Biometrika (0000), **00**, 0, *pp.* 1–10
© 0000 Biometrika Trust
Printed in Great Britain

Advance Access publication on 000

A Moving Average Cholesky Factor Model in Covariance Modeling for Longitudinal Data

BY WEIPING ZHANG

Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, China

zwp@ustc.edu.cn

AND CHENLEI LENG

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Republic of Singapore

stalc@nus.edu.sg

SUMMARY

We propose new regression models for parameterising covariance structures in longitudinal data analysis. Using a novel Cholesky factor, the entries in this decomposition have a moving average and log innovation interpretation and are modeled as linear functions of covariates. We propose efficient maximum likelihood estimates for joint mean-covariance analysis based on this decomposition and derive the asymptotic distributions of the coefficient estimates. Furthermore, we study a local search algorithm, computationally more efficient than traditional all subset selection, based on BIC for model selection, and show its model selection consistency. Thus, a key conjecture made by Pan & MacKenzie (2003) is verified. We demonstrate the finite-sample performance of the proposed method via analysis of the data on CD4 trajectories and through simulations.

Some key words: BIC; Longitudinal data analysis; Maximum likelihood estimation; Model selection; Modified Cholesky decomposition; Moving average.

1. INTRODUCTION

Longitudinal data, with repeated measurements collected from the same subject, are frequently encountered. A variety of regression models for the mean analysis have been studied (Diggle et al., 2002). Recently, regression analysis of the covariance structure, aiming at providing parsimonious models for characterising the dependence structure among repeated measurements, has attracted increasing attention. Pourahmadi (1999, 2000) first introduced a modified Cholesky decomposition to factor the inverse covariance matrix. An attractive property of this decomposition is to provide unconstrained parametrisation for the positive definite covariance matrix. More importantly, the entries in this decomposition can be interpreted as autoregressive parameters and log innovation variances in a time series context. Regression models can then be applied to these entries in a manner similar to the mean model, thus permitting parsimonious characterisation of the covariance structure. See Pan & MacKenzie (2003), Ye & Pan (2006), Pourahmadi (2007) and Leng et al. (2010) for related discussion.

In this paper, we use a new Cholesky factor for analyzing the within-subject variation by decomposing the covariance matrix rather than its inverse. The entries in this decomposition are moving average parameters and log innovation variances. Thus, covariance modeling is brought closer to time series analysis, for which the moving average model may provide an alternative, equally powerful and parsimonious representation. We propose new regression models for the mean-covariance analysis in this decomposition, and show that the maximum likelihood estimates are asymptotically normal and fully efficient. Furthermore, the resulting mean, the moving average coefficients and the log innovation coefficients are asymptotically independent. This result leads to a computational strategy which reduces the complexity of the traditional BIC-based model selection method using all subset selection. We rigorously establish the consistency of this model selection strategy. Our result can be used to prove a conjecture of Pan & MacKenzie (2003) that their model selection algorithm is consistent. Rothman et al. (2010) used the proposed decomposition in conjunction with regularisation to analyze large dimensional covariance matrices when a banded structure exists. Our model is more general.

2. THE MODEL AND THE ESTIMATION METHOD

Denote the response vector of the ith subject by $y_i = (y_{i1}, \dots, y_{im_i})^{\mathrm{T}}$ $(i = 1, \dots, n)$, whose components are observed at times $t_i = (t_{i1}, \dots, t_{im_i})^{\mathrm{T}}$. We assume that the response vector is normally distributed as $y_i \sim N(\mu_i, \Sigma_i)$. By allowing m_i and t_{ij} to be subject specific, our approach can handle datasets that are observed at irregular times and are highly unbalanced.

To parameterise Σ_i, Pourahmadi (1999) first proposed to decompose it as $T_i \Sigma_i T_i^{\mathrm{T}} = D_i$. The lower triangular matrix T_i is unique with 1's on its diagonal and the below-diagonal entries of T_i are the negative autoregressive parameters ϕ_{ijk} in the model $y_{ij} - \mu_{ij} = \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik}) + \epsilon_{ij}$. The diagonal entries of D_i are the innovation variances as $\sigma_{ij}^2 = \mathrm{var}(\epsilon_{ij})$.

By setting $L_i = T_i^{-1}$, a lower triangular matrix with 1's on its diagonal, we can write $\Sigma_i = L_i D_i L_i^{\mathrm{T}}$. The entries l_{ijk} in L_i can be interpreted as the moving average coefficients in

$$y_{ij} - \mu_{ij} = \sum_{k=1}^{j-1} l_{ijk}\varepsilon_{ik} + \varepsilon_{ij},\ (j = 2, \dots, m_i), \tag{1}$$

where $\varepsilon_{i1} = y_{i1} - \mu_{i1}$ and $\varepsilon_i \sim N(0, D_i)$ for $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im_i})^{\mathrm{T}}$. The parameters l_{ijk} and $\log(\sigma_{ij}^2)$ are unconstrained.

