Journal of Statistical Planning and Inference 142 (2012) 743–751

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A cautionary note on generalized linear models for covariance of unbalanced longitudinal data

Jianhua Z. Huang [a,*], Min Chen [b], Mehdi Maadooliat [a], Mohsen Pourahmadi [a]

[a] Department of Statistics, Texas A&M University, United States
[b] ExxonMobil Biomedical Sciences, Inc., United States

ARTICLE INFO

Article history:
Received 15 November 2010
Received in revised form
20 September 2011
Accepted 22 September 2011
Available online 1 October 2011

Keywords:
Cholesky decomposition
Missing data
Joint mean–covariance modeling

ABSTRACT

Missing data in longitudinal studies can create enormous challenges in data analysis when coupled with the positive-definiteness constraint on a covariance matrix. For complete balanced data, the Cholesky decomposition of a covariance matrix makes it possible to remove the positive-definiteness constraint and use a generalized linear model setup to jointly model the mean and covariance using covariates (Pourahmadi, 2000). However, this approach may not be directly applicable when the longitudinal data are unbalanced, as coherent regression models for the dependence across all times and subjects may not exist. Within the existing generalized linear model framework, we show how to overcome this and other challenges by embedding the covariance matrix of the observed data for each subject in a larger covariance matrix and employing the familiar EM algorithm to compute the maximum likelihood estimates of the parameters and their standard errors. We illustrate and assess the methodology using real data sets and simulations.

1. Introduction

To cope with the positive-definiteness constraint, the modified Cholesky decomposition has been introduced as a tool for reparameterization of the covariance matrix in longitudinal studies (Pourahmadi, 1999, 2000). The entries of the lower triangular matrix and the diagonal matrix from the modified Cholesky decomposition have interpretations as auto-regressive coefficients and prediction variances when regressing a measurement on its predecessors. This unconstrained reparameterization and its statistical interpretability makes it easy to incorporate covariates in covariance modeling and to cast the joint modeling of mean and covariance into the generalized linear model framework. The methodology has proved to be useful in recent literature; see for example, Pourahmadi and Daniels (2002), Pan and MacKenzie (2003), Ye and Pan (2006), Daniels (2006), Huang et al. (2006), Levina et al. (2008), Yap et al. (2009), and Lin and Wang (2009).

However, it encounters the problem of incoherency of the (auto)regression coefficients and innovation variances across the subjects when the longitudinal data are unbalanced and covariates are used. Unfortunately, this problem has not been noticed or pointed out explicitly in the literature. Although covariates have been used in Pourahmadi (1999) for modeling balanced data, the coherency consideration suggests that care must be taken when the data are unbalanced. In fact, the formulations in Pourahmadi and Daniels (2002) and the subsequent papers are suitable only when the missing data are dropouts, where for a subject the missingness occurs from certain time point to the end of the study. In general, as we

[*] Corresponding author.
E-mail addresses: jianhua@stat.tamu.edu (J.Z. Huang), min.chen@exxonmobil.com (M. Chen), madoliat@stat.tamu.edu (M. Maadooliat), pourahm@stat.tamu.edu (M. Pourahmadi).

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.09.011

illustrate by an example in Section 2, a coherent system of regressions based on the modified Cholesky decomposition may not exist if there are intermittent missing values.

In this paper, we propose to handle both dropouts and intermittent missing values using an incomplete data model and the EM algorithm (Dempster et al., 1977; Jennrich and Schluchter, 1986) when the data are missing at random (Rubin, 1976). Our incomplete data framework assumes that a fixed number of measurements are to be collected at a common set of times for all subjects with a common “grand covariance matrix” Σ, but since not all responses are observed for all subjects, a generic subject i’s measurements will have a covariance matrix Σ_i which is a principal minor of Σ. Since the covariance model for Σ is built from measurements at a common set of times, the incoherency problem is completely avoided. A “generalized EM algorithm” (*in which we try to increase the objective function in the “M” step rather than maximizing it*) is then developed to deal with the missing data in the context of the modified Cholesky decomposition and to compute the maximum likelihood estimates.

