Biometrika (2000), **87,** 2, *pp.* 425–435

Printed in Great Britain

Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix

BY MOHSEN POURAHMADI

Division of Statistics, Northern Illinois University, DeKalb, Illinois 60115, U.S.A.

pourahm@math.niu.edu

SUMMARY

The positive-definiteness constraint is the most awkward stumbling block in modelling the covariance matrix. Pourahmadi's (1999) unconstrained parameterisation models covariance using covariates in a similar manner to mean modelling in generalised linear models. The new covariance parameters have statistical interpretation as the regression coefficients and logarithms of prediction error variances corresponding to regressing a response on its predecessors. In this paper, the maximum likelihood estimators of the parameters of a generalised linear model for the covariance matrix, their consistency and their asymptotic normality are studied when the observations are normally distributed. These results along with the likelihood ratio test and penalised likelihood criteria such as BIC for model and variable selection are illustrated using a real dataset.

Some key words: Asymptotic normality; Cholesky decomposition; Fisher information; Newton–Raphson algorithm; Unconstrained parameterisation; Variable selection and diagnostics.

1. INTRODUCTION

This follow-up paper to Pourahmadi (1999) continues the author's programme to apply the familiar iterative three-stage statistical model-fitting process to covariance matrices. The goal is to develop a flexible, systematic and data-based methodology for modelling covariance matrices similar to the generalised linear model framework for mean modelling. The success of the latter is partly because a link function is used to induce unconstrained parameterisation of the mean vector.

Unfortunately, most approaches to modelling covariance matrices do not heed adequately the positive-definiteness constraint. However, there is recent progress in unconstrained reparameterisation using either the matrix-logarithm or variants of the Cholesky decomposition of a covariance matrix (Leonard & Hsu, 1992; Chiu, Leonard & Tsui, 1996; Pinheiro & Bates, 1996; Pourahmadi, 1999). These references along with Diggle, Liang & Zeger (1994, Ch. 4, 5), provide good reviews of the literature on modelling covariance matrices. Using the fact that the matrix-logarithm of a positive-definite matrix is symmetric but otherwise unconstrained, Chiu et al. (1996) defined the class of matrix-logarithm covariance models by $\log \Sigma = \alpha_1 U_1 + \ldots + \alpha_q U_q$, where the U_i's are known symmetric matrices and the α_i's are unconstrained. The α_i's, however, do not always have simple statistical interpretation. Pourahmadi (1999) shows that the modified Cholesky decomposition of Σ^{-1} offers a simple unconstrained and statistically meaningful reparameterisation of the covariance matrix. In fact, for a random vector $y = (y_1, \ldots, y_n)'$ with

mean vector μ and positive-definite covariance matrix $\Sigma = (\sigma_{kl})$, there are a unique unit lower triangular matrix T with 1's as diagonal entries and a unique diagonal matrix D with positive diagonal entries such that

$$T\Sigma T' = D \text{ or } \Sigma^{-1} = T'D^{-1}T. \tag{1}$$

Fortunately, it is easy to interpret T and D statistically: the below-diagonal entries of T are the negatives of the coefficients of $\hat{y}_t = \mu_t + \sum_{j=1}^{t-1} \phi_{tj}(y_j - \mu_j)$, the linear least-squares predictor of y_t based on its predecessors $y_{t-1}, \ldots, y_1$, and the diagonal entries of D are the prediction error variances $\sigma_t^2 = \text{var}(y_t - \hat{y}_t)$, for $1 \leqslant t \leqslant n$. Since ϕ_{tj} and $\log \sigma_t^2$ are unconstrained, we may model them in terms of covariates. To this end, for $t = 1, \ldots, n$ and $j = 1, \ldots, t-1$, consider the models

$$\mu_t = x_t'\beta, \quad \log \sigma_t^2 = z_t'\lambda, \quad \phi_{tj} = z_{tj}'\gamma, \tag{2}$$

where x_t, z_t and z_{tj} are $p \times 1$, $q \times 1$ and $d \times 1$ vectors of known covariates, and $\beta = (\beta_1, \ldots, \beta_p)'$, $\lambda = (\lambda_1, \ldots, \lambda_q)'$ and $\gamma = (\gamma_1, \ldots, \gamma_d)'$ are parameters for the means, variances and correlations of y, respectively. Note that the last two equations in (2) provide a generalised linear model for a covariance matrix with a link function $g(\Sigma) = 2I - T - T' + \log D$ (McCullagh & Nelder, 1989; Pourahmadi, 1999, p. 680). Our goal here is to study the maximum likelihood estimator of the parameters in (2).