Since the main difference between our decomposition and that in Pourahmadi (1999) is whether to use T_i or its inverse, it is helpful to examine these two decompositions for commonly used covariance matrices. If Σ is a compound symmetry p by p matrix given by $\sigma^2\{(1-\rho)I + \rho J\}$ where J is a matrix of ones, then the decomposition in Pourahmadi (1999) gives $\phi_{jk} = \rho\{1 + (j-2)\rho\}^{-1}$, while for our decomposition $l_{jk} = \rho\{1 + (k-1)\rho\}^{-1}$. If Σ is an AR(1) matrix with elements $\Sigma_{ij} = \sigma^2(\rho^{|i-j|})_{i,j=1}^{p}$, then Pourahmadi's decomposition gives $\phi_{j,j-1} = \rho, \phi_{jk} = 0, k < j-1$ and ours gives $l_{jk} = \rho^{|j-k|}$.

To parsimoniously parameterise the mean-variance structure in terms of covariates, we impose the regression models

$$g(\mu_{ij}) = x_{ij}^{\mathrm{T}}\beta,\ l_{ijk} = z_{ijk}^{\mathrm{T}}\gamma,\ \log(\sigma_{ij}^2) = h_{ij}^{\mathrm{T}}\lambda,$$

motivated by Pourahmadi (1999, 2000) and Pan & MacKenzie (2003). Here $g(\cdot)$ is a monotone and differentiable known link function, and x_{ij}, z_{ijk} and h_{ij} are $p \times 1$, $q \times 1$ and $d \times 1$ vectors of covariates, respectively. The covariates x_{ij} and h_{ij} are those used in regression analysis, while z_{ijk} is usually taken as a polynomial of time difference $t_{ij} - t_{ik}$ or that of time dependent covariates. Later, we shall refer to the three regression models collectively as moving average models, and the regression models in Pourahmadi (1999, 2000) as autoregressive models. For Gaussian data, we use the identity function for $g(\cdot)$.

Write the twice negative log-likelihood function, up to a constant, as

$$A(\beta,\gamma,\lambda) = -2l(\beta,\gamma,\lambda) = \sum_{i=1}^{n} \log|L_i D_i L_i^{\mathrm{T}}| + \sum_{i=1}^{n} r_i^{\mathrm{T}} L_i^{-\mathrm{T}} D_i^{-1} L_i^{-1} r_i$$

for $r_{ij} = y_{ij} - \mu_{ij}$. By taking partial derivatives of $A(\beta,\gamma,\lambda)$ with respect to these parameters respectively, the maximum likelihood estimating equations become

$$\begin{aligned} U_1(\beta;\gamma,\lambda) &= \sum_{i=1}^{n} X_i^{\mathrm{T}} \Delta_i \Sigma_i^{-1} \{y_i - \mu(X_i\beta)\} = 0, \\ U_2(\gamma;\beta,\lambda) &= \sum_{i=1}^{n} \left(\frac{\partial \varepsilon_i^{\mathrm{T}}}{\partial \gamma}\right) D_i^{-1} \varepsilon_i = 0, \\ U_3(\lambda;\beta,\gamma) &= \sum_{i=1}^{n} H_i^{\mathrm{T}} (D_i^{-1} f_i - 1_{m_i}) = 0. \end{aligned} \tag{2}$$

Here $\Delta_i = \Delta_i(X_i\beta) = \mathrm{diag}\{\dot{g}^{-1}(x_{ij}^{\mathrm{T}}\beta),\ldots,\dot{g}^{-1}(x_{im_i}^{\mathrm{T}}\beta)\}$, $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse of the link function $g^{-1}(\cdot)$ and we have used the notation $\mu(\cdot) = g^{-1}(\cdot)$, $\varepsilon_i = (\varepsilon_{i1},\ldots,\varepsilon_{im_i})^{\mathrm{T}}$ with $\varepsilon_{ij} = r_{ij} - \sum_{k=1}^{j-1} l_{ijk}\varepsilon_{ik}$. Also $\partial\varepsilon_i^{\mathrm{T}}/\partial\gamma$ is a $q \times m_i$ matrix with the first column zero and the jth ($j > 2$) column $\partial\varepsilon_{ij}/\partial\gamma = -\sum_{k=1}^{j-1}(\varepsilon_{ik} z_{ijk} + l_{ijk}\partial\varepsilon_{ik}/\partial\gamma)$; $H_i = (h_{i1}^{\mathrm{T}},\ldots,h_{im_i}^{\mathrm{T}})^{\mathrm{T}}$, $f_i = (f_{i1},\ldots,f_{im_i})^{\mathrm{T}}$ with $f_{ij} = \varepsilon_{ij}^2$ and 1_{m_i} is a vector of 1's. The parameters ε_{ij} and $\partial\varepsilon_{ij}/\partial\gamma$ are defined recursively as is usual in moving average models.

Since the solutions of β, γ and λ satisfy the equations in (2), these parameters are solved iteratively by fixing the others. An application of the quasi-Fisher scoring algorithm on (2) directly yields the numerical solutions for these parameters. Details on the expectations of the Hessian are discussed in the Supplementary material. More specifically, the algorithm works as follows.

