1-1-2013

Robust Estimation of the Correlation Matrix of Longitudinal Data

Mehdi Maadooliat
Marquette University, mehdi.maadooliat@marquette.edu

Mohsen Pourahmadi
Texas A & M University

Jianhua Z. Huang
Texas A & M University

Accepted version. *Statistics and Computing*, Vol. 23, No. 1 (January 2013): 17-28. DOI.

Statistics and Computing manuscript No.
(will be inserted by the editor)

Robust Estimation of the Correlation Matrix of Longitudinal Data

Mehdi Maadooliat, Mohsen Pourahmadi, and Jianhua Z. Huang

Abstract We propose a double-robust procedure for modeling the correlation matrix of a longitudinal dataset. It is based on an alternative Cholesky decomposition of the form $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{L}\boldsymbol{L}^\top\boldsymbol{D}$ where $\boldsymbol{D}$ is a diagonal matrix proportional to the square roots of the diagonal entries of $\boldsymbol{\Sigma}$ and $\boldsymbol{L}$ is a unit lower-triangular matrix determining solely the correlation matrix. The first robustness is with respect to model misspecification for the innovation variances in $\boldsymbol{D}$, and the second is robustness to outliers in the data. The latter is handled using heavy-tailed multivariate t-distributions with unknown degrees of freedom. We develop a Fisher scoring algorithm for computing the maximum likelihood estimator of the parameters when the nonredundant and unconstrained entries of $(\boldsymbol{L}, \boldsymbol{D})$ are modeled parsimoniously using covariates. We compare our results with those based on the modified Cholesky decomposition of the form $\boldsymbol{L}\boldsymbol{D}^2\boldsymbol{L}^\top$ using simulations and a real dataset.

Keywords cholesky decomposition; correlation modelling; multivariate t; robust estimation

1 Introduction

Longitudinal data arise frequently in the biomedical, epidemiological and social sciences, where subjects are measured repeatedly over time and the observations on the same subject are intrinsically correlated (Diggle et al. 2002). The technique of generalized estimating equations (GEE) introduced in Liang & Zeger (1986) is widely used when the focus is on modeling the mean. In GEE and many of its extensions, in the interest of expediency, parsimony and ensuring the positive-definiteness of the estimated correlation matrix, it is common to pick a *working correlation* matrix, from a long menu of structured correlation matrices. Although consistency of the estimators of the mean parameters is not affected, misspecification of the correlation may result in a great loss of efficiency (Wang & Carey 2003) and may lead to invalid inferences (Cannon et al. 2001, Carroll 2003). The correlation matrix itself might be of scientific interest (Diggle & Verbyla 1998) in which case it is desirable to develop a bona fide data-based framework for modeling correlation matrices following the familiar three stages of model formulation, estimation and diagnostics in the modeling process for the mean vector (McCullagh & Nelder 1989). Attempts to develop such methods have been made in recent years by Chiu et al. (1996), Pourahmadi (1999, 2000), Pan & MacKenzie (2003), Ye & Pan (2006), Lin & Wang (2009), Leng et al. (2010) and references therein, using the spectral and Cholesky decompositions of covariance matrices, respectively.

A methodology based on the modified Cholesky decomposition (M.CD) of the covariance matrix $\boldsymbol{\Sigma}$ of a random vector $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ has proved quite successful for longitudinal data in the sense that the positive-definiteness of the estimated covariance is guaranteed and parsimony can be achieved using covariates. However, it seems for historical reasons the focus has been mostly on specific transitional models of autoregressive (AR) type for the actual successive measurements on a subject:

$$y_t = \phi_{t,t-1}y_{t-1} + \ldots + \phi_{t,1}y_1 + \epsilon_t, \qquad t = 1, 2, \ldots, p, \quad (1)$$

where the $\phi_{t,j}$'s are the so-called generalized autoregressive parameters (GARPs) with $\phi_{1,0} = 0$, and ϵ_t's are the prediction errors or innovations with $Var(\epsilon_t) = \sigma_t^2$; see Pourahmadi (1999, 2000), Pan & MacKenzie (2003), Ye & Pan (2006), Lin & Wang (2009), and Leng et al. (2010). Although the idea of inverting the AR model

M. Maadooliat, M. Pourahmadi, and J. Z. Huang
Department of Statistics, Texas A&M University
E-mail: madoliat@stat.tamu.edu

(1) and writing it as a moving average (MA) of the actual response in terms of the present and past innovations was mentioned in (Pourahmadi 2001, Sec. 3.5), the idea and its potentials have not been pursued vigorously in the literature of longitudinal and correlated data. Given the duality and synergy between the AR and MA models in the theory of finite parameter stationary time series (Brockwell & Davis 1991), one would expect a level of similar fruitful connections to exist between such type of models for nonstationary longitudinal data. For example, inverting (1) gives rise to the generalized moving average parameters (GMAPs) which are known (Pourahmadi 2001, Sec. 3.5; Rothman et al. 2010) to be useful in parsimonious modeling and guaranteeing the positive-definiteness of $\boldsymbol{\Sigma}$ itself. These models, whether of AR or MA type, lead to a factorization of the form $\boldsymbol{\Sigma}^{\pm} = \boldsymbol{L}\boldsymbol{D}^2\boldsymbol{L}^{\top}$, where $\boldsymbol{\Sigma}^{\pm}$ indicates either the covariance or the inverse covariance matrix, and $\boldsymbol{L}, \boldsymbol{D}$ are generic unit lower triangular and diagonal matrices, respectively. Since $\boldsymbol{D}^2$ is trapped in the middle, the correlation matrix corresponding to $\boldsymbol{\Sigma}^{\pm}$ depends on the innovation variances represented by the diagonal entries of $\boldsymbol{D}^2$, and hence is not necessarily *robust* to their model misspecifications.

By contrast, there is an alternative Cholesky decomposition (A.CD), due to Chen & Dunson (2003), which is of the generic form $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{L}\boldsymbol{L}^{\top}\boldsymbol{D}$ with the diagonal matrix $\boldsymbol{D}$ of innovation standard deviations placed outside. Consequently, such factorization amounts to directly modeling the covariance matrix but in a manner that its estimated correlation matrix $\boldsymbol{R}$ does not depend on the quality of modeling and estimation of the innovation variances σ_t^2's, see (3). In other words, estimation of $\boldsymbol{R}$ is robust to misspecification of models for σ_t^2's, the component shared by both the M.CD and A.CD. Beside this basic observation, not much is known about the consequences of using A.CD in modeling covariance and correlation matrices other than Chen & Dunson (2003), and Cai et al. (2006) in the context of random-effects selection. This factorization is more closely related to the MA representation of a "standardized" version of repeated measures on a subject, see (2), Pourahmadi (2007) and Rothman et al. (2010).

In this paper, our primary objective is to study some of the consequences of modeling the components of the A.CD factorization on estimating the correlation matrix of longitudinal data. The secondary objective is to have procedures for estimating correlation matrices that are robust to outliers. We use the multivariate t distributions with ν the degrees of freedom unknown, as a model for the data and focus on accurate estimation of the *df*.

We point out some other structural, computational and statistical differences that exist between the M.CD in Pourahmadi (2000) and the A.CD in Chen & Dunson (2003). For example, recognizing that the M.CD and A.CD of a covariance matrix correspond to AR and MA representations of the underlying nonstationary longitudinal data (Pourahmadi 2001, Sec. 3.5; Pourahmadi 2007, Rothman et al. 2010), therefore one expects more computational difficulties in computing the MLE of the parameters of the A.CD than those from M.CD (Brockwell & Davis 1991, Chaps 5 and 9). In the A.CD framework the focus is on modeling the covariance matrix, while it is common to think of the M.CD framework as being related to modeling the precision matrix (inverse covariance matrix). However, recently Rothman et al. (2010) have proposed sparse estimation of $\boldsymbol{\Sigma}$ itself based on its M.CD and a related regression/MA interpretation of the entries of the factors. They show that there are significant structural and computational differences when working with $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{-1}$ and their respective correlation matrices. A somewhat surprising result is that banding the Cholesky factor of the precision matrix coincides with constrained maximum likelihood, but banding the Cholesky factor of the covariance matrix itself does not. Such results are based on some interesting relationships between zero patterns of covariance matrices and their Cholesky factors. For example, the Cholesky factor of either the covariance matrix or its inverse is k-banded if and only if the corresponding matrix itself is k-banded, see Propositions 1-3 in Rothman et al. (2010).

