Simultaneous Modelling of the Cholesky Decomposition of Several Covariance Matrices

M. Pourahmadi
Division of Statistics
Northern Illinois University
DeKalb, IL 60115, USA
pourahm@math.niu.edu

M.J. Daniels & T. Park
Department of Statistics
University of Florida
Gainesville, FL 32611-8545, USA
mdaniels@stat.ufl.edu
tpark@stat.ufl.edu

Summary

A method for simultaneous modelling of the Cholesky decomposition of several covariance matrices is presented. We highlight the conceptual and computational advantages of the unconstrained parameterization of the Cholesky decomposition and compare the results with those obtained using the classical spectral (eigenvalue) and variance-correlation decompositions. All these methods amount to decomposing complicated covariance matrices into "dependence" and "variance" components, and then modelling them virtually separately using regression techniques. The entries of the "dependence" component of the Cholesky decomposition have the unique advantage of being unconstrained so that further reduction of the dimension of its parameter space is fairly simple. Normal theory maximum likelihood estimates for complete and incomplete data are presented using iterative methods such as EM (Expectation-Maximization) algorithm and their improvements. These procedures are illustrated using a dataset from a growth hormone longitudinal clinical trial.

Key Words: Common principal components; Longitudinal data; Maximum likelihood estimation; Missing data; Spectral decomposition; Variance-Correlation decomposition.

1 Introduction

Virtually all areas of classical multivariate statistics involve estimation of a single $p \times p$ covariance matrix with as many as $^{p(p+1)}/_{2}$ parameters (Anderson, 2003). Many modern applications instead require dealing with several $p \times p$ covariance matrices $\Sigma_1, \cdots, \Sigma_c$, corresponding to c separate groups of multivariate observations, where both c and p are potentially large. Often there are not enough data to adequately estimate a separate Σ_i for each group, but if Σ_i's share some common features, they can be estimated more efficiently by pooling the data. Prominent examples of this pooling

phenomenon include model-based principal component analysis (Flury, 1984, 1988); model-based cluster analysis and discriminant analysis (Murtagh and Raftery, 1984; Banfield and Raftery, 1993), longitudinal data analysis (Diggle et al. 2002), and multivariate volatility in finance (Bollerslev, Engle and Woodridge, 1988; Engle, 2002) where the number of covariances to be estimated could be as large as the number of observations.

Some of the most commonly used methods for handling several covariance matrices in the literature of multivariate statistics, the biomedical sciences, and financial econometrics are based on the spectral decomposition (Flury, 1984, 1988; Boik, 2002; Fraley and Raftery, 2002), the variance-correlation decomposition (Manly and Rayner, 1987; Barnard, McCulloch and Meng, 2000), and multivariate generalized autoregressive conditionally heteroscedastic (GARCH) models (Bollerslev, 1990; Engle, 2002). It is conceivable that a framework like Nelder and Wedderburn's (1972) generalized linear models (GLM) could be used to compare, unify and possibly generalize the above approaches to covariance modelling. Some of the powerful principles of the GLM are: (i) the use of *link functions* leading to unconstrained and interpretable parameters, (ii) the use of *linear predictors* to gauge the covariates effect additively and (iii) the use of a *likelihood method* for estimation of the parameters (Pourahmadi, 2000).

Even unequal Σ_i's may share certain common features (components). A natural way to search for common features is to decompose complicated covariance matrices into simpler "dependence" and "variance" components and scan these for simplifying patterns such as equality of the "dependence" components across groups. Three of the most popular approaches in increasing order of adherence to the GLM principles, employ the variance-correlation, spectral (eigenvalue) and Cholesky decompositions of covariance matrices. In this paper, we show the distinguished role of the latter in providing unconstrained reparameterization and a systematic data-based statistical procedure for parsimonious modelling of several covariance matrices. While the entries of the correlation and orthogonal matrices appearing in the variance-correlation and spectral decompositions are always constrained, those appearing in the unit lower triangular matrix of the Cholesky

decomposition, referred to as the generalized autoregressive parameters (GARP), are always unconstrained (Pourahmadi, 1999, 2000). Consequently, computing the maximum likelihood estimates (MLE) of the Cholesky decomposition involves unconstrained optimization, unlike the algorithms needed for estimation with the other two decompositions; see Flury and Gautschi (1986), Barnard et al. (2000) and Boik (2002).

The outline of the paper is as follows. The three decompositions of covariance matrices and the corresponding hierarchies of nested models are introduced in Section 2. Algorithms for computing the normal theory MLE of the parameters under the common correlation matrices, principal components, and GARP (with extensions) are presented in Section 3. For various hierarchies, we also partition the likelihood ratio statistic for testing the equality of c covariance matrices:

$$T = \sum_{i=1}^{c} n_i \log \frac{|S|}{|S_i|}, \tag{1}$$

where S_i is the sample covariance matrix of a sample of size $n_i + 1$ from the ith population and S is the pooled covariance matrix of all c samples (Anderson, 2003, Chap. 10). Section 4 develops the EM algorithm for computing the MLE of parameters of the Cholesky decomposition when data are incomplete. Application of these hierarchies in longitudinal clinical trials is detailed in Section 5 and illustrated using the data from a growth hormone clinical trial (Kiel et al.,1998). Section 6 concludes the paper.

2 The Three Decompositions and Parameterizations

In multivariate statistics (Anderson, 2003) and in the context of modelling common features of Σ_i's among c groups, the variance-correlation, spectral, and Cholesky decompositions are used frequently. They have the advantage of being familiar, providing interpretable parameterizations in some situations and giving rise to hierarchies of nested models. However, only the Cholesky decomposition provides a simple unconstrained parameterization with a fine enough hierarchy to allow models with any number of parameters from 1 to ${}^{cp(p+1)}/_{2}$ as in the GLM for a mean vector.

We start with the familiar variance-correlation decomposition given by

$$\Sigma_i = D_i R_i D_i, \tag{2}$$

where $D_i = diag\left(\sqrt{\sigma_{i11}}, \cdots, \sqrt{\sigma_{ipp}}\right)$ is a diagonal matrix whose diagonal entries are the square-roots of those of Σ_i and R_i is the corresponding correlation matrix. Manly and Rayner (1987) introduce a hierarchy and a corresponding ANOVA-type partition of (1) which we rely on as our road map in this paper. Their hierarchy has four coarse levels:

(M1) *equality*, $\Sigma_1 = \cdots = \Sigma_c$ with $p(p+1)/2$ parameters;

(M2) *proportionality*, $\Sigma_i = \rho_i \Sigma_1, i = 2, \cdots, c$ with $p(p+1)/2 + c - 1$ parameters;

(M3) *common correlation matrices*, $R_i \equiv R$, with $pc + p(p-1)/2$ parameters and

(M4) *arbitrary covariance matrices* with $cp(p+1)/2$ parameters.

