ARMA Cholesky factor models for the covariance matrix of linear models

Keunbaik Lee\textsuperscript{a,}\textsuperscript{*}, Changryong Baek\textsuperscript{a}, Michael J. Daniels\textsuperscript{b,}\textsuperscript{c}

\textsuperscript{a} Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Republic of Korea
\textsuperscript{b} Department of Statistics & Data Sciences, University of Texas at Austin, Austin, TX, United States
\textsuperscript{c} Department of Integrative Biology, University of Texas at Austin, Austin, TX, United States

A R T I C L E   I N F O

Article history:
Received 25 August 2016
Received in revised form 22 February 2017
Accepted 6 May 2017
Available online xxxx

Keywords:
Cholesky decomposition
Longitudinal data
Heteroscedastic

A B S T R A C T

In longitudinal studies, serial dependence of repeated outcomes must be taken into account to make correct inferences on covariate effects. As such, care must be taken in modeling the covariance matrix. However, estimation of the covariance matrix is challenging because there are many parameters in the matrix and the estimated covariance matrix should be positive definite. To overcome these limitations, two Cholesky decomposition approaches have been proposed: modified Cholesky decomposition for autoregressive (AR) structure and moving average Cholesky decomposition for moving average (MA) structure, respectively. However, the correlations of repeated outcomes are often not captured parsimoniously using either approach separately. In this paper, we propose a class of flexible, nonstationary, heteroscedastic models that exploits the structure allowed by combining the AR and MA modeling of the covariance matrix that we denote as ARMACD.

We analyze a recent lung cancer study to illustrate the power of our proposed methods. © 2017 Elsevier B.V. All rights reserved.

1. Introduction

In a longitudinal study, an outcome variable is observed repeatedly over a period of time. Since the repeated observations are not independent, their dependence must be taken into account to make proper inference on covariate effects in terms of correct standard errors (Diggle et al., 2002). In addition, in the presence of missing data (which is common in longitudinal studies), correctly modeling the covariance matrix is necessary to avoid bias in (mean) covariate effects (Daniels and Hogan, 2008). However, estimation of the covariance matrix is challenging because there are many parameters and the estimated covariance matrix needs to be positive definite. To remove these restrictions, simple structures for the covariance matrix are often assumed such as exchangeable or autoregressive order one. However, such statistical approaches do not permit more general forms of the correlation structure and cannot explain heterogeneous covariances (Pourahmadi, 1999). To overcome these limitations, joint mean–variance modeling approaches have been proposed for the covariance matrix (Pourahmadi, 1999; Chiu et al., 1996; Zhang and Leng, 2012; Hoff and Niu, 2012) and the correlation matrix (Daniels and Pourahmadi, 2009; Zhang et al., 2015). In this paper, we focus on modeling of the covariance matrix. Pourahmadi (1999) and Zhang and Leng (2012) proposed Cholesky-type decomposition approaches for modeling of covariance matrix of longitudinal data, the modified Cholesky decomposition and the moving average Cholesky decomposition, respectively.

In the modified Cholesky decomposition (MCD), the inverse of the covariance matrix is decomposed into generalized autoregressive parameters (GARPs) and innovation variances (IVs); the positive-definite restriction corresponds to the IVs...
being positive (Pourahmadi, 1999, 2000). Bayesian modeling using the MCD was proposed in Daniels and Pourahmadi (2002) and Daniels and Zhao (2003). Pan and MacKenzie (2003, 2006) extended the MCD to deal with unbalanced longitudinal data and proposed the BIC as the optimal approach to identify the degree of the mean, innovation variance and GAR parameters under an assumption that each follows a polynomial model (with the GAR parameters also assumed to just depend on lag). Lee et al. (2012) and Lee (2013) used the MCD to model the random effects covariance matrix in GLMMs and Lee and Sung (2014) extended these to marginalized random effects models (Heagerty, 1999).

In the moving average Cholesky decomposition (MACD), the covariance matrix is decomposed into generalized moving average parameters (GMAPs) and IVs (Zhang and Leng, 2012). This approach uses the moving average parameterization of the covariance matrix as an alternative to the autoregressive one. Similar to the MCD, the covariance matrix is positive definite when the IVs are positive. Lee and Yoo (2014) used this decomposition for modeling of the random effects covariance matrix to analyze long series of longitudinal binary data.

When a high-order AR structure of the covariance matrix is required to capture the dependence structure, practitioners often consider autoregressive moving average (ARMA) models (Judge et al., 1980). The advantage of the ARMA models is that the models provide for a wide variety of structures in the covariance matrix but can be specified using a small number of parameters (Rochen and Helms, 1989). Rochen (1992) extended Rochen and Helms (1989) model, which had a homogeneous ARMA covariance matrix, to allow for a heterogeneous ARMA covariance matrix. In this paper, we also consider the heterogeneous ARMA covariance matrix by ‘combining’ the MCD and MACD to create a more flexible decomposition of the covariance matrix. This new decomposition will still provide simple conditions for positive definiteness (in terms of the IVs) and allow flexible non-stationary and heteroscedastic models that can be more parsimonious than models based on either the MCD or MACD alone. We emphasize the models proposed here are much more general (and flexible) than stationary ARMA models used in time series; we make connections to ‘standard’ ARMA models to emphasize that we try to exploit structure on both autoregressive and moving average type parameters in a non-stationary setting.

This paper is organized as follows. In Section 2, we propose covariance models using the ARMA Cholesky decomposition (ARMACD). We show that the resulting maximum likelihood estimator of the mean and the parameters of the covariance matrix are consistent and asymptotically normally distributed. In Section 3, we examine the bias of mean parameter estimates to misspecification of covariance matrix with and without missing data in simulation studies. In Section 4, we apply our proposed models to data from a recent lung cancer study. Finally, we summarize and propose future work in Section 5.

2. ARMA Cholesky factor models for covariance matrix

We propose a new class of models for the covariance matrix in longitudinal data that relies on the parameters in the new ARMACD.

2.1. Proposed model

Let \( y_i = (y_{i1}, \ldots, y_{in})^T \) be the longitudinal response vector of the \( i \)th subject \((i = 1, \ldots, N)\); \( y_i \) is the response at time \( g_{it} \) for subject \( i \); we assume the observation times are equally spaced, so without loss of generality, we can set \( g_{it} = t \). We assume that the responses for different subjects are independent. Let \( x_{it} \) indicate covariates corresponding to \( y_{it} \). We assume the \( y_{it} \) follows the linear model which is given by

\[
y_{it} - x_{it}^T \beta = \sum_{l=1}^{t-1} \phi_{lt} (y_{it-l} - x_{it-l}^T \beta) + \sum_{j=1}^{t-1} l_{it-t-j} e_{it-j} + e_{it},
\]

(1)

where \( \beta \) is a \( p \times 1 \) coefficients vector of \( x_{it}, e_{i1} = y_{i1} - x_{i1}^T \beta, e_i = (e_{i1}, \ldots, e_{in})^T, e_i \sim N(0, D_i) \), and \( D_i = \text{diag}(\sigma_{i1}^2, \ldots, \sigma_{in}^2) \). The parameters \( \phi_{lt} \) and \( l_{it-t-j} \) are generalized autoregressive parameters (GARPs), generalized moving average parameters (GMAPs), and innovation variances (IVs), respectively. The GARPs and GMAPs are unconstrained. We implicitly assume that model (1) is well-defined.

We can rewrite (1) in matrix form as

\[
T_i(y_i - X_i \beta) = L_ie_i
\]

(2)

where \( T_i \) is a unit lower triangular matrix having ones on its diagonal and \(-\phi_{li,t}\) at its \((t, j)\)th position for \( j < t \), and \( L_i \) is a unique lower triangular matrix having ones on its diagonal and \( l_{i,t-j} \) at its \((t, j)\)th position for \( j < t \), and \( X_i = (x_{i1}, \ldots, x_{in})^T \) is the design matrix. In this paper we assume that \( X_i \) is of full rank. From (2), we have

\[
T_i \Sigma_i T_i^T = L_i D_i L_i^T
\]

(3)

where \( \Sigma_i \) is the covariance matrix for \( y_i \).