The outline of the paper is as follows. Section 2 deals with the maximum likelihood estimation of mean and covariance, i.e. variance and correlation, parameters for normal data. The loglikelihood function has three distinct representations corresponding to the three submodels for mean, variance and correlation in (2), and surprisingly it is quadratic in the correlation parameters γ as it is in the means. Consequently, closed-form solution of the likelihood equation is possible for these parameters when λ is fixed. The likelihood equation for the latter, however, is nonlinear, and an iterative Newton–Raphson method is developed with computational complexity similar to that for the joint modelling of mean and variance heterogeneity (Smyth, 1989; Verbyla, 1993). Strong consistency and asymptotic normality of the maximum likelihood estimators are studied. The conditions and results are simpler than, but similar in spirit to, those in Chiu et al. (1996) for matrix-logarithm covariance models. The methodology, along with the likelihood ratio test and penalised likelihood criteria such as BIC for model and variable selection, is illustrated in § 3 using Kenward's (1987) cattle data.

A limitation of our approach based on (1) is its implicit assumption of an ordering in the responses. While this ordering is natural in the longitudinal setting, it may not be the case in some other situations. Of course, a reparameterisation of the covariance matrix not depending on the coordinate system is desirable. A good example of this is Chiu et al. (1996). However, comparing their transformation with (1) suggests that there might be a trade-off between coordinate-free parameterisation and statistical interpretability of the ensuing parameters. In the absence of natural ordering in the response one may rely either on the qualitative ordering proposed by Brown, Le & Zidek (1994, p. 88) in the context of specifying priors for Σ or decomposing the joint distribution of a set of random variables into a running sequence of conditional distributions as in graphical models (Cox & Wermuth, 1996, Ch. 3).

2. THE LIKELIHOOD FUNCTION AND PARAMETER ESTIMATION

2·1. *The likelihood function*

For clarity, simplicity of presentation and notation we consider only standard multivariate data or balanced longitudinal data (Diggle et al., 1994, p. 16). For $i = 1, 2, \ldots, m$, we

assume that $y_i \sim N(X_i\beta, \Sigma)$ are independent n-vectors, where y_i may stand for the n repeated measures on the ith subject, X_i its $n \times p$ design matrix, i.e. covariates, and β its $p \times 1$ vector of mean parameters. We assume throughout that the components T and D of the covariance matrix Σ are modelled as in (2).

It is known (Pourahmadi, 1999) that the loglikelihood function has three representations corresponding to the three submodels in (2):

$$\begin{aligned} -2L(\beta, \lambda, \gamma) &= m \log|\Sigma| + \sum_{i=1}^{m} (y_i - X_i\beta)'\Sigma^{-1}(y_i - X_i\beta) \\ &= m \sum_{t=1}^{n} \log \sigma_t^2 + \sum_{t=1}^{n} \frac{\mathrm{RSS}_t}{\sigma_t^2} \\ &= m \sum_{t=1}^{n} \log \sigma_t^2 + \sum_{i=1}^{m} \{r_i - Z(i)\gamma\}'D^{-1}\{r_i - Z(i)\gamma\}, \end{aligned} \tag{3}$$

where $r_i = y_i - X_i\beta = (r_{it})_{t=1}^n$, and RSS_t and $Z(i)$, both depending on the r_i's, are defined below. The $n \times d$ matrix $Z(i)$ is defined by

$$Z(i) = (z(i, 1), \ldots, z(i, n))', \quad z(i, t) = \sum_{j=1}^{t-1} r_{ij} z_{tj}, \tag{4}$$

where z_{tj} is the $d \times 1$ vector of covariates associated with the ϕ_{tj}.

The matrices W_t and W defined below are used in the Newton–Raphson algorithm and the asymptotic distribution of the estimator of γ. From (4) we have that

$$E\{z(i, t)z'(i, t)\} = \sum_{k=1}^{t-1}\sum_{l=1}^{t-1} E(r_{ik}r_{il})z_{tk}z_{tl}' = \sum_{k=1}^{t-1}\sum_{l=1}^{t-1} \sigma_{kl}z_{tk}z_{tl}' = W_t, \tag{5}$$

$$W = E\{Z'(i)D^{-1}Z(i)\} = \sum_{t=1}^{n} \sigma_t^{-2} E\{z(i, t)z'(i, t)\} = \sum_{t=1}^{n} \sigma_t^{-2} W_t. \tag{6}$$

Finally, we note that $z(i, 1) = 0$ so that the first row of $Z(i)$ is zero and hence $W_1 = 0$.

To define RSS_t appearing in (3), let $\hat{r}_{it}$ be the predictor of r_{it} based on its predecessors $r_{i,t-1}, \ldots, r_{i1}$ and let $r(t) = (r_{it})_{i=1}^m$ be the vector of centred observations made at the tth occasion on all m subjects and $\hat{r}(t) = (\hat{r}_{it})_{i=1}^m$. Then $\mathrm{RSS}_t = \sum_{i=1}^m (r_{it} - \hat{r}_{it})^2$ is indeed the sum of squared prediction errors or the residual sum of squares from the analysis of covariance of $r(t)$ with $r(t-1), \ldots, r(1)$ as covariates (Kenward, 1987). It follows from (2) and (4) that $\hat{r}_{it} = z'(i, t)\gamma$ and hence

$$\mathrm{RSS}_t = \sum_{i=1}^{m} \{r_{it} - z'(i, t)\gamma\}^2. \tag{7}$$

This representation is useful when computing its derivatives with respect to γ, but the former is more convenient for numerical computation, finding the moments and distribution of the random vector $R = (\mathrm{RSS}_1, \ldots, \mathrm{RSS}_n)'$. Note that the matrix T creates the vector of successive prediction errors for any random vector, with mean zero and covariance Σ, so that $Tr_i = r_i - \hat{r}_i$, and hence the entries to Tr_i are independent random variables. Some useful consequences of this property of T are summarised in the following lemma.

