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.*
[b] *ExxonMobil Biomedical Sciences, Inc.*

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.

Keywords: Cholesky decomposition, missing data, joint mean-covariance modeling

1. Introduction

To cope with the positive-definiteness constraint, the modified Cholesky decomposition has been introduced as a tool for reparameterization of the

Email addresses: jianhua@stat.tamu.edu (Jianhua Z. Huang), min.chen@exxonmobil.com (Min Chen), madoliat@stat.tamu.edu (Mehdi Maadooliat), pourahm@stat.tamu.edu (Mohsen Pourahmadi)

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 autoregressive 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 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 Choleskey 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}, $j = 1, \cdots, m_i$, 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

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

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

$$\begin{aligned} z_{itj} &= (1, (t_{it} - t_{ij}), (t_{it} - t_{ij})^2, \ldots, (t_{it} - t_{ij})^{q-1})', \\ u_{it} &= (1, t_{it}, t_{it}^2, \ldots, t_{it}^{r-1}). \end{aligned} \qquad (2)$$

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), equ. (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's consider the simple model, $y_{it} = \phi y_{it-1} + \epsilon_{it}$, for $t = 2, 3, 4$ with $y_{i1} = \epsilon_{i1}$ and $\epsilon_i \sim N_4(0, I)$. Thus for a completely observed

subject $D = I_4$ with the following structures for T and Σ:

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

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 3rd row and column of Σ, similarly Σ_2 is obtained by deletion of the 2nd 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:

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

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

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 measurement at time 3 is regressed on that at time 1. More precisely, it is evident that $\phi_{121} = \phi$ while $\phi_{221} = \phi^2$. Similar results and statements hold for the innovation variances. This difference in the values of the regression coefficients or the lack of coherence implies that a naive approach that simply relates the ϕ_{itj} to covariates may not be prudent when the data are unbalanced. Numerical examples in Section 4 will show that the use of the naive method in the presence of the incoherency problem can lead to serious bias and tremendously high risk in covariance estimation.

3. The Incomplete Data Model and The EM Algorithm

Let y_i be an $m_i \times 1$ vector containing the responses for subject i, where $i = 1, \ldots, n$. The y_i are assumed to follow the model

$$y_i = X_i\beta + e_i,$$

where X_i is an $m_i \times p$ known matrix of covariates, β is a $p \times 1$ vector of unknown regression parameters, and e_i is an $m_i \times 1$ vector of errors. The

e_i's are distributed as $N(0, \Sigma_i)$ individually and are independent of each other. We assume that e_i is a sub-vector of a larger $m \times 1$ vector e_i^* that corresponds to the same set of m observation times $t_1, \ldots, t_m$, for all i. This model assumption is valid in a typical setting of longitudinal data when the measurements are collected at the same set of scheduled time points for all subjects although for a particular subject, the measurements at some time points may be missing (Jennrich and Schluchter, 1986). When the time points are totally irregular across subjects, one can approximate this setting by binning the time points.

Under the above model assumptions, Σ_i is a sub-matrix of $\Sigma_i^* = \mathrm{var}(e_i^*)$. According to the modified Cholesky decomposition, there exists a unique lower triangular matrix T_i with 1's as main diagonal entries and a unique diagonal matrix D_i with positive diagonal entries such that $T_i \Sigma_i^* T_i' = D_i$. The below-diagonal entries of T_i are the negatives of the autoregressive coefficients, ϕ_{itj}, in $\hat{e}_{it}^* = \sum_{j=1}^{t-1} \phi_{itj} e_{ij}^*$, the linear least squares predictor of e_{ij}^* based on its predecessors $e_{i(t-1)}^*, \ldots, e_{i1}^*$. The diagonal entries of D_i are the innovation variances $\sigma_{it}^2 = \mathrm{var}(e_{it}^* - \hat{e}_{it}^*)$, where $1 \le t \le m$ and $1 \le i \le n$. The parameters ϕ_{itj} and $\log \sigma_{it}^2$ are unconstrained and are modeled as in (1) with the same parameters γ and λ for all subjects. We assume that there is no missing value in these covariates, which is the case if they only depend on baseline covariates and scheduled observation times.