Theorem 1. \( \Sigma_i \) in (3) is positive definite if the diagonal elements of \( D_i \) (i.e., the IVs) are all positive.

Please cite this article in press as: Lee, K., et al., ARMA Cholesky factor models for the covariance matrix of linear models. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.05.001.
The GARP$s$, GMAP$s$, and IV$s$ can be modeled using time and/or subject-specific covariate vectors $w_{i,j}$ and $h_{i,t}$ by setting

$$\phi_{i,j} = w_{i,j}^T \alpha, \quad h_{i,t} = z_{i,t}^T \gamma,$$

$$\log(\sigma^2_{i,j}) = h_{i,t}^T \lambda,$$

where $\alpha$, $\gamma$, and $\lambda$ are $a \times 1$, $b \times 1$, and $c \times 1$ vectors of unknown parameters, respectively. Note that we jointly model the mean and covariance structures of responses in terms of the generalized linear models from (1) and (4).

The design vectors $w_{i,j}$, $z_{i,j}$, and $h_{i,t}$ are used to model the GARP/GMAP/IV parameters. These design vectors can include subject-specific covariates (so heteroscedastic regression) and/or incorporate structure. For the latter, $h_{i,t} = (1, t)$ corresponds to the innovation variances change (log-)linearly with time and corresponds to a nonstationary covariance structure. Parsimonious higher lag models can be specified by including time lag, $|t-j|$ in the design vectors $w_{i,j}$ and/or $z_{i,j}$. Finally, structured nonstationary (in the GARP and GMAP) models can be specified; for example, a nonstationary first order moving average process could be specified using $z_{i,t} = (1, t, \ldots, (t-j) \times 1)$. Overall, the covariance matrix with these models (4) includes heteroscedastic and nonstationary processes for $y_{i,t}$.

Note that the parameters in the decomposition are not identifiable without any structure of the covariance matrix. When there are no subject-specific covariates in the covariance matrix, the matrices $T$ and $L$ in (2) have $T(T - 1)/2 + T(T - 1)/2$ parameters where $T = \max(n_i)$. However, with a specific structure of the covariance matrix such as those based on the ARMA Cholesky structure, both matrices can be identified and the identifiability is easily assessed by checking invertibility of the Hessian matrix discussed in Section 2.3. Also note that our ARMACD models have the parameters of the MCD and MACD as special cases; so our model extends previous models and provides a unified framework for modeling the covariance structure of linear models in longitudinal studies.

The advantage of ARMA modeling in time series analysis is well known. For example, parsimony in parameterization makes the model interpretation easy and provides stable estimation of parameters. In addition, it also provides better forecasting performance than competing higher order AR or MA model. For example, Hansen and Lunde (2005) pointed out that including both AR and MA structures in the financial GARCH model leads to no worse performance than other sophisticated models in forecasting. Along similar lines, we expect ARMACD also to enjoy many useful advantages of ARMA modeling. In addition, in our setting of longitudinal data, we can construct more complex (but still parsimonious) ‘ARMA’ types models that allow nonstationary and heteroscedasticity as described above.

## 2.2. Maximum likelihood estimation

We derive the likelihood function for the model specified in Section 2.1. Let $\theta = (\beta^T, \alpha^T, \gamma^T, \lambda^T)$. The likelihood function is a product of multivariate normal probability densities,

$$L(\theta; y) = \prod_{i=1}^{N} (2\pi)^{-n_i/2} |\Sigma_i|^{-1/2} \exp \left\{ -\frac{1}{2} (y_i - X_i \beta)^T \Sigma_i^{-1} (y_i - X_i \beta) \right\}$$

$$= \prod_{i=1}^{N} (2\pi)^{-n_i/2} \left( \prod_{t=1}^{n_i} \sigma^2_{i,t} \right)^{-1/2} \exp \left\{ -\frac{1}{2} (y_i - X_i \beta)^T T_i^{-T} L_i^{-1} L_i^{-T} T_i (y_i - X_i \beta) \right\}.$$ 

(5)

Recall from (3) that $\Sigma_i = T_i^{-1} L_i D_i L_i^T T_i^{-T}$.

The log likelihood is given by

$$\log L(\theta; y) = \sum_{i=1}^{N} \left\{ -\frac{n_i}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{n_i} h_{i,t}^T \lambda - \frac{1}{2} \sum_{t=1}^{n_i} r_i (r_i - W_i \alpha)^T L_i^{-T} D_i^{-1} L_i^{-T} r_i (r_i - W_i \alpha) \right\},$$

where $r_i = y_i - X_i \beta$ and

$$W_i = \begin{pmatrix}
0^T \\
(y_{i1} - x_{i1}^T \beta) w_{11}^T \\
\vdots \\
(y_{iT} - x_{iT}^T \beta) w_{T1}^T \\
\sum_{j=1}^{n_i - 1} (y_{ij} - x_{ij}^T \beta) w_{1j}^T \\
\vdots \\
\sum_{j=1}^{n_i - 1} (y_{ij} - x_{ij}^T \beta) w_{Tj}^T 
\end{pmatrix}$$

with $0^T$ being $1 \times 1$ vector of zeros.
Maximizing the log-likelihood with respect to \( \theta \) yields the likelihood equations
\[
\sum_{i=1}^{N} \frac{\partial \log L(\theta; y_i)}{\partial \theta} = 0.
\]

where
\[
\frac{\partial \log L(\theta; y_i)}{\partial \beta} = X_i^T \Sigma_i^{-1} (y_i - X_i \beta),
\]
(6)
\[
\frac{\partial \log L(\theta; y_i)}{\partial \alpha} = W_i^T L_i^{-T} D_i^{-1} L_i^{-1} (r_i - W_i \alpha),
\]
(7)
\[
\frac{\partial \log L(\theta; y_i)}{\partial \gamma_j} = -\frac{1}{2} (y_i - X_i \beta)^T \left( T_i^T \frac{\partial L_i}{\partial \gamma_j} D_i^{-1} L_i^{-1} T_i + T_i^T L_i^{-1} D_i^{-1} \frac{\partial L_i}{\partial \gamma_j} T_i \right) (y_i - X_i \beta),
\]
(8)
\[
\frac{\partial \log L(\theta; y_i)}{\partial \lambda_l} = -\frac{1}{2} \left\{ \sum_{t=1}^{n} h_{it} + (y_i - X_i \beta)^T T_i^T L_i^{-1} \frac{\partial D_i^{-1}}{\partial \lambda_l} L_i^{-1} T_i (y_i - X_i \beta) \right\},
\]
(9)

with
\[
\frac{\partial L_i}{\partial \gamma_j} = \begin{pmatrix}
0 & 0 & 0 & \cdots & 0 \\
\xi_{1j2} & 0 & 0 & \cdots & 0 \\
\xi_{2j2} & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\xi_{n_{1j2}} & \xi_{n_{2j2}} & \xi_{n_{3j2}} & \cdots & 0
\end{pmatrix},
\]
\[
\frac{\partial \gamma_j}{\partial \lambda_l} = \text{diag} \left\{ \frac{h_{1j1}}{\sigma_{1j1}^2}, \frac{h_{1j2}}{\sigma_{1j2}^2}, \ldots, \frac{h_{njn}}{\sigma_{njn}^2} \right\},
\]

for \( j = 1, \ldots, b \) and \( l = 1, \ldots, c \).