LEMMA 1. *With notation as in* (1), (3), (4) *and* (7) *we have*:
(a) $(r_{it} - \hat{r}_{it})_{t=1}^n = Tr_i \sim N(0, D)$, *for* $i = 1, 2, \ldots, m$;
(b) $E(\mathrm{RSS}_t) = m\sigma_t^2$, $E(\mathrm{RSS}_t \times \mathrm{RSS}_s) = m^2\sigma_t^2\sigma_s^2$, *for* $s \neq t$, *and* $E(\mathrm{RSS}_t)^2 = (2m + m^2)\sigma_t^4$;
(c) $\mathrm{RSS}_t/\sigma_t^2 \sim \chi_m^2$, *for* $t = 1, 2, \ldots, n$;

(d) *with* $r=(r_1', \ldots, r_m')'$ *and* $1_n=(1, \ldots, 1)'$ *we have* $R=(Tr')^{(2)}1_n$, *where, for a matrix* $A=(a_{ij})$, $A^{(2)}=(a_{ij}^2)$.

In view of (c), for given β and γ, the second representation in (3) is the loglikelihood for a variance model with RSS_t as the response, and corresponds to a generalised linear model with gamma errors and known scale parameter equal to 2 (Smyth, 1989). Part (d) suggests a simple procedure for computing R.

2·2. *The score function and the Fisher information*

This section is devoted to computing the score function, the Hessian matrix and the Fisher information for the parameters. It is convenient to write the $p+q+d$ parameters of model (2) as $\theta=(\beta', \lambda', \gamma')=(\beta', \alpha')'$ and to partition the score function, the Hessian matrix, the Fisher information and its inverse as

$$U(\theta)=\frac{\partial L(\theta)}{\partial \theta}=(U_1'(\beta), U_2'(\lambda), U_3'(\gamma))', \quad H=(H_{ij})_{i,j=1}^3, \quad I_\theta=(I_{ij}(\theta))_{i,j=1}^3, \quad I_\theta^{-1}=(I^{ij}(\theta)),$$

respectively.

We compute the components of I_θ as the negative of the expected value of $\partial U/\partial\theta$. Computation of H_{23} and I_{23} are more challenging and are discussed in the Appendix. It follows from (3), after some algebra, that

$$\begin{aligned} U_1(\beta)&=\sum_{i=1}^m X_i'\Sigma^{-1}r_i, \quad I_{11}=\sum_{i=1}^m X_i'\Sigma^{-1}X_i,\\ U_2(\lambda)&=\tfrac{1}{2}Z'(D^{-1}R-m1_n), \quad I_{22}=\tfrac{1}{2}mZ'Z,\\ U_3(\gamma)&=\sum_{i=1}^m Z'(i)D^{-1}\{r_i-Z(i)\gamma\}, \quad I_{33}=mW, \end{aligned} \tag{8}$$

where $Z=(z_1, \ldots, z_n)'$ is the design matrix for the $\log\sigma_t^2$'s in (2). Note that U_1 and U_3 are linear in β and γ so that no iteration is needed for solving these score equations if λ is fixed.

Next, we compute H_{22} and the entries of H_{12} and H_{13}, and show that I_{12} and I_{13} are zero matrices so that the mean parameter β and the covariance parameters λ, γ are orthogonal. Since $\sum$ depends on λ and γ, we have

$$\frac{\partial U_1}{\partial\lambda_j}=\sum_{i=1}^m X_i'\left(\frac{\partial\Sigma^{-1}}{\partial\lambda_j}\right)r_i, \quad \frac{\partial U_1}{\partial\gamma_j}=\sum_{i=1}^m X_i'\left(\frac{\partial\Sigma^{-1}}{\partial\gamma_j}\right)r_i,$$

$$H_{22}=-\frac{1}{2}\sum_{t=1}^m \sigma_t^{-2}\,\text{RSS}_t z_t z_t', \quad H_{33}=-\sum_{i=1}^m Z'(i)D^{-1}Z(i).$$

Since $E(r_i)=0$, it follows that

$$I_{12}=E\left(-\frac{\partial U_1}{\partial\lambda}\right)=0, \quad I_{13}=E\left(-\frac{\partial U_1}{\partial\gamma}\right)=0.$$

The matrix H and the Fisher information I_θ can now be constructed.

2·3. *The maximum likelihood estimators of* β *and* γ

Since the score functions U_1 and U_3 in (8) are linear in β and γ, the maximum likelihood estimators of β and γ have closed forms and hence various aspects of their finite sample

distributions can be assessed through either theoretical calculations or numerical simulations provided that λ is fixed. This is particularly attractive for the new correlation parameter γ about which and its estimator little is currently known. The score function U_2 is nonlinear in λ and an iterative method for computing the maximum likelihood estimator of λ is given in § 2·4.

