 여기서 $I_{|t-j|=1}$는 시차가 1인 경우에만 1이고 나머지는 모두 0인 함수이다. 그러면 $\phi_{i,t,j}$는 t와 j의 차가 1인 경우에만 γ가 되고, 나머지는 모두 0이 된다. 따라서 AR(1)의 구조를 가지게 된다. 그리고 $h_{i,t} = (1, \text{SEX}_i)$이라고 가정하자. 여기서 SEX_i는 i번째 개체의 성별을 나타내고 남자일 때 1이고 여자일 때 0으로 가정한다. 그러면 $\log \sigma_{i,t}^2$는 성별에 따라 다른 값을 가지게 된다. 따라서 전체 공분산행렬은 성별에 따라 달라지게 된다. 따라서 모수들이 어떠한 제약이 없기 때문에 공변량($w_{i,t,j}$, $h_{i,t}$)의 선택으로 임의효과 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).

수정 콜레스키분해는 처음에는 경시적 연속형 자료분석에서 공분산행렬의 모형화를 위하여 처음 제안되었다 (Pourahmadi, 1999, 2000). 그리고 식 (3.5)에서 제시된 일반화 선형모형을 이용하여 양정치성을 보였고, 그리고 그것의 점근적 정규성(asymptotic normality)과 직교성(orthogonality)을 증명하였다 (Pourahmadi, 2007). 이러한 특성들은 일반화 선형혼합모형의 임의효과 공분산행렬의 모형화에도 적용되었고, 모수의 추정은 베이지안 방법을 이용한 MCMC로 실현되었다 (Lee 등, 2012).

3.2. 공분산 행렬의 이동평균 콜레스키 분해

일반적으로 앞의 수정 콜레스키분해를 이용한 자기회귀 모형의 경우 반복 수가 증가함에 따라 그 차수를 높여야 함을 알 수 있다. 이 경우 모수 γ의 수가 차수와 비례해서 함께 증가함을 알 수 있다. 그리고 수정 콜레스키방법에 의한 공분산행렬의 추정은 식 (3.4)로 부터 공분산행렬의 역행렬을 분해해서 모형화 하는 경향이 있다. 이 절에서는 공분산행렬을 직접 분해하는 방법을 제시한다. Zhang과 Leng (2012)은 임의효과 공분산행렬의 이동평균 콜레스키 분해(moving average Cholesky decomposition)를 제안하였다. 우리는 여기서 일반화 선형혼합모형의 임의효과 공분산행렬에 이동평균 콜레스키 분해를 적용한 Lee와 Yoo (2014)의 모형을 제시한다. 임의효과의 분포는 앞의 수정 콜레스키 분해에서와 같이 $b_i \sim N(0, \Sigma_i)$를 가정한다. 그리고 b_i를 아래와 같이 분해한다.

$$b_{i,1} = e_{i,1}, \tag{3.6}$$

$$b_{i,t} = \sum_{j=1}^{t-1} l_{i,t,j} e_{i,j} + e_{i,t}, \quad \text{for } t = 2, \ldots, n_i, \tag{3.7}$$

여기서 $e_i = (e_{i,1}, \ldots, e_{i,n_i})^T$는 수정 콜레스키 분해와 같이 $e_i \overset{indep}{\sim} N(0, D_i)$를 가정한다. 식 (3.6)과 (3.7)을 행렬 형식으로 표현하면 다음과 같다.

$$b_i = L_i e_i,$$

여기서

$$L_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{i,2,1} & 1 & 0 & \cdots & 0 \\ l_{i,3,1} & l_{i,3,2} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{i,n_i,1} & l_{i,n_i,2} & l_{i,n_i,3} & \cdots & 1 \end{pmatrix}.$$

l_{itj}는 일반화 이동평균모수(generalized moving average parameters; GMAPs)이고, 수정 콜레스키 분해와 같이 혁신분산이 양수이면 Σ_i는 양정치성을 만족한다. 임의효과 공분산행렬 Σ_i은

$$\Sigma_i = L_i D_i L_i^T. \tag{3.8}$$

만약 $L_i = T_i^{-1}$로 설정하면 식 (3.8)과 식 (3.4)는 동일하게 된다.