From (6), we have the maximum likelihood estimator (MLE) for \( \beta \) which is given by
\[
\hat{\beta} = \left( \sum_{i=1}^{N} X_i^T \Sigma_i^{-1} X_i \right)^{-1} \sum_{i=1}^{N} X_i^T \Sigma_i^{-1} y_i,
\]
where \( \hat{\Sigma}_i = \Sigma_i(\hat{\alpha}; \hat{\gamma}, \hat{\lambda}) \) and (\( \hat{\alpha}, \hat{\gamma}, \hat{\lambda} \)) are the ML estimator of \( (\alpha, \gamma, \lambda) \). From (7), we also have the MLE for \( \alpha \) which is given by
\[
\hat{\alpha} = \left( \sum_{i=1}^{N} \hat{W}_i^T \hat{L}_i^{-T} \hat{D}_i^{-1} \hat{L}_i^{-1} \hat{W}_i \right)^{-1} \sum_{i=1}^{N} \hat{W}_i^T \hat{L}_i^{-1} \hat{D}_i^{-1} \hat{L}_i^{-1} \hat{r}_i,
\]
where \( \hat{W}_i = W_i(\hat{\beta}), \hat{L}_i = L_i(\hat{\gamma}), \hat{D}_i = D_i(\hat{\lambda}), \) and \( \hat{r}_i = y_i - X_i \hat{\beta}. \) Because we do not have closed forms of the solutions of \( \gamma \) and \( \lambda \) from (8) and (9), we use an iterative procedure to find the MLE.

Since the analytic forms of second derivatives of the observed data log-likelihood in (8) and (9) are not available in closed form, we use quasi-Newton methods to solve the likelihood equations. Let \( \omega = (\gamma^T, \lambda^T) \). Then the \( (c+1) \)th iteration \( \omega^{(c+1)} \) is updated using
\[
\omega^{(c+1)} = \omega^{(c)} + \left[ H(\omega^{(c)}; y) \right]^{-1} \frac{\partial \log L}{\partial \omega^{(c)}},
\]

where
\[
\frac{\partial \log L}{\partial \omega} = \left( \sum_{i=1}^{N} \frac{\partial \log L(\theta; y_i)}{\partial \theta^T}, \sum_{i=1}^{N} \frac{\partial \log L(\theta; y_i)}{\partial \lambda} \right)^T,
\]
\[
H(\omega; y) = \text{diag} \left\{ E \left( -\frac{\partial^2 \log L}{\partial \gamma \partial \gamma^T} \right), E \left( -\frac{\partial^2 \log L}{\partial \lambda \partial \lambda^T} \right) \right\},
\]

Please cite this article in press as: Lee, K., et al., ARMA Cholesky factor models for the covariance matrix of linear models. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.05.001.
with
\[
E \left( -\frac{\partial^2 \log L}{\partial \gamma' \partial \gamma} \right) = \frac{1}{2} \text{tr} \left( B_i \Sigma_i \right),
\]
\[
E \left( -\frac{\partial^2 \log L}{\partial \lambda' \partial \lambda} \right) = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{n_i} h_{iit} h_{igt},
\]
(10)

having
\[
B_i = T_i^T \frac{\partial^2 L^{-T}}{\partial \gamma' \partial \gamma} D_i^{-1} L_i^{-1} T_i + T_i^T \frac{\partial L_i^{-1}}{\partial \gamma} D_i^{-1} \frac{\partial L_i^{-1}}{\partial \gamma} T_i + T_i^T \frac{\partial L_i^{-T}}{\partial \gamma} D_i^{-1} \frac{\partial L_i^{-T}}{\partial \gamma} T_i + T_i^T L_i^{-T} D_i^{-1} \frac{\partial^2 L_i^{-1}}{\partial \gamma' \partial \gamma} T_i.
\]

### 2.3. Asymptotic properties

Since we use maximum likelihood for estimation, the resulting estimators are efficient. Let \( \theta = (\beta^T, \omega^T, \lambda^T)^T \), where \( \omega = (\alpha, \gamma) \), and \( \theta_0 = (\beta_0^T, \omega_0^T, \lambda_0^T)^T \) is the true value of \( \theta \). Also let \( I(\theta_0) = \text{diag}(I_1, I_2, I_3) \) where
\[
I(\beta) = \sum_{i=1}^{N} X_i \Sigma_i^{-1} X_i,
\]
\[
I(\omega) = \begin{pmatrix}
    E \left( -\frac{\partial^2 \log L}{\partial \alpha' \partial \alpha} \right) \\
    E \left( -\frac{\partial^2 \log L}{\partial \gamma' \partial \alpha} \right) \\
    E \left( -\frac{\partial^2 \log L}{\partial \gamma' \partial \gamma} \right)
\end{pmatrix},
\]
\[
I(\lambda) = E \left( -\frac{\partial^2 \log L}{\partial \lambda' \partial \lambda} \right).
\]

Finally, \( E \left( -\frac{\partial^2 \log L}{\partial \alpha' \partial \beta} \right) = \sum_{i=1}^{N} E \left[ W_i^T L_i^{-T} D_i^{-1} L_i^{-1} W_i \right], \)
\[
E \left( -\frac{\partial^2 \log L}{\partial \gamma' \partial \alpha} \right) = \frac{1}{2} \sum_{i=1}^{N} \text{tr} \left( A_i \Sigma_i \right), \)
\[
E \left( -\frac{\partial^2 \log L}{\partial \gamma' \partial \gamma} \right) = \frac{1}{2} \sum_{i=1}^{N} \text{tr} \left( D_i C_i \right)
\]
each are given in (10), where
\[
A_i = \frac{\partial T_i^T}{\partial \alpha} \frac{\partial L_i^{-T}}{\partial \gamma'_a} D_i^{-1} L_i^{-1} T_i + \frac{\partial T_i^T}{\partial \alpha} L_i^{-T} D_i^{-1} \frac{\partial L_i^{-1}}{\partial \gamma_a} T_i + T_i^T \frac{\partial L_i^{-T}}{\partial \gamma_a} D_i^{-1} \frac{\partial L_i^{-T}}{\partial \gamma_a} T_i + T_i^T L_i^{-T} D_i^{-1} \frac{\partial^2 L_i^{-1}}{\partial \gamma'_a \partial \gamma_a} T_i,
\]
\[
C_i = \frac{\partial L_i^T}{\partial \gamma} \left( \frac{\partial L_i^{-1}}{\partial \gamma} \right)' D_i^{-1} L_i + \frac{\partial L_i^{-1}}{\partial \gamma'} D_i^{-1} \frac{\partial L_i^{-1}}{\partial \gamma} + \frac{\partial L_i^{-1}}{\partial \gamma} D_i^{-1} \frac{\partial L_i^{-1}}{\partial \gamma} + L_i^{-T} D_i^{-1} \frac{\partial}{\partial \gamma} \left( \frac{\partial L_i^{-1}}{\partial \gamma} \right) L_i.
\]

To establish the theoretical properties, we assume the following regularity conditions hold.

**C1** The dimensions \( p, a, b, c \) of covariates \( x_i, u_{i,j}, z_{i,g}, \) and \( h_{it} \), and \( \max n_i \) are fixed.

**C2** The true value \( \theta_0 = (\beta_0^T, \alpha_0, \gamma_0, \lambda_0^T)^T \) is in the interior of the parameter space \( \Theta \), which is a compact subset of \( \mathbb{R}^{p+a+b+c} \).

**C3** The covariates \( x_i, u_{i,j}, z_{i,g}, \) and \( h_{it} \) are all bounded, meaning that all elements of the vectors are bounded.

**C4** When \( N \to \infty \), \( I(\theta_0)/N \) converges to a positive definite matrix \( \mathcal{I}(\theta_0) \).

Then we have the following results:

**Theorem 2.** Under regularity conditions given in (C1)–(C4), the following three results hold as \( N \to \infty \):

1. \( E \left( -\frac{\partial^2 \log L}{\partial \alpha' \partial \beta} \right) \to 0 \), \( E \left( -\frac{\partial^2 \log L}{\partial \gamma' \partial \beta} \right) \to 0 \), \( E \left( -\frac{\partial^2 \log L}{\partial \alpha \partial \beta} \right) \to 0 \), \( E \left( -\frac{\partial^2 \log L}{\partial \gamma \partial \beta} \right) \to 0 \);

2. The maximum likelihood estimator \( \hat{\theta} = (\hat{\beta}^T, \hat{\alpha}^T, \hat{\gamma}^T, \hat{\lambda}^T)^T \) is strongly consistent for \( \theta_0 = (\beta^T, \alpha^T, \gamma^T, \lambda^T)^T \);

3. The maximum likelihood estimator is asymptotically normal
\[
\sqrt{N} \left( \hat{\theta} - \theta_0 \right) \to^d N(0, \mathcal{I}^{-1}(\theta_0)),
\]
(11)

where \( \mathcal{I}(\theta_0) = \text{diag}(\mathcal{I}(\beta_0), \mathcal{I}(\alpha_0), \mathcal{I}(\gamma_0)) \) with \( \mathcal{I}(\beta_0) \in \mathbb{R}^p, \mathcal{I}(\alpha_0) \in \mathbb{R}^{a+b}, \mathcal{I}(\lambda_0) \in \mathbb{R}^c \).