1. Initialize the parameters as $\beta^{(0)}$, $\gamma^{(0)}$ and $\lambda^{(0)}$. Set $k = 0$.
2. Compute Σ_i using $\gamma^{(k)}$ and $\lambda^{(k)}$. Update β by

$$\beta^{(k+1)} = \beta^{(k)} + \left[\left(\sum_{i=1}^{n} X_i^{\mathrm{T}} \Delta_i \Sigma_i^{-1} \Delta_i X_i\right)^{-1} \sum_{i=1}^{n} X_i^{\mathrm{T}} \Delta_i \Sigma_i^{-1} \{y_i - \mu_i(X_i\beta)\}\right]\Bigg|_{\beta=\beta^{(k)}}. \tag{3}$$

3. Given $\beta = \beta^{(k+1)}$ and $\lambda = \lambda^{(k)}$, update γ via

$$\gamma^{(k+1)} = \gamma^{(k)} - \left[\left\{\sum_{i=1}^{n} G_i(\gamma)\right\}^{-1} U_2(\gamma;\beta,\lambda)\right]\Bigg|_{\gamma=\gamma^{(k)}}, \tag{4}$$

where

$$G_i(\gamma) = \sum_{j=2}^{m_i} \frac{1}{\sigma_{ij}^2} \left(Z_{ij} + \sum_{k=1}^{j-1} a_{ijk} Z_{ik}\right) D_i \left(Z_{ij} + \sum_{k=1}^{j-1} a_{ijk} Z_{ik}\right)^{\mathrm{T}},$$

$Z_{ij} = (z_{ij1},\ldots,z_{ij(j-1)},0,\ldots,0)$ is a $q \times m_i$ matrix and a_{ijk} is the (j,k)th element of L_i^{-1}. The $d \times m_i$ matrix $Z_{i1} = 0$.

4. Given $\beta = \beta^{(k+1)}$ and $\gamma = \gamma^{(k+1)}$, update λ using

$$\lambda^{(k+1)} = \lambda^{(k)} + \left(\sum_{i=1}^{n} H_i^{\mathrm{T}} H_i \right)^{-1} U_3(\lambda; \beta, \gamma) \Big|_{\lambda = \lambda^{(k)}}. \qquad (5)$$

5. Set $k \leftarrow k + 1$ and repeat Steps 2–5 until a prespecified convergence criterion is met.

This algorithm converges only to a local optimum which depends critically on the initial values. A natural starting value for β is to use identity matrices for the variance matrices Σ_i in (3). Then we initiate γ in (4) assuming $D_i = I_i$. It is not difficult to see that these initial estimates are $\sqrt{n}$-consistent. From the theoretical analysis in Theorem 1 in Section 3 and the proofs in the Supplementary material, the negative log-likelihood function is asymptotically convex around a small neighborhood of the true parameters. This ensures that asymptotically, the final estimates obtained by this iterative algorithm, denoted as $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\lambda}$, are the global optima and are more efficient than the initial values. For our data analysis and simulation studies, convergence was usually obtained within ten iterations.

3. ASYMPTOTIC PROPERTIES AND MODEL SELECTION

3·1. *Asymptotic properties*

Since we use maximum likelihood for estimation, the resulting estimators are efficient. To formally establish the theoretical properties, we impose the following regularity conditions.

Condition A1: The dimensions p, q and d of covariates x_{ij}, z_{ijk} and h_{ij} are fixed; $n \to \infty$ and $\max_i m_i$ is bounded.

Condition A2: The true value $\theta_0 = (\beta_0^{\mathrm{T}}, \gamma_0^{\mathrm{T}}, \lambda_0^{\mathrm{T}})^{\mathrm{T}}$ is in the interior of the parameter space Θ that is a compact subset of $\mathbb{R}^{p+q+d}$.

Condition A3: When $n \to \infty$, $I(\theta_0)/n$ converges to a positive definite matrix $\mathcal{I}(\theta_0)$.

Conditions A1 and A2 are standard in longitudinal data analysis. The asymptotic property of the maximum likelihood estimation involves the negative of the expected Hessian matrix $I(\theta) = -E(\partial^2 l / \partial\theta\partial\theta^{\mathrm{T}})$ with $\theta = (\beta^{\mathrm{T}}, \gamma^{\mathrm{T}}, \lambda^{\mathrm{T}})^{\mathrm{T}}$, where the expectation is conditional on the covariates x_{ij}, z_{ijk} and h_{ij}. Condition A3 is standard in regression analysis. Formally, we have the following asymptotic results for the maximum likelihood estimates.

THEOREM 1. *If $n \to \infty$ and regularity conditions A1–A3 hold, then: (a) the maximum likelihood estimator $(\hat{\beta}^{\mathrm{T}}, \hat{\gamma}^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ is strongly consistent for the true value $(\beta_0^{\mathrm{T}}, \gamma_0^{\mathrm{T}}, \lambda_0^{\mathrm{T}})^{\mathrm{T}}$; and (b) $(\hat{\beta}^{\mathrm{T}}, \hat{\gamma}^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ is asymptotically distributed as*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\left[0, \{\mathcal{I}(\theta_0)\}^{-1}\right]$$

where $\mathcal{I}(\theta_0) = diag(\mathcal{I}_{11}, \mathcal{I}_{22}, \mathcal{I}_{33})$ is a block diagonal matrix with $\mathcal{I}_{11} \in \mathbb{R}^{p \times p}$, $\mathcal{I}_{22} \in \mathbb{R}^{q \times q}$ and $\mathcal{I}_{33} \in \mathbb{R}^{d \times d}$.