The MLE of the parameters under (M3) is reviewed in Section 3.1.

Flury's (1984, 1988, Chap. 7) slightly finer hierarchy is based on the spectral decomposition of the covariance matrices:

$$\Sigma_i = P_i \Lambda_i P_i', i = 1, \cdots, c, \tag{3}$$

where P_i's are orthogonal matrices and $\Lambda_i =$ diag $(\lambda_{i1}, \cdots, \lambda_{ip})$ with λ_{ij} standing for the jth eigenvalue of Σ_i. His hierarchy replaces the (M3) by the following three variants of the *common principal components* (CPC):

(M′3) CPC, $P_i \equiv P$ for all i, with $d_3' = pc + p(p-1)/2$ parameters;

(M′4) CPC (q), *partial CPC of order* $q (1 \leq q \leq p-2)$ where the first q columns of P_i's are the same, with $d_3' + d_4'$ parameters and $d_4' = \frac{1}{2}(c-1)(p-q)(p-q-1)$;

(M′5) CS(q), *common space of order* q where the first q eigenvectors of Σ_i span the *same subspace* as those of Σ_1 with $d_3' + d_4' + \frac{1}{2}(c-1)q(q-1)$ parameters.

We review the MLE of the parameters under (M′3) in Section 3.1. Note that in the decompositions (2)-(3), the "dependence" components are a correlation matrix R_i and an orthogonal matrix P_i, respectively, and hence their elements are subject to constraints.

Next, a more flexible hierarchy among several covariance matrices is introduced using their modified Cholesky decompositions:

$$T_i \Sigma_i T_i' = \nu_i. \tag{4}$$

Here the "dependence" component T_i, a unit lower triangular matrix, has unconstrained entries with statistical interpretation as the *generalized autoregressive parameters* (GARP) and the entries of $\nu_i = \text{ diag } (\nu_{i1}^2, \cdots, \nu_{ip}^2)$ are the corresponding *innovation (residual) variances* (Pourahmadi, 1999). More concretely, let $Y = (Y_1, \cdots, Y_p)$ be a generic random vector with mean zero and positive-definite covariance matrix Σ. Let $\hat{Y}_j$ stand for the linear least-squares predictor of Y_j based on its predecessors $Y_{j-1}, \cdots, Y_1$ and ε_j be its prediction error:

$$\hat{Y}_j = \sum_{\ell=1}^{j-1} \phi_{j,\ell} Y_\ell, \varepsilon_j = Y_j - \hat{Y}_j = Y_j - \sum_{\ell=1}^{j-1} \phi_{j,\ell} Y_\ell, j = 1, \cdots, p, \tag{5}$$

where the regression coefficients $\phi_{j,\ell}$'s are unconstrained and the variances $\nu_j^2 = \text{ var } (\varepsilon_j)$are non-negative. Evidently, the successive prediction errors are uncorrelated, so that with $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_p)$ we have cov $(\varepsilon) = \text{ diag } (\nu_1^2, \cdots, \nu_p^2) = \nu$ and (5) can be written in matrix form

$$\varepsilon = TY, \tag{6}$$

where T is a unit lower triangular matrix with $-\phi_{j,\ell}$ in the (j, ℓ)th position for $2 \leq j \leq p, \ell = 1, \cdots, j-1$. Consequently, from (6) we obtain $T\Sigma T' = \nu$, i.e. the matrix T diagonalizes the covariance matrix Σ as in (4). It is clear that this decomposition depends on the ordering of the components of Y; thus, it is well suited to data that have ordered responses, such as longitudinal data.

Analogues of (M′3)-(M′5) for the decomposition (4) can be defined with the same number of parameters by imposing a suitable hierarchy on T_i's:

(M″3) *Common* GARP, $T_i \equiv T$;

(M″4) *Common GARP of order* q, where the first q subdiagonals of T_i's are common.

(M″5) *Common GARP of dimension* r, where certain r entries of T_i's are common.

A notable disadvantage of the first two classes of hierarchies is that the number of covariance parameters from one level of hierarchy to the next increases not by one, but by a multiple of $c-1$. Boik's (2002, 2003) spectral models attempt to provide a more "gradual" parameterization of the pair of matrices $(P_i, \Lambda_i), i = 1, \cdots, c$. However, the unconstrained nature of GARPs make them ideal for introducing finer hierarchies whereby the number of parameters increases by one when going from one level to the next as in the following model (also motivated by the Growth Hormone trial discussed in Section 5):

(M″6) *Common GARP of variable dimension* r, where r entries of the T_i's are common across all groups and the other $p(p-1)/2 - r$ entries can either be distinct or common across some subset(s) of the groups.

Other advantages of the GARP hierarchies over the competing correlation and spectral models include both computational aspects and asymptotic theory (both discussed in Section 3.3).

Further reduction of the dimension of the parameter space of $\{\Sigma_i\}_{i=1}^c$ is achieved by imposing restrictions on the "variance" components in (2)-(4). For example, adopting *multiplicative variance models* for these matrices, namely

$$\sigma_{ij} = \alpha_i \sigma_{1j}; \lambda_{ij} = \beta_i \lambda_{1j}; \nu_{ij} = \gamma_i \nu_{1j}, i = 1, \cdots, c, j = 1, \cdots, p, \tag{7}$$

will reduce the number of "variance" parameters from pc to $2p-1$. Furthermore, it is evident that (7) coupled with either common correlation matrices, CPC (Flury, 1988, p. 103), or common GARP amounts to the class of *proportional covariance matrices* (M2). We study more general log-linear models for "variances" and highlight their roles in reducing the dimension of the overall parameter space.

In this paper, in addition to presenting the details of using the GARP to model covariances

across groups, we also develop a new algorithm for the common correlation models (M3) which allows for log-linear variance models.