**Proof.** See Appendix A.
We note that (C4) in the above theorem corresponds to parameter identifiability. (C4) will hold if the following three conditions hold:

1. the design matrix for the mean parameters is full column rank
2. the design matrix for the (innovation) variance parameters is full column rank
3. \( I(\omega_0) \) is invertible.

For the third condition, if there are just GMAP or just GARP parameters and the corresponding design matrix is full column rank, then \( I(\omega_0) \) is invertible. When the models include both GMAP and GARP parameters, the necessary and sufficient condition is less clear.

Notice that result 1 in Theorem 2 implies that \( \hat{\beta}, \hat{\omega} = (\hat{\alpha}^T, \hat{\gamma}^T)^T \), and \( \hat{\lambda} \) are asymptotically independent. Since \( \hat{\beta}, \hat{\alpha}, \hat{\gamma}, \) and \( \hat{\lambda} \) are consistent estimators, \( \hat{\beta} \) in the asymptotic covariance is consistently estimated by a block diagonal matrix with block components

\[
\hat{\beta} = N^{-1} I(\hat{\beta}), \\
\hat{\omega} = N^{-1} I(\hat{\omega}), \\
\hat{\lambda} = N^{-1} I(\hat{\lambda}).
\]

However, we note that the observed information matrix is not orthogonal unless all the \( n_i \)'s are equal. In Section 3, we examine the operating characteristics of our models in the common scenario of dropout for longitudinal data, for which the orthogonality in Theorem 2 between the mean and dependence parameters does not hold since the \( n_i \)'s are not equal (Daniels and Hogan, 2008). Next, we suggest a graphical approach to find a reasonable structured model.

2.4. Model building and selection

In this section we explore the covariance structure using a graphical method which extends that in Pourahmadi (1999) and Daniels and Pourahmadi (2002). We first assume a particular mean structure \((x_i^T \beta)\). Then we propose a procedure to explore the covariance structure using the following steps.

Step 1 Compute the MLEs of \( \beta \) and \( \Sigma \) (assuming no structure), \( \hat{\beta} \) and \( \hat{\Sigma} \) respectively. Then compute the estimated GARPs from \( \hat{\Sigma} \) using the following formula (Pourahmadi, 1999):

\[
\phi_t = \Sigma_t^{-1} \sigma_t,
\]

where \( \Sigma_t \) is the \((t-1) \times (t-1)\) leading principal submatrix of \( \Sigma \) and \( \sigma_t \) is the column vector composed of the first \( t-1 \) entries of the \( t \)th column of \( \Sigma \). Now plot regressograms as in Pourahmadi (1999) and Daniels and Pourahmadi (2002) to find parsimonious structure. These would be plots of the GARP s versus lag and/or plots of particular lag GARPs versus time. See the implementation in the data example in Section 4.

Step 2 Based on the structure chosen from the exploratory analysis of the GARP in Step 1, compute the MLE of \( \beta \) and \( \Sigma \). Here we denote the MLE of \( \Sigma \) under this structure as \( \hat{\Sigma} \). Compute \( \hat{T}_i \) from this and create transformed data, \( \hat{T}_i(y_i - x_i^T \beta) \).

Now compute the MLE of \( \beta \) and the new \( \Sigma \) using the transformed data.

Compute the estimated GMAPs from the MLE of the new \( \Sigma \) by calculating the new \( \hat{T}_i \) from the MLE of the new \( \Sigma \) and then setting \( \hat{L}_i = \hat{T}_i^{-1} \) to obtain the estimated GMAPs.

Plot regressograms similar to Step 1, to explore parsimonious structures for the GMAPs.

Step 3 Refit the model using the parsimonious structures identified for \( \hat{T}_i \) and \( \hat{L}_i \) in Steps 1 and 2. Plot the diagonal elements of \( D \) versus time to explore the structure in the IVs (versus time).

Note that we can pick a covariance structure heuristically using above procedure. Then we can more formally select the best model using likelihood ratio tests and/or penalized model selection criteria such as AIC and BIC. And also note that the exploratory analysis can ‘choose’ slightly different models if Steps 1 and 2 are switched.

3. Simulation study

We conducted several simulation studies to examine the bias, efficiency, and coverage in estimating the coefficients in the mean model in the settings of both true and misspecified structures of the covariance matrix and under both ignorable (MAR) missingness and complete data.

Study 1 We simulated longitudinal data under our proposed models (1) and (4) with two covariates, group (2 levels) and \( c_{it} = (t - 10/2) + \epsilon_{it}, \epsilon_{it} \sim N(0, 0.5^2) \). The model was specified as

\[
y_{it} - x_i^T \beta = \phi_{it,t-1}(y_{it-1} - x_i^T \beta) + l_{it,t-1}^\epsilon_{it-1} + \epsilon_{it},
\]

for \( t = 1, \ldots, 10 \) with \( x_i^T \beta = \beta_0 + \beta_1 \text{Group}_1 + \beta_2 c_{it} + \beta_3 \text{Group}_1 \times c_{it} \). The true mean parameter values were \((\beta_0, \beta_1, \beta_2, \beta_3) = (0.1, -0.1, 0.1, 0.1)\); the ‘group’ covariate, Group, equals 0 or 1 with probability 0.5, \( \phi_{ij} = \alpha \times I_{i | j = 1}, l_{ij} = \gamma \times I_{i | j = 1}, \) and \( \log \sigma_{ij}^2 = \lambda \) with \((\alpha, \gamma, \lambda) = (0.7, 0.7, 3)\).
1.00, which demonstrates the potential utility of the AIC for model selection (at least in these simulation settings). The proportion for Model 1 was 0.11, 0.10, 0.10 in Model 1, 0.08, 0.08, 0.08 in Model 2, and 0.08, 0.08, 0.08 in Model 3. The proportion for Model 1 was 0.10, 0.10, 0.10 in Model 1, 0.09, 0.09, 0.09 in Model 2, and 0.09, 0.09, 0.09 in Model 3. The proportion for Model 1 was 0.09, 0.09, 0.09 in Model 1, 0.09, 0.09, 0.09 in Model 2, and 0.09, 0.09, 0.09 in Model 3. The proportion for Model 1 was 0.08, 0.08, 0.08 in Model 1, 0.08, 0.08, 0.08 in Model 2, and 0.08, 0.08, 0.08 in Model 3. The proportion for Model 1 was 0.07, 0.07, 0.07 in Model 1, 0.07, 0.07, 0.07 in Model 2, and 0.07, 0.07, 0.07 in Model 3. The proportion for Model 1 was 0.06, 0.06, 0.06 in Model 1, 0.06, 0.06, 0.06 in Model 2, and 0.06, 0.06, 0.06 in Model 3. The proportion for Model 1 was 0.05, 0.05, 0.05 in Model 1, 0.05, 0.05, 0.05 in Model 2, and 0.05, 0.05, 0.05 in Model 3. The proportion for Model 1 was 0.04, 0.04, 0.04 in Model 1, 0.04, 0.04, 0.04 in Model 2, and 0.04, 0.04, 0.04 in Model 3. The proportion for Model 1 was 0.03, 0.03, 0.03 in Model 1, 0.03, 0.03, 0.03 in Model 2, and 0.03, 0.03, 0.03 in Model 3. The proportion for Model 1 was 0.02, 0.02, 0.02 in Model 1, 0.02, 0.02, 0.02 in Model 2, and 0.02, 0.02, 0.02 in Model 3. The proportion for Model 1 was 0.01, 0.01, 0.01 in Model 1, 0.01, 0.01, 0.01 in Model 2, and 0.01, 0.01, 0.01 in Model 3. The proportion for Model 1 was 0.00, 0.00, 0.00 in Model 1, 0.00, 0.00, 0.00 in Model 2, and 0.00, 0.00, 0.00 in Model 3. The proportion for Model 1 was 0.00, 0.00, 0.00 in Model 1, 0.00, 0.00, 0.00 in Model 2, and 0.00, 0.00, 0.00 in Model 3.