It follows that $\hat{\beta}, \hat{\gamma}$ and $\hat{\lambda}$ are asymptotically independent. Following (2), it is shown in the Supplementary material that the block diagonal components of $I(\theta)$ satisfy

$$I_{11}(\theta) = \sum_{i=1}^{n} X_i^{\mathrm{T}} \Delta_i \Sigma_i^{-1} \Delta_i X_i, \qquad I_{33}(\theta) = \frac{1}{2} \sum_{i=1}^{n} H_i^{\mathrm{T}} H_i,$$

$$I_{22}(\theta) = \sum_{i=1}^{n} \sum_{j=2}^{m_i} \frac{1}{\sigma_{ij}^2} \left(Z_{ij} + \sum_{k=1}^{j-1} a_{ijk} Z_{ik} \right) D_i \left(Z_{ij} + \sum_{k=1}^{j-1} a_{ijk} Z_{ik} \right)^{\mathrm{T}}.$$

Since $\hat{\beta}, \hat{\gamma}$ and $\hat{\lambda}$ are consistent estimators for θ_0, $\mathcal{I}$ in the asymptotic covariance matrix is consistently estimated by a block diagonal matrix with block components

$$\hat{\mathcal{I}}_{ii} = n^{-1} I_{ii}(\hat{\theta}) \quad (i = 1, 2, 3). \tag{6}$$

3·2. *Model selection*

A standard method to choose the optimal model for the mean and covariance structures is based on the Bayesian information criterion, BIC. We discuss a computationally efficient algorithm which gives consistent models.

For notational purposes, we use the generic notation $S = (S_\beta, S_\gamma, S_\lambda)$ to denote an arbitrary candidate model where $S_\beta = \{j_1, \ldots, j_{p^*}\}$ includes $X_{ij_1}, \ldots, X_{ij_{p^*}}$ as the relevant predictors to model the mean; $S_\gamma = \{k_1, \ldots, k_{q^*}\}$ and $S_\lambda = \{l_1, \ldots, l_{d^*}\}$ are similarly defined. The true model is denoted as $S^o = (S^o_\beta, S^o_\gamma, S^o_\lambda)$ where, for example, S^o_β is the vector consisting of all the non-zero coefficients of β_0. We define the family of overfitted models as $\mathcal{S}^+ = (\mathcal{S}^+_\beta, \mathcal{S}^+_\gamma, \mathcal{S}^+_\lambda)$ and that of underfitted models as $\mathcal{S}^- = (\mathcal{S}^-_\beta, \mathcal{S}^-_\gamma, \mathcal{S}^-_\lambda)$. Thus, for any $S_\beta \in \mathcal{S}^+_\beta$, we have $S^o_\beta \subset S_\beta$ and for any $S_\beta \in \mathcal{S}^-_\beta$, we have $S^o_\beta \not\subset S_\beta$, and so on. Let $|S|$ denote the size of the model S, that is, $|S| = p^* + q^* + d^*$. Next, define $\beta_{S_\beta} = (\beta_{j_1}, \ldots, \beta_{j_{p^*}})^T$, and similarly, γ_{S_γ}, λ_{S_λ} and $\theta_S = (\beta^T_{S_\beta}, \gamma^T_{S_\gamma}, \lambda^T_{S_\lambda})^T$. With this notation, we define

$$\mathrm{BIC}(S_\beta, S_\gamma, S_\lambda) = -\frac{2}{n} l(\hat{\theta}_S) + |S| \frac{\log(n)}{n}, \tag{7}$$

where $\hat{\theta}_S$ is the maximum likelihood estimate of θ for the model S. Shao (1997) and Shi and Tsai (2002) demonstrated that (7) can identify the true model consistently, if a finite dimension true model exists and the predictor dimension is fixed. To use it for model selection, we apply all subset selection by fitting 2^{p+q+d} models and choose the model that gives the minimum value.

Here we study a computing algorithm which drastically reduces the complexity of all subset selection, motivated by the asymptotic independence of $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\lambda}$ in Theorem 1. Similar to Pan and Mackenzie (2003), we propose a search strategy for finding the optimum model by using the following searches involving the likelihood obtained by saturating the parameter sets in pairs:

$$\min_{S_\beta \in \mathcal{S}^+_\beta \cup \mathcal{S}^-_\beta} \mathrm{BIC}(S_\beta, F_\gamma, F_\lambda), \quad \min_{S_\gamma \in \mathcal{S}^+_\gamma \cup \mathcal{S}^-_\gamma} \mathrm{BIC}(F_\beta, S_\gamma, F_\lambda), \quad \min_{S_\lambda \in \mathcal{S}^+_\lambda \cup \mathcal{S}^-_\lambda} \mathrm{BIC}(F_\beta, F_\gamma, S_\lambda),$$

where $F_\beta, F_\gamma,$ and F_λ denote the full models for the mean, the moving average and the log innovations respectively. Thus, we only need to apply all subset selection for a particular set of coefficients in β, γ or λ, by using full models for the other two sets of coefficients. This strategy requires us to compare the BIC values of $2^p + 2^q + 2^d$ models, which is computationally much less demanding than all subset selection. The model selection consistency of this algorithm is established in the following theorem.