Setting the score functions U_1 and U_3 to zero we obtain

$$\begin{aligned} \hat{\beta} = \hat{\beta}(\Sigma) &= \left(\sum_{i=1}^{m} X_i' \Sigma^{-1} X_i \right)^{-1} \sum_{i=1}^{m} X_i' \Sigma^{-1} y_i, \\ \hat{\gamma} = \hat{\gamma}(\beta, D) &= \left\{ \sum_{i=1}^{m} Z'(i) D^{-1} Z(i) \right\}^{-1} \sum_{i=1}^{m} Z'(i) D^{-1} r_i, \end{aligned} \tag{9}$$

where by $\hat{\gamma} = \hat{\gamma}(\beta, D)$ we mean an estimator of γ assuming β and D are known. This convention is used throughout. Note the striking similarities between the forms of the two estimators. In a sense, $\hat{\gamma}$ is easier to work with than $\hat{\beta}$ computationally since D is diagonal. On the other hand, while the exact distribution of $\hat{\beta}$ is known and given by $\hat{\beta} \sim N(\beta, I^{11})$, that of $\hat{\gamma}$ is unknown. However, it can be expressed in terms of functionals of certain quadratic and bilinear forms in normal random variables. In fact, using (4) for any i, we have

$$\begin{aligned} Z(i)' D^{-1} r_i &= \sum_{t=1}^{n} \sigma_t^{-2} r_{it} z(i,t) = \sum_{t=1}^{n} \sigma_t^{-2} \sum_{j=1}^{t-1} (r_{it} r_{ij}) z_{tj}, \\ Z(i)' D^{-1} Z(i) &= \sum_{t=1}^{n} \sigma_t^{-2} z(i,t) z(i,t)' = \sum_{t=1}^{n} \sigma_t^{-2} \sum_{k=1}^{t-1} \sum_{l=1}^{t-1} (r_{ik} r_{il}) z_{tk} z_{tl}'. \end{aligned} \tag{10}$$

These formulae are useful for the numerical calculation of $\hat{\gamma}$ as well as for simulating data to assess its finite sample distribution. The following simple examples illuminate what is needed.

Example 1. (a) Consider $\phi_{tj} = \gamma_1$, for $t = 2, \ldots, n$ and $j = 1, \ldots, t-1$. Then $d = 1$, $z_{tj} = 1$ and, from (9) and (10), the maximum likelihood estimator of γ_1 is given by

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^{m} \sum_{t=1}^{n} \sigma_t^{-2} \sum_{k=1}^{t-1} r_{it} r_{ik}}{\sum_{i=1}^{m} \sum_{t=1}^{n} \sigma_t^{-2} \sum_{k=1}^{t-1} \sum_{l=1}^{t-1} r_{ik} r_{il}}.$$

(b) Consider $\phi_{tj} = \gamma_2 (t-j)^{-1}$, for $t = 2, \ldots, n$ and $j = 1, 2, \ldots, t-1$. Then $d = 1$, $z_{tj} = (t-j)^{-1}$ and, from (9) and (10), the maximum likelihood estimator of γ_2 is given by

$$\hat{\gamma}_2 = \frac{\sum_{i=1}^{m} \sum_{t=1}^{n} \sigma_t^{-2} \sum_{k=1}^{t-1} r_{it} r_{ik} (t-j)^{-1}}{\sum_{i=1}^{m} \sum_{t=1}^{n} \sigma_t^{-2} \sum_{k=1}^{t-1} \sum_{l=1}^{t-1} r_{ik} r_{il} (t-j)^{-2}}.$$

2·4. *The Newton–Raphson algorithm*

The iterative Newton–Raphson algorithm updates current values $\tilde{\theta}$ to $\hat{\theta}$ using

$$\hat{\theta} = \tilde{\theta} - H^{-1}(\tilde{\theta}) U(\tilde{\theta}). \tag{11}$$

The Fisher scoring algorithm replaces the Hessian matrix H in (11) by its expectation I_θ. Since I_{12} and I_{13} are zero matrices, in this approach $\hat{\beta}$ and $\hat{\alpha}$ are obtained by solving separate equations, and the former is a generalised least squares estimator. An iterative Fisher scoring method for obtaining $\hat{\beta}$ and $\hat{\alpha}$, using an inner loop, Step 3, is as follows.

Step 1. Select an initial value $\tilde{\beta}$ for β.

Step 2. Compute $S = m^{-1}\sum_{i=1}^{m}(y_i - X_i\tilde{\beta})(y_i - X_i\tilde{\beta})'$ and its factors $\tilde{T}$ and $\tilde{D}$ in (1) to be used as initial values for T and D in the next step.