To compute the maximum likelihood estimator, we use an iterative EM algorithm for the incomplete data model (Dempster et al., 1977; Jennrich and Schluchter, 1986). The algorithm consists of two parts. The first part applies the generalized least squares solution to update β:

$$\tilde{\beta} = \Big(\sum_{i=1}^{n} X_i' \Sigma_i^{-1} X_i \Big)^{-1} \Big(\sum_{i=1}^{n} X_i' \Sigma_i^{-1} y_i \Big), \tag{3}$$

which is obtained by maximizing the likelihood function with respect to β while holding γ and λ fixed at their current values. The second part comprises one iteration of a generalized EM algorithm to update λ and γ, using e_i^* as complete data and sub-vectors $e_i = y_i - X_i \beta$ as observed data, and assuming β is equal to its current value. The algorithm iterates between the two parts until convergence.

The details of the EM algorithm for estimating the parameters and the asymptotic inference are given in the Appendix. For unbalanced normally distributed longitudinal data, Theorem 1 of Holan and Spinka (2007) estab-

lishes the consistency and asymptotic normality of the maximum likelihood estimator of $\theta = (\beta', \gamma', \lambda')'$ under some mild regularity conditions.

4. Data Analysis

We use two real data sets to illustrate and assess the performance of our approach.

4.1. The Fruit Fly Data

The "fruit fly mortality" (FFM) data (Zimmerman and Núñez Antón, 2010) are age-specific measurements of mortality for 112 cohorts of a common fruit fly, "Drosphila melanoster". Everyday, dead flies were counted for each cohort, and these counts were pooled into 11 five-day intervals. The raw mortality rate was recorded as $-\log\{N(t+1)/N(t)\}$, where $N(t)$ is the number of live flies in the cohort at the beginning of time t ($t = 0, 1, \ldots, 10$). For unknown reason 22% of the data were missing. The raw mortality rate were log-transformed to ensure that the responses are close to be normally distributed. To apply our method, we implicitly assume that the data are missing at random. This assumption is fairly common for the likelihood method to work and is made by Zimmerman and Núñez Antón (2010), but little is known about the missing data mechanism for this data set.

To formulate model (1) for the FFM data, we rely on the regressograms of the sample covariance introduced in Pourahmadi (1999). Generally, to come up with a good replacement for the sample covariance matrix in presence of missing data, one may construct the sample covariance matrix based on the pairwise complete observations, but this estimator may lack the positive definiteness property. An alternative solution is to fit saturated models to the matrices T and D, based on our proposed EM algorithm to obtain a raw estimate, denoted by $\widehat{\Sigma}_s$ where "s" stands for saturated. Now, the regressograms of $\widehat{\Sigma}_s$ can be used to choose the order of polynomials for modeling the Cholesky factors of covariance matrix. Figure 1, shows the sample regressograms of $\widehat{\Sigma}_s$ which suggest cubic polynomial models for the autoregressive coefficients and the log-innovation variances. Next, we compare the fitted results using the EM algorithm and the naive method for this dataset. When applying the EM and the naive algorithm, the design matrices for the regressions were constructed using (2) with $q = r = 4$. It is clear from Figure 1 that there is a difference between the two fits particularly for the innovation variances, but since we do not have the complete data, it is impossible to

know which one is better. This issue is settled in the next two subsections using a complete real data set and simulations.