3 Model Estimation: The Likelihood Procedures

The three decompositions of covariance matrices lead to simpler covariance structures by reducing the high number of parameters. So far as estimation is concerned, perhaps the most steady progress has been made using the spectral decomposition in the context of principal components analysis (Anderson, 2003, Chap. 11) and variants of Flury's (1984, 1988) CPC; see the introduction of Boik (2002) for an excellent review. However, the orthonormality of the eigenvectors makes them awkward to model in terms of covariates, and MLE requires optimization procedures capable of handling orthogonality constraints. In sharp contrast, for the Cholesky decomposition these tasks are relatively easy and, in fact, closed-formula for the MLE of common GARPs can be derived. For the ease of reference and comparison, we start with a brief overview of the MLE for normal-theory covariance matrices and then present the MLE for common correlation matrices, CPC, and common GARPs in the next three subsections.

Throughout the paper we assume that the p-variate random vectors $Y_{i\ell}, i = 1, \cdots, c, \ell = 1, \cdots, n_i$ are independent, with $Y_{i\ell}$ distributed as $N(X_i\alpha, \Sigma_i)$; we assume that $\min_i n_i > p$ and that all Σ_i are strictly positive definite. For convenience, denote by S_i the "sample" covariance matrix for the ith sample: $S_i = S_i(\alpha) = \frac{1}{n_i}\sum_{\ell=1}^{n_i}(Y_{i\ell} - X_i\alpha)(Y_{i\ell} - X_i\alpha)'$, where the data are centered by the unknown mean vector $X_i\alpha$. Then the likelihood function of $\Sigma_1, \cdots, \Sigma_c$, and α is given by

$$L(\Sigma_1, \cdots, \Sigma_c, \alpha) = C \prod_{i=1}^{c} |\Sigma_i|^{-n_i/2} etr(-\frac{n_i}{2}\Sigma_i^{-1} S_i(\alpha)),$$

where C does not depend on the parameters and etr stands for the exponential function of the trace. Thus the log-likelihood is

$$l(\Sigma_1, \cdots, \Sigma_c, \alpha) = \sum_{i=1}^{c} \left[-\frac{n_i}{2} \log |\Sigma_i| - \frac{n_i}{2} tr(\Sigma_i^{-1} S_i(\alpha)) \right], \tag{8}$$

up to an additive constant. Since parsimonious modeling of both the mean vectors and covariance

matrices is becoming increasingly important in longitudinal data analysis (Carroll, 2003) and other areas of application, we include α in our estimation algorithms.

3.1 MLE of Common Correlation Matrices

When the mean and variance parameters are unrestricted, the MLEs for the common correlation model (M3) can be obtained using a simple iterative algorithm developed by Manly and Rayner (1987). Unfortunately, this algorithm cannot be easily generalized to common correlation models with restricted variance parameters. In addition to the multiplicative variance models in (7), commonly-used variance models from various application areas include log-linear models (Barnard et al. 2000), univariate GARCH models (Bollerslev, 1990) and specific variance functions suggested by GLM.

Throughout this section, the variances will be assumed to follow general log-linear models $\log \sigma_{ij}^2 = Z_{ij}\gamma, j = 1, \ldots, p$. For notational convenience, let

$$Z_i = \begin{bmatrix} Z_{i1} \\ \vdots \\ Z_{ip} \end{bmatrix}, \qquad i = 1, \ldots, c$$

(that is, the row vectors $Z_{ij}, j = 1, \ldots, p$, stacked into a matrix), and let

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_c \end{bmatrix}.$$

Bollerslev (1990) pursues an alternative computational approach in the context of multivariate time series models. This approach is applicable if the matrices Z_i can be partitioned as $Z_i = [I_p \, \tilde{Z}_i]$, where I_p is the $p \times p$ identity matrix. (The matrices Z_i can be coerced to this form via a linear reparameterization of γ if $\text{range}(Z) \supset 1_c \otimes I_p$, where 1_c is the c-dimensional ones vector and $\otimes$ denotes the Kronecker product.) Let $\gamma = (\gamma_1', \gamma_2')'$ be the corresponding decomposition of γ, and define the "standardized" residuals $\tilde{\epsilon}_{i\ell}(\alpha, \gamma_2) = \tilde{V}_i(\gamma_2)^{-1/2}(Y_{i\ell} - X_i\alpha)$, where $\tilde{V}_i(\gamma_2) = \exp(\text{diag}(\tilde{Z}_i\gamma_2))$. Then an appropriate counterpart to Bollerslev's equation (7), in which R and γ_1

are 'removed' from the log-likelihood, is the profile log-likelihood (up to an additive constant)

$$l(\alpha, \gamma_2) = -\frac{1}{2}\sum_{i=1}^{c} n_i \log|\tilde{V}_i(\gamma_2)| \; - \; \frac{n}{2}\log\left|\sum_{i=1}^{c}\sum_{\ell=1}^{n_i}\tilde{\epsilon}_{i\ell}(\alpha,\gamma_2)\tilde{\epsilon}_{i\ell}(\alpha,\gamma_2)'\right|, \tag{9}$$

where $n = \sum_{i=1}^{c} n_i$. The maximization of this profile log-likelihood over α and γ_2 can proceed using any suitable unconstrained optimization method to obtain the MLEs $\hat{\alpha}$ and $\hat{\gamma}_2$. Subsequently, the MLEs $\hat{R}$ and $\hat{\gamma}_1$ for R and γ_1 can be obtained from the variance-correlation decomposition

$$\frac{1}{n}\sum_{i=1}^{c}\sum_{\ell=1}^{n_i}\tilde{\epsilon}_{i\ell}(\hat{\alpha},\hat{\gamma}_2)\tilde{\epsilon}_{i\ell}(\hat{\alpha},\hat{\gamma}_2)' = \hat{V}^{1/2}\hat{R}\hat{V}^{1/2},$$

where $\hat{V} = \exp(\mathrm{diag}(\hat{\gamma}_1))$.

The structural assumption on the Z_i matrices above is unduly restrictive. For instance, it would preclude a model that specified the variances of two different components to be equal within a group. Unfortunately, no convenient analytical reduction for removing the correlation parameters is available in the general case. Even in the simple case where $c = 1$ and $p = 2$, finding the MLE of the correlation parameter for fixed values of the means and variances requires the solution of a cubic equation (Kendall and Stuart, 1967, Example 18.3). The common-correlation MLEs presented in Section 5 were obtained by applying a general Newton-based algorithm simultaneously on all parameters (α, γ, and off-diagonal elements of R), with a trust-region restriction (Fletcher, 1987) incorporated to ensure that the matrix R is always positive-definite and the likelihood is monotone increasing.

Parsimonious representation of a single correlation matrix via its spectral decomposition has recently been proposed in Boik (2003), as an adaptation of his earlier models for a covariance matrix (Boik, 2002).

