

ULRR

Modelling covariance structures for multivariate longitudinal data

Item Type	Meetings and Proceedings
Authors	Xu, Jing;Mackenzie, Gilbert
Citation	Proceedings of the 25th International Workshop on Statistical Modelling;
Publisher	IWSM
Download date	2026-04-17 21:41:06
Item License	https://creativecommons.org/licenses/by-nc-sa/1.0/
Link to Item	https://hdl.handle.net/10344/2807

Modelling covariance structures for multivariate longitudinal data (An oral presentation)

Jing Xu¹ and, Gilbert MacKenzie²

¹ Centre of Biostatistics, University of Limerick, Ireland.

² ENSAI, Rennes, France

Email: jing.xu@ul.ie gilbert.mackenzie@ul.ie

Abstract: The analysis of multivariate longitudinal data can be challenging because of the existence of correlations between multiple time-dependent responses repeated over time. Therefore, one major task in analyzing such data is to model efficiently the covariance matrices $cov(y_i) = \Sigma_i$ for $i = 1, \dots, n$ subjects. In this paper, we develop a data-driven method to model the covariance structures. Thereby, constrained and hard-to-model parameters of Σ_i are *traded-in* for unconstrained and interpretable parameters. Estimates of these parameters, together with the parameters in the mean, are obtained by maximum likelihood approach, and the large-sample asymptotic properties are derived. A real-life example is given to illustrate the method introduced.

Keywords: multivariate longitudinal data; marginal models; covariance modelling; block triangular factorization; matrix logarithm

1 Introduction

In many epidemiological studies and clinical trials, subjects are measured on several occasions with regard to a collection of response variables. Analysis of such multivariate longitudinal data involves modelling the joint evolution of the response variables over time. Consider, as an example, a study of anaemia in pregnancy (McMullan, et al. 2003) carried out in Belfast. A total of 263 patients had three visits to the clinic. For them, two blood measurements, Erythropoietin (Epo) and Haemoglobin (Hb), were taken throughout pregnancy. There are many similar examples: Chapman et al. (2003), Thiebaut et al.(2002),and Newsom (2002).

However, the analysis of such multivariate longitudinal data is complicated by: a) the correlation between the responses at each time point, b) the correlation within separate responses over time, and c) the cross-correlation between different responses at different times. Therefore, one major task in analyzing these data is to model the covariance matrices $cov(y_i) = \Sigma_i$

for $i = 1, \dots, n$ subjects. Several approaches have been developed: doubly multivariate models (DMM) analysis (Timm, 1980), multivariate repeated measurement models with a Kronecker product covariance structure (Galecki, 1994), multivariate mixed models (Jones 1993) and a structural equation modelling approach (Hatcher, 1998). In this paper, we developed a data-driven method to model the covariance structures. We extend the idea of covariance modelling (Pourahmadi 1999) for traditional univariate longitudinal data to the multivariate case by using the block triangular factorization of Σ_i . This new method maintains most of the nice properties enjoyed by the univariate case, i.e, the decomposition is unique, positive definiteness of Σ_i is guaranteed, the new parameters are unconstrained and have useful statistical interpretations.

2 Covariance modelling

For simplicity, the method is presented for the bivariate case in the rest of paper, although it can be extended straight-forwardly to the multivariate case.

2.1 Block triangular factorization of Σ

Let $y_{ij} = (y_{ij}^{(1)}, y_{ij}^{(2)})'$ present the observations of two response variables for the i -th individual at j -th time point ($i = 1, \dots, n; j = 1, \dots, m$). Further let $y_i = (y'_{i1}, \dots, y'_{im})'$. Denote the covariance matrix of $cov(y_i)$ by Σ . Here we assume that the covariance matrices of y_i are homogeneous across subjects. Noting that Σ is positive definite, Σ can be factorized block-triangularly as (see Hamilton 1994)

$$T\Sigma T' = D, \quad \text{or} \quad \Sigma^{-1} = T'D^{-1}T,$$

where T is a block lower triangular with 2×2 identity matrices as diagonal entries and D is a block-diagonal matrix with positive definite 2×2 matrices as diagonal entries. It is easily seen that Σ is positive definite if and only if D is positive definite and the decomposition is unique, which has the following statistical interpretation: the block matrices, denoted by $\Theta_{i,j}$, as the below-diagonal entries of T are the negatives of the coefficient matrices of $\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \Theta_{j,k}(y_{ik} - \mu_{ik})$, the linear least-squares predictor of y_{ij} based on its predecessors y_{ij-1}, \dots, y_{i1} , and the block diagonal entries, denoted by D_j , of D are the prediction error covariances $D_j = cov(y_{ij} - \hat{y}_{ij})$, for $1 \leq j \leq m$. With this decomposition, the $\frac{1}{2}2m(2m+1)$ constrained and hard-to-model parameters of Σ can be traded in for the $\frac{1}{2}2m(2m+1)$ unconstrained and interpretable parameters $\Theta_{j,k}$, $\log D_j$ (see subsection 2.2) for $1 \leq j \leq m$ and $1 \leq k \leq j-1$. We refer to the new parameters $\Theta_{j,k}$'s and D_j 's as the autoregressive coefficient matrices and the innovation covariance matrices of Σ .