The outline of the paper is as follows: In Section 2, M.CD and A.CD are reviewed along with the statistical interpretations of the entries of their Cholesky decompositions. Section 3 discusses the multivariate t-distribution and the MLE of its parameters with a particular focus on the orthogonality of the parameters estimate. Section 4 illustrates the methodology using a real dataset, and assess its performance using a simulation experiment. Section 5 concludes the paper.

2 M.CD and A.CD of a Covariance Matrix

In this section, we review properties of two distinct Cholesky decompositions of the positive-definite covariance matrix of a longitudinal dataset, and discuss their roles in estimating the correlation matrix.

It is known that any $p \times p$ positive-definite covariance matrix can be factorized as $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^{\top}$, referred to as its standard Cholesky decomposition, where $\boldsymbol{C}$ is a unique lower triangular matrix with positive diagonal entries. What are the statistical relevance of the diagonal and subdiagonals entries of $\boldsymbol{C}$? Letting

$\boldsymbol{D} = \operatorname{diag}(c_{11}, \ldots, c_{pp})$, this factorization can take the following two distinct forms depending on whether the matrix $\boldsymbol{D}$ is inserted between the two lower triangular matrices or outside.

The M.CD for $\boldsymbol{\Sigma}$ keeps $\boldsymbol{D}^2$ inside:

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{C}^\top = \boldsymbol{L}\boldsymbol{D}^2\boldsymbol{L}^\top,$$

where $\boldsymbol{L} = \boldsymbol{C}\boldsymbol{D}^{-1}$ is a “standardized” version of $\boldsymbol{C}$, dividing each column by its diagonal entry. Defining $\boldsymbol{T} = \boldsymbol{L}^{-1}$, it is known (Pourahmadi 1999) that the entries of $\boldsymbol{T}$ and $\boldsymbol{D}^2$, respectively, are negative of the GARPs in (1) and the prediction error variances σ_t^2's, when a measurement is regressed on its predecessors. Details of formulating parsimonious models using graphical tools like regressograms and estimating the ensuing parameters of $\boldsymbol{T}$ and $\boldsymbol{D}$ are given in Pourahmadi (1999).

The A.CD in Chen & Dunson (2003) keeps $\boldsymbol{D}$ outside:

$$\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{D}^{-1}\boldsymbol{D} = \boldsymbol{D}\boldsymbol{L}\boldsymbol{L}^\top\boldsymbol{D},$$

where now $\boldsymbol{L} = \boldsymbol{D}^{-1}\boldsymbol{C}$ is obtained from $\boldsymbol{C}$ using a slightly different “standardization”, namely dividing each row of $\boldsymbol{C}$ by its diagonal entries. In Pourahmadi (2001, 2007), the statistical interpretation of entries of $\boldsymbol{L}$ is given as the moving average coefficients when a standardized measurement is regressed on its past and present innovations, see also Rothman et al. (2010). Let $(y_1, \ldots, y_p)^\top$ be a zero mean random vector with covariance matrix $\boldsymbol{\Sigma}$. Denote $\boldsymbol{L}_{p\times p} = (\theta_{tj})$ and $\boldsymbol{D}_{p\times p} = \operatorname{diag}(\sigma_t)$. It's clear that $\boldsymbol{D}^{-1}\boldsymbol{y}$ has the covariance $\boldsymbol{L}\boldsymbol{L}^\top$. More precisely, defining $\boldsymbol{\epsilon} = (\boldsymbol{D}\boldsymbol{L})^{-1}\boldsymbol{y}$, it follows that $\operatorname{cov}(\boldsymbol{\epsilon}) = \boldsymbol{I}_p$ and then $\boldsymbol{D}^{-1}\boldsymbol{y} = \boldsymbol{L}\boldsymbol{\epsilon}$, from which we obtain a variable-order, varying-coefficients moving average representation for the standardized y_t/σ_t as:

$$y_t/\sigma_t = \epsilon_t + \sum_{j=1}^{t-1} \theta_{tj}\epsilon_j. \tag{2}$$

From (2), for any $1 \le s, t \le p$, with $s \wedge t = \min\{s, t\}$, it follows that

$$\operatorname{cov}(y_s, y_t) = \sigma_s\sigma_t \sum_{j=1}^{s\wedge t} \theta_{tj}\theta_{sj},$$

so that the correlation between y_s and y_t given by

$$\operatorname{corr}(y_s, y_t) = \frac{\sum_{j=1}^{s\wedge t} \theta_{sj}\theta_{tj}}{\sqrt{\left(\sum_{j=1}^{s} \theta_{sj}^2 \sum_{j=1}^{t} \theta_{tj}^2\right)}}, \tag{3}$$

is solely determined by the $\boldsymbol{L}$ matrix. This property is a great motivation for modeling a correlation matrix using A.CD, so that it is robust to model misspecifications for the innovation variances, $\sigma_t^2,\ t = 1, \ldots, p$.

3 MLEs for the A.CD Model: The Multivariate t_ν

The assumption of multivariate normality commonly made for the vector of repeated measures on a subject may not be tenable in many practical situations when outliers exist or the underlying data exhibit heavy-tails. In this situation, a number of authors have used the multivariate t-distribution for robust estimation of the parameters of general linear models (Zellner 1976, Lange et al. 1989); Lin & Wang (2009) has used it for robust estimation under the M.CD decomposition. Robust estimation for linear mixed models using the multivariate t-distribution has been studied by Welsh & Richardson (1997) and Pinheiro et al. (2001).

In the sequel, for $i = 1, \ldots, n$, we assume that the vector of repeated measures on the i-th subject $\boldsymbol{y}_i \sim t(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \nu)$. This means that the p-dimensional vector $\boldsymbol{y}_i$ is following a multivariate t-distribution with degrees of freedom (df) ν, location vector $\boldsymbol{\mu}_i$ and scale matrix $\boldsymbol{\Sigma}$ with the probability density function given as:

$$f(\boldsymbol{y}_i|\boldsymbol{\mu}_i\ ,\ \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\Big(\dfrac{\nu+p}{2}\Big)}{\Gamma\Big(\dfrac{\nu}{2}\Big)(\pi\nu)^{p/2}}|\boldsymbol{\Sigma}|^{-1/2} \times \left(1 + \frac{(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)}{\nu}\right)^{-(\nu+p)/2},$$

where ν is a positive real number. For $\nu > 1$ the mean vector is defined to be $\boldsymbol{\mu}_i$, the covariance matrix exists for $\nu > 2$ and is equal to $\dfrac{\nu}{\nu-2}\boldsymbol{\Sigma}$.

Following the general approach in Pourahmadi (2000), Lin & Wang (2009) we model $\boldsymbol{\mu}_i, \boldsymbol{L} = (\theta_{tj})$ and $\boldsymbol{D} = \operatorname{diag}(\sigma_t)$ as:

$$\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta}, \quad \theta_{tj} = d(\boldsymbol{z}_{tj}, \boldsymbol{\gamma}), \quad \log\sigma_t = v(\boldsymbol{z}_t, \boldsymbol{\lambda}), \tag{4}$$

where $d(\cdot,\cdot)$, $v(\cdot,\cdot)$ are known functions, $\boldsymbol{X}_i$, $\boldsymbol{z}_{tj}$ and $\boldsymbol{z}_t$ are $p \times m$, $d \times 1$ and $q \times 1$ matrices of covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top, \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)^\top$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^\top$ are parameters of the mean, log-innovation and the moving average parameters $\boldsymbol{y}$ in the A.CD, respectively. When $d(\cdot,\cdot)$, $v(\cdot,\cdot)$ are polynomials, we use the notation $\operatorname{Poly}(d, q)$ as a shorthand for two distinct polynomials of degrees d, q in the lagged times $(t-j)$ and t for θ_{tj} and $\log\sigma_t$, respectively. Specifically, in this case the covariates $\boldsymbol{z}_t$ and $\boldsymbol{z}_{tj}$ are of the form:

$$\boldsymbol{z}_{tj} = (1, (t-j), \ldots, (t-j)^d)^\top, \qquad j = 1, \ldots, t-1,$$
$$\boldsymbol{z}_t = (1, t, \ldots, t^q)^\top, \qquad t = 1, \ldots, p.$$

For example, in most of our simulation work we use $\operatorname{Poly}(3,3)$ as models for the components of $\boldsymbol{L}, \boldsymbol{D}$.