Step 3. For the inner loop, compute $\tilde{\alpha} = (\tilde{\lambda}', \tilde{\gamma}')'$ by solving the last two equations in (8) using the Newton–Raphson iterative method with Fisher scoring. At convergence, compute $D(\tilde{\lambda})$, $T(\tilde{\gamma})$ and $\tilde{\Sigma}^{-1} = T(\tilde{\gamma})'D^{-1}(\tilde{\lambda})T(\tilde{\gamma})$.

Step 4. Update the value $\tilde{\beta}$ using

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

Step 5. Stop the process if $\hat{\beta} \doteq \tilde{\beta}$ and take $\hat{\beta}$ as an estimate of β. The estimates of α, T, D and Σ are given by $\hat{\alpha} = \tilde{\alpha}$, $\hat{T} = T(\tilde{\gamma})$, $\hat{D} = D(\tilde{\lambda})$ and $\hat{\Sigma} = \tilde{\Sigma}$. Otherwise, repeat Steps 2–4 replacing $\tilde{\beta}$ by $\hat{\beta}$.

A convenient initial value for β is its ordinary least-squares estimate. We note that the last quadratic form in (3) can be interpreted as the weighted least-squares criterion for estimating the parameters β and γ of the dynamic linear model

$$y_i = X_i\beta + Z(i)\gamma + \varepsilon_i \quad (i = 1, \ldots, m), \tag{12}$$

where $\mathrm{cov}(\varepsilon_i) = D$. The phrase dynamic linear model is appropriate since the $Z(i)$ depends on the response y_i. This new set-up suggests an iteratively reweighted least-squares method for estimating $(\beta', \gamma')'$ and the parameters λ of D, a familiar approach in the context of generalised linear models (McCullagh & Nelder, 1989). A referee has pointed out that various modifications of the maximum likelihood method such as profile, modified profile and integrated likelihood (Leonard, 1982; Berger, Liseo & Wolpert, 1999) could produce less complicated solutions by exploiting the simple structure of (3). These alternatives, a Bayesian analysis and the numerical properties of the above Newton–Raphson algorithm are currently under study for a follow-up paper.

2·5. *Asymptotic distribution of the estimators*

In this section the consistency and asymptotic normality of the maximum likelihood estimators of β, λ and γ are presented under some mild regularity conditions. Throughout this section we assume that

(i) model (2) is correct,
(ii) the parameter spaces for β, λ and γ are compact subspaces of R^p, R^q and R^d, respectively, and
(iii) $\theta_0 = (\beta_0', \lambda_0', \gamma_0')'$ the true value of $\theta = (\beta', \lambda', \gamma')'$ is in the interior of the parameter space for θ.

Our proof, which is simpler than but in the spirit of Chiu et al. (1996), is sketched in the Appendix.

THEOREM 1. *Suppose that the design matrices in* (2) *and* X_i *are all bounded componentwise, i.e. all of their components are bounded by a single finite real number, and that*

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m} X_i'\Sigma X_i$$

exists and is finite. Then

(a) *the maximum likelihood estimator* $\hat{\theta} = (\hat{\beta}', \hat{\lambda}', \hat{\gamma}')'$ *is strongly consistent for* $\theta_0 = (\beta_0', \lambda_0', \gamma_0')'$;

(b) *the maximum likelihood estimator* $\hat{\theta}$ *is asymptotically normally distributed, with*

$$\sqrt{m}\begin{bmatrix} \hat{\beta} - \beta_0 \\ \hat{\lambda} - \lambda_0 \\ \hat{\gamma} - \gamma_0 \end{bmatrix} \to N(0, I_{\theta_0}^{-1})$$

in distribution as $m \to \infty$.

From the block-diagonal form of I_θ, it follows immediately that $\hat{\beta}$ and $\hat{\alpha} = (\hat{\lambda}', \hat{\gamma}')'$ are asymptotically independent. Since $(\hat{\beta}', \hat{\alpha}')'$ is a consistent estimator for θ_0, the asymptotic covariance matrix $I^{11}(\theta_0)$ of $\hat{\beta}$ can be estimated by $(m^{-1}\sum_{i=1}^{m} X_i'\hat{\Sigma}^{-1}X_i)^{-1}$, where $\hat{\Sigma} = \Sigma(\hat{\alpha})$. Similarly, the asymptotic covariance matrix of $\hat{\alpha}$ can be estimated. In the same vein, the empirical Fisher information matrix

$$\left(-\frac{1}{m}\frac{\partial^2 L(\theta)}{\partial\theta\,\partial\theta'}\bigg|_{\theta=\hat{\theta}} \right)^{-1}$$

can be used to approximate the asymptotic covariance matrix of $\hat{\theta}$.

3. The cattle data

Kenward's (1987) cattle study deals with cattle receiving two treatments A and B for intestinal parasites. They were weighed $n = 11$ times over a 133-day period. The first 10 measurements on each animal were made at two-week intervals and the final measurement was made after one week. Measurement times were common across animals and are rescaled to $t = 1, 2, \ldots, 10, 10{\cdot}5$; no observation was missing. Of 60 cattle, $m = 30$ received treatment A and the other 30 received treatment B. Zimmerman & Núñez-Antón (1997) rejected the equality of the two within treatment-group covariance matrices using the classical likelihood ratio test. Thus, it is advisable to study each treatment group's covariance matrix separately; here we report our results for the group A cattle.