일반화 이동평균모수와 혁신분산은 공변량 $z_{i,t,j}$와 $h_{i,t}$을 이용하여 다음과 같이 모형화할 수 있다.

$$l_{i,t,j} = z_{i,t,j}^T \gamma, \qquad \log \sigma_{i,t}^2 = h_{i,t}^T \lambda,$$

γ와 λ는 $a \times 1$와 $b \times 1$의 각각의 모르는 모수벡터이다. 여기서 수정 콜레스키분해 방법에서 처럼 벡터 $z_{i,t,j}$와 $h_{i,t}$는 개체-특정적 공변량이고, 이를 이용하여 임의효과 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 임의효과 공분산행렬은 항상 양정치성을 만족한다.

Zhang과 Leng (2012)는 경시적 연속형자료 분석을 위한 이동평균 콜레스키 분해에서 모수들의 직교성과 점근적 정규성을 보였다. Lee와 Yoo (2014)는 여기서 제시된 모형을 제안했고, 베이지안 방법을 이용하여 모수들을 추정하였다.

3.3. 부분자기 상관계수행렬을 이용한 상관행렬

앞의 수정/이동평균 콜레스키 분해는 임의효과 공분산행렬의 모형화를 위하여 제안되었다. 하지만 이 방법은 상관계수행렬에 그대로 사용할 수는 없다. 왜냐하면 상관계수행렬의 경우 주대각의 요소들이 1이어야 한다는 제약조건 때문이다. 따라서 앞에서 공분산행렬의 모형화에서의 제약조건과 상관계수행렬의 제약조건을 동시에 만족하면서 추정하는 방법으로 부분 자기상관계수를 이용한 모형화가 제안되었다 (Daniels와 Pourahmadi, 2009). 이 방법은 Lee 등 (2013)에 의해서 주변화모형으로 확장되었다. 여기서 그 방법을 좀 더 상세하게 소개하겠다.

2절에서 제시된 임의효과의 분포를 다변량 정규분포로 가정하며, 그 공분산행렬을 다음 같이 분해한다.

$$\Sigma_i = C_i^{\frac{1}{2}} R_i C_i^{\frac{1}{2}},$$

여기서 $C_i^{1/2}$는 주대각행렬로 주대각 요소는 임의효과 b_{it}의 표준편차를 요소로 한다. 따라서 $C_i^{1/2} = \text{diag}\{\sigma_{i,1,1}, \ldots, \sigma_{i,n_i,n_i}\}$이며 $\sigma_{i,t,t}^2 = \text{Var}(b_{i,t})$이다. 그리고 R_i는 임의효과 b_i의 상관계수행렬로 다음과 같다.

$$R_i = \begin{pmatrix} 1 & \rho_{i,1,2} & \cdots & \rho_{i,1,n_i} \\ \rho_{i,1,2} & 1 & \cdots & \rho_{i,2,n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{i,1,n_i} & \rho_{i,2,n_i} & \cdots & 1 \end{pmatrix},$$

여기서 $\rho_{i,j,t} = \mathrm{corr}(Y_{i,j}, Y_{i,t})$이다.

$b_{i,t}$의 표준편차 $\sigma_{i,t,t}$의 모형화는 로그선형모형을 통하여 모형화하며 다음과 같이 제시된다.

$$\log \sigma_{i,t,t} = z_{i,t}^T \alpha, \tag{3.9}$$

여기서 α는 $m \times 1$ 모수벡터이고, $z_{i,t}$는 개체-특정적 공변량이다.

상관계수행렬 R_i의 각 원소들은 양정치성을 만족하면서 주대각원소가 1이어야 함으로 그 추정이 쉽지 않다. 따라서 상관계수행렬을 직접 추정하는 대신에 상관계수행렬 R_i와 일대일 대응하는 부분 자기상관행렬 Π_i를 먼저 추정하고, 그리고 그것에 대응하는 상관계수행렬로 변환하여 추정한다. 이를 설명하기 위하여 부분 자기상관을 다음과 같이 설명한다.

$$\pi_{i,j,t} = \mathrm{corr}\left(b_{i,j}, b_{i,t} | b_{j,h}, j < h < t\right),$$

여기서 $\pi_{i,j,t}$는 $b_{i,j}$와 $b_{i,t}$의 조건부 자기상관계수로 그 사이에 있는 임의효과인 $b_{i,h}$는 주어진 상태이다. 부분 자기상관행렬 Π_i과 상관계수행렬 R_i의 관계를 우선 설명한다. $R[j : j+l]$은 상관행렬 R_i의 j번째 행과 열에서부터 $j+l$번째의 행과 열까지의 부분행렬을 아래와 같이 제시한다.

$$R[j : j+l] = \begin{pmatrix} 1 & r_1^T(j,l) & \rho_{j,j+l} \\ r_1(j,l) & R_2(j,l) & r_3(j,l) \\ \rho_{j,j+l} & r_3^T(j,l) & 1 \end{pmatrix},$$

여기서 $r_1^T(j,l) = (\rho_{i,j,j+1}, \ldots, \rho_{i,j,j+l-1})$, $r_3^T(j,l) = (\rho_{i,j+l,j+1}, \ldots, \rho_{i,j+l,j+l-1})$와 $R_2(j,l)$는 상관행렬 R_i에서 $j+1$부터 $j+l-1$까지의 원소들이다.