Assuming $m, q,$ and d are known, let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\lambda}^\top, \nu)^\top$ be the partitioned vector of all parameters in the model, then the log-likelihood function $\ell(\boldsymbol{\theta})$

is:

$$\ell(\boldsymbol{\theta}) = n\Bigg(\log \mathit{\Gamma}\Big(\frac{\nu+p}{2}\Big) - \log \mathit{\Gamma}\Big(\frac{\nu}{2}\Big) - \frac{p}{2}\log(\pi\nu)\Bigg)$$

$$- \frac{n}{2}\log|\boldsymbol{D}^2| - \frac{1}{2}(\nu+p)\sum_{i=1}^{n}\log\Big(1 + \frac{\mathit{\Delta}_i(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda})}{\nu}\Big),$$

where $\mathit{\Delta}_i(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda}) := (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$. We suppress its arguments and use the abbreviation $\mathit{\Delta}_i$ in the sequel.

3.1 Maximum Likelihood Estimation Using Fisher Scoring

In this section, we study some computational and statistical implications of using covariates in the parsimonious modeling of $\boldsymbol{L}$ in (4) as compared to the same approach in modeling $\boldsymbol{T}$ in the M.CD approach studied in Pourahmadi (2000), Lin & Wang (2009). Similar to the M.CD models, it turns out that there is no closed-form solution for the MLEs of A.CD models, thus iterative algorithms like the Newton-Raphson or Fisher scoring as in Pourahmadi (2000) and Lin & Wang (2009) will be developed here.

The Fisher scoring algorithm is developed in this subsection. For the partitioning of $\boldsymbol{\theta}$ as above, the blocks of the score function $U(\boldsymbol{\theta}) = \big(U^\top(\boldsymbol{\beta}), U^\top(\boldsymbol{\gamma}), U^\top(\boldsymbol{\lambda}), U(\nu)\big)^\top$ can be obtained and simplified as:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n}\omega_i \boldsymbol{X}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{r}_i,$$

$$U(\gamma_r) = \mathrm{tr}\Bigg((\boldsymbol{T}\boldsymbol{D}^{-1})\Big(\sum_{i=1}^{n}\omega_i\boldsymbol{S}_i\Big)(\boldsymbol{T}\boldsymbol{D}^{-1})^\top \boldsymbol{T}\boldsymbol{L}_{\gamma_r}\Bigg),$$

$$U(\lambda_s) = \mathrm{tr}\Bigg(\bigg(\Big(\sum_{i=1}^{n}\omega_i\boldsymbol{S}_i\Big)\boldsymbol{\Sigma}^{-1} - n\boldsymbol{I}\bigg)\boldsymbol{D}^{-1}\boldsymbol{D}_{\lambda_s}\Bigg),$$

$$U(\nu) = \frac{1}{2}\sum_{i=1}^{n}\Bigg(\phi\Big(\frac{\nu+p}{2}\Big) - \phi\Big(\frac{\nu}{2}\Big) - \frac{p}{\nu}$$

$$- \log\Big(1 + \frac{\mathit{\Delta}_i}{\nu}\Big) + \frac{\omega_i}{\nu}\mathit{\Delta}_i\Bigg),$$

where $r = 1, \ldots, d$, $s = 1, \ldots, q$, $\boldsymbol{L}_{\gamma_r} = \dfrac{\partial}{\partial\gamma_r}\boldsymbol{L}$, $\boldsymbol{D}_{\lambda_s} = \dfrac{\partial}{\partial\lambda_s}\boldsymbol{D}$, $\omega_i = \dfrac{\nu+p}{\nu+\mathit{\Delta}_i}$, $\boldsymbol{r}_i = (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$, $\boldsymbol{S}_i = \boldsymbol{r}_i\boldsymbol{r}_i^\top$ and $\phi(x) = \frac{d}{dx}\log\mathit{\Gamma}(x)$.

Now, we have the necessary ingredients to present the Fisher information in terms of the blocks of a partitioned 4×4 matrix corresponding to $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}$, and ν. The blocks of the Fisher information that involve $\boldsymbol{\beta}$ (the location parameter) are as follows:

$$\boldsymbol{I}_{11}(\boldsymbol{\beta}) = -\mathbb{E}(\ell_{\boldsymbol{\beta}\boldsymbol{\beta}}) = \frac{\nu+p}{\nu+p+2}\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i,$$

$$\boldsymbol{I}_{12}(\boldsymbol{\beta},\boldsymbol{\gamma}) = -\mathbb{E}(\ell_{\boldsymbol{\beta}\boldsymbol{\gamma}}) = \boldsymbol{0},$$

$$\boldsymbol{I}_{13}(\boldsymbol{\beta},\boldsymbol{\lambda}) = -\mathbb{E}(\ell_{\boldsymbol{\beta}\boldsymbol{\lambda}}) = \boldsymbol{0},$$

$$\boldsymbol{I}_{14}(\boldsymbol{\beta},\nu) = -\mathbb{E}(\ell_{\boldsymbol{\beta}\nu}) = \boldsymbol{0}.$$

In addition, we obtain other blocks of the Fisher information matrix using Proposition 4 of Lange et al. (1989). We state two versions of the result corresponding to the parameterizations based on M.CD and A.CD.

Let $\boldsymbol{\varphi}$ denote a generic parametrization of either $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$ for the p-variate t_ν distribution with the scale matrix $\boldsymbol{\Sigma}$, the contribution of a single observation to the Fisher information block for the scale parameter and the degrees of freedom are as follows:

$$I_{i,j}(\boldsymbol{\varphi}) = \frac{1}{2(\nu+p+2)}\Big[(\nu+p)\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\varphi_i}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\varphi_j}\Big)$$

$$-\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\varphi_i}\Big)\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\varphi_j}\Big)\Big]$$

$$= \frac{1}{2(\nu+p+2)}\Big[(\nu+p)\mathrm{tr}\Big(\boldsymbol{\Sigma}{\boldsymbol{\Sigma}^{-1}}_{\varphi_i}\boldsymbol{\Sigma}{\boldsymbol{\Sigma}^{-1}}_{\varphi_j}\Big)$$

$$-\mathrm{tr}\Big(\boldsymbol{\Sigma}{\boldsymbol{\Sigma}^{-1}}_{\varphi_i}\Big)\mathrm{tr}\Big(\boldsymbol{\Sigma}{\boldsymbol{\Sigma}^{-1}}_{\varphi_j}\Big)\Big],$$

$$I_i(\boldsymbol{\varphi},\nu) = -\frac{1}{(\nu+p+2)(\nu+p)}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\varphi_i}\Big)$$

$$= -\frac{1}{(\nu+p+2)(\nu+p)}\mathrm{tr}\Big(\boldsymbol{\Sigma}{\boldsymbol{\Sigma}^{-1}}_{\varphi_i}\Big).$$

The equations involving $\boldsymbol{\Sigma}_{\boldsymbol{\varphi}}$ $\left(\text{i.e. } \dfrac{\partial\boldsymbol{\Sigma}}{\partial\boldsymbol{\varphi}}\right)$ are useful for the A.CD model, while those involving ${\boldsymbol{\Sigma}^{-1}}_{\boldsymbol{\varphi}}$ $\left(\text{i.e. } \dfrac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\boldsymbol{\varphi}}\right)$ can be used for modelling $\boldsymbol{\Sigma}^{-1}$. In the application to model (4), $\boldsymbol{\varphi}^\top = (\boldsymbol{\gamma}^\top, \boldsymbol{\lambda}^\top)^\top$ is for parameterizing the scale matrix.