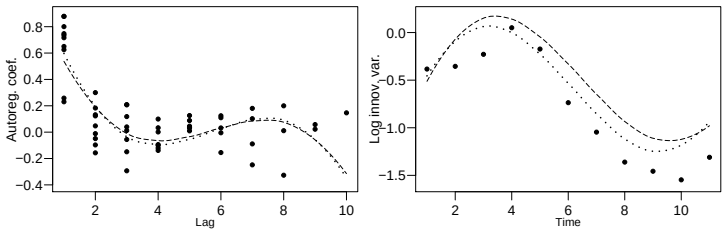

Figure 1: Fitted cubic polynomial models on the FFM data. Dotted and dashed curves represent respectively results by applying the EM and the naive algorithm. Circles represent the regressogram for the saturated model of covariance matrix, based on the EM algorithm.

4.2. The Cattle Data

The Kenward (1987)'s cattle data were obtained by randomly assigning thirty cattles to two treatment groups, A and B. The weights of the cattles were recorded 11 times over a 133-day period. Similar to Pourahmadi (1999), we considered cubic fits for the Cholesky factors of the covariance matrix for the treatment A group. The fitted cubic polynomials are shown as solid lines in Figure 2.

We randomly removed four observations from each subject and applied both our EM algorithm and the naive algorithm on the incomplete data. The data removing and model fitting process was repeated ten times. It is clear from Figure 2 that the fits using the EM algorithm on the incomplete data are less variable and closer to that obtained from the complete data, while the same cannot be said about the results from the naive algorithm.

4.3. Simulation

We considered five different setups for our simulation study. In each setup we ran the simulation $N = 200$ times, and in each run we simulated $n = 30$ subjects with $m = 11$ time points. In the first three setups, the data are drawn from multivariate normal distributions with parameters specified as follows:

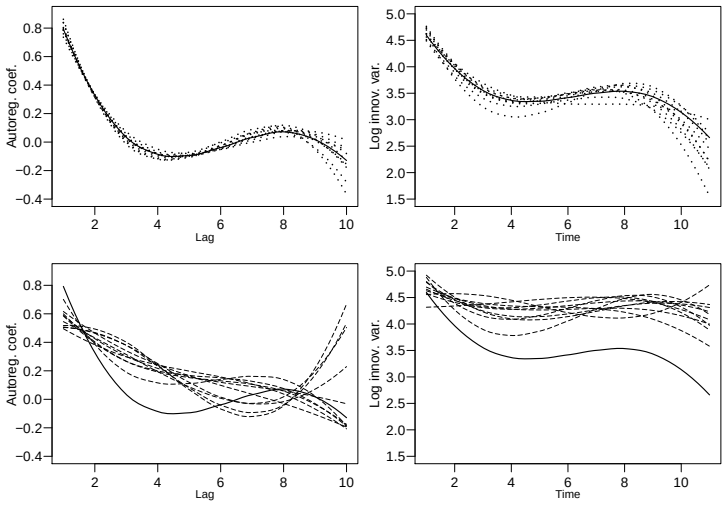

Figure 2: Fitted cubic polynomial models on the cattle data. Solid curves represent the fitted model from the complete data. Dotted and dashed curves represent respectively results for ten simulation runs by applying the EM (top panels) and the naive algorithm (bottom panels) on the incomplete data.

1. Σ_1 (the covariance matrix) and μ_1 (the mean vector) are obtained from applying the EM algorithm to the FFM data;

2. Σ_2 and μ_2 are the ML estimates based on the complete cattle data;

3. Σ_3 is an 11×11 matrix where the $(i,j)^{\text{th}}$ element $\sigma_{ij} = \min(i,j)$, and the mean vector μ_3 has all entries zero.

For setups 1 and 2, we fit cubic polynomials to both the autoregressive coefficients and the log-innovation variances. For setup 3, we considered a cubic fit for autoregressive coefficients, but a linear fit for log-innovation variances. In setup 3, the polynomials are not the true models and only serve as approximations.

For each of the simulation setups, we generated the data from the multivariate normal distribution with mean μ_j and variance Σ_j, and randomly removed four out of eleven observations from each subject. Figures 3, 4 and 5 compare the results of the EM algorithm and the naive method for each setup for ten randomly selected simulation runs. It is apparent that while

the naive method produces misleading results, the EM algorithm yields more reasonable results.