For estimating the covariance structure of a dataset it is generally believed (Diggle et al., 1994, p. 64) that a sensible strategy is to use an over-elaborate or saturated model for the mean response profile. Thus, assuming a saturated mean model with $n = 11$ parameters, that is $\mu = (\mu_1, \ldots, \mu_{11})'$, we identify models for the 11×11 covariance matrix Σ of the data using the regressograms (Pourahmadi, 1999) and compute the maximum likelihood estimate of their parameters.

Let S be the sample covariance matrix of the group A cattle. Given its factors $\tilde{T}$ and $\tilde{D}$ as in (1), plots of $\tilde{\phi}_{tj}$ against j and $\log\tilde{\sigma}_t^2$ against t, called its regressograms, are extremely helpful in identifying parsimonious models for T and $\log D$ or Σ. Pourahmadi (1999) identified and estimated the following cubic polynomials for $t = 1, 2, \ldots, 11$ and $j = 1, \ldots, t-1$:

$$\begin{aligned} \log\tilde{\sigma}_t^2 &= 3{\cdot}37 - 1{\cdot}47t + 0{\cdot}24t^2 - 0{\cdot}93t^3 + \varepsilon_{tv}, \\ \tilde{\phi}_{tj} &= 0{\cdot}18 - 1{\cdot}7(t-j) + 1{\cdot}64(t-j)^2 - 1{\cdot}11(t-j)^3 + \varepsilon_{tjc}. \end{aligned} \tag{13}$$

This shows how a data-based covariance modelling procedure can model a 66-parameter covariance matrix parsimoniously using just 8 unconstrained parameters.

The estimates in (13) are the ordinary least-squares estimates of the parameters. In the rest of this section we compute the maximum likelihood estimates of the parameters and some of its nested submodels. To minimise notation, we use Poly(q, d) as a shorthand for polynomial models in t and $t-j$ of degree q for $\log \sigma_t^2$ and degree d for ϕ_{tj}, respectively. Note that a Poly(q, d) model for Σ has $q+d+2$ parameters. We have used S-Plus (Specter, 1994) for computations where polynomial regressions are fitted using orthogonal design matrices. Thus, the coefficients here are those corresponding to the orthogonal design matrices.

While modelling variances or entries of $\log D$ is perceived important in the literature of correlated data, the same importance is not accorded to modelling correlations or the entries of T. To emphasise the importance of modelling the latter, we fit Poly$(3, d)$ models for $d = 0, 1, 2, 3$ to Σ and compute the maximum likelihood estimates of their parameters, the maximised loglikelihood functions and the corresponding BIC values. These values are summarised in Tables 1 and 2, where Poly(3) is a cubic in t for the nonredundant elements in a diagonal Σ. Recall that BIC, a penalised likelihood criterion, for a model selection situation is defined as

$$\text{BIC} = -\frac{2}{m} L_{\max} + p \frac{\log m}{m},$$

where m is the sample size, $L_{\max}$ is the maximised loglikelihood for the model under consideration and p is its number of parameters. Smaller values of BIC are associated with better-fitting models.

Table 1. *Values of* $L_{\max}$, *number of parameters and* BIC *for several models. The last four rows are from Zimmerman & Núñez-Antón* (1997)

Model	$L_{\max}$	Number of parameters	BIC
Unstructured Σ	−1019·69	66	75·35
Poly(3, 3)	−1049·01	8	70·84
Poly(3, 2)	−1080·08	7	72·80
Poly(3, 1)	−1131·61	6	76·09
Poly(3, 0)	−1215·35	5	81·59
Poly(3)	−1377·43	4	92·28
Unstructured AD(2)	−1035·98	30	72·47
Structured AD(2)	−1054·13	8	71·18
Stationary AR(2)	−1062·89	3	71·20
Structured AD(2) with $\lambda_1 = \lambda_2 = 1$	−1054·20	6	70·96

By scanning the $L_{\max}$ and BIC columns in Table 1, we see that the cost of not modelling T properly is evident in both the decrease in $L_{\max}$ and the increase in the BIC values. The last four rows of Table 1, from Zimmerman & Núñez-Antón (1997), are included here for ease of comparison and reference. An unstructured antedependence model of order 2, or an AD(2) model for short, corresponds to (2) with $\phi_{tj} = 0$ for $t-j > 2$, which has 30 parameters, 11 on the diagonal of D and $10+9 = 19$ on the first two subdiagonals of T. These can be reduced by introducing the following structure (Zimmerman & Núñez-Antón, 1997): if $t_1 < t_2 < \ldots < t_n$ are the measurement times for an arbitrary subject, then

Table 2. *Maximum likelihood estimates of the variance parameters λ and correlation parameters γ for several nested* Poly(q, d) *models for the covariance matrix of the group* A *cattle data*