Once the information matrix is computed, the iterative Fisher scoring algorithm can be used to compute the MLE of the parameters by updating the current value of $\tilde{\boldsymbol{\theta}}$ to $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} + I^{-1}(\tilde{\boldsymbol{\theta}})U(\tilde{\boldsymbol{\theta}}).$$

Note that when using linear link functions for $d(\cdot,\cdot)$, and $v(\cdot,\cdot)$ in (4), simpler structures for the score function and the Fisher information will result. Also, when $\nu \to \infty$, the results in this section reduce to those for an iterative procedure for computing the MLEs of the A.CD model parameters under the multivariate normal setup.

Computation and the form of the entries of the Fisher information matrix are slightly different for A.CD and M.CD and are summarized in the following two subsections.

3.2 Fisher Information Matrix for A.CD

As an immediate consequence of the results given in subsection 3.1 we obtain the Fisher information blocks for the parameters of the components of the scale matrix and the degrees of freedom for the A.CD model.

$$\begin{aligned}
I_{22,rs}(\boldsymbol{\gamma}) &= \frac{(\nu+p)n}{\nu+p+2}\mathrm{tr}(\boldsymbol{L}_{\gamma_r}\boldsymbol{L}_{\gamma_s}^\top\boldsymbol{T}^\top\boldsymbol{T}),\\
I_{33,rs}(\boldsymbol{\lambda}) &= \Bigg[\mathrm{tr}\big(\boldsymbol{L}\boldsymbol{L}^\top\boldsymbol{D}_{\lambda_r}\boldsymbol{\varSigma}^{-1}\boldsymbol{D}_{\lambda_s}+\boldsymbol{D}^{-2}\boldsymbol{D}_{\lambda_r}\boldsymbol{D}_{\lambda_s}\big)\\
&\qquad -\frac{2\mathrm{tr}(\boldsymbol{D}^{-1}\boldsymbol{D}_{\lambda_r})\mathrm{tr}(\boldsymbol{D}^{-1}\boldsymbol{D}_{\lambda_s})}{(\nu+p)}\Bigg]\frac{n(\nu+p)}{\nu+p+2},\\
I_{44}(\nu) &= \frac{n}{4}\Bigg[\psi\Big(\frac{\nu}{2}\Big)-\psi\Big(\frac{\nu+p}{2}\Big)\\
&\qquad -\frac{2p(\nu+p+4)}{\nu(\nu+p)(\nu+p+2)}\Bigg],\\
I_{23,rs}(\boldsymbol{\gamma},\boldsymbol{\lambda}) &= \frac{(\nu+p)n}{\nu+p+2}\mathrm{tr}(\boldsymbol{D}\boldsymbol{L}_{\gamma_r}\boldsymbol{L}^\top\boldsymbol{D}_{\lambda_s}\boldsymbol{\varSigma}^{-1}),\\
I_{24,r}(\boldsymbol{\gamma},\nu) &= 0,\\
I_{34,s}(\boldsymbol{\lambda},\nu) &= -\frac{2n}{(\nu+p+2)(\nu+p)}\mathrm{tr}(\boldsymbol{D}^{-1}\boldsymbol{D}_{\lambda_s}),
\end{aligned}$$

where $\psi(x) = \frac{d^2}{dx^2}\log\varGamma(x)$ stands for the trigamma function.

Letting $\nu \to \infty$ in the above identities, we obtain the corresponding results for the multivariate normal model where the log-likelihood function $\ell(\theta)$, up to an additive constant is:

$$-\frac{2}{n}\ell(\theta) = \log|\boldsymbol{D}^2|+n^{-1}\sum_{i=1}^{n}\varDelta_i = \sum_{t=1}^{p}\log\sigma_t^2+\mathrm{tr}\boldsymbol{S}\boldsymbol{\varSigma}^{-1}.$$

The score function and the Fisher information for the multivariate normal distribution is easy to obtain by considering the following facts and substituting in the previous results:

$$\omega_i \to 1, \quad \frac{(\nu+p)n}{\nu+p+2} \to n, \quad \frac{2n}{\nu+p+2} \to 0,$$

and $\sum_{i=1}^{n}\omega_i\boldsymbol{S}_i = n\boldsymbol{S}$, where $\boldsymbol{S} = n^{-1}\sum_{i=1}^{n}\boldsymbol{r}_i\boldsymbol{r}_i^\top$.

3.3 Comparison with the Fisher Information Matrix for M.CD

In this section, we find the Fisher information matrix for the M.CD and compare it with that for the A.CD models. For simplicity, we use the same notation for the information matrices corresponding to A.CD and M.CD. Using the result given in subsection 3.1, the entries of the Fisher information associated to the scale parameter and the degrees of freedom for the M.CD model are:

$$\begin{aligned}
I_{22,rs}(\boldsymbol{\gamma}) &= \frac{(\nu+p)n}{\nu+p+2}\mathrm{tr}(\boldsymbol{T}_{\gamma_r}^\top\boldsymbol{D}^{-2}\boldsymbol{T}_{\gamma_s}\boldsymbol{\varSigma}),\\
I_{33,rs}(\boldsymbol{\lambda}) &= \frac{n}{2(\nu+p+2)}\Bigg[(\nu+p)\mathrm{tr}(\boldsymbol{D}^{-4}\boldsymbol{D}^2{}_{\lambda_r}\boldsymbol{D}^2{}_{\lambda_s})\\
&\qquad -\mathrm{tr}(\boldsymbol{D}^{-2}\boldsymbol{D}^2{}_{\lambda_r})\mathrm{tr}(\boldsymbol{D}^{-2}\boldsymbol{D}^2{}_{\lambda_s})\Bigg],\\
I_{23,rs}(\boldsymbol{\gamma},\boldsymbol{\lambda}) &= 0,\\
I_{24,r}(\boldsymbol{\gamma},\nu) &= 0,\\
I_{34,s}(\boldsymbol{\lambda},\nu) &= -\frac{n}{(\nu+p+2)(\nu+p)}\mathrm{tr}(\boldsymbol{D}^{-2}\boldsymbol{D}^2{}_{\lambda_s}).
\end{aligned}$$

Comparing similar entries in the two sections, it is evident that their forms and values are quite different for the A.CD and M.CD models even for general link functions $d(\cdot,\cdot)$, $v(\cdot,\cdot)$. However, some notable and computationally useful differences are singled out below:

1. The parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are asymptotically orthogonal in the M.CD, but not in the A.CD. It is known that for the multivariate normal distribution, the $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are asymptotically orthogonal in the M.CD model (Ye & Pan 2006, Holan & Spinka 2007), but not in the A.CD model (Pourahmadi 2007). Here we have shown the same to be true for the multivariate t_ν setup. Our finding is different from that in Lin & Wang (2009), p. 3016. Under the M.CD and multivariate t-distribution, Lin & Wang (2009) showed $I_{23,rs}(\boldsymbol{\gamma},\boldsymbol{\lambda})$ to be nonzero, and hence $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are not asymptotically orthogonal.
2. The parameters ν and $\boldsymbol{\gamma}$ are asymptotically orthogonal in both the A.CD and M.CD models, this is not the case for ν and $\boldsymbol{\lambda}$, the parameters of the innovation variance.
3. Since $\boldsymbol{D} = \mathrm{diag}(\sigma_t)$ is a diagonal matrix, letting $\log(\sigma_t) := \boldsymbol{z}_t^\top\boldsymbol{\lambda}$, the derivative of $\boldsymbol{D}$ with respect to λ_s is $\boldsymbol{D}_{\lambda_s} = (\boldsymbol{Z}_{D,s})\boldsymbol{D}$, where
$$\boldsymbol{Z}_{D,s} = \mathrm{diag}(z_{1,s},\ldots,z_{p,s}), \qquad s = 1,\ldots,q.$$
Thus, replacing the matrix $\boldsymbol{D}^{-1}\boldsymbol{D}_{\lambda_s}$ by $\boldsymbol{Z}_{D,s}$ using the above results will lead to simpler forms for parts of the score function and the Fisher information that involve $\boldsymbol{\lambda}$. Also, using the log-linear models

for the innovation standard deviation, both M.CD and A.CD models will have the same quantity for $I_{34}(\boldsymbol{\lambda}, \nu)$.