3.2 MLE of PC Models

For a single covariance matrix a complete theory of MLE for its eigenvectors and eigenvalues for the saturated case has been available for a while (Anderson, 2003, Chap. 11). Its analogue for several covariance matrices was developed later by Flury (1986).

3.2.1 MLE of CPC Models

Assume that the hypothesis of common principal components holds, i.e. (M′3) is satisfied with $P_i \equiv P = (\beta_1, \beta_2, \cdots, \beta_p)$ where β_j is the jth column of P. The MLEs of the α, β_j's and λ_{ij}'s are then obtained by maximizing

$$l(\beta_1, \cdots, \beta_p, \lambda_{11}, \cdots, \lambda_{cp}, \alpha) = \sum_{i=1}^{c} \sum_{j=1}^{p} \left[-\frac{n_i}{2} \log \lambda_{ij} - \frac{n_i}{2} \beta_j' S_i(\alpha) \beta_j / \lambda_{ij} \right],$$

subject to the orthonormality constraint on β_j's:

$$\beta_j' \beta_\ell = \delta_{j,\ell}, \quad j \geq \ell = 1, \cdots, p. \tag{10}$$

This can be formulated as an (unconstrained) optimization problem by using Lagrange multipliers. Following the derivation of Flury (1984), but additionally considering estimation of the regression parameters, we obtain the following likelihood equations:

$$\begin{aligned}
&\alpha = \left(\sum_{i=1}^{c} n_i X_i' \Sigma_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{c} n_i X_i' \Sigma_i^{-1} \bar{Y}_i \right), \\
&\lambda_{ij} = \beta_j' S_i(\alpha) \beta_j, && i = 1, \cdots, c, j = 1, \cdots, p, \\
&\beta_\ell' \left(\sum_{i=1}^{c} n_i \frac{\lambda_{i\ell} - \lambda_{ij}}{\lambda_{i\ell} \lambda_{ij}} S_i(\alpha) \right) \beta_j = 0. && \ell, j = 1, \cdots, p, \ell \neq j
\end{aligned} \tag{11}$$

An iterative procedure for solving the last two equations in (11) was developed by Flury and Gautschi (1986). Noniterative estimators of β_j's are given by Krzanowski (1984) as the orthonormalized eigenvectors of the sum of the sample covariance matrices.

Substituting the expression for λ_{ij} in the log-likelihood and dropping irrelevant additive constants yields the profile log-likelihood in $P = (\beta_1, \ldots, \beta_p)$ and α:

$$l(\beta_1, \ldots, \beta_p, \alpha) = -\frac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{p} n_i \log \beta_j' S_i(\alpha) \beta_j. \tag{12}$$

Optimization over P may proceed using any of several specialized algorithms for optimization over orthogonal matrices (see, for example, Edelman, Arias, and Smith (1998)). Simultaneously maximizing over α might require an extension of current methods, e.g., . the Fisher scoring algorithm in Boik (2002).

Extending the common principal component model to allow models for the eigenvalues λ_{ij} is possible, but direct log-linear models of the form $\log \lambda_{ij} = Z_{ij}\gamma$ are unsatisfactory for this purpose because they allow no control over the ordering of the eigenvalues. For instance, although a log-linear model would allow specification that two eigenvalues in a particular group are equal, it would not allow specification that these two were the largest eigenvalues for that group, rather than some other pair. Two alternatives that allow for ordering are

$$\lambda_{ij} = \sum_r u_{ijr} \exp Z_{ir}\gamma \qquad \text{and} \qquad \log \lambda_{ij} = \sum_r u_{ijr} \exp Z_{ir}\gamma, \tag{13}$$

where the matrices $U_i = [u_{ijr}]$, $i = 1, \ldots c$, allow for specification of the order of the eigenvalues. Details can be found in Boik (2002, Sec. 2.2).

3.2.2 Other PC Models

Algorithms to fit models (M′4) and (M′5) are given in Flury (1988). In addition, Boik's (2002) spectral models subsume most of these earlier extensions of CPC and other models. These models attempt to parameterize the matrices $(P_i, \Lambda_i), i = 1, \cdots, c$, with more flexibility including various models for eigenvalue 'sharing' across groups (for example, equality, proportionality, and equal volume) and 'sharing' of spaces with the eigenvectors. A Fisher scoring algorithm is proposed for optimization. Despite the flexibility of such models, the optimization algorithms still tend to be overly complex as compared to the simple algorithms for the GARP models that will be discussed in Section 3.3.

3.3 MLE for GARP Models

For a single covariance matrix the theory of MLE for GARPs and innovation variances is developed in Pourahmadi (1999, 2000). Their analogues and ramifications for several covariance matrices will be developed next.

3.3.1 MLE for Common GARPs

In analogy with the estimation of common correlation and CPC reviewed above, we compute the MLE of common *generalized autoregressive parameters* (GARP) when (4) is satisfied with $T_i \equiv T =$

$(\tilde{T}_1, \tilde{T}_2, \cdots, \tilde{T}_p)$ where $\tilde{T}_j$ is the jth column of T and $\nu_i = diag(\nu_{i1}^2, \cdots, \nu_{ip}^2)$ is a diagonal matrix of *innovation variances* (IV) changing across the c populations. First, we allow the nonredundant entries of T and ν_i's to remain unstructured; then, in Section 3.3.2, we discuss the structured case. For normal populations, the likelihood equations for α and ν_{ij}^2's are similar to those in (11), but the equation for the nonredundant and unconstrained parameters of T denoted by $\Phi = (\phi_{21}, \phi_{31}, \phi_{32}, \cdots, \phi_{p,p-1})'$ is much simpler with a closed-form solution resembling that of a weighted least-squares problem (see (16) below).