Based on this specification, the observed dropout rates were approximately 50%.

**Table 1** presents the mean point estimates and the percent relative biases (RB) of the parameters. In the presence of MAR missingness, the estimates in Model 1 had smaller biases than those in Models 2 and 3. Coverage probabilities for all parameters for Model 1 were approximately 95%. In particular, the percent relative biases and coverage probabilities of $\beta_0$ and $\beta_2$ show that the true model (Model 1) fits much better than Models 2 and 3. These results demonstrate that the incorrect covariance structure can induce potentially large biases in mean parameters. To assess the accuracy of our standard errors (cf. Section 2.2), we compared the sample standard deviation (SD) of the 500 parameter estimates to the sample average (SE) of the 500 standard errors using (11); they were very close for the true model which suggests the standard error in (11) is accurate. We also calculated the estimated covariance matrix using the mean of 500 fitted values of $\alpha$, $\gamma$, and $\lambda$ given in Table 1. Then we calculated the sum of absolute differences (SAD) between the estimated and true covariance matrices. The SAD for Model 1 was the smallest among the three models (as expected). In addition, we calculated AICs for all three models and the proportion of times each model had the minimum AIC among the three models. The proportion for Model 1 was 1.00, which demonstrates the potential utility of the AIC for model selection (at least in these simulation settings).

When there were no missing data, the estimates were essentially unbiased for the three models. The coverage probabilities of mean parameters in Models 1 and 2 were all close to 95% with Model 2 somewhat conservative. However, the coverage probabilities in Model 3 show significant undercoverage. As we expected, the SAD for Model 1 was the smallest and the proportion of times the AIC chose Model 1 was 1.00.

**Study 2.** We also generated 500 longitudinal data sets with a sample size of 200 using model (13) with a heterogeneous ARMA(1,1) structure of the covariances matrix by setting $\phi_{ij} = \alpha \times I_{i-j=1}$, $\lambda_{ij} = \gamma \times I_{i-j = 1}$, and $\log \sigma_{R}^2 = \lambda_0 + \lambda_1$ Group, with $(\alpha, \gamma, \lambda_0, \lambda_1) = (.7, .7, .7, .1)$. We again fit three models. Model 1 is true model (a heterogeneous ARMA(1,1) structure of the covariance matrix); Model 2 has an AR(1) structure of the covariance matrix; Model 3 has an MA(1) structure of the covariance matrix. **Table 2** shows the mean point estimates of the parameters for the three models as well as standard errors, relative bias, and coverage.

Similar to Study 1, the estimates in Model 1 had smaller biases than those in Models 2 and 3 when there were MAR missingness. Coverage probabilities for all parameters for Model 1 were all approximately 95% whereas coverage for Models 2 and 3 had some substantial undercoverage. The percent relative biases of $\beta_0$ and $\beta_2$ in Models 2 and 3 were large.

---

**Table 1**
Simulation results for parameter estimates in Study 1. Bias of Models 1–3 for sample sizes of 200 and $T=10$ for 500 simulated data sets under misspecification of the dependence. Displayed is the average regression coefficient estimate (Mean), the percent relative bias (RB) ($|\hat{\beta} - \beta|/\beta \times 100$), the average standard error (SE), the sample standard deviation (SD) of 500 estimates, the coverage probabilities (COVER), sum of absolute difference (SAD) between the estimated and true covariance matrices, and the proportion of times that AIC chose the specified model as best.

<table>
<thead>
<tr>
<th>MAR</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Complete</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean SE SD</td>
<td>RB COVER</td>
<td></td>
<td>Mean SE SD</td>
<td>RB COVER</td>
<td></td>
<td>Mean SE SD</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>0.15 (0.13)</td>
<td>50.0</td>
<td>0.23 (0.14)</td>
<td>130.0</td>
<td>0.06 (0.06)</td>
<td>40.0</td>
<td>0.10 (0.10)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.10 (0.18)</td>
<td>0.0</td>
<td>-0.10 (0.20)</td>
<td>0.0</td>
<td>-0.09 (0.21)</td>
<td>10.0</td>
<td>-0.10 (0.14)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>1.00 (0.02)</td>
<td>10.0</td>
<td>0.13 (0.03)</td>
<td>30.0</td>
<td>0.08 (0.04)</td>
<td>20.0</td>
<td>0.10 (0.02)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.71 (0.02)</td>
<td>1.4</td>
<td>0.89 (0.04)</td>
<td>27.1</td>
<td>0.70 (0.02)</td>
<td>0.0</td>
<td>0.87 (0.02)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.70 (0.02)</td>
<td>0.0</td>
<td>0.96 (0.01)</td>
<td>37.1</td>
<td>0.70 (0.02)</td>
<td>0.0</td>
<td>0.95 (0.01)</td>
</tr>
<tr>
<td>SAD</td>
<td>12.89</td>
<td>32.16</td>
<td>104.18</td>
<td>35.27</td>
<td>103.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(AIC)</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

We generated 500 simulated data sets each with a sample size of 200 under complete data or under MAR missingness. We fit 3 models. Model 1 is the true model (with ARMA(1,1) structure of the covariance matrix); Model 2 has an AR(1) structure of the covariance matrix; Model 3 has an MA(1) structure of the covariance matrix. For missingness, we specified the following MAR dropout model,

$logit(p_{\text{dropout}} = 1 | \text{dropout} \geq 1) = -2.3 + 0.3 Y_{it-1}$.

Based on this specification, the observed dropout rates were approximately 50%.
Table 2
Simulation results for parameter estimates in Study 2. Bias of Models 1–3 for sample sizes of 200 and \( T = 10 \) for 500 simulated data sets under misspecification of the dependence. Displayed is the average regression coefficient estimate (Mean), the percent relative bias (RB) \((\hat{\beta} - \beta)/\beta \times 100\), the sample standard deviation (SD) of 500 estimates, the coverage probabilities (COVERAGE), sum of absolute difference (SAD) between the estimated and true covariance matrices, and the proportion of times that AIC chose the specified model as best.