4. After implementing both models and exploring the performance of the algorithms under different conditions, on average we obtained less number of iterations and faster convergence rate in the Fisher scoring algorithm for the M.CD models over the A.CD models in both normal and multivariate t_ν setup. This could be due to the fact that we deal with the covariance matrix in the A.CD and the precision matrix in the M.CD models, and we know that in the likelihood formulation and the Fisher scoring algorithm the precision matrix is the quantity that is involved directly, and that could lead to the faster convergence rate of M.CD.

4 Data Analysis

In this section, we compare the robustness and capabilities of the A.CD and M.CD for modeling various correlation structures using simulated and real data. We denote the M.CD and A.CD when used in conjunction with the multivariate normal and t distributions as M.CD.N and A.CD.N, M.CD.T and A.CD.T, respectively.

We compare estimators of correlation matrices using the following two loss functions and their corresponding risks:

$$\varDelta_1(\boldsymbol{R}, \boldsymbol{G}) = \mathrm{tr}\boldsymbol{R}^{-1}\boldsymbol{G} - \log|\boldsymbol{R}^{-1}\boldsymbol{G}| - n,$$

$$\text{and} \quad \varDelta_2(\boldsymbol{R}, \boldsymbol{G}) = \mathrm{tr}(\boldsymbol{R}^{-1}\boldsymbol{G} - \boldsymbol{I})^2,$$

where $\boldsymbol{R}$ is the target correlation matrix and $\boldsymbol{G}$ is another positive-definite correlation matrix of the same size. The loss $\varDelta_1(\boldsymbol{R}, \boldsymbol{G})$ is known as the entropy loss and $\varDelta_2(\boldsymbol{R}, \boldsymbol{G})$ as the quadratic loss. Both of these loss functions are 0, when $\boldsymbol{G} = \boldsymbol{R}$ and positive, when $\boldsymbol{G} \neq \boldsymbol{R}$. Their corresponding risk functions are:

$$R_i(\boldsymbol{R}, \boldsymbol{G}) = E_R\{\varDelta_i(\boldsymbol{R}, \boldsymbol{G})\}, \qquad i = 1, 2.$$

An estimator $\hat{\boldsymbol{R}}$ is better than $\tilde{\boldsymbol{R}}$, if its associated risk is smaller, that is, $R_i(\boldsymbol{R}, \hat{\boldsymbol{R}}) < R_i(\boldsymbol{R}, \tilde{\boldsymbol{R}})$.

4.1 Simulation

We fix the true parameters (mean, covariance/correlation matrix) for the simulation setup using those of the well-known Kenward (1987)'s cattle data. Here the weight of thirty cattle were recorded 11 times over a 133-day period, the dataset has been analyzed by several authors Zimmerman & Núñez Antón (2009). As in Pourahmadi (1999), cubic polynomials were fitted to the Cholesky factors $\boldsymbol{T}, \boldsymbol{D}$ of the sample covariance matrix of the treatment A of the cattle data.

For simulating data, we construct two true 11×11 covariance matrices corresponding to those of the cattle data fitted with M.CD.N-Poly$(3, 3)$ and A.CD.N-Poly$(3, 3)$ denoted by $\boldsymbol{\varSigma}_{\text{mcd}}$ and $\boldsymbol{\varSigma}_{\text{acd}}$, respectively. Thus, the true covariance (correlation) matrices are known and correspond to the above fits.

We generated $m = 100$ datasets from a multivariate t-distribution with the mean vector equal to the sample mean of the cattle data and the scale matrix equal to $\boldsymbol{\varSigma}_{\text{mcd}}$ and $\boldsymbol{\varSigma}_{\text{acd}}$, respectively, and for the following combinations of (ν, n): "$\nu = 4, 50$"(df) , "$n = 25, 100$"(sample sizes). We calculated the entropy and quadratic risks after fitting M.CD.N, M.CD.T, A.CD.N and A.CD.T using the Fisher scoring algorithm described in Section 3. In each iteration, after updating the estimates of the ν and the covariance structure, we obtain the updated estimate of the mean parameters using the weighted least square. Note that here we fit cubic polynomials both to the GARPs (GMAPs) and the log-innovation variances, the same models as their true counterparts. The results in Table 1(a) show that the risks in the third and forth columns are much smaller than those in the first two columns of both panels. This indicates the improved performance of M.CD over A.CD, when the data are actually generated from the same M.CD covariance (correlation) structure. Furthermore, in the left panel corresponding to $\nu = 4$, a smaller degrees of freedom, the risks for M.CD.T and A.CD.T are much smaller than M.CD.N and A.CD.N, and this difference disappears, as expected, for $\nu = 50$. Similar statements can be made about the results in Table 1(b) where the data are generated using the A.CD covariance structure, but now one can see that the first two columns of the two panels are smaller than their counterparts in the last two columns. In summary, the simulation results reported in Table 1 show the importance of knowing the structure of the underlying covariance matrix, where the M.CD works better for datasets coming from M.CD structure, and the A.CD fits the covariance matrix better if the data is coming from an A.CD structure.

Next, the theoretical result in Chen & Dunson (2003) and Section 2 suggest that the estimate of the correlation matrix is robust to model misspecification of the innovation variances when using the A.CD. To verify this empirically, we rely on the same dataset used for the simulations in Table 1, but for log-innovation variances we fit a linear structure rather than the true cubic polynomial. The impact of this innovation variance misspecification on estimating the correlation matrix can

Table 1 (a). Simulating data from $\boldsymbol{\Sigma}_{\text{mcd}}$ and fitting Poly$(3, 3)$ (cubic fit for innovation variance). Values within parentheses are empirical standard errors.

		Simulating from $\boldsymbol{\Sigma}_{\text{mcd}}$							
		$\nu = 4$				$\nu = 50$			
	Risk type	A.CD.T	A.CD.N	M.CD.T	M.CD.N	A.CD.T	A.CD.N	M.CD.T	M.CD.N
n=25	Entropy	0.8379	1.1563	0.4009	0.7870	0.9169	0.9334	0.5043	0.5144
		(0.4901)	(0.8780)	(0.3581)	(0.8900)	(0.4661)	(0.4563)	(0.3958)	(0.4043)
	Quadratic	2.4038	3.5832	0.9126	1.9400	2.5983	2.6308	1.1184	1.1392
		(2.0487)	(3.9302)	(1.2406)	(2.8187)	(1.7479)	(1.7201)	(1.1642)	(1.1911)
n=100	Entropy	0.6206	0.7224	0.1215	0.2653	0.6555	0.6544	0.1118	0.1124
		(0.1591)	(0.2539)	(0.0920)	(0.2827)	(0.1857)	(0.1825)	(0.1189)	(0.1205)
	Quadratic	1.7016	2.0171	0.2490	0.5677	1.8151	1.8118	0.2470	0.2480
		(0.5412)	(0.8908)	(0.1976)	(0.6303)	(0.6137)	(0.6037)	(0.3166)	(0.3227)

(b). Simulating data from $\boldsymbol{\Sigma}_{\text{acd}}$ and fitting Poly$(3, 3)$ (cubic fit for innovation variance)