Model		$j=1$	$j=2$	$j=3$	$j=4$
Poly(3, 3)	λ	3·52	−1·14	0·30	−0·85
	γ	0·18	−1·71	1·64	−1·11
Poly(3, 2)	λ	3·71	−0·69	0·54	−0·62
	γ	0·18	−1·71	1·64	
Poly(3, 1)	λ	4·02	−0·22	0·63	−0·39
	γ	0·18	−1·71		
Poly(3, 0)	λ	4·53	0·12	0·59	0·23
	γ	0·18			
Poly(3)	λ	5·51	1·50	−0·17	0·04

for $i = 3, \ldots, n$ and $j = 1, 2$ set

$$\phi_{ij} = \phi_j^{t_i^{\lambda_j} - t_{i-1}^{\lambda_j}}, \quad \sigma_i^2 = \sigma^2,$$

where ϕ_1, ϕ_2, λ_1, λ_2, ϕ_{21}, σ_1^2, σ_2^2 and σ^2 are the new parameters of T and D. Judging from the BIC values, the Poly(3, 3) model is clearly the model of choice for Σ which happens to be close to the structured AD(2) model with $\lambda_1 = \lambda_2 = 1$ (Zimmerman & Núñez-Antón, 1997).

Alternatively, nested hypotheses about model parameters can be tested using likelihood ratio tests. For example, let L_1 denote the maximised loglikelihood for the Poly(3, 3) model for Σ and let L_0 denote the maximised loglikelihood for the submodel Poly(3, 2). We can test the null hypothesis that the submodel holds by comparing $2(L_1 - L_0) = 62{\cdot}14$ to the appropriate percentage point of the chi-squared distribution with $v = 1$ degree of freedom. The null hypothesis is clearly rejected so that the third power of $t - j$ is kept in (13).

From Table 2, it is evident that there are considerable changes in the magnitude and sign of the variance parameters λ, as one moves away along $d = 3, 2, 1, 0$ from the Poly(3, 3) model which is preferred by the BIC and the likelihood ratio test. However, the values of the correlation parameters γ do not change, because the design matrix employed in our calculations is orthogonal. The variances of the respective components of $\hat{\lambda}$ and $\hat{\gamma}$ for the final Poly(3, 3) model are

$$(0{\cdot}00606, 0{\cdot}06667, 0{\cdot}06667, 0{\cdot}06667, 0{\cdot}00001, 0{\cdot}00345, 0{\cdot}01085, 0{\cdot}02244)/30.$$

Acknowledgement

I would like to thank the editor and referees for their concrete suggestions that improved the presentation of the paper considerably. The work was supported by grants from the National Security Agency and the National Science Foundation.

APPENDIX

Proofs

Computations of H_{23} and I_{23}. It is evident from U_2 in (8) that we need to compute $\partial R/\partial\gamma$ and hence $\partial \mathrm{RSS}_t/\partial\gamma$ for $t = 1, 2, \ldots, n$, along with their expected values. From (7) we have, for $t = 1, 2, \ldots, n$,

$$\frac{\partial \mathrm{RSS}_t}{\partial\gamma} = 2\sum_{i=1}^{m}\{-r_{it}z(i,t) + z(i,t)z'(i,t)\gamma\}, \tag{A1}$$

and using (5) we obtain

$$E\left(\frac{\partial \mathrm{RSS}_t}{\partial\gamma}\right) = 2m\left(-\sum_{k=1}^{t-1}\sigma_{tk}z_{tk} + \sum_{k=1}^{t-1}\sum_{l=1}^{t-1}\sigma_{kl}z_{tk}z'_{tl}\gamma\right). \tag{A2}$$

Substituting $z'_{tl}\gamma$ by ϕ_{tl} in (A2) and collecting similar terms we get

$$E\left(\frac{\partial \mathrm{RSS}_t}{\partial\gamma}\right) = 2m\sum_{k=1}^{t-1}\left(\sum_{l=1}^{t-1}\sigma_{kl}\phi_{tl} - \sigma_{tk}\right)z_{tk} = -2m\sum_{k=1}^{t-1}a_{kt}z_{tk} = -2mb_t, \tag{A3}$$

where a_{kt} is the (k, t)th entry of the matrix $A = \Sigma T'$, and, for $t = 1, \ldots, n$,

$$b_t = \sum_{k=1}^{t-1}a_{kt}z_{tk},$$

which is similar to $z(i, t)$ in (4) but with r_{ij} replaced by a_{kt}. Thus, from (8), (A1) and (A3), we get

$$H_{23} = \frac{1}{2}Z'D^{-1}\frac{\partial R}{\partial\gamma}, \quad I_{23} = E\left(-\frac{\partial U_2}{\partial\gamma}\right) = mZ'D^{-1}B,$$

where $B = (b_1, \ldots, b_n)'$ resembles the definition of $Z(i)$ in (4).

Sketch of proof of Theorem 1. Our proof is essentially the same as the proofs of Theorems 1 and 2 in Chiu et al. (1996). Thus, we point out only those differences in computing certain moments that are mostly due to our different but certainly smoother reparameterisation of Σ. A more complete proof of the theorem is available from the author upon request.