THEOREM 2. *If the conditions in Theorem 1 hold, we have that as $n \to \infty$,*

$$\mathrm{pr}\Big\{\min_{S_\beta \in \mathcal{S}_\beta^+ \cup \mathcal{S}_\beta^-} \textsc{bic}(S_\beta, F_\gamma, F_\lambda) > \textsc{bic}(S_\beta^o, F_\gamma, F_\lambda)\Big\} \to 1,$$

$$\mathrm{pr}\Big\{\min_{S_\gamma \in \mathcal{S}_\gamma^+ \cup \mathcal{S}_\gamma^-} \textsc{bic}(F_\beta, S_\gamma, F_\lambda) > \textsc{bic}(F_\beta, S_\gamma^o, F_\lambda)\Big\} \to 1,$$

$$\mathrm{pr}\Big\{\min_{S_\lambda \in \mathcal{S}_\lambda^+ \cup \mathcal{S}_\lambda^-} \textsc{bic}(F_\beta, F_\gamma, S_\lambda) > \textsc{bic}(F_\beta, F_\gamma, S_\lambda^o)\Big\} \to 1.$$

Pan & MacKenzie (2003) used a similar algorithm for selecting polynomial orders in Pourahmadi's model and gave empirical evidence of its success. They conjectured that their algorithm is model selection consistent. Due to the asymptotic independence of the maximum likelihood estimates in their model, their conjecture follows directly from the proof of Theorem 2.

4. DATA ANALYSIS AND SIMULATIONS

4·1. *CD4 cell data*

We apply the proposed estimation method to the CD4 cell study previously analyzed by Zeger & Diggle (1994) and Ye & Pan (2006). This dataset comprises CD4 cell counts of 369 HIV-infected men, a total of 2376 values, measured at different times for each individual, over a period of approximately eight and a half years. The number of measurements for each individual varies from 1 to 12 and the time points are not equally spaced, so this is a highly unbalanced dataset. We use square root transformation of the response (Zeger & Diggle, 1994) to relate the CD4 counts to six covariates: time since seroconversion t_{ij}; age relative to arbitrary origin x_{ij1}; packs of cigarettes smoked per day x_{ij2}; recreational drug use x_{ij3}; number of sexual partners x_{ij4}; and mental illness score x_{ij5}. To model jointly the mean and covariance structures, we use the mean model (Diggle et al., 2002)

$$y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + \beta_5 x_{ij5} + f(t_{ij}) + e_{ij},$$

where $f(t) = \beta_0 + \beta_6 t_+ + \beta_7 t_+^2$ with $t_+ = tI(t > 0)$. We use cubic polynomials to model the log-innovation variance and the moving average parameters, that is, $h_{ij} = (1, t_{ij}, t_{ij}^2, t_{ij}^3)^{\mathrm{T}}$ and $z_{ijk} = \{1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2, (t_{ij} - t_{ik})^3\}^{\mathrm{T}}$. Our model yields the following estimated mean parameters, with standard errors as subscripts, $\beta_0 = 28{\cdot}9_{0{\cdot}4}$, $\beta_1 = 0{\cdot}019_{0{\cdot}030}$, $\beta_2 = 0{\cdot}65_{0{\cdot}12}$, $\beta_3 = 0{\cdot}60_{0{\cdot}31}$, $\beta_4 = 0{\cdot}062_{0{\cdot}037}$, $\beta_5 = -0{\cdot}040_{0{\cdot}013}$, $\beta_6 = -4{\cdot}53_{0{\cdot}26}$, $\beta_7 = 0{\cdot}58_{0{\cdot}06}$, and estimated moving average coefficients and log innovation coefficients $\gamma_0 = 0{\cdot}60_{0{\cdot}05}$, $\gamma_1 = -0{\cdot}20_{0{\cdot}08}$, $\gamma_2 = 0{\cdot}056_{0{\cdot}033}$, $\gamma_3 = -0{\cdot}0054_{0{\cdot}0042}$; $\lambda_0 = 3{\cdot}30_{0{\cdot}04}$, $\lambda_1 = -0{\cdot}13_{0{\cdot}03}$, $\lambda_2 = -0{\cdot}031_{0{\cdot}012}$, and $\lambda_3 = 0{\cdot}0080_{0{\cdot}0032}$.

We now use the model selection method proposed in Section 3 to select the optimal model. Table 1 lists BIC values for the selected models by our local search strategy and all subset selection. The coefficients of the time dependent polynomial $f(t)$ are all significant, which agrees with Diggle et al. (2002). For the moving average coefficients, Table 1 shows that the moving average parameter l_{ijk} can be represented by a linear function of the time lag $t_{ij} - t_{ik}$. Table 1 also gives the all subset selection result by fitting 2^{p+q+d} models and shows that the optimal model coincides with the final model chosen by our proposed computing method.