부분 상관계수 $R[j : j+l]$의 요소들을 이용하여 $\pi_{i,j,j+l}$와 $\rho_{i,j,j+l}$의 관계는 다음과 같다.

$$\pi_{i,j,j+l} = \frac{\rho_{i,j,j+l} - r_1^T(j,l) R_2(j,l)^{-1} r_3(j,l)}{\left[1 - r_1^T(j,l) R_2(j,l)^{-1} r_1(j,l)\right]^{\frac{1}{2}} \left[1 - r_3^T(j,l) R_2(j,l)^{-1} r_3(j,l)\right]^{\frac{1}{2}}}, \tag{3.10}$$

여기서 $\pi_{i,j,j+1} = \rho_{i,j,j+1}$이다. 위의 식 (3.10)을 다음과 같은 식으로 다시 표현할 수 있다.

$$\rho_{i,j,j+l} = r_1^T(j,l) R_2(j,l)^{-1} r_3(j,l) + D_{jl} \pi_{i,j,j+l}, \tag{3.11}$$

여기서 $D_{j,l} = [1 - r_1^T(j,l) R_2(j,l)^{-1} r_1(j,l)]^{1/2} [1 - r_3^T(j,l) R_2(j,l)^{-1} r_3(j,l)]^{1/2}$이고, $\rho_{i,j,j+1} = \rho_{i,j,j+1}$.

식 (3.11)을 이용하여 부분자기 상관행렬 Π_i은 다음과 같다.

$$\Pi_i = \begin{pmatrix} 1 & \pi_{i,1,2} & \pi_{i,1,3} & \cdots & \pi_{i,1,n_i-1} & \pi_{i,1,n_i} \\ \pi_{i,1,2} & 1 & \pi_{i,2,3} & \cdots & \pi_{i,2,n_i-1} & \pi_{i,2,n_i} \\ \pi_{i,1,3} & \pi_{i,2,3} & 1 & \cdots & \pi_{i,3,n_i-1} & \pi_{i,3,n_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \pi_{i,1,n_i-1} & \pi_{i,2,n_i-1} & \pi_{i,3,n_i-1} & \cdots & 1 & \pi_{i,n_i-1,n_i} \\ \pi_{i,1,n_i} & \pi_{i,2,n_i} & \pi_{i,3,n_i} & \cdots & \pi_{i,n_i-1,n_i} & 1 \end{pmatrix}.$$

위의 각 $\pi_{i,j,j+l}$는 -1과 1 사이에 존재하고, Π_i는 양정치성의 조건은 필요 없다. 수정/이동평균 콜레스키 분해방법에서처럼 자기상관행렬에서도 n_i가 증가하면서 행렬의 모수들의 개수도 증가한다. 따라서 피셔의 Z변환(Fisher's Z-transformation)을 이용하여 자기상관계수를 아래와 같이 모형화 한다.

$$z(\pi_{i,j,l}) \equiv \frac{1}{2} \log\left(\frac{1 + \pi_{i,j,l}}{1 - \pi_{i,j,l}}\right) = w_{i,j,l}^T \gamma, \tag{3.12}$$

여기서 γ는 $q \times 1$의 모수벡터이며, $w_{i,j,l}$은 수정 콜레스키 분해방법에서 제시된 것과 같다. 따라서 $w_{i,j,l}$를 AR(p)를 만족시키는 공변량을 사용하면, 그것에 의해 만들어진 상관계수행렬은 AR(p)구조를 가지게 된다.

부분 자기상관을 이용한 상관계수행렬은 항상 양정치성을 만족하게 되며, 그리고 로그선형모형 (3.9)와 피셔 Z변환을 이용한 모형화 (3.12)를 통하여 공분산행렬의 모수의 개수를 줄일 수 있다. Daniels와 Pourahmadi (2009)는 경시적 연속형 자료의 분석을 위하여 부분자기상관계수를 이용하였고, 베이지안 추론을 이용하여 모수들을 추정하였다. Lee 등 (2013)는 주변화모형에서 부분자기상관을 이용하여 이변량 경시적 순서화자료를 분석하였다.