		Simulating from $\boldsymbol{\Sigma}_{\text{acd}}$							
		$\nu = 4$				$\nu = 50$			
	Risk type	A.CD.T	A.CD.N	M.CD.T	M.CD.N	A.CD.T	A.CD.N	M.CD.T	M.CD.N
n=25	Entropy	0.3460	0.7072	0.5866	0.9780	0.3583	0.3641	0.6337	0.6516
		(0.3597)	(0.9609)	(0.3735)	(0.9712)	(0.2682)	(0.2813)	(0.3476)	(0.3820)
	Quadratic	0.8250	1.8045	1.0928	2.1981	0.7807	0.7846	1.0805	1.0848
		(1.2135)	(3.6170)	(1.1615)	(4.6778)	(0.7516)	(0.7764)	(0.5850)	(0.5876)
n=100	Entropy	0.0917	0.2681	0.3283	0.5041	0.0826	0.0849	0.3116	0.3152
		(0.0747)	(0.4659)	(0.1573)	(0.4378)	(0.0807)	(0.0830)	(0.1597)	(0.1597)
	Quadratic	0.1813	0.6898	0.5320	0.9643	0.1694	0.1750	0.5146	0.5204
		(0.1495)	(2.1286)	(0.2159)	(1.7977)	(0.1819)	(0.1883)	(0.2425)	(0.2425)

be seen in Table 2. More precisely, we observe the followings:

1. Comparing the first two A.CD columns of Table 1 with the first two columns of Tables 2 in both panels, shows that the correlation estimation is robust to the model misspecification for innovation variances. This conclusion seems to be independent of the structure of the covariance matrix used for the simulation ($\boldsymbol{\Sigma}_{\text{mcd}}$ or $\boldsymbol{\Sigma}_{\text{acd}}$).
2. The last two M.CD columns of Table 1(a) (Simulation from $\boldsymbol{\Sigma}_{\text{mcd}}$) have smaller risks compare to the last two columns of Table 2(a) in both panels. This confirms that the correlation estimation is not robust to the model misspecification for innovation variances in the M.CD structure.

Finally, we undertook a simulation study to examine the performance and flexibility of the proposed A.CD.T approach. The main objective is to study the robustness or sensitivity to the true distribution. For example, it is important to know when data are from a t distribution, how bad the M.CD or A.CD will perform when we use the normal distribution to estimate the parameters, and vice versa? For the sake of diversity, now the true parameters are set to be those of the tumor data (discussed in subsection 4.2) analyzed next and fitted with A.CD.T-Poly$(3, 3)$, except that the *df* is specified at two different settings. For the *df*'s, we take a low value ($\nu = 4$) corresponding to heavy-tailed distributions and a high value ($\nu = 50$) corresponding to near normality. The two sample sizes were from small ($n = 25$) to a relatively large ($n = 100$). Simulations were run with $m = 500$ replications for each combination of ν and n and each simulated data set was fitted under A.CD.T and A.CD.N scenarios. The detailed numerical results, including the average ML estimates for the fixed effects, the moving average parameters and the scale innovation variances, the average of maximized log-likelihood values $\ell_{\max}$, the average of associated BIC values and the median estimates for the *df*, together with their standard errors in parentheses, are summarized in Table 3. It shows that for smaller ν the point estimators of the parameters under the A.CD.T and A.CD.N scenarios are generally the same, but their SE's differ with the normal distributions leading to larger SE's. Furthermore, the estimated *df* has a downward bias for the smaller sample size $n = 25$.

4.2 The tumor growth data

We apply our methodology to the *in vivo* growth of lung tumor for the control group of 22 xenografted nude mice, which has been also analyzed in Lin & Wang (2009) using M.CD.T. Figure 1 shows the profile plot of the logarithm of tumor growth volumes over an un-

Table 2 (a). Simulating data from $\boldsymbol{\Sigma}_{\text{mcd}}$ and fitting Poly$(3, 1)$ (linear fit for innovation variance). Values within parentheses are empirical standard errors.

		Simulating from $\boldsymbol{\Sigma}_{\text{mcd}}$							
		$\nu = 4$				$\nu = 50$			
	Risk type	A.CD.T	A.CD.N	M.CD.T	M.CD.N	A.CD.T	A.CD.N	M.CD.T	M.CD.N
n=25	Entropy	0.8651	1.1367	0.7890	1.0247	0.9580	0.9846	0.8655	0.8867
		(0.4200)	(0.7285)	(0.3993)	(0.7108)	(0.4393)	(0.4437)	(0.3905)	(0.3985)
	Quadratic	2.4347	3.3864	2.1492	2.9354	2.6603	2.7305	2.3016	2.3523
		(1.7953)	(3.1821)	(1.5791)	(2.9867)	(1.6925)	(1.7229)	(1.4214)	(1.4603)
n=100	Entropy	0.6523	0.7618	0.5892	0.6865	0.6739	0.6742	0.6072	0.6071
		(0.1467)	(0.2444)	(0.1431)	(0.2362)	(0.1449)	(0.1429)	(0.1470)	(0.1473)
	Quadratic	1.7762	2.0849	1.5803	1.8342	1.8601	1.8593	1.6486	1.6466
		(0.5336)	(0.8488)	(0.4979)	(0.7583)	(0.5209)	(0.5124)	(0.5060)	(0.5064)

(b). Simulating data from $\boldsymbol{\Sigma}_{\text{acd}}$ and fitting Poly$(3, 1)$ (linear fit for innovation variance)

		Simulating from $\boldsymbol{\Sigma}_{\text{acd}}$							
		$\nu = 4$				$\nu = 50$			
	Risk type	A.CD.T	A.CD.N	M.CD.T	M.CD.N	A.CD.T	A.CD.N	M.CD.T	M.CD.N
n=25	Entropy	0.3942	0.7456	0.4181	0.7617	0.4552	0.4665	0.4592	0.4712
		(0.2791)	(0.8003)	(0.2773)	(0.8229)	(0.3258)	(0.3577)	(0.3159)	(0.3406)
	Quadratic	0.8791	1.8830	0.8315	1.8184	0.7768	0.7818	0.7841	0.7868
		(0.4917)	(2.4344)	(0.5498)	(2.6867)	(0.5345)	(0.5773)	(0.5383)	(0.5585)
n=100	Entropy	0.1289	0.3237	0.2158	0.4358	0.1276	0.1323	0.1995	0.2042
		(0.1548)	(0.8596)	(0.1412)	(0.8468)	(0.1383)	(0.1390)	(0.1234)	(0.1242)
	Quadratic	0.2295	0.7257	0.3456	0.7676	0.2030	0.2102	0.3294	0.3364
		(0.2135)	(1.7008)	(0.1913)	(1.7750)	(0.1978)	(0.1982)	(0.1780)	(0.1778)

Table 3 Average estimates for $\boldsymbol{\gamma}, \boldsymbol{\lambda}, \ell_{\max}$ and BIC and the median estimate for ν based on 500 replications. Values within parentheses are empirical standard errors.