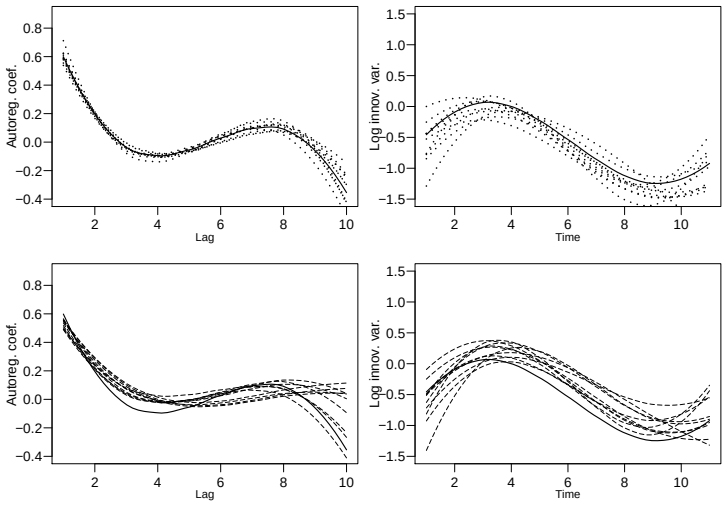

Figure 3: Simulation. Solid curves represent the model for Σ_1. Dotted and dashed curves represent respectively results for ten randomly selected simulation runs by applying the EM (top panels) and the naive algorithm (bottom panels) on the incomplete data.

To compare the performance of two methods in estimating the covariance matrix, we considered the risks of parameter estimation for the following two loss functions

$$\Delta_1(\Sigma, G) = \mathrm{tr}\Sigma^{-1}G - \log|\Sigma^{-1}G| - n \quad \text{and} \quad \Delta_2(\Sigma, G) = \mathrm{tr}(\Sigma^{-1}G - I)^2,$$

where Σ is the true covariance matrix and G is a positive-definite matrix with the same size. $\Delta_1(\Sigma, G)$ is known as the entropy loss and $\Delta_2(\Sigma, G)$ is called the quadratic loss. Both of these loss functions are 0 when $G = \Sigma$ and positive when $G \neq \Sigma$. The corresponding risk functions can be defined as

$$R_i(\Sigma, G) = E_\Sigma\{\Delta_i(\Sigma, G)\}, \qquad i = 1, 2.$$

The estimator $\hat{\Sigma}$ is better than $\tilde{\Sigma}$ for Σ, if its associated risk function is smaller, that is, $R_i(\Sigma, \hat{\Sigma}) < R_i(\Sigma, \tilde{\Sigma})$. Table 1 shows that the naive method produces substantially larger risks than the EM algorithm, which is anticipated because of the incoherency problem of the naive method discussed in Section 2.

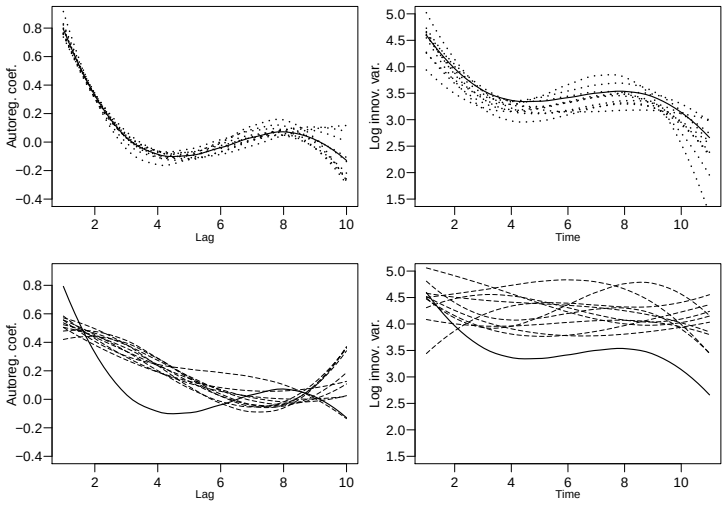

Figure 4: Simulation. Solid curves represent the model for Σ_2. Dotted and dashed curves represent respectively results for ten randomly selected simulation runs by applying the EM (top panels) and the naive algorithm (bottom panels) on the incomplete data.