2.2 Matrix logarithm of D

Since D is positive definite, i.e., all the diagonal entries D_1, \dots, D_m are positive definite, the matrix logarithm of D_j can now be defined by

$$A_j = \log D_j$$

basing on the spectral decomposition of D_j . The positive definiteness of D_j is guaranteed by the definition of matrix exponential (Chiu et al. 1996).

2.3 Linear covariance models

Since $\Theta_{j,k}$ and $\log D_j$ are unconstrained, we may model them in terms of covariates. For example, the polynomials of time and lag. The new parameters in the linear models for $\Theta_{j,k}$ and $\log D_j$ are denoted by the unknown vectors γ and λ .

3 Maximum Likelihood Estimation

3.1 Estimation of parameters

In the marginal linear regression models with normal distributed responses, the estimates of the parameters γ and λ in the covariance matrices, together with the parameters, denoted by β , in the mean part, can be obtained by maximum likelihood approach. The log-likelihood of β , γ and λ , given y_1, \dots, y_n , satisfies

$$2 \log \ell(\beta, \gamma, \lambda | y_1, \dots, y_n) = -mn \log(2\pi) - n \log |D| - \sum_{i=1}^n r_i' T' D^{-1} T r_i, \quad (1)$$

where $r_i = y_i - X_i \beta$ and X_i is the design matrix in the regression model. Fixing γ and λ in (1) creates the weighted least squares solution of β is

$$\tilde{\beta} = \left\{ \sum_{i=1}^n X_i' \Sigma^{-1} X_i \right\}^{-1} \sum_{i=1}^n X_i' \Sigma^{-1} y_i. \quad (2)$$

Secondly, given β and λ , the solution of the first derivative of γ is

$$\tilde{\gamma} = \left\{ \sum_{i=1}^n Z_i^*{}' D^{-1} Z_i^* \right\}^{-1} \sum_{i=1}^n Z_i^*{}' D^{-1} r_i, \quad (3)$$

where $Z_i^* = (r_{i1}^*, \dots, r_{iq}^*)$ with $r_{il}^* = U_l^* r_i$. Here $U_l^* (l = 1, \dots, q)$ are the block lower triangular matrices with off-diagonal matrices $U_{jkl} (k < j, j = 2, \dots, m)$ and zero matrices as diagonal entries.

Denote the $(2j-1)$ -th and $2j$ -th elements of the vector $T r_i$ by the 2×1 vector $e_{ij} = r_{ij} - \hat{r}_{ij}$ with $\hat{r}_{ij} = \sum_{k=1}^{j-1} \Theta_{j,k} r_{ik}$ for $j = 1, \dots, m$. By the

definition of matrix exponential and logarithm, the log-likelihood function excluding the constant becomes

$$2 \log \ell(\beta, \gamma, \lambda | y_1, \dots, y_n) \sim -mn \sum_{l=1}^d \lambda_l \text{tr}(\bar{V}_{\cdot l}) - n \sum_{j=1}^m \text{tr}\{B_j \exp(-A_j)\}, \quad (4)$$

where $\bar{V}_{\cdot l} = \sum_{j=1}^m V_{jl}/m$ and $B_j = \sum_{i=1}^n e_{ij} e'_{ij}/n$.

The first and second order of derivatives of $\log \ell$ with respect to λ are derived by applying the directional derivative of the matrix exponential (Bellman 1970) to the Taylor series expansion of function (4) with respect to λ . Fixed β and γ , the solution of the estimation equation for λ can be obtained by the Newton-Raphson iterations. we denote it by $\tilde{\lambda}$.

The iterative procedure proceeds within (2), (3) and (4) by initializing at $\Sigma = I_m$ where I_m is a $m \times m$ identity matrix and iterating until convergence to obtain the ML estimator $(\hat{\beta}', \hat{\gamma}', \hat{\lambda}')'$ simultaneously.

3.2 Asymptotic properties

Briefly speaking, under some necessary regularity conditions, the ML estimators $\hat{\theta} = (\hat{\beta}', \hat{\gamma}', \hat{\lambda}')'$ is strongly consistent for the true value $\theta_0 = (\beta'_0, \gamma'_0, \lambda'_0)'$ and the ML estimator $\hat{\theta}$ has an asymptotically normal distribution.

4 Application: a study of anemia in pregnancy

An observational study was carried out at the Mater Infirmary Hospital in Belfast to investigate changes occurring in two blood measurements, erythropoietin Epo and haemoglobin Hb, through out pregnancy. Some 263 patients were recruited, and three blood samples of Hb and Epo were taken at booking-in, 28-32 weeks and 38 weeks, though these times were variable and there were missing visits. We fitted a marginal bi-variate model to the two variables. We modelled the mean and covariance linearly in terms of covariates. Details of the analysis will be shown in the main paper.

References

- Chiu, T.Y.M., Leonard, T., and Tsui, K-W. (1996). The Matrix-Logarithmic Covariance Model. *Journal of the American Statistical Association*. **91**, 198-210.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.
- Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*. **86**, 677-690.