Param.	True Param.	n=25				n=100			
		$\nu = 4$		$\nu = 50$		$\nu = 4$		$\nu = 50$	
		A.CD.T	A.CD.N	A.CD.T	A.CD.N	A.CD.T	A.CD.N	A.CD.T	A.CD.N
γ_0	0.9318	0.9293	0.9261	0.9381	0.9373	0.9326	0.9288	0.9268	0.9268
		(0.0085)	(0.0115)	(0.0081)	(0.0081)	(0.0040)	(0.0060)	(0.0037)	(0.0038)
γ_1	0.0962	0.1230	0.1442	0.1533	0.1538	0.1106	0.1220	0.0979	0.0988
		(0.0200)	(0.0275)	(0.0183)	(0.0183)	(0.0091)	(0.0141)	(0.0086)	(0.0086)
γ_2	0.0898	0.1012	0.1119	0.1116	0.1101	0.0985	0.1020	0.0935	0.0942
		(0.0100)	(0.0127)	(0.0095)	(0.0095)	(0.0047)	(0.0072)	(0.0045)	(0.0045)
γ_3	0.3041	0.3087	0.3095	0.3076	0.3063	0.3076	0.3053	0.3011	0.3015
		(0.0054)	(0.0074)	(0.0052)	(0.0052)	(0.0027)	(0.0043)	(0.0025)	(0.0025)
λ_0	-1.6379	-1.6706	-1.7175	-1.7031	-1.6693	-1.6919	-1.6668	-1.6690	-1.6468
		(0.0044)	(0.0064)	(0.0022)	(0.0021)	(0.0020)	(0.0037)	(0.0011)	(0.0010)
λ_1	-0.5685	-0.5801	-0.5989	-0.5844	-0.5821	-0.5760	-0.5799	-0.5721	-0.5725
		(0.0071)	(0.0099)	(0.0068)	(0.0069)	(0.0033)	(0.0050)	(0.0033)	(0.0033)
λ_2	0.5416	0.5241	0.5121	0.5280	0.5264	0.5396	0.5307	0.5375	0.5376
		(0.0058)	(0.0075)	(0.0055)	(0.0055)	(0.0026)	(0.0045)	(0.0026)	(0.0027)
λ_3	-0.2795	-0.2766	-0.2745	-0.2812	-0.2822	-0.2800	-0.2776	-0.2774	-0.2774
		(0.0044)	(0.0061)	(0.0040)	(0.0040)	(0.0021)	(0.0031)	(0.0021)	(0.0021)
ν		4.1040	.	34.5796	.	4.0612	.	49.7055	.
		(0.2006)	.	(2.8122)	.	(0.0329)	.	(2.7051)	.
$\ell_{\max}$		121.67	89.560	76.112	75.110	451.18	297.40	275.4831	273.40
		(1.3179)	(1.9102)	(0.6318)	(0.6385)	(2.4676)	(4.4406)	(1.2285)	(1.2430)
BIC		-7.0301	-4.5897	-3.3851	-3.4337	-8.0566	-5.0270	-4.5426	-4.5470
		(0.1054)	(0.1528)	(0.0505)	(0.0511)	(0.0494)	(0.0888)	(0.0246)	(0.0249)

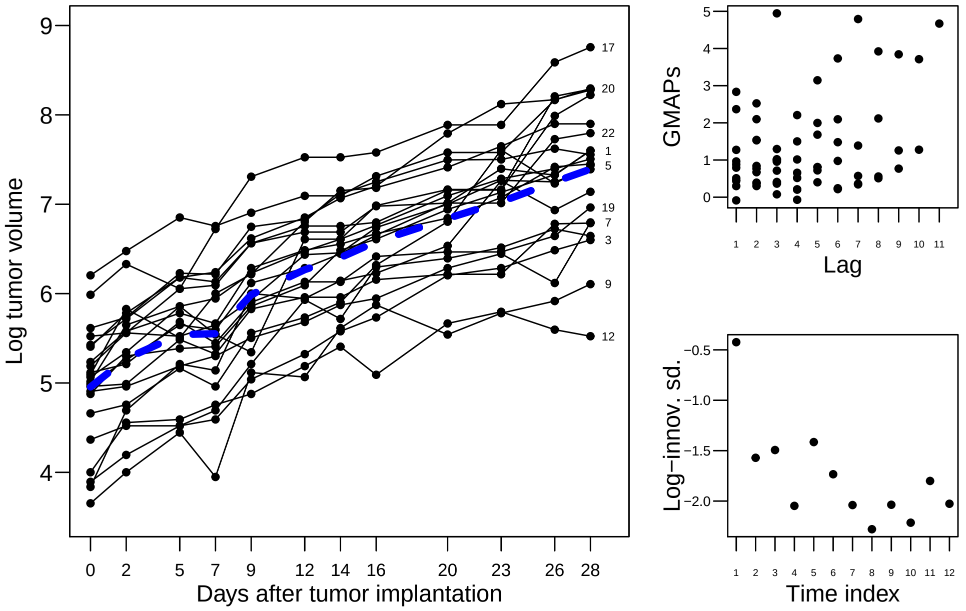

Fig. 1 Profile plot of the tumor data (left panel), where the least square estimates of the saturated model for the mean function is shown with dashed line. The plots of GMAPs (upper right panel) and log-innovation standard deviation (lower right panel).

equally spaced 28-day period for the 22 mice, together with the sample regressograms of the generalized moving average parameters (GMAPs), and the sample innovation standard deviations. It should be noted that our analysis is based on the saturated model for the mean function, where a separate parameter for the mean response at each time has been considered (Pourahmadi 2000; Diggle et al. 2002, p. 65; Pan & MacKenzie 2003). In fact, following the analysis of Lin & Wang (2009) and using the design matrix for the mean response to be $\boldsymbol{X}_i = [\mathbf{1}\ \boldsymbol{k}]$, where $\mathbf{1} = (1, 1, ..., 1)^\top$, $\boldsymbol{k} = (0, 1, 2.5, 3.5, 4.5, 6, 7, 8, 10, 11.5, 13, 14)^\top$, the optimization procedure using the Newton-Raphson algorithm for the A.CD.T model will converge only to a local maximum which depends noticeably on the choice of the initial values. However, using the saturated mean model the algorithm converges to the global maximum for both A.CD.T and A.CD.N. We fit the tumor data using A.CD.N and A.CD.T for various choices of the degrees of the Poly(d, q) models. The values of $\ell_{\max}$, together with the corresponding number of parameters and BIC values for selected pairs(d, q) are listed in Table 4. Judging from the BIC values, Poly$(6, 5)$ is the best and also Poly$(3, 5)$ is relatively parsimonious and a competitive choice for both A.CD.N and A.CD.T models. Table 5 shows the ML estimates and the associated standard errors for the best two fitting A.CD.N and A.CD.T. It is noteworthy that the estimates of the *df* for the two fitted A.CD.T are somewhat small, suggesting that the error distribution has a larger tail than the normal distribution, which confirms the finding of Lin & Wang (2009). Finally note that, based on the different interpretation of A.CD and M.CD parameters, the GMAPs and GARPs are not comparable.

5 Conclusions

We have established the role of an alternative Cholesky decomposition of the covariance matrix of a longitudinal dataset in providing robust estimator of its correlation matrix. Depending on the true structure of the underlying covariance matrix, whether it is from M.CD or A.CD models, the respective model will outperform the other in obtaining an efficient estimator for the covariance structure. Robustness to outliers is handled using heavy-tailed multivariate t-distributions with unknown degrees of freedom. Simulations and a real data example confirm the benefit of using the multivariate t-distribution to obtain a relatively more robust estimate of the parameters.

Newton-Raphson algorithm with Fisher scoring for computing the maximum likelihood estimators of the parameters of the alternative Cholesky decomposition turns out to be more complicated than the standard

Table 4 Comparison of $\ell_{\max}$, number of parameters, and BIC values for some Poly(d, q) choices of A.CD.N and A.CD.T models.