Note that our fitting procedure is based on maximizing the normal likelihood. To evaluate its performance when the normality assumption is violated, we considered the following two setups:

4. Multivariate skew-normal (SN) distribution (Azzalini and Capitanio, 1999, eq. 9, pg. 584), with covariance matrix Σ_2, location vector μ_2, and skewness parameter $\alpha = 4 * \mathbf{1}$, where $\mathbf{1}$ is a vector of ones.

5. Multivariate t distribution (Kotz and Nadarajah, 2004, pg. 1), with 4 degrees of freedom, scale matrix $(1/2)\Sigma_2$ and location vector μ_2. The covariance matrix of the distribution is Σ_2.

The number of subjects, the number of time points each subject, the missing data generating mechanism, and the number of simulation runs are the same as in the first three setups. Results are presented in Table 1 and lead to the same conclusion as in the multivariate normal simulations, that is, the EM approach produces substantially smaller risks than the naive approach.

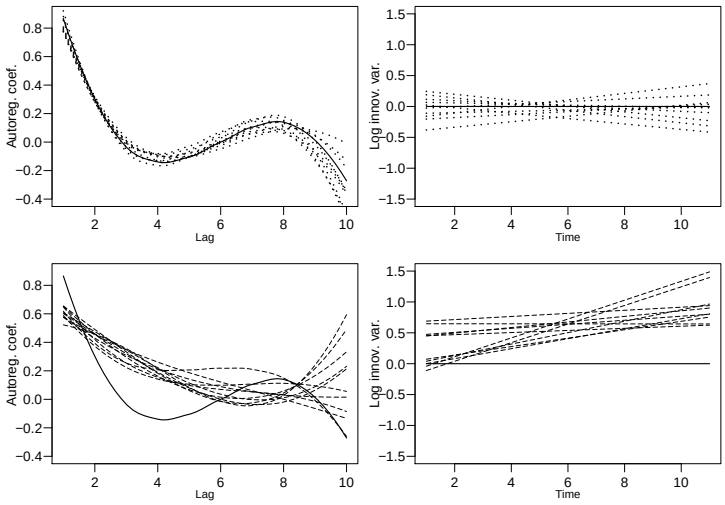

Figure 5: Simulation. Solid curves represent the model for Σ_3. Dotted and dashed curves represent respectively results for ten randomly selected simulation runs by applying the EM (top panels) and the naive algorithm (bottom panels) on the incomplete data.

Acknowlegement

Huang and Pourahmadi were partially supported by NSF of the US. Huang was also supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

Appendix

The EM Algorithm

The E-step of the generalized EM algorithm relies on the specifics of the parameterization of the modified Cholesky decomposition of the covariance matrix. Minus twice the log likelihood function for complete data, except for a constant, is given by

$$-2l = \sum_{i=1}^{n} (\log |\Sigma_i^*| + e_i^{*'} \Sigma_i^{*-1} e_i^*) = \sum_{i=1}^{n} \{\log |\Sigma_i^*| + \mathrm{tr}(\Sigma_i^{*-1} V_i)\}, \qquad (4)$$

where $V_i = e_i^* e_i^{*'}$. Let Q be the expected log likelihood given the observed data and the current parameter values. Denote $\widehat{V_i} = E(e_i^* e_i^{*'} | e_i)$, whose

Table 1: Comparison of the EM algorithm and the naive method in three simulation setups. The numbers in each cell are the mean and SE (in parathesis) based on 200 simulation runs.