We compare our approach with the autoregressive decomposition models in Pourahmadi (1999) and Pan & MacKenzie (2003). We use the same full models for the mean, the autoregressive and the log-innovation coefficients. We apply the proposed computing approach for model selection and the results are presented in Table 1. First, our model selection algorithm gives the

Table 1. *Model selection results for the CD4 data using the proposed search strategy*

	MA			AR	
	Model	BIC		Model	BIC
Full model		38·80	Full model		39·03
Mean	$(\beta_0, \beta_2, \beta_5, \beta_6, \beta_7)$	38·78	Mean	$(\beta_0, \beta_2, \beta_5, \beta_6, \beta_7)$	39·00
Moving Average	(γ_0, γ_1)	38·79	Autoregressive	$(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$	39·03
Log Innovation	$(\lambda_0, \lambda_1, \lambda_2, \lambda_3)$	38·80	Log Innovation	$(\lambda_0, \lambda_1, \lambda_3)$	39·03
Best Subset	$(\beta_0, \beta_2, \beta_5, \beta_6, \beta_7)$ (γ_0, γ_1) $(\lambda_0, \lambda_1, \lambda_2, \lambda_3)$	38·76	Best Subset	$(\beta_0, \beta_2, \beta_5, \beta_6, \beta_7)$ $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ $(\lambda_0, \lambda_1, \lambda_3)$	39·00

MA, moving average decomposition; AR, autoregressive decomposition

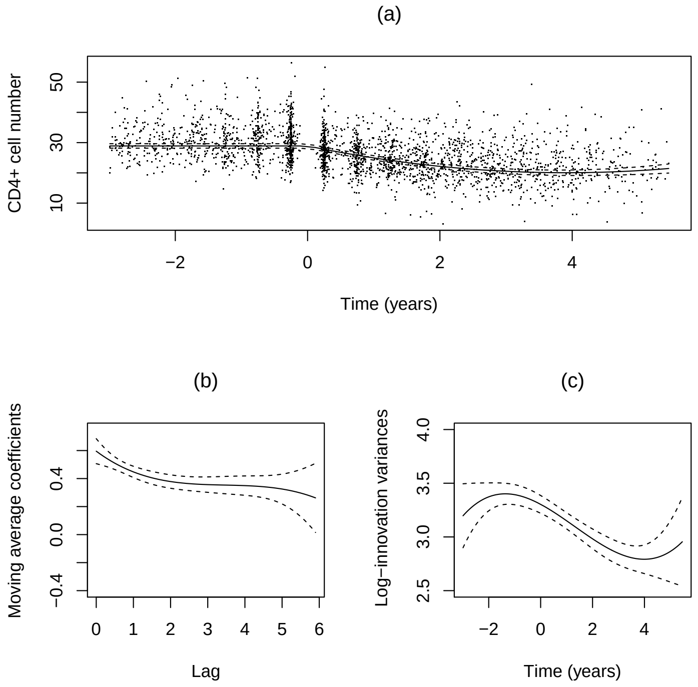

Fig. 1. Results for the CD4 data: (a) fitted polynomial mean against time, (b) generalized moving average parameters against lag, and (c) log innovation variances against time. Dashed curves represent asymptotic 95% pointwise confidence intervals.

same final model as that using all subset selection, whether the moving average or the autoregressive model is used. Second, both models give the same final mean model, with cigarette smoking, mental illness score and a polynomial function of the time as the important covariates. Third, the full model for the autoregressive decomposition has a larger BIC value, indicating that the new moving average decomposition may have certain advantages for this dataset. The final chosen model with our approach has 11 coefficients while the autoregressive decomposition has 12 with a larger BIC, so our model is more parsimonious with a better fit.

Figure 1 displays the fitted curves for the polynomial for the mean, moving average as a function of the time lag and the log innovation variances. The mean trajectory is consistent with that in Zeger & Diggle (1994) and Leng et al. (2010). Figure 1(b) plots the estimated moving average parameters l_{ijk} against the time lag between measurements in the same subject, which, according to our model, is simply a cubic polynomial and can be simplified to a linear function of the time lag. This plot indicates that the moving average parameter l_{ijk} is close to 0·6 if the time difference $t_{ik} - t_{ij}$ is small, and gradually decreases when time difference increases. This is reasonable because observations closer in time would be more correlated as seen from (1). On the other hand, Table 1 shows that the autoregressive model would need a cubic polynomial to represent the autoregressive parameters in terms of time lag. These observations agree with those in Ye & Pan (2006) and Leng et al. (2010).

To compare these two optimal models, we apply leave-one-subject-out cross-validation to assess the predictive performance in terms of $\sum_{i=1}^m \|y_i - \hat{y}_i\|/m$ and $\sum_{i=1}^m \|\Sigma_i - \hat{\Sigma}_i\|/m$, where $\hat{y}_i$ and $\hat{\Sigma}_i$ represent the predicted response and covariance matrix for subject i and Σ_i is the empirical covariance matrix. The moving average decomposition yields 13·68 and 244·03 for these two quantities, while the autoregressive decomposition gives 13·66 and 250·58 respectively. The proposed decomposition gives prediction accuracy in estimating the response similar to the autoregressive decomposition, while outperforming the latter in modeling the covariance structure.