From (4), because ν_i is diagonal, it follows that

$$\log|\Sigma_i| = \sum_{j=1}^{p} \log \nu_{ij}^2, i = 1, \cdots, c,$$

and

$$\begin{aligned} tr(\Sigma_i^{-1} S_i) &= tr(T' \nu_i^{-1} T S_i) = tr(\nu_i^{-1} T S_i T') \\ &= \sum_{j=1}^{p} \tilde{T}_j' S_i \tilde{T}_j / \nu_{ij}^2. \end{aligned} \tag{14}$$

Therefore, (8) reduces to

$$\ell(\Sigma_1, \cdots, \Sigma_c, \alpha) = \sum_{i=1}^{c} \sum_{j=1}^{p} \left(-\frac{n_i}{2} \log \nu_{ij}^2 - \frac{n_i}{2} \tilde{T}_j' S_i \tilde{T}_j / \nu_{ij}^2 \right), \tag{15}$$

which can be minimized by computing its partial derivatives with respect to α, ν_{ij}^2 and the nonredundant entries of T. Setting these to zero yield the first equation in (11) for α, and (for details, see Appendix A)

$$\begin{aligned} \hat{\nu}_{ij}^2 &= \tilde{T}_j' S_i \tilde{T}_j, \quad i = 1, \cdots, c, j = 1, \cdots, p, \\ \hat{\Phi} &= \left[\sum_{i=1}^{c} n_i \sum_{\ell=1}^{n_i} \mathbf{Y}_{i\ell}' \nu_i^{-1} \mathbf{Y}_{i\ell} \right]^{-1} \left[\sum_{i=1}^{c} n_i \sum_{\ell=1}^{n_i} \mathbf{Y}_{i\ell}' \nu_i^{-1} y_{i\ell} \right] \end{aligned} \tag{16}$$

where

$$y_{i\ell} = Y_{i\ell} - X_i \alpha = (y_{i\ell 1}, \cdots, y_{i\ell p})'$$

is the vector of regression residuals and the matrix

$$\mathbf{Y}_{i\ell} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ y_{i\ell 1} & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & y_{i\ell 1} & y_{i\ell 2} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & 0 & y_{i\ell 1} & \cdots & y_{i\ell,p-1} \end{pmatrix},$$

is of size $p \times \frac{p(p-1)}{2}$. Furthermore, it follows from (14) and the first equation in (16) that

$$tr(\Sigma_i^{-1} S_i) = p, i = 1, \cdots, c.$$

Using the likelihood equations (16) one can devise an iterative three-step method for computing the MLE of α, IVs ν_{ij}^2 and GARPs ϕ_{ij}. For instance, under the assumption of common GARPs, a vector of initial values for Φ can be obtained by suitably stacking up the nonredundant entries of the matrix T_0 obtained from the modified Cholesky decomposition of $\sum_{i=1}^{c} S_i$. Using this, an initial value for α and the first equation in (16) one obtains an estimate of IV's, and iterates until a convergence criterion is met.

Although the last formula in (16) seems to require inversion of a matrix of order $^{p(p-1)}/_2$, its block diagonal structure can be exploited to save computation time. Specifically,

$$\sum_{i=1}^{c} n_i \sum_{\ell=1}^{n_i} \mathbf{Y}'_{i\ell} \nu_{\ i}^{-1} \mathbf{Y}_{i\ell} = diag(B_2, \cdots, B_p), \tag{17}$$

where

$$B_t = \sum_{i=1}^{c} n_i \sum_{\ell=1}^{n_i} \nu_{it}^{-2} y'_{i\ell(t)} y_{il(t)},$$

and

$$y_{i\ell(t)} = (y_{i\ell 1}, \cdots, y_{i\ell,t-1}).$$

Thus, in computing $\hat{\Phi}$ the largest linear system to be solved is of order $p-1$.

3.3.2 MLE of Structured "Dependence" and "Variance" Parameters

A natural way to reduce the number of covariance parameters is to use covariates and develop models for the "dependence" and "variance" components of the three decompositions. Some early examples based on the variance-correlation decomposition include the theory of multivariate GARCH in finance (Bollerslev, 1990; Engle 2002) and the general location models (GLOM) where the multiplicative variance models (7) were used by Liu and Rubin (1998) and log-linear variance models were proposed by Barnard et al. (2000), but not fitted. Boik's (2002) spectral models for covariance matrices appear to be the first to use regression models for the diagonal entries of the Λ_i's.

Note that modelling the "dependence" components in decompositions (2)-(3) is difficult because of the positive-definiteness and orthonormality constraints on R and P, respectively. In sharp contrast, since the GARPs in (4) are unconstrained, covariates can be used to model the "dependence" component with relative ease (Pourahmadi, 1999, 2000; Pourahmadi and Daniels, 2002). Furthermore, graphical diagnostics for model-formulation based on the regressogram are available. For a single generic covariance matrix one may identify models of the form

$$\phi_{tj} = z'_{tj}\delta,\ \log \nu_t^2 = z'_t\lambda, \tag{18}$$

where z_{tj} and z_t are vectors of covariates and δ and λ are $q_1 \times 1$ and $q_2 \times 1$ vectors of unknown parameters for the "dependence" and "variance" components. Computation of MLE of δ and λ and their asymptotic properties are fully studied in Pourahmadi (2000).

Interestingly, the estimation results presented in Section 3.3.1 for (M$''$3) correspond to the extreme case of the linear models for $\log \nu_{ij}^2$ and ϕ_{tj} in (18) where $q_1 = \frac{p(p-1)}{2}, q_2 = p$ and the covariates are suitable columns of the identity matrices of sizes q_1 and q_2. Thus, the computational and asymptotic results in Pourahmadi (2000) hold verbatim for this special case under mild regularity conditions. In addition, various models suggested by (M$''$4)-(M$''$6) for several covariance matrices can also be written in the above form:

$$\phi_{i,tj} = z'_{i,tj}\delta,\ \log \nu_{i,t}^2 = z'_{i,t}\lambda, i = 1, \cdots, c, \tag{19}$$

with smaller q_1, q_2 and possibly nontrivial covariates representing various group conditions. As such, MLE of the parameters in (19) and their asymptotic properties can be obtained by adapting the techniques from Pourahmadi (2000). As an alternative computational approach to obtain the MLE of the parameters in (19), one could use the iteratively reweighted least squares (IRLS) algorithm in Daniels and Pourahmadi (2002).

3.4 Likelihood Ratio Tests

When the null hypothesis of equality of $\Sigma_1, \cdots, \Sigma_c$ is rejected the group-specific covariance matrices could still have certain common features (components). This possibility can be assessed using the

likelihood ratio test and a hierarchy of flexible covariance models. The likelihood ratio test statistic for comparing two nested models (1 and 2) within the correlation, PC, or GARP hierarchies is

$$X^2 = -2 \log \frac{L(\hat{\Sigma}_1^{(1)}, \cdots, \hat{\Sigma}_c^{(1)}, \hat{\alpha}^{(1)})}{L(\hat{\Sigma}_1^{(2)}, \cdots, \hat{\Sigma}_c^{(2)}, \hat{\alpha}^{(2)})}. \tag{20}$$

In the PC and GARP hierarchies, when the variance parameters are unrestricted, this can be simplified to

$$-2 \log \frac{K \prod_{i=1}^{c} \exp(-pn_i/2) |\hat{\Sigma}_i^{(1)}|^{-n_1/2}}{K \prod_{i=1}^{c} \exp(-pn_i/2) |\hat{\Sigma}_i^{(2)}|^{-n_1/2}} = \sum_{i=1}^{c} n_i \log \frac{|\hat{\Sigma}_i^{(1)}|}{|\hat{\Sigma}_i^{(2)}|}, \tag{21}$$

where the maximum likelihood estimators of the covariance matrices $\Sigma_1, \cdots, \Sigma_c$ of models 1 and 2 above are computed using the methods described in Sections 3.2–3.3.