Model	Method	Entropy risk	Quadratic risk
Normal(μ_1, Σ_1)	EM	0.62(0.028)	1.06(0.047)
	Naive	3.04(0.166)	20.24(2.477)
Normal(μ_2, Σ_2)	EM	0.69(0.030)	1.20(0.062)
	Naive	49.81(1.928)	2752.49(255.769)
Normal(μ_3, Σ_3)	EM	1.05(0.029)	2.27(0.085)
	Naive	27.94(1.146)	841.66(87.257)
SN$(\mu_2, \Sigma_2, \alpha)$	EM	0.99(0.029)	1.42(0.054)
	Naive	21.08(1.016)	602.80(59.077)
$t_4(\mu_2, (1/2)\Sigma_2)$	EM	1.95(0.289)	12.16(5.708)
	Naive	56.15(3.497)	4901.98(890.565)

computation is detailed at the end of this paragraph. Then

$$-2Q = \sum_{i=1}^{n} \{\log|\Sigma_i^*| + \mathrm{tr}(\Sigma_i^{*-1}\widehat{V}_i)\}. \tag{5}$$

We now give two expressions of $-2Q$ that are useful in deriving the steps of the EM-algorithm. Define $\mathrm{RS}_{it} = (e_{it}^* - \sum_{j=1}^{t-1} e_{ij}^* z_{itj}'\gamma)^2$ and $\widehat{\mathrm{RS}}_{it} = E(\mathrm{RS}_{it}|e_i)$. The modified Cholesky decomposition $T_i\Sigma_i^*T_i' = D_i$ can be used (Pourahmadi, 2000) to get

$$-2Q = \sum_{i=1}^{n}\sum_{t=1}^{m}\Big(\log\sigma_{it}^2 + \frac{\widehat{\mathrm{RS}}_{it}}{\sigma_{it}^2}\Big). \tag{6}$$

For $t > 1$, denote $Z_{it}' = (z_{it1}, \ldots, z_{it(t-1)})$, and let $\widehat{V}_{itt} = \widehat{V}_i[t,t]$, $\widehat{V}_{it}^{(t-1)} = \widehat{V}_i[1{:}(t-1), t]$, $\widehat{V}_i^{(t-1)} = \widehat{V}_i[1{:}(t-1), 1{:}(t-1)]$ be sub-matrices of $\widehat{V}_i$. We also make the convention that $\widehat{V}_{i1}^{(0)} = 0$ and $\widehat{V}_i^{(0)} = 0$. Using the fact that $\widehat{\mathrm{RS}}_{it}$ is the (t,t)-th element of the matrix $T_i\widehat{V}_iT_i'$, we obtain from (6) that

$$-2Q = \sum_{i=1}^{n}\sum_{t=1}^{m}\Big(\log\sigma_{it}^2 + \frac{\widehat{V}_{itt}}{\sigma_{it}^2}\Big) + \sum_{i=1}^{n}\sum_{t=1}^{m}\sigma_{it}^{-2}(-2\gamma'Z_{it}'\widehat{V}_{it}^{(t-1)} + \gamma'Z_{it}'\widehat{V}_i^{(t-1)}Z_{it}\gamma). \tag{7}$$

The calculation of $\widehat{V}_i$ is as follows. Note that $\widehat{V}_i = E(e_i^* e_i^{*\prime}|e_i) = \widehat{e_i^*}\widehat{e_i^*}' + \mathrm{var}(e_i^*|e_i)$ with $\widehat{e_i^*} = E(e_i^*|e_i)$. Write

$$e_i^* = \begin{pmatrix} e_i \\ e_i^+ \end{pmatrix} \sim N(0, \Sigma_i^*), \qquad \Sigma_i^* = \begin{pmatrix} \Sigma_{i11}^* & \Sigma_{i12}^* \\ \Sigma_{i21}^* & \Sigma_{i22}^* \end{pmatrix}.$$