<table>
<thead>
<tr>
<th>MAR</th>
<th>Mean ( SE_{SD} )</th>
<th>Model 1</th>
<th>RB COVERAGE</th>
<th>Model 2</th>
<th>RB COVERAGE</th>
<th>Model 3</th>
<th>RB COVERAGE</th>
<th>Complete</th>
<th>Mean ( SE_{SD} )</th>
<th>Model 1</th>
<th>RB COVERAGE</th>
<th>Model 2</th>
<th>RB COVERAGE</th>
<th>Model 3</th>
<th>RB COVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>( (0.1) )</td>
<td>0.14</td>
<td>40.0</td>
<td>0.11</td>
<td>0.94</td>
<td>0.22</td>
<td>120.0</td>
<td>0.06</td>
<td>-40.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>( (-0.1) )</td>
<td>-0.09</td>
<td>-10.0</td>
<td>-0.09</td>
<td>-10.0</td>
<td>-0.09</td>
<td>-10.0</td>
<td>-0.14</td>
<td>0.13 ( 0.14 )</td>
<td>-0.10</td>
<td>0.0</td>
<td>-0.10</td>
<td>0.0</td>
<td>-0.11</td>
<td>1.0</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>( (0.1) )</td>
<td>0.11</td>
<td>10.0</td>
<td>0.03</td>
<td>0.93</td>
<td>0.13</td>
<td>30.0</td>
<td>0.08</td>
<td>-30.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>( (0.1) )</td>
<td>0.10</td>
<td>0.0</td>
<td>0.01</td>
<td>0.94</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>( (0.7) )</td>
<td>0.71</td>
<td>1.4</td>
<td>0.89</td>
<td>0.95</td>
<td>0.02</td>
<td>27.1</td>
<td>0.70</td>
<td>0.0</td>
<td>0.70</td>
<td>0.0</td>
<td>0.87</td>
<td>24.3</td>
<td>0.70</td>
<td>0.0</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>( (0.7) )</td>
<td>0.70</td>
<td>0.0</td>
<td>0.96</td>
<td>0.99</td>
<td>0.01</td>
<td>37.1</td>
<td>0.70</td>
<td>0.0</td>
<td>0.70</td>
<td>0.0</td>
<td>0.97</td>
<td>38.6</td>
<td>0.70</td>
<td>0.0</td>
</tr>
<tr>
<td>( \lambda_0 )</td>
<td>( (0.2) )</td>
<td>0.20</td>
<td>0.0</td>
<td>0.47</td>
<td>0.75</td>
<td>0.57</td>
<td>133.0</td>
<td>0.60</td>
<td>200.0</td>
<td>0.20</td>
<td>0.0</td>
<td>0.48</td>
<td>140.0</td>
<td>0.62</td>
<td>210.0</td>
</tr>
<tr>
<td>( \lambda_1 )</td>
<td>( (0.1) )</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.10</td>
<td>0.0</td>
<td>0.11</td>
<td>0.0</td>
</tr>
<tr>
<td>SAD</td>
<td></td>
<td>12.47</td>
<td>46.62</td>
<td>99.40</td>
<td>9.94</td>
<td>39.14</td>
<td>99.30</td>
<td>1.00</td>
<td>0.0</td>
<td>1.00</td>
<td>0.0</td>
<td>0.00</td>
<td>0.0</td>
<td>0.00</td>
<td>0.0</td>
</tr>
<tr>
<td>P(AIC)</td>
<td></td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.0</td>
<td>1.00</td>
<td>0.0</td>
<td>0.00</td>
<td>0.0</td>
<td>0.00</td>
<td>0.0</td>
</tr>
</tbody>
</table>

The SAD for Model 1 was again the smallest among the three models. In addition, proportion of times Model 1 had the minimum AIC was 1.00.

When there were no missing data, the estimates were essentially unbiased for the three models and results were similar to those in Study 1. However, there was slight overcoverage for Model 2 and significant undercoverage for Model 3 due to the misspecified covariance structure. The SAD for Model 1 was again the smallest. The proportion of times Model 1 had the minimum AIC remained at 1.00.

We performed additional simulations to check the flexibility of ARMACD model (results are in the Web Appendix). We simulated data under an AR(1) structure of covariance matrix without MAR dropout, and then fitted an ARMA(1,1) model. We saw the average point estimates of the mean parameters, GARF, GMAP, and IV parameters which were very close to true values. Type I error for \( H_0: \gamma = 0 \) (AR(1) parameter is zero) was zero when AR(1) was the true model. We also conducted the same simulation under a MA(1) structure of covariance matrix and then fitted an ARMA(1,1) model. We saw similar results.

Overall, the simulation results emphasize the importance of correctly modeling the covariance structure on estimation of the mean parameters in the presence of ignoring missing data. In addition, the ARMACD models appear to have reasonable operating characteristics when simpler nested models are the true models. Finally, in the limited simulations, the LRT seems to work well (though a bit conservative) in testing nested ARMACD models.

4. Example
4.1. Lung cancer study

Lung cancer is a leading cause of cancer-related deaths world-wide, with a 5-year survival of less than 15%, because most patients are diagnosed with advanced stage disease. Recently, a longitudinal study on lung cancer was designed as a prospective open-label randomized non-comparative parallel study in a single institution (Kim et al., 2012). A total of 96 patients with lung cancer were randomly assigned to one of two treatments (GEFITINIB or ERLOTINIB). The main goal of the study was to evaluate the response rate (the percentage of patients whose cancer shrinks or disappears after treatment) for each arm. The overall response rates in the GEFITINIB and ERLOTINIB arms were 47.9% (95% CI, 33.8–62.0%) and 39.6% (95% CI, 25.7–53.4%), respectively (Kim et al., 2012); these were not significantly different. Given the non-significant treatment effect on the response rate, it was of interest to see if there was a negative impact of treatments on a patient’s quality of life (QOL).

The European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Lung Cancer (EORTC QLQ-LC13) is considered as a standard instrument to assess the quality of life of lung cancer patients. The EORTC QLQ-LC13 is a 13-item lung cancer-specific questionnaire module comprising both multi-item and single-item measures of lung quality of life.
Table 3
Means of GARP and GMAP by lag for the lung cancer study.

<table>
<thead>
<tr>
<th>LAG</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>GARP</td>
<td>0.463</td>
<td>0.036</td>
<td>0.040</td>
<td>0.009</td>
<td>0.062</td>
<td>−0.044</td>
<td>0.013</td>
<td>0.003</td>
<td>−0.007</td>
<td>−0.001</td>
<td>0.018</td>
<td>0.001</td>
</tr>
<tr>
<td>GMAP</td>
<td>0.449</td>
<td>0.397</td>
<td>0.254</td>
<td>0.152</td>
<td>0.104</td>
<td>0.104</td>
<td>0.041</td>
<td>−0.075</td>
<td>−0.093</td>
<td>−0.084</td>
<td>−0.053</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. GARP versus lag for the lung cancer study.

cancer-associated symptoms (i.e. coughing, hemoptysis, dyspnea and pain) and side-effects from conventional chemo- and radiotherapy (i.e. hair loss, neuropathy, sore mouth and dysphagia). Patients completed the questionnaire before receiving the first dose of treatment at baseline, on day 1 of each subsequent 29-day cycle (4 weeks), and at the end of the study. Each questionnaire was originally scored on a 4-point scale ranging from normal, ‘Not at All’ (1) to abnormal, ‘Very Much’ (4). In this paper, we focus on sum of the 13-item questionnaire module (SQOL).

Covariates considered included type of treatment (ARM = 1 for GEFTINIB, 0 for ERLOTINIB) and week number, TIME (0, 4, 8, . . . , 92); TIME was re-scaled by dividing by 10. We assumed the missing responses (mostly due to dropout) were MAR in these analyses. The Akaike Information Criteria (AIC) (Akaike, 1974) and likelihood ratio tests were used as the model selection criteria.

4.2. Model fit

We explore the covariance structure using the procedure in Section 2.4. From Step 1, we graph the GARP related to lag (Fig. 1). We also calculated means of GARP by lag (Table 3). These suggest that the GARP decreased approximately quadratically in lag. So we considered a quadratic structure in lag for GARP. We also plotted the GMAP from Step 2 (Fig. 2) and calculated the means of GMAP by lag (Table 3). These suggest that the GMAP also decreased roughly quadratically in lag. Finally, we calculated IVs from Step 3 and examined the plot of IVs versus time (Fig. 3). It suggested a quadratic trend in time.

Based on the exploratory analysis described in the previous paragraph, we considered various covariance structures. In particular, we fit the five models for \( \Sigma \) specified in Table 4. We use the following notation: ARMA-?? where the ‘?’ correspond to the polynomial in the GARP, GMAP, and IV respectively or the AR and MA order, respectively. So an ARMA-QQQ would correspond to a quadratic in lag for the GARP and GMAP and quadratic in time for the IV. And an ARMA-20Q would correspond to an AR(2) model with a quadratic in time for the IV.