The null distribution of (20) for testing within the correlation, PC, or GARP hierarchies is asymptotically χ^2 with degrees of freedom equal to the difference between the number of covariance parameters in the two models. The number of parameters in the correlation and PC models is given in Section 2. For the GARP hierarchy, the number of parameters is $(c-1)p(p-1)/2$ for (M″3), $(c-1)(p-q-1)(p-q-2)/2$ for (M″4), and $(c-1)(p(p-1)/2-r)$ for (M″5); the number of parameters in (M″6) is difficult to write in general form. For comparing non-nested models between hierarchies, we can compare the maximized log-likelihoods directly (with appropriate modifications for differing numbers of parameters).

4 Incomplete Data and the EM Algorithm

It is common for multivariate responses to have missing components, especially in longitudinal settings (Diggle et al., 2002). The EM algorithm is often used to fill in the missing data and obtain valid inferences (Little and Rubin, 2002). We will detail some of the specifics for the EM algorithm here for the general case of $Y_{it} \sim N(X_i\alpha, \Sigma_i)$, where $i = 1, \ldots, c$, $t = 1, \ldots, n_i$ and Y_{it} is of dimension p. We also point out that the rate of convergence of the EM algorithm depends on the fraction of missing observations; for details, see Chapter 3 in Schafer (1997).

We will provide details for the case of data that is missing completely at random (MCAR) or missing at random (Little and Rubin, 2002). Both these types of missing data are termed ignorable since the missing data mechanism need not be specified (i.e., it can be "ignored"). However, these results will easily generalize to non-ignorable missingness often addressed explicitly by specifying a selection model (Diggle and Kenward, 1994) or implicitly through pattern mixture models (Little, 1994). For the latter, the procedures for modelling across groups discussed in Section 3 could be used to model the covariance across the missing data patterns.

The EM algorithm is composed of an expectation and a maximization step. The expectation step involves taking the expectation over the distribution of the missing data, conditional on the observed data, of the log likelihood. In our setting, this involves expectations of Y_{it} and $Y_{it}Y_{it}'$. The latter can be written as $E[Y_{it,mis}|Y_{it,obs}]E[Y_{it,mis}|Y_{it,obs}]' + C_{it}$, where $C_{it} = Var[Y_{it,mis}|Y_{it,obs}]$. The maximization step involves maximizing the expected log likelihood over α and the parameters of Σ_i, $i = 1, \ldots, c$. This maximization can proceed by iterating between maximizing over α and Σ_i, using

$$-\frac{1}{2}tr[\sum_{i=1}^{c}\hat{\Sigma}_i^{-1}\sum_{t=1}^{n_i}(Y_{it} - X_i\alpha)(Y_{it} - X_i\alpha)'] \tag{22}$$

for α (which results in the generalized least squares estimate for α) and

$$-\sum_{i=1}^{c}\frac{n_i}{2}log|\Sigma_i| - \frac{1}{2}tr[\sum_{i=1}^{c}\Sigma_i^{-1}C_i^{\star}] \tag{23}$$

for the Σ_i, where $C_i^{\star} = \sum_{t=1}^{n_i}[(Y_{it} - X_i\hat{\alpha})(Y_{it} - X_i\hat{\alpha})' + C_{it}]$. In these, the missing values have been filled in (during the E-step of the algorithm). The maximization over Σ_i will proceed as in Sections 3.1 and 3.2 for the common correlation and CPC models. However, for the GARP (Section 3.3) models, the maximization routines outlined need to be altered. The key idea behind the maximization routine for the GARP models (see Appendix) involves using two representations of the exponential terms in the likelihood: $tr(\Sigma_i^{-1}S_i)$ and $(Y_{it} - X_i\alpha)'\Sigma_i^{-1}(Y_{it} - X_i\alpha)$. Unfortunately, in the M step of the EM algorithm for estimating the components of Σ_i, we can only express the expectation of the log likelihood in the former form.

We provide details on a modified maximization step for the covariance matrix parameters for the GARP models in the next subsection.

4.1 M-step for GARP

First, we rewrite (23) as

$$-\sum_{i=1}^{c}\frac{n_i}{2}\sum_{j=1}^{p}\log\nu_{ij}^2-\frac{1}{2}tr[\sum_{i=1}^{c}\Sigma_i^{-1}C_i^{\star}]. \tag{24}$$

To maximize over the parameters of Σ_i, we iterate between maximizing over the innovation variances, ν_{ij}^2 and the GARP parameters, $\phi_{i,tj}$. For the former, note that $tr(\Sigma_i^{-1}C_i^{\star})=tr(\nu_i^{-1}T_iC_i^{\star}T_i')$. Define $G_i=T_iC_i^{\star}T_i'$, with jk element g_{ijk}. It is easy to show that the gradient of the expected log likelihood with respect to ν_{ij}^2 is proportional to

$$-\frac{1}{(\nu_{ij}^2)^2}g_{ijj}+\frac{n_i}{\nu_{ij}^2}. \tag{25}$$

When ν_{ij}^2 does not depend on covariates, we obtain $\hat{\nu}_{ij}^2=\frac{1}{n_i}g_{ijj}$, otherwise a simple Newton-Raphson algorithm can be implemented employing the 2nd derivatives. When some of the ν_{ij}^2 are shared across some of the groups, (25) can be modified by 'summing' over i.