4. 결론

이 논문에서 우리는 경시적 범주형 자료분석을 위한 일반화 선형혼합모형을 고찰하였고, 이 모형에서 임의효과 공분산행렬의 모수추정에 대한 방법들을 고찰하였다. 주로 사용되는 방법으로 수정/이동평균 콜레스키분해 방법과 부분 자기상관행렬을 이용한 방법을 제시하였다. 수정/이동평균 콜레스키분해 방법을 이용한 모형화는 일반화 자기회귀/이동평균 모수에 선형회귀모형을 그리고 혁신분산에 로그선형모형을 적용하여 원하는 차수의 AR/MA구조를 가지면서 이공분산성을 가지는 행렬을 추정할 수 있었다. 부분 자기상관행렬을 이용한 방법에서도 부분 자기상관에 피셔의 Z변환을 이용한 모형을 적용하고, 표준편차에 로그선형모형을 적용함으로써 원하는 차수의 AR구조의 이분산성을 가지는 공분산행렬을 추정할 수 있었다. 여기서 제시된 모든 방법들은 임의효과 공분산/상관계수행렬이 항상 양정치성을 만족하게 추정할 수 있고, 행렬의 모수의 개수도 줄일 수 있게 되었다.

임의효과 공분산행렬의 추정을 위한 수정 콜레스키분해 방법은 각각의 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 고려할 수 있다. 이 경우 수정 콜레스키 방법에서 차수가 큰 경우 이를 이용한 통계적 모형의 비효율성을 극복할 수 있다. 이러한 모형들을 앞으로의 연구과제로 생각할 수 있다.

References

Agresti, A. (2013). *Categorical Data Analysis*, 3rd Edition, Wiley.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 125–134.

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

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

Daniels, M. J. and Pourahmadi, M. (2009). Modeling covariance matrices via partial autocorrelations, *Journal of Multivariate Analysis*, **100**, 2352–2363.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd Edition, Analysis of Longitudinal Data.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data, *Biometrics*, **55**, 688–698.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data, *Biometrics*, **58**, 342–351.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models, *Biometrika*, **88**, 973–985.

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

Lee, K., Daniels, M. and Joo, Y. (2013). Flexible marginalized models for bivariate longitudinal ordinal data, *Biostatistics*, **14**, 462–476.

Lee, K. and Yoo, J. K. (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.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal analysis using generalized linear models, *Biometrika*, **73**, 13–22.

Pan, J. and Mackenzie, G. (2003). On modeling mean-covariance structure in longitudinal studies, *Biometrika*, **90**, 239–244.

Pan, J. and Mackenzie, G. (2006). Regression models for covariance structures in longitudinal studies, *Statistical Modeling*, **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.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters, *Biometrika*, **94**, 1006–1013.

Pourahmadi, M. and Daniels, M. J. (2002). Dynamic conditionally linear mixed models for longitudinal data, *Biometrics*, **58**, 225–231.

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

일반화 선형혼합모형의 임의효과 공분산행렬을 위한 모형들의 조사 및 고찰

김지영[a] · 이근백[a,1]

[a]성균관대학교 통계학과

(2015년 3월 13일 접수, 2015년 3월 30일 수정, 2015년 3월 30일 채택)

요 약

일반화 선형혼합모델은 일반적으로 경시적 범주형 자료를 분석하는데 사용된다. 이 모델에서 임의효과는 반복 측정치들의 시간에 따른 의존성을 설명한다. 임의효과 공분산행렬의 추정은 여러가지 제약조건들 때문에 쉽지 않은 문제이다. 제약조건으로는 행렬의 모수들의 수가 많으며, 또한 추정된 공분산행렬은 양정치성을 만족하여야 한다. 이러한 제한을 극복하기 위해, 임의효과 공분산행렬의 모형화를 위한 여러가지 방법이 제안되었다: 수정 단냠레스키분해, 이동평균 단냠레스키분해와 부분 자기상관행렬을 이용한 방법이 있다. 이 논문에서 위의 제안된 방법들을 소개한다.

주요용어: 경시적 자료, 임의효과 공분산행렬, 수정한 단냠레스키 분해, 이동평균 콜레스키 분해, 부분 자기상관행렬.

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

[1]교신저자: (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: keunbaik@skku.edu