We used R version 3.1.2 to fit the models and R code is available in supplementary materials (see Appendix B). We initialize the fitting algorithm by setting \( \Sigma_i = \sigma^2I \) and using ordinary linear model estimation methods using R to obtain \( \theta^{(0)} \). We choose \( \alpha^{(0)} = 0, \gamma^{(0)} = 0, \) and \( \lambda^{(0)} = 0 \) (i.e., \( \Sigma_i = I \)) as starting values for \( \alpha, \gamma, \) and \( \lambda \), respectively. We iterate until the sum of absolute differences in the parameters is less than \( 10^{-4} \). Computational burden was not high. Computing time on an Intel
Core i5-2450M CPU with a 2.50 GHz processor took about 20, 60, 300 s for the ARMA-20Q (ARMA-02Q), the ARMA-11Q, ARMA-QQQ, respectively.

4.3. Results

Maximized loglikelihoods and AICs are presented in Table 5. The model ARMA-QQQ fits better than ARMA-QCQ using LRTs ($p$-value = 0.655). The AIC for the model ARMA-11Q was the smallest of the models considered and is the model we use for inference.

Table 6 presents maximum likelihood estimates for models ARMA-20Q, ARMA-02Q, ARMA-11Q, ARMA-QQQ, and ARMA-QCQ. In the model ARMA-11Q, the parameter of GARP was not significant ($\hat{\alpha}_0 = -0.03$, SE = 0.02, $p$-value = 0.134) and that of GMAP was significant ($\hat{\gamma}_0 = 0.52$, SE = 0.04, $p$-value < 0.001). The parameters for the linear and the quadratic...
Fig. 4. Fitted expected values and 95% confidence intervals under the two treatment arms (Gefitinib (circle) and Erlotinib (dark square)).

Table 4
Models for $\phi_{itj}$, $l_{itj}$, and $\log \sigma^2_{it}$ in the lung cancer example.

<table>
<thead>
<tr>
<th>Model</th>
<th>GARP</th>
<th>GMAP</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMA-20Q</td>
<td>$\phi_{itj} = \alpha_0 I(</td>
<td>t-j</td>
<td>=1) + \alpha_1 I(</td>
</tr>
<tr>
<td>ARMA-02Q</td>
<td>$\phi_{itj} = \alpha_0 I(</td>
<td>t-j</td>
<td>=1)$</td>
</tr>
<tr>
<td>ARMA-QQQ</td>
<td>$\phi_{itj} = \alpha_0 + \alpha_1</td>
<td>t-j</td>
<td>+ \alpha_2</td>
</tr>
<tr>
<td>ARMA-QCQ</td>
<td>$\phi_{itj} = \alpha_0 + \alpha_1</td>
<td>t-j</td>
<td>+ \alpha_2</td>
</tr>
</tbody>
</table>

Table 5
Maximized loglikelihoods and AICs for the models considered.

<table>
<thead>
<tr>
<th>Model</th>
<th>ARMA-20Q</th>
<th>ARMA-02Q</th>
<th>ARMA-11Q</th>
<th>ARMA-QQQ</th>
<th>ARMA-QCQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. loglik.</td>
<td>$-2040.5$</td>
<td>$-2129.2$</td>
<td>$-2034.5$</td>
<td>$-2032.8$</td>
<td>$-2032.7$</td>
</tr>
<tr>
<td>AIC</td>
<td>$4099.0$</td>
<td>$4276.4$</td>
<td>$4087.0$</td>
<td>$4091.6$</td>
<td>$4093.4$</td>
</tr>
</tbody>
</table>

of time were significant ($\hat{\lambda}_1 = -0.43$, SE = 0.08, p-value < 0.001; $\hat{\lambda}_2 = 0.05$, SE = 0.01, p-value < 0.001). It indicates that $\log(IV)$ decreased quadratically over time. The parameters of coefficients of covariates in the mean model were not significant. Thus treatment did not appear to differentially impact the mean. Note that there were fairly big differences in the estimated coefficients of covariates in the mean models in Table 6. The estimated parameters for the effect of treatment $\hat{\beta}_1$ in ARMA-02Q, ARMA-QQQ, and ARMA-QCQ were noticeably different from those in the other models. This is consistent with the simulations that show that a misspecified covariance matrix results in biased estimates of parameters in mean model when there is missing data.

Fig. 4 presents the fitted expected values and their 95% confidence intervals under the model ARMA-11Q. Since two confidence intervals were overlapping, there was not much evidence of any difference between the two arms.

5. Conclusions

We have proposed linear models with a covariance matrix that is modeled using the parameters of a new ARMA Cholesky decomposition (ARMACD). These models allow for nonstationary and heteroscedasticity and are more flexible (and potentially more parsimonious) than models using only the parameters of the Modified Cholesky decomposition or Moving Average Cholesky decomposition.

Simulation studies showed the importance of correctly modeling the covariance structure on mean parameters in the presence of ignorable missing data. We also confirmed that our proposed model worked well even when the true data were generated from models with a simpler structure.

Analysis of lung cancer data shows that the covariance matrix was homoscedastic with the ARMA(1,1) structure having innovation variance parameters quadratic in time. There was no significant difference of quality of life between two
treatment arms. Fairly big differences in the estimated covariates effects in the mean models indicate that a misspecified covariance matrix results in biased estimates of parameters in mean model. Future work includes exploring robustness of inference to non-Gaussian continuous outcomes, automated model selection with this class of models, and extensions to non-continuous data by using this class of models for the random effects covariance matrix in generalized linear mixed models (GLMMs).

Acknowledgments

We would like to thank Drs. Myung-Ju Ahn and Insuk Sohn of Samsung Medical Center in South Korea for providing the data and for their help in data collection and clarifying some issues with data. 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, NRF-2016R1D1A1B03930343, NRF-2017R1A1A1A05000831). MJD was partially supported by US NIH grants CA85295 and CA183854.

Appendix A

Proof of Theorem.

1. The derivatives of (6) with respect to $\alpha$, $\gamma$, and $\lambda$, respectively, have the following forms.

$$
\frac{\partial^2 \log L(\theta; y_i)}{\partial \alpha_j \partial \beta^T} = X_i^T \frac{\partial}{\partial \alpha_j} \Sigma_i^{-1}(y_i - \mu_i(\beta)), \text{ for } j = 1, \ldots, a; \\
\frac{\partial^2 \log L(\theta; y_i)}{\partial \gamma_j \partial \beta^T} = X_i^T \frac{\partial}{\partial \gamma_j} \Sigma_i^{-1}(y_i - \mu_i(\beta)), \text{ for } j = 1, \ldots, b; \\
\frac{\partial^2 \log L(\theta; y_i)}{\partial \lambda_j \partial \beta^T} = X_i^T \frac{\partial}{\partial \lambda_j} \Sigma_i^{-1}(y_i - \mu_i(\beta)), \text{ for } j = 1, \ldots, c.
$$

The expectations of each of these are clearly zero since $E[Y_i - \mu_i(\beta)] = 0$.

To examine the other partial derivatives, we can rewrite the log likelihood as follows,

$$
\log L(\theta; y) = \sum_{i=1}^{N} \left\{ -\frac{n_i}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{n_i} \beta_i^T \Sigma^{-1} \beta_i - \frac{1}{2} \beta_i^T D^{-1} e_i \right\}.
$$

Using this form, we take a derivative with respect to $\gamma_j$ for $j = 1, \ldots, c$ to obtain

$$
\frac{\partial \log L(\theta; y)}{\partial \gamma_j} = -\sum_{i=1}^{N} \frac{\partial e_i^T}{\partial \gamma_j} D_i^{-1} e_i,
$$

Table 6

Maximum likelihood estimates for models ARMA-20Q, ARMA-02Q, ARMA-11Q, ARMA-QQQ, and ARMA-QCQ fit to the lung cancer data. Parameter estimates with standard errors in the parentheses.