The standard results for multivariate normal distributions give that

$$E(e_i^*|e_i) = \begin{pmatrix} I \\ \Sigma_{i21}^* \Sigma_{i11}^{*-1} \end{pmatrix} e_i, \qquad \mathrm{var}(e_i^*|e_i) = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{i22}^* - \Sigma_{i21}^* \Sigma_{i11}^{*-1} \Sigma_{i12}^* \end{pmatrix}. \quad (8)$$

Using (6) and (7), the update of γ and λ proceeds as follows. For fixed λ, $-2Q$ is a quadratic form in γ and is minimized by

$$\tilde{\gamma} = \Big(\sum_{i=1}^n \sum_{t=1}^m \sigma_{it}^{-2} Z_{it}' \widehat{V}_i^{(t-1)} Z_{it} \Big)^{-1} \sum_{i=1}^n \sum_{t=1}^m \sigma_{it}^{-2} Z_{it}' \widehat{V}_{it}^{(t-1)}. \quad (9)$$

For fixed γ, optimization of $-2Q$ over λ does not have a closed-form expression and we resort to the Newton-Raphson algorithm. Since $\log \sigma_{it}^2 = u_{it}'\lambda$, simple calculation yields

$$\frac{\partial Q}{\partial \lambda} = -\frac{1}{2} \sum_{i=1}^n \sum_{t=1}^m \Big(1 - \frac{\widehat{\mathrm{RS}}_{it}}{\sigma_{it}^2} \Big) u_{it}$$

and

$$\frac{\partial^2 Q}{\partial \lambda \partial \lambda'} = -\frac{1}{2} \sum_{i=1}^n \sum_{t=1}^m \frac{\widehat{\mathrm{RS}}_{it}}{\sigma_{it}^2} u_{it} u_{it}'.$$

The Newton-Raphson algorithm updates the current values $\lambda^{(0)}$ to $\lambda^{(1)}$ using

$$\lambda^{(1)} = \lambda^{(0)} + \Delta\lambda, \qquad \Delta\lambda = -\Big(\frac{\partial^2 Q}{\partial \lambda \partial \lambda'} \Big)^{-1} \frac{\partial Q}{\partial \lambda}. \quad (10)$$

For the generalized EM algorithm, we don't need to do a full iteration of the Newton-Raphson. We only need to make sure that $Q(\lambda)$ increases at each iteration, using partial stepping such as step-halving if necessary. Step-halving works as follows. If $Q(\lambda^{(1)}) \leq Q(\lambda^{(0)})$, we replace $\Delta\lambda$ by its half in the update $\lambda^{(1)} = \lambda^{(0)} + \Delta\lambda$, and continue doing so until $Q(\lambda^{(1)}) > Q(\lambda^{(0)})$.

The steps of the algorithm are summarized as follows:

(i) Initialization: set $\Sigma_i^* = I$, $i = 1, \ldots, n$.

(ii) Using the current estimates of γ and λ (or Σ_i^* in the first iteration), compute the updated estimate $\tilde{\beta}$ of β using equation (3).

(iii) Compute $\widehat{V}_i$, $i = 1, \ldots, n$, where the relevant conditional expectations are calculated using (8).

(iv) Using the current estimates of β and λ, update γ using (9).

(v) Using the current estimates of β and γ, update the current estimate $\lambda^{(0)}$ to $\lambda^{(1)}$ using one step of Newton-Raphson as (10). Use step-halving to guarantee that the criterion is increased.

(vi) Iterate (ii)–(v) until convergence.