For the GARP parameters, a closed form solution can be obtained by recognizing that the gradient is a linear function of the GARPs. Denote the jkth element of $C_i^{\star}$ as c_{ikj}. The relevant pieces of the expected log likelihood with respect to the GARP parameters, can be written as

$$\sum_{i=1}^{c}\sum_{t=2}^{p}\frac{1}{\nu_{it}^2}\sum_{j=1}^{t}\sum_{k=1}^{t}\phi_{i,tj}\phi_{i,tk}c_{ikj} \tag{26}$$

where $\phi_{i,tt}=-1$. For illustration, we consider the common GARP model, where $\phi_{i,tj}\equiv\phi_{tj}$, so we only need to compute derivatives with respect to ϕ_{tj}. In addition, the block diagonal structure in (17) implies that we can estimate each set of GARP, i.e., $\phi_{(t)}=(\phi_{t1},\ldots,\phi_{t,t-1})'$, independently. The first derivative with respect to ϕ_{tj} can be expressed as

$$2\sum_{i=1}^{c}\frac{1}{\nu_{it}^2}[\sum_{k=1}^{t}\phi_{tk}c_{ikj}]. \tag{27}$$

After setting this to zero and some algebra, we can show that $\phi_{(t)} = A_t^{-1} C_{t,\star}$, where the jth row of A_t is

$$A_{tj} = (2\sum_{i=1}^{c} \frac{1}{\nu_{it}^2} c_{i1j}, 2\sum_{i=1}^{c} \frac{1}{\nu_{it}^2} c_{i2j}, \ldots, 2\sum_{i=1}^{c} \frac{1}{\nu_{it}^2} c_{i,t-1,j}), \tag{28}$$

and $C_{t,\star}$ is a $(t-1)$ dimensional vector with jth component $2\sum_{i=1}^{c} \frac{1}{\nu_{it}^2} c_{itj}$.

Extension to the more general GARP models is straightforward. Writing the GARP parameters, $\phi_{i,tj}$ as in (19), it is easy to show that $\delta = A^{-1}b$, where $A = \sum_{i=1}^{c} \sum_{t=2}^{p} \frac{1}{\nu_{it}^2} \sum_{j=1}^{t-1} \sum_{k=1}^{t-1}$ $c_{ikj}[Z_{i,tj} Z'_{i,tk} + Z_{i,tk} Z'_{i,tj}]$ and $b = \sum_{i=1}^{c} \sum_{t=2}^{p} \frac{1}{\nu_{it}^2} 2 \sum_{k=1}^{t-1} c_{itk} Z_{i,tk}$. The orthogonality of the common GARP model is lost when a structure is put on the GARP as in Section 3.3.2.

4.2 Likelihood Ratio Test

In the context of incomplete data, the relevant likelihood is the observed data likelihood. Thus, for incomplete data, the ratio (20) of the complete data likelihoods will be replaced by the ratio of the observed data log likelihoods and will not take the simple form given in the last line of (21). As an example, in the setting of monotone missing data, the maximized likelihood will take the form

$$L(\hat{\Sigma}_1, \cdots, \hat{\Sigma}_c, \hat{\alpha}) = K \prod_{i=1}^{c} \prod_{j=1}^{n_i} \exp\{-(1/2)(Y_{p_{ij}} - X_{p_{ij}}\hat{\alpha})' \hat{\Sigma}_i(p_{ij})^{-1} (Y_{p_{ij}} - X_{p_{ij}}\hat{\alpha})\} |\hat{\Sigma}_i(p_{ij})|^{-1/2} \tag{29}$$

where p_{ij} is the number of observed responses, taking values $1, \ldots .p$ for individual j in group i, $\Sigma_i(p_{ij})$ is the upper p_{ij} dimensional block of Σ_i, $Y_{p_{ij}}$ is a vector composed of the first p_{ij} components of Y_{ij} and $X_{p_{ij}}$ is a matrix consisting of the first p_{ij} rows of X_{ij}.

5 Application: Longitudinal Clinical Trials

An important application of simultaneous modelling of covariance matrices across groups is in longitudinal clinical trials. The main inferential question of interest is often whether the longitudinal trajectories (or some function of them) differ across treatments. However, little attention is typically given to the covariance matrix itself (or particular components) differing across treatments. The covariance matrix is typically assumed constant (across treatments), especially if the sample sizes

per treatment are not large, or the entire covariance matrix may be allowed to differ across groups (in larger sample sizes). In the longitudinal setting, a natural compromise would be to allow only particular components to vary across treatments and the hierarchy of GARP models proposed here would be well-suited for this situation. As an illustration of a typical setting where not all components vary across treatments, consider the marginal variance of the response at baseline (the first time point), σ_1^2. Due to randomization, we might expect the marginal variance at baseline to be the same across treatment groups, but after baseline, the variability and dependence within treatments might differ across treatments.

Obviously, by carefully modelling commonality of components of the covariance matrix across groups, we will obtain both more precise and accurate inferences. In addition, in the presence of MAR or non-ignorable missing data and dropouts, incorrectly modelling the covariance matrix can result in biased inferences on the mean (trajectory) parameters as the information matrix for α and Σ will no longer be orthogonal. So, to properly integrate over (and/or impute) the missing values, the covariance structure within treatments needs to be correctly specified (Daniels and Hogan, working paper). The GARP models also would allow a sensible parameterization on which to conduct sensitivity analyses in the presence of informative dropout, particularly in the context of pattern mixture models (Daniels and Hogan, 2000); this will be reported on elsewhere.

Example. *Growth Hormone Longitudinal Clinical trial*

We illustrate the application of our methodology to data from a recent longitudinal clinical trial of growth hormone for maintaining muscle strength in the elderly. Details of the trial can be found in Kiel et al. (1998). Previous analyses of this trial is reported in Daniels and Hogan (2000). One hundred sixty subjects entered the trial and were randomized into one of four treatment groups: placebo (P), growth hormone only (GH), exercise plus placebo (EP), and exercise plus growth hormone (EGH). The placebo and growth hormone treatments were administered daily via injections. Various muscle strength measures were recorded at baseline (0 months), 6 months, and 12 months. For this analysis, we will focus on mean quadriceps strength as the outcome of interest.

The dropout rates in the four treatment groups were 11/41 (P), 13/41 (GH), 9/40 (EP), and 16/38 (EGH).

We will conduct our analysis under an assumption of random dropout (MAR) for illustration of the methods here. We note, however, to conduct a sensitivity analysis under informative dropout, the GARP models would be quite useful, especially in the context of pattern mixture models (Little, 1994).

Let $Y_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})$ denote the vector of longitudinal responses for subject $j = 1, \ldots, n_i$ in treatment group $i = 1, \ldots, 4$. We assume $Y_{ij} \sim N(\mu_i, \Sigma_i)$ and consider the correlation and GARP models of Section 2 for modelling Σ_i across the treatment groups. We do not report on the CPC models from Section 2 as they are somewhat less interpretable in the present longitudinal context.