<table>
<thead>
<tr>
<th></th>
<th>ARMA-20Q</th>
<th>ARMA-02Q</th>
<th>ARMA-11Q</th>
<th>ARMA-QQQ</th>
<th>ARMA-QCQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (Int.)</td>
<td>19.76 (0.54)</td>
<td>19.55 (0.46)</td>
<td>19.69 (0.52)</td>
<td>19.69 (0.51)</td>
<td>19.71 (0.51)</td>
</tr>
<tr>
<td>$\beta_1$ (Arm)</td>
<td>-0.70 (0.76)</td>
<td>-0.86 (0.66)</td>
<td>-0.66 (0.73)</td>
<td>-0.74 (0.71)</td>
<td>-0.75 (0.72)</td>
</tr>
<tr>
<td>$\beta_2$ (Time)</td>
<td>-0.14 (0.18)</td>
<td>-0.23 (0.15)</td>
<td>-0.08 (0.17)</td>
<td>-0.10 (0.18)</td>
<td>-0.10 (0.18)</td>
</tr>
<tr>
<td>$\beta_3$ (Arm × Time)</td>
<td>0.37 (0.27)</td>
<td>0.43 (0.22)</td>
<td>0.31 (0.26)</td>
<td>0.31 (0.26)</td>
<td>0.31 (0.26)</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>0.54 (0.03)</td>
<td>-0.03 (0.02)</td>
<td>-0.23 (0.17)</td>
<td>-0.20 (0.19)</td>
<td>0.08 (0.01)</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.17 (0.09)</td>
<td>0.07 (0.02)</td>
<td>0.00 (0.01)</td>
<td>0.03 (0.08)</td>
<td>0.00 (0.01)</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.33 (0.04)</td>
<td>0.57 (0.03)</td>
<td>0.52 (0.04)</td>
<td>0.65 (0.11)</td>
<td>0.63 (0.12)</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.45 (0.04)</td>
<td>0.05 (0.18)</td>
<td>0.10 (0.25)</td>
<td>-0.09 (0.06)</td>
<td>-0.05 (0.11)</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>2.79 (0.10)</td>
<td>2.79 (0.10)</td>
<td>2.81 (0.10)</td>
<td>2.80 (0.10)</td>
<td>2.80 (0.10)</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>-0.41 (0.08)</td>
<td>-0.26 (0.08)</td>
<td>-0.43 (0.08)</td>
<td>-0.42 (0.08)</td>
<td>-0.42 (0.08)</td>
</tr>
<tr>
<td>$\lambda_0$</td>
<td>0.05 (0.01)</td>
<td>0.03 (0.01)</td>
<td>0.05 (0.01)</td>
<td>0.05 (0.01)</td>
<td>0.05 (0.01)</td>
</tr>
</tbody>
</table>

* indicates significance with 95% confidence level.
where \( \frac{\partial \mathbf{e}_t^T}{\partial \gamma_j} \) is a vector with components \( \frac{\partial e_k}{\partial \gamma_j} = - \sum_{k=1}^{t-1} (z_{ik} e_{ik} + l_{ik} \frac{\partial e_k}{\partial \gamma_j}) \) for \( t = 1, \ldots, n_i \). Now taking another derivative with respect to \( \lambda_j \), we obtain

\[
\frac{\partial^2 \log L(\theta; y)}{\partial \lambda_j \partial \gamma_j} = - \sum_{i=1}^{N} \frac{\partial \mathbf{e}_t^T}{\partial \gamma_j} \frac{\partial D_i^{-1} e_t}{\partial \lambda_j} = - \sum_{i=1}^{N} \sum_{t=1}^{n_i} \frac{\partial e_t}{\partial \gamma_j} \frac{\partial}{\partial \lambda_j} \left( \frac{1}{\sigma_n^2} \right) e_t.
\]

Since \( \frac{\partial e_t}{\partial \gamma_j} \) consists of \( e_k \) for \( k = 1, \ldots, t - 1 \), \( e_t \) and \( \frac{\partial e_t}{\partial \gamma_j} \) are independent. Thus, \( E \left( - \frac{\partial^2 \log L(\theta; y)}{\partial \lambda_j \partial \gamma_j} \right) = 0 \).

Finally, we have

\[
\frac{\partial^2 \log L}{\partial \lambda_j \partial \alpha_j} = - \frac{1}{2} \sum_{i=1}^{N} \log 2 \pi - \frac{1}{2} \sum_{i=1}^{N} h_i^T \lambda - \frac{1}{2} \left( y_i - x_i \beta \right)^T \Sigma_i^{-1} (y_i - x_i \beta).
\]

When \( \theta = \theta_0 \), the mean and variance of \( l_i \) are respectively

\[
E_0(l_i) = - \frac{n_i}{2} \log (2 \pi) - \frac{1}{2} E_0 \left\{ \sum_{i=1}^{n_i} h_i^T \lambda - \frac{1}{2} (y_i - x_i \beta)^T \Sigma_i^{-1} (y_i - x_i \beta) \right\}
\]

\[
= - \frac{n_i}{2} \log (2 \pi) - \frac{1}{2} E_0 \left\{ \sum_{i=1}^{n_i} h_i^T \lambda - \frac{1}{2} E \left\{ (y_i - x_i \beta_0)^T \Sigma_i^{-1} (y_i - x_i \beta_0) \right\} \right\}
\]

\[
= - \frac{n_i}{2} \log (2 \pi) - \frac{1}{2} E_0 \sum_{i=1}^{n_i} h_i^T \lambda - \frac{1}{2} E \left\{ (x_i \beta - x_i \beta_0)^T \Sigma_i^{-1} (x_i \beta - x_i \beta_0) \right\}
\]

\[
= - \frac{n_i}{2} \log (2 \pi) - \frac{1}{2} E_0 \sum_{i=1}^{n_i} h_i^T \lambda - \frac{1}{2} \text{tr}(\Sigma_i^{-1} \Sigma_0) - \frac{1}{2} (x_i \beta - x_i \beta_0)^T \Sigma_i^{-1} (x_i \beta - x_i \beta_0),
\]

\[
\text{var}_0(l_i) = \frac{1}{4} \text{var}_0 \left\{ (y_i - x_i \beta)^T \Sigma_i^{-1} (y_i - x_i \beta) \right\}
\]

\[
= \frac{1}{4} \text{tr} \left\{ (\Sigma_i^{-1} \Sigma_0)^2 \right\} + (x_i \beta - x_i \beta_0)^T \Sigma_i^{-1} \Sigma_0 i \Sigma_i^{-1} (x_i \beta - x_i \beta_0),
\]

where \( \Sigma_i = T_i^{-1} L_i D_i L_i^T T_i^{-T} \) and \( \Sigma_0 = \Sigma(\theta_0) \). It follows from the compactness of the parameter space in (C2) and boundedness of the covariates in (C3) that \( \text{var}_0(l_i) \leq K \) for all \( i \) where \( K \) is a finite constant; here, we also need a fixed max \( n_i \) in (C1) to satisfy Lemma 2 in Chiu et al. (1996). By Kolmogorov’s strong law of large numbers, we have that

\[
\frac{1}{N} \sum_{i=1}^{N} l_i - \frac{1}{N} \sum_{i=1}^{N} E_0(l_i) \rightarrow 0 \tag{C4}
\]

almost surely. Notice that the above constant \( K \) is independent of \( \theta \) and that

\[
-2 \sum_{i=1}^{N} E_0(l_i)/N = \frac{1}{N} \left[ \sum_{i=1}^{N} n_i \log (2 \pi) + \sum_{i=1}^{N} n_i h_i^T \lambda + \sum_{i=1}^{N} \text{tr}(\Sigma_i^{-1} \Sigma_0) \right.
\]

\[
+ \sum_{i=1}^{N} (x_i \beta - x_i \beta_0)^T \Sigma_i^{-1} (x_i \beta - x_i \beta_0) \].
\]

From the proof of Theorem 1 in Chiu et al. (1996), it is easy to show the consistency of \( \hat{\theta} \) using (C2) and (C3).

3. The proof here is essentially the same as that of Theorem 2 in Chiu et al. (1996). Note that Theorem 2 of Chiu et al. (1996) also relies on results in Roussas (1968), and requires several regularity conditions for asymptotic normality. These are all satisfied from (C2)–(C4).
Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2017.05.001.

References