Asymptotic Inference

The asymptotic covariance matrix of the parameters can be computed after the EM algorithm following Oakes (1999). The observed information of (γ, λ) evaluated at $(\tilde{\gamma}, \tilde{\lambda})$ can be approximated by

$$\begin{pmatrix} \sum_{i=1}^m S(\tilde{\gamma}; y_i) S^t(\tilde{\gamma}; y_i) & \sum_{i=1}^m S(\tilde{\gamma}; y_i) S^t(\tilde{\lambda}; y_i) \\ \sum_{i=1}^m S(\tilde{\lambda}; y_i) S^t(\tilde{\gamma}; y_i) & \sum_{i=1}^m S(\tilde{\lambda}; y_i) S^t(\tilde{\lambda}; y_i) \end{pmatrix}, \tag{11}$$

where

$$S(\tilde{\gamma}; y_i) = \frac{\partial Q_i}{\partial \gamma}\bigg|_{\gamma=\tilde{\gamma}} = \sum_{t=1}^m \sigma_{it}^{-2} (Z_{it}' \widehat{V}_{it}^{(t-1)} - Z_{it}' \widehat{V}_i^{(t-1)} Z_{it}\gamma)\bigg|_{\gamma=\tilde{\gamma}}$$

and

$$S(\tilde{\lambda}; y_i) = \frac{\partial Q_i}{\partial \lambda}\bigg|_{\lambda=\tilde{\lambda}} = -\frac{1}{2} \sum_{t=1}^m \Big(1 - \frac{\widehat{\mathrm{RS}}_{it}}{\sigma_{it}^2}\Big) u_{it}\bigg|_{\lambda=\tilde{\lambda}},$$

where Q_i is the term in Q corresponding to subject i. The asymptotic covariance matrix of the maximum likelihood estimate $(\hat{\gamma}, \hat{\lambda})$ is obtained as the inverse of the observed information matrix (11), evaluated at the estimated parameter values. Since $\hat{\beta}$ and $(\hat{\gamma}, \hat{\lambda})$ are asymptotically independent, the asymptotic covariance matrix of $\hat{\beta}$ is estimated by $(\sum_{i=1}^m X_i' \widehat{\Sigma}^{-1} X_i)^{-1}$.

References

Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61, 579–602.

Daniels, M.J., 2006. Bayesian modeling of several covariance matrices and some results on propriety of the posterior for linear regression with correlated and/or heterogeneous errors. Journal of Multivariate Analysis 97, 1185–1207.

Dempster, A., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. JRSSB 39, 1–38.

Holan, S., Spinka, C., 2007. Maximum likelihood estimation for joint mean-covariance models from unbalanced repeated-measures data. Statistics & Probability Letters 77, 319–328.

Huang, J.Z., Liu, N., Pourahmadi, M., Liu, L., 2006. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika 93, 85–98.

Jennrich, R.I., Schluchter, M.D., 1986. Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42, 805–820.

Kenward, M.G., 1987. A method for comparing profiles of repeated measurements. *Applied Statistics* 36, 296–308.

Kotz, S., Nadarajah, S., 2004. Multivariate T-Distributions and Their Applications. Cambridge University Press.

Levina, E., Rothman, A., Zhu, J., 2008. Sparse estimation of large covariance matrices via a nested lasso penalty. The Annals of Applied Statistics 2, 245–263.

Lin, T.I., Wang, Y.J., 2009. A robust approach to joint modeling of mean and scale covariance for longitudinal data. Journal of Statistical Planning and Inference 139, 3013–3026.

Oakes, D., 1999. Direct calculation of the information matrix via the em algorithm. Journal of the Royal Statistical Society, Series B 61, 479–482.

Pan, J.X., MacKenzie, G., 2003. On modelling mean-covariance structures in longitudinal studies. Biometrika 90, 239–244.

Pourahmadi, M., 1999. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. Biometrika 86, 667–690.

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

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

Rubin, D., 1976. Inference and missing data (with discussion). Biometrika 63, 581–592.

Yap, J.S., Fan, J., Wu, R., 2009. Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. Biometrics 65, 1068–1077.

Ye, H., Pan, J.X., 2006. Modelling of covariance structures in generalised estimating equations for longitudinal data. Biometrika 93, 927–994.

Zimmerman, D.L., Núñez Antón, V., 2010. Antedependence Models for longitudinal Data. Chapman & Hall / CRC Press, New York.