2. The incoherency problem in incomplete longitudinal data

Assume that the vector of repeated measures y_i of subject i collected at completely irregular times $t_{ij}, j=1,\ldots,m_i$, follows a zero mean multivariate normal distribution with covariance matrix Σ_i. The modified Cholesky decomposition gives $T_i\Sigma_i T_i' = D_i$, where T_i is a lower triangular matrix whose below-diagonal entries are the negatives of the autoregressive coefficients, ϕ_{itj}, in $\hat{y}_{it} = \sum_{j=1}^{t-1}\phi_{itj}y_{ij}$, and D_i is a diagonal matrix whose diagonal entries σ_{it}^2’s are the innovation variances of the autoregressions. A generalized linear model for Σ_i can be built for each subject by relating the autoregressive parameters ϕ_{itj} and the log innovation variances $\log \sigma_{it}^2$ to some covariates as

$$\phi_{itj} = z_{itj}'\gamma_i \quad \text{and} \quad \log(\sigma_{it}^2) = u_{it}'\lambda_i, \ 1 \le j \le t-1, \ 1 \le t \le m_i, \tag{1}$$

where z_{itj} and u_{it} are covariates for covariance matrices, and $\gamma_i \in R_i^q$ and $\lambda_i \in R_i^r$ are the corresponding regression parameters which have different dimensions for different subjects. The covariates in (1) are usually of the form

$$z_{itj} = (1,(t_{it}-t_{ij}),(t_{it}-t_{ij})^2,\ldots,(t_{it}-t_{ij})^{q-1})',$$

$$u_{it} = (1,t_{it},t_{it}^2,\ldots,t_{it}^{r-1}). \tag{2}$$

This general form gives rise to the following two statistical problems:

- Estimation of γ_i and λ_i based on a single vector y_i is impossible unless m_i is large or a sort of stationarity assumption is imposed. In other words, one cannot borrow strength from other subjects.
- Even if these parameters are assumed the same for all subjects so that one may borrow strength from other subjects, there remains a problem of interpretation or incoherency of the parameters.

The next example shows the incoherency problem, when the data are unbalanced. It seems Pourahmadi and Daniels (2002), Eq. (4), is the first place where this problem was encountered and not addressed properly. Another source is Lin and Wang (2009) and the references therein. For ease of reference we call such a method the naive method in what follows.

Example. Let us consider the simple model, $y_{it} = \phi y_{it-1} + \varepsilon_{it}$, for $t=2,3,4$ with $y_{i1} = \varepsilon_{i1}$ and $\varepsilon_i \sim N_4(0,I)$. Thus for a completely observed subject $D = I_4$ with the following structures for T and Σ:

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\phi & 1 & 0 & 0 \\ 0 & -\phi & 1 & 0 \\ 0 & 0 & -\phi & 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1+\phi^2 & \phi^2+\phi^3 & \phi^3+\phi^4 \\ \phi^2 & \phi^2+\phi^3 & 1+\phi^2+\phi^4 & \phi+\phi^3+\phi^5 \\ \phi^3 & \phi^3+\phi^4 & \phi+\phi^3+\phi^5 & 1+\phi^2+\phi^4+\phi^6 \end{pmatrix}.$$

Now, consider two subjects where Subject 1 has three measurements at times 1, 2, 4 and Subject 2 has measurements at times 1, 3, 4. It is straightforward to obtain Σ_1 by deletion of the third row and column of Σ, similarly Σ_2 is obtained by deletion of the second row and column of the Σ. Now by using the modified Cholesky decomposition, one can obtain T_i and D_i for $i=1,2$ as follows:

$$T_1 = \begin{pmatrix} 1 & 0 & 0 \\ -\phi & 1 & 0 \\ 0 & -\phi^2 & 1 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1+\phi^2 \end{pmatrix},$$

$$T_2 = \begin{pmatrix} 1 & 0 & 0 \\ -\phi^2 & 1 & 0 \\ 0 & -\phi & 1 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1+\phi^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Although both ϕ_{i21} can be interpreted as the coefficient when regressing the second measurement on the first, they actually take different values: For Subject 1, the measurement at time 2 is regressed on that at time 1, but for Subject 2, the