Poly(d,q)	# of param.		$\ell_{\max}$		BIC	
	A.CD.N	A.CD.T	A.CD.N	A.CD.T	A.CD.N	A.CD.T
(1,1)	4	5	-26.94	-9.575	4.697	3.259
(1,2)	5	6	-14.64	-0.134	3.719	2.541
(1,3)	6	7	-14.21	1.661	3.821	2.519
(1,4)	7	8	-13.70	3.041	3.915	2.534
(1,5)	8	9	-10.17	5.541	3.734	2.447
(1,6)	9	10	-10.15	5.596	3.873	2.582
(2,1)	5	6	-24.36	-8.929	4.603	3.341
(2,2)	6	7	-14.17	1.244	3.817	2.556
(2,3)	7	8	-14.01	2.168	3.943	2.613
(2,4)	8	9	-13.48	3.577	4.036	2.625
(2,5)	9	10	-10.09	5.982	3.868	2.547
(2,6)	10	11	-10.07	6.021	4.007	2.684
(3,1)	6	7	-22.01	-7.462	4.530	3.348
(3,2)	7	8	-13.10	2.239	3.860	2.606
(3,3)	8	9	-11.68	5.364	3.872	2.463
(3,4)	9	10	-10.44	8.812	3.899	2.290
(3,5)	10	11	-7.911	10.42	3.810	*2.284*
(3,6)	11	12	-7.909	10.43	3.951	2.424
(4,1)	7	8	-22.01	-7.362	4.670	3.479
(4,2)	8	9	-13.06	2.428	3.998	2.730
(4,3)	9	10	-11.65	5.822	4.010	2.562
(4,4)	10	11	-10.00	8.928	4.000	2.420
(4,5)	11	12	-6.377	11.21	3.811	2.353
(4,6)	12	13	-6.328	11.23	3.947	2.492
(5,1)	8	9	-21.96	-7.316	4.807	3.616
(5,2)	9	10	-13.06	2.768	4.138	2.839
(5,3)	10	11	-11.62	6.695	4.147	2.623
(5,4)	11	12	-9.914	9.848	4.133	2.477
(5,5)	12	13	-4.365	13.98	3.769	2.241
(5,6)	13	14	-4.242	14.34	3.898	2.349
(6,4)	12	13	-7.730	12.34	4.075	2.391
(6,5)	13	14	-2.415	16.81	3.732	**2.125**
(6,6)	14	15	-2.247	16.84	3.857	2.263
(7,4)	13	14	-7.481	12.40	4.193	2.526
(7,5)	14	15	-2.297	16.81	3.862	2.266
(7,6)	15	16	-2.145	16.84	3.989	2.404

Table 5 Parameter estimates for the best two Poly(d, q) choices of A.CD.N and A.CD.T

	Poly$(6,5)$				Poly$(3,5)$			
	A.CD.N		A.CD.T		A.CD.N		A.CD.T	
	MLE	SE	MLE	SE	MLE	SE	MLE	SE
γ_0	1.0026	0.1828	0.9755	0.1957	0.9393	0.1722	0.9292	0.1845
γ_1	0.4299	0.4335	0.0853	0.4538	0.3374	0.4037	0.0841	0.4290
γ_2	0.1969	0.2155	0.1426	0.2372	0.1463	0.2007	0.1666	0.2260
γ_3	0.4648	0.1451	0.5574	0.1585	0.2430	0.1137	0.3917	0.1270
γ_4	0.3352	0.1172	0.2655	0.1207	.	.	.	.
γ_5	0.1627	0.0922	0.1616	0.0926	.	.	.	.
γ_6	-0.1458	0.0679	-0.1693	0.0679	.	.	.	.
λ_0	-1.4098	0.0435	-1.7097	0.1003	-1.3890	0.0435	-1.6733	0.0981
λ_1	-0.7341	0.1505	-0.5697	0.1597	-0.6866	0.1501	-0.5311	0.1584
λ_2	0.5473	0.1204	0.6631	0.1271	0.5148	0.1244	0.6080	0.1308
λ_3	-0.1595	0.0961	-0.2767	0.1001	-0.1500	0.1010	-0.2727	0.1049
λ_4	-0.1174	0.0850	-0.2006	0.0849	-0.0436	0.0864	-0.1522	0.0870
λ_5	-0.2492	0.0752	-0.2164	0.0731	-0.1348	0.0716	-0.1080	0.0711
ν	.	.	3.4490	1.1747	.	.	3.6626	1.2688

Cholesky decomposition. This computational complexity is comparable to maximum likelihood estimation of parameters of the moving average models from time series analysis.

We would like to point out that, although often useful for restricting the influence of outliers, the use of the multivariate t-distribution alone does not necessarily guarantee robustness. Deriving robustness characteristics such as breakdown points and influence functions for the t-distribution-based approach is an open problem. On the other hand, the connection of variants of the Cholesky decomposition with AR and MA models as discussed in Section 1 of this paper suggests that various robust time series methods as reviewed in Chapter 8 of Maronna et al. (2006) could be extended for robust estimation of the correlation/covariance matrices. Systematic development of such extensions is left for future research.

Acknowledgements

We would like to thank an AE and the referees for their constructive comments, Dr. T.-I. Lin for providing us the tumor growth data. The work of the second author was partially supported by the NSF grant DMS-0906252, and that of the third was partially supported by grants from NCI (CA57030), NSF (DMS-0907170), and by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

References

Brockwell, P. J. & Davis, R. A. (1991), *Time Series: Theory and Methods*, second edn, Springer.

Cai, B., Dunson, D. B. & Gladen, T. B. (2006), 'Bayesian covariance selection in generalized linear mixed models', *Biometrics* **62**, 446–457.

Cannon, M. J., Warner, L., Taddei, J. A. & Kleinbaum, D. G. (2001), 'What can go wrong when you assume that correlated data are independent: an illustration from the evaluation of a childhood health intervention in brazil.', *Stat Med* **20**(9-10), 1461–7.

Carroll, R. J. (2003), 'Variances are not always nuisance parameters', *Biometrics* **59**(2), 211–220.

Chen, Z. & Dunson, D. B. (2003), 'Random effects selection in linear mixed models', *Biometrics* **59**(4), 762–769.

Chiu, T. Y. M., Leonard, T. & Tsui, K.-W. (1996), 'The matrix-logarithmic covariance model', *Journal of the American Statistical Association* **91**(433), pp. 198–210.

Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. (2002), *Analysis of Longitudinal Data*, 2 edn, Oxford University Press, USA.

Diggle, P. J. & Verbyla, A.-n. P. (1998), 'Nonparametric estimation of covariance structure in longitudinal data', *Biometrics* **54**(2), 401–415.

Holan, S. & Spinka, C. (2007), 'Maximum likelihood estimation for joint mean-covariance models from unbalanced repeated-measures data', *Statistics & Probability Letters* **77**(3), 319–328.

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

Lange, K. L., Little, R. J. A. & Taylor, J. M. G. (1989), 'Robust statistical modeling using the t distribution', *Journal of the American Statistical Association* **84**(408), pp. 881–896.

Leng, C., Zhang, W. & Pan, J. (2010), 'Semiparametric mean-covariance regression analysis for longitudinal data', *Journal of the American Statistical Association* **105**(489), 181–193.

Liang, K. Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

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**(9), 3013 – 3026.

Maronna, R., Martin, R. & Yohai, V. (2006), *Robust statistics: theory and methods*, Wiley series in probability and statistics, J. Wiley.

McCullagh, P. & Nelder, J. A. (1989), *Generalized linear models (Second edition)*, London: Chapman & Hall.

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

Pinheiro, J. C., Liu, C. & Wu, Y. N. (2001), 'Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution', *Journal of Computational and Graphical Statistics* **10**(2), pp. 249–276.

Pourahmadi, M. (1999), 'Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation', *Biometrika* **86**(3), 677–690.

Pourahmadi, M. (2000), 'Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix', *Biometrika* **87**(2), 425–435.

Pourahmadi, M. (2001), *Foundations of Time Series Analysis and Prediction Theory*, Wiley, New York.

Pourahmadi, M. (2007), 'Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance correlation parameters', *Biometrika*

94(4), 1006–1013.

Rothman, A. J., Levina, E. & Zhu, J. (2010), 'A new approach to Cholesky-based covariance regularization in high dimensions', *Biometrika* **97**(3), 539–550.

Wang, Y.-G. & Carey, V. (2003), 'Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance', *Biometrika* **90**(1), 29–41.

Welsh, A. & Richardson, A. (1997), Approaches to the robust estimation of mixed models, *in* G. Maddala & C. Rao, eds, 'Robust Inference', Vol. 15 of *Handbook of Statistics*, Elsevier, pp. 343 – 384.

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

Zellner, A. (1976), 'Bayesian and non-bayesian analysis of the regression model with multivariate student-t error terms', *Journal of the American Statistical Association* **71**(354), pp. 400–405.

Zimmerman, D. L. & Núñez Antón, V. (2009), *Antedependence Models for longitudinal Data*, Chapman & Hall / CRC Press, New York.