4·2. *Simulation studies*

In this section we investigate the finite sample performance of the proposed estimation and inference methods. For each setup, we generate 1000 data sets. We consider sample sizes $n = 100,\ 200$ or 500. Each subject is measured m_i times with $m_i - 1 \sim Binomial(11, 0{\cdot}8)$, and then the measurement times t_{ij} are generated from the uniform distribution on the unit interval. This results in differential total numbers of repeated measurements m_i between subjects.

Study 1. The data sets are generated from the model

$$y_{ij} = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2 + x_{ij3}\beta_3 + e_{ij},\ (i = 1, \ldots, n; \quad j = 1, \ldots, n_i),$$

while the moving average coefficients and log innovation variances are generated by

$$l_{ijk} = \gamma_0 + z_{ijk1}\gamma_1 + z_{ijk2}\gamma_2 + z_{ijk3}\gamma_3,\ \log(\sigma_{ij}^2) = \lambda_0 + h_{ij1}\lambda_1 + h_{ij2}\lambda_2 + h_{ij3}\lambda_3,$$

where $x_{ij} = (x_{ij1}, x_{ij2}, x_{ij3})^{\mathrm{T}}$ is generated from a multivariate normal distribution with mean zero, marginal variance 1 and all correlations 0·5. Motivated by the CD4 data analysis, we take $h_{ij} = x_{ij}$ and $z_{ijk} = \{1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2, (t_{ij} - t_{ik})^3\}^{\mathrm{T}}$. Table 2 shows the accuracy of the estimated parameters in terms of their mean absolute biases. All the biases are small, especially when n is large. To evaluate the inference procedure, we compare the sample standard deviation of 1000 parameter estimates to the sample average of 1000 standard errors using formula (6). Table 2 demonstrates that they are quite close, especially for large n. This indicates that the standard error formula works well.

Study 2. We use the settings in Study 1 to assess the consistency of the model selection method in Theorem 2 by setting $\beta = (1, -0{\cdot}5, 0, 0)^{\mathrm{T}}$, $\gamma = (-0{\cdot}3, 0{\cdot}2, 0, 0)^{\mathrm{T}}$ and $\lambda = (-0{\cdot}3, 0{\cdot}5, 0, 0)^{\mathrm{T}}$. Table 3 shows the empirical percentage of the models which are incorrectly selected over 1000 replications. These results show that under various sample sizes, the proposed model selection method has the desired performance.

Study 3. We use the settings in Study 1 to compare the two factorisations under different data generating process. The main measurements for comparison are differences between the fitted mean $\hat{\mu}_i$ and the true mean μ_i, and the fitted covariance matrix $\hat{\Sigma}_i$ to the true Σ_i. In particular, we define two relative errors as $\mathrm{ERR}(\hat{\mu}_i) = \|\hat{\mu}_i - \mu_i\|/\|\mu_i\|$ and $\mathrm{ERR}(\hat{\Sigma}_i) = \|\hat{\Sigma}_i - \Sigma_i\|/\|\Sigma_i\|$.

We compute the averages of these two quantities for 1000 replications with $n = 100$ or 200. Table 4 gives the averages for the moving average decomposition and autoregressive decomposi-

Table 2. *Simulation results for the estimation of parameters in Study 1* (all the values are multiplied by a factor 10^2)

	True value	$n = 100$ MAB	$n = 100$ SE$_{\text{SD}}$	$n = 200$ MAB	$n = 200$ SE$_{\text{SD}}$	$n = 500$ MAB	$n = 500$ SE$_{\text{SD}}$
β_0	1	0·40	$0{\cdot}48_{0{\cdot}51}$	0·28	$0{\cdot}34_{0{\cdot}36}$	0·17	$0{\cdot}21_{0{\cdot}21}$
β_1	$-0{\cdot}5$	1·16	$1{\cdot}36_{1{\cdot}46}$	0·74	$0{\cdot}94_{0{\cdot}93}$	0·48	$0{\cdot}60_{0{\cdot}61}$
β_2	0	1·15	$1{\cdot}36_{1{\cdot}43}$	0·77	$0{\cdot}95_{0{\cdot}95}$	0·48	$0{\cdot}60_{0{\cdot}61}$
β_3	0·5	1·12	$1{\cdot}35_{1{\cdot}43}$	0·76	$0{\cdot}96_{0{\cdot}95}$	0·46	$0{\cdot}60_{0{\cdot}58}$
γ_0	$-0{\cdot}3$	0·57	$0{\cdot}64_{0{\cdot}71}$	0·38	$0{\cdot}45_{0{\cdot}47}$	0·24	$0{\cdot}29_{0{\cdot}29}$
γ_1	0·2	2·71	$3{\cdot}07_{3{\cdot}40}$	1·87	$2{\cdot}17_{2{\cdot}34}$	1·05	$1{\cdot}37_{1{\cdot}34}$
γ_2	0	3·03	$3{\cdot}35_{3{\cdot}78}$	1·95	$2{\cdot}36_{2{\cdot}45}$	1·22	$1{\cdot}49_{1{\cdot}53}$
γ_3	0·5	6·56	$7{\cdot}49_{8{\cdot}33}$	4·57	$5{\cdot}27_{5{\cdot}69}$	2·62	$3{\cdot}31_{3{\cdot}28}$
λ_0	$-0{\cdot}3$	3·77	$4{\cdot}52_{4{\cdot}48}$	2·68	$3{\cdot}20_{3{\cdot}29}$	1·62	$2{\cdot}02_{2{\cdot}01}$
λ_1	0·5	4·58	$5{\cdot}55_{5{\cdot}73}$	3·20	$3{\cdot}92_{3{\cdot}95}$	1·94	$2{\cdot}48_{2{\cdot}43}$
λ_2	0·4	4·46	$5{\cdot}54_{5{\cdot}63}$	3·06	$3{\cdot}92_{3{\cdot}89}$	1·94	$2{\cdot}47_{2{\cdot}44}$
λ_3	0	4·62	$5{\cdot}54_{5{\cdot}72}$	3·10	$3{\cdot}92_{3{\cdot}88}$	2·02	$2{\cdot}47_{2{\cdot}54}$