(a) Let $L_i = L_i(\theta) = \log f(y_i; \theta)$ be the log of the density of y_i. Then, ignoring the constant $\frac{1}{2}n\log 2\pi$, we obtain

$$L_i = -\frac{1}{2}\sum_{t=1}^{n}\log\sigma_t^2 - \frac{1}{2}(y_i - X_i\beta)'\Sigma^{-1}(y_i - X_i\beta) = -\frac{1}{2}\left(\sum_{t=1}^{n}z_t\right)'\lambda - \frac{1}{2}r_i'\Sigma^{-1}r_i.$$

Next, we compute $E_0(L_i)$ and $V_0(L_i)$, the mean and variance of L_i when $\theta = \theta_0$. Note that, since

$$r_i = y_i - X_i\beta = y_i - X_i\beta_0 + X_i(\beta_0 - \beta), \quad E_0(r_i) = X_i(\beta_0 - \beta),$$

we have

$$E_0(r_ir_i') = \Sigma_0 + X_i(\beta_0 - \beta)(\beta_0 - \beta)'X_i',$$

where $\Sigma_0 = \Sigma(\theta_0)$. Thus, using $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ and known results on expectation and variance of quadratic forms of normal random variables, we obtain

$$E_0(L_i) = -\frac{1}{2}\left(\sum_{t=1}^{n}z_t\right)'\lambda - \frac{1}{2}\mathrm{tr}\,\Sigma^{-1}\Sigma_0 - \frac{1}{2}(\beta_0 - \beta)'X_i'\Sigma^{-1}X_i(\beta_0 - \beta),$$

$$V_0(L_i) = \frac{1}{4}\{\mathrm{tr}(\Sigma^{-1}\Sigma_0)^2 + 2(\beta_0 - \beta)'X_i'\Sigma^{-1}\Sigma_0\Sigma^{-1}X_i(\beta_0 - \beta)\}.$$

We recall that $\Sigma^{-1} = T'D^{-1}T$, $\Sigma_0 = T_0'^{-1}D_0T_0^{-1}$, where T, D, T_0 and D_0 have entries as in (2) with

$\alpha = (\lambda', \gamma')'$ and $\alpha_0 = (\lambda_0', \gamma_0')'$ as their parameters and z_t's and z_{tj}'s as their covariates. It follows from the compactness of these parameter spaces and boundedness of their covariates along with that of X_i that $V_0(L_i) \leqslant K$, for all i.

The rest of the proof of (a) is essentially the same, though its details are much simpler than the proof of Theorem 1 in Chiu et al. (1996, p. 207). The following identity is useful in verifying the equicontinuity of the sequence $\{m^{-1}\sum_{i=1}^{m} E_0(L_i)\}$ in θ as well as providing a glimpse of its limit $K_0(\theta)$ as $m \to \infty$:

$$\frac{-2}{m}\sum_{i=1}^{m} E_0(L_i) = \left(\sum_{t=1}^{n} z_t\right)' \lambda + \operatorname{tr}\Sigma^{-1}\Sigma_0 + \operatorname{tr}\left\{\frac{1}{m}\sum_{i=1}^{m} X_i(\beta_0 - \beta)(\beta_0 - \beta)'X_i'\right\}\Sigma^{-1}.$$

The proof of (b) is essentially the same as that of Theorem 2 in Chiu et al. (1996).

References

Berger, O. J., Liseo, B. & Wolpert, R. (1999). Integrated likelihood methods for eliminating nuisance parameters (with Discussion). *Statist. Sci.* **14**, 1–28.

Brown, P. J., Le, N. D. & Zidek, J. V. (1994). Inference for a covariance matrix. In *Aspects of Uncertainty*, Ed. P. R. Freeman and A. F. M. Smith, pp. 77–90. Chichester: John Wiley.

Chiu, T. Y. M., Leonard, T. & Tsui, K. W. (1996). The matrix-logarithm covariance model. *J. Am. Statist. Assoc.* **91**, 198–210.

Cox, D. R. & Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation.* London: Chapman and Hall.

Diggle, P. J., Liang, K. Y. & Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

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

Leonard, T. (1982). Comments on 'A simple predictive density function' by M. Lejeune and G. D. Faulkenberry. *J. Am. Statist. Assoc.* **77**, 657–8.

Leonard, T. & Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix. *Ann. Statist.* **20**, 1669–96.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

Pinheiro, J. D. & Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statist. Comp.* **6**, 289–96.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–90.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *J. R Statist. Soc.* B **51**, 47–60.

Specter, P. (1994). *An Introduction to S and S-Plus.* Belmont, CA: Duxbury.

Verbyla, A. P. (1993). Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *J. R. Statist. Soc.* B **55**, 493–508.

Zimmerman, D. L. & Núñez-Antón, V. (1997). Structured antedependence models for longitudinal data. In *Modeling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, Springer Lecture Notes in Statistics, No. 122, Ed. T. G. Gregoire et al., pp. 63–76. New York: Springer-Verlag.

[*Received April* 1999. *Revised November* 1999]