MAB, the estimated mean absolute bias; SD, the sample standard deviation of 1000 estimates; SE, the average of standard error

Table 3. *Percentage of incorrectly selected models for Study 2*

n	Mean	Moving Average	Log Innovation
100	5·7	5·2	5·9
200	4·4	4·8	4·5
500	2·8	3·2	2·3

Table 4. *Study 3: Average of relative errors* $\text{ERR}(\hat{\mu}) = \sum_{l=1}^{n} \text{ERR}(\hat{\mu}_l)/n$ *and* $\text{ERR}(\hat{\Sigma}) = \sum_{l=1}^{n} \text{ERR}(\hat{\Sigma}_l)/n$ *for* $n = 100$ *or* 200

	Fit	MA		AR	
True	n	$\text{ERR}(\hat{\mu}) \times 10^2$	$\text{ERR}(\hat{\Sigma}) \times 10^2$	$\text{ERR}(\hat{\mu}) \times 10^2$	$\text{ERR}(\hat{\Sigma}) \times 10^2$
MA	100	1·72	5·74	8·66	47·09
	200	1·15	4·09	6·10	45·43
AR	100	3·48	31·36	2·98	10·33
	200	2·43	30·27	2·04	7·26

MA, moving average factorisation; AR, autoregressive factorisation

tion, when the data are either generated from our model or the model in Pourahmadi (1999). For the latter, instead of using the model in Study 1 for l_{ijk}, we use this model for ϕ_{ijk}. We see that when the true covariance matrix follows the moving average structure, the errors in estimating μ and Σ both increase when incorrectly decomposing the covariance matrix using the autoregressive structure, and vice versa. However, for this simulation study, model mis-specification seems to affect the moving average decomposition to a lesser degree.

5. Discussion

We have proposed new models based on a Cholesky decomposition with moving average interpretation as an alternative to the autoregressive models in Pourahmadi (1999). Which one is preferred is likely to be data dependent. In practice, we may rely on a combination of graphical tools such as regressograms (Pourahmadi, 1999), suitable for balanced data sets, and numerical tools such as cross-validation to choose an appropriate factorisation and parameterisation. An illustration of the former is presented in the Supplementary material. The latter was demonstrated in the CD4 data analysis as well as in the Supplementary material. If a clear trend is spotted in the sample regressogram, the corresponding factorisation may be preferred. Quantitatively, we can always employ cross-validation for comparing the predictive performance for estimating the mean and the observed covariance. A more accurate prediction is an indication to use the corresponding decomposition.

With autoregressive and moving average models available to parameterise the covariance matrix, it is of interest to unite the two by studying the autoregressive moving average model. Furthermore, when nonlinearity arises, more flexible models such as semiparametric mean-covariance models can be considered (Fan et al., 2007; Fan & Wu, 2008; Leng et al., 2010).

Acknowledgments

We thank the Editor, as associate editor and two referees for their insightful comments and constructive suggestions. Zhang is supported by the National Science Foundation of China. Leng is supported by National University of Singapore grants.

Supplementary material

Supplementary material available at *Biometrika* online includes an analysis of a cattle data set, the derivation of the Hessian matrix, and the proofs of Theorem 1 and 2.

References

Diggle, P. J., Heagerty P. J., Liang, K. Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press. Oxford, UK.

Fan, J., Huang, T. & Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Am. Statist. Assoc.* **102**, 632–641.

Fan, J. & Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *J. Am. Statist. Assoc.* **103**, 1520–1533.

Leng, C., Zhang, W. & Pan, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *J. Am. Statist. Assoc.* **105**, 181–193.

Pan, J. & Mackenzie, G. (2003). Model selection for joint 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**, 677–90.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425–35.

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

Rothman, A. J., Levina, E. & Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97**, 539–550.

Shao, J. (1997) An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 221–264.

Shi, P. & Tsai, C. L. (2002) Regression model selection — a residual likelihood approach. *J. R. Statist. Soc. B* **64**, 237–252.

Ye, H. & Pan, J. (2006). Modeling covariance structures in generalized estimating equations for longitudinal data. *Biometrika* **93**, 927–941.

Zeger, S. L. & Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–699.

[*Received July* 2010. *Revised* *** 2011]