Groups	T			D
1	1	0	0	622.14
	-0.97	1	0	453.18
	-0.45	-0.65	1	175.73
2	1	0	0	497.97
	-0.90	1	0	150.38
	-0.26	-0.61	1	72.71
3	1	0	0	668.17
	-0.88	1	0	168.09
	-0.21	-0.59	1	65.64
4	1	0	0	560.76
	-0.73	1	0	172.79
	0.01	-0.78	1	120.54

Table 1: GARP and IV parameters for the growth hormone data fitting a distinct Σ for each of the four treatment groups.

Tables 1 and 2 show the GARP-IV and the correlation-variance components, by fitting the multivariate normal model under MAR, separately for each treatment group, using the EM algorithm. Such decompositions can help elucidate local differences in the dependence/variance structure that can be hard to detect when just looking at the estimated variances and covariances. They can

also provide intuition into how the dependence/variance actually differs by choosing parameterizations/decompositions most appropriate for the application. The estimated ϕ_{31} look very different in treatment groups 1 and 4 versus their values in groups 2 and 3. In addition, the innovation variances at time 2 and time 3 in treatment group 1 are much larger than the corresponding variances in the other three treatment groups. We will fit the common GARP model below, but it appears that a specialized model of the class ($M''6$) will be more suitable (see Table 3). The common correlation model would appear to be reasonable with the only correlation appearing to differ significantly across the four groups being the correlation between 0 and 12 months in treatment group 4, which is much lower than the other correlations. In terms of the variances, the marginal variance at 6 and 12 months in treatment group 1 appear much larger than the corresponding variances in the other three treatment groups.

Groups	R			D
1	1	0.75	0.81	622.1
	0.75	1	0.89	1038.6
	0.81	0.89	1	1094.7
2	1	0.85	0.85	498.0
	0.85	1	0.91	556.5
	0.85	0.91	1	456.8
3	1	0.87	0.86	668.2
	0.87	1	0.92	690.3
	0.86	0.92	1	485.3
4	1	0.80	0.66	560.8
	0.80	1	0.84	474.8
	0.66	0.84	1	405.8

Table 2: Correlation and variance estimates for the growth hormone data fitting a distinct Σ for each of the four treatment groups.

Table 4 gives the log likelihoods and number of parameters to conduct likelihood ratio tests within the classes of GARP and correlation models, respectively. A test of common Σ vs. unrestricted $\Sigma_i, i = 1, \ldots, 4$ is rejected, $X^2 = 34.2$ on 18 degrees of freedom. This is not surprising

Groups	T			D
1	1	0	0	σ_1^2
	ϕ_{21}	1	0	$\sigma_{1,2}^2$
	$\phi_{1,31}$	ϕ_{32}	1	$\sigma_{1,3}^2$
2	1	0	0	σ_1^2
	ϕ_{21}	1	0	σ_2^2
	ϕ_{31}	ϕ_{32}	1	σ_3^2
3	1	0	0	σ_1^2
	ϕ_{21}	1	0	σ_2^2
	ϕ_{31}	ϕ_{32}	1	σ_3^2
4	1	0	0	σ_1^2
	$\phi_{4,21}$	1	0	σ_2^2
	$\phi_{4,31}$	ϕ_{32}	1	σ_3^2

Table 3: Specialized GARP/IV model, $(T_i, \nu_i)^\star$

given the treatment specific GARP/IV and Corr/D given in Tables 1 and 2. A LRT of common GARP versus unrestricted Σ_i does not reject the common GARP model ($X^2 = 10.2$ on 9 degrees of freedom). Neither does a LRT of common correlations versus unrestricted Σ_i reject the common correlation model ($X^2 = 8.8$ on 9 degrees of freedom). The log likelihood of the specialized GARP model (Table 3), labelled as $(T_i, \nu_i)^\star$, was almost 4 units larger than that of the common GARP model, even though it has 4 fewer parameters. This appears to be the best fitting model of all considered.

We also examine the estimated means for the treatment groups at month 12 under the different models for Σ_i to examine the importance of correctly modelling Σ_i in the presence of missing data in this example. We point out that under MAR missing data, Σ_i, impacts the estimates of the mean parameters, even under a saturated mean model as was fit here. This can be clearly seen in the E-step of the EM algorithm as described in Section 4. Table 5 shows the month 12 means under several models for Σ. The most obvious differences are between the month 12 means for treatment group 1, which differ by as much as 2.3 units over the four models; in addition, the standard errors

can differ by more than 30%. The amount of difference in means and standard errors ultimately depends on two features of the distributions: 1) how much the Σ_i differ across groups and 2) how much the means before dropout differ among those who complete the study and those who drop out (Daniels and Hogan, working paper).

Model	log likelihood	No. of parameters in Σ_i's
Σ_i	-1664.0	24
$\Sigma_i = \Sigma$	-1681.1	6
$R_i = R$	-1668.4	15
$T_i = T$	-1669.1	15
$(T_i, \nu_i)^\star$	-1665.4	11

Table 4: Log likelihoods for several GARP and correlation models

	treatment group			
Model	1	2	3	4
Σ_i	78.9 (7.05)	65.1 (3.9)	72.7 (4.0)	63.1 (3.8)
$\Sigma_i = \Sigma$	81.2 (5.10)	65.1 (4.4)	72.7 (4.3)	62.7 (4.5)
$R_i = R$	79.3 (7.3)	65.1 (3.7)	72.7 (3.6)	62.7 (4.3)
$(T_i, \nu_i)^\star$	79.2 (6.7)	65.1 (4.0)	72.7 (4.0)	62.9 (3.7)

Table 5: Month 12 means (standard errors).

6 Discussion and Future Work

Because it does not require special constraints, the Cholesky decomposition for parsimonious modelling of several covariance matrices offers a fairly straightforward modelling and estimation procedure relative to the alternative variance-correlation and spectral decompositions. Flexibility of this procedure is demonstrated with a dataset from a growth hormone clinical trial which has a high percentage of missing values. A suitable EM algorithm for modelling the Cholesky factors in the presence of missing values is developed.

A drawback of modelling the Cholesky factors is that this parameterization depends on the ordering of the data. Longitudinal data poses no problem because it has a natural ordering. For unordered data, one strategy is to find the ordering of the data that is most consistent with the models under consideration. To avoid the evaluation of all $p!$ possible orderings, a sequential

approach might be used. For example, consider this algorithm for fitting the common GARP model. Step 1: Fit all simple linear regressions of Y_j on Y_k for $k \neq j$ for each group. Choose the pair whose regression is closest to a single common GARP (i.e.,$\phi_{i,jk} = \phi_{jk}$). Step 2: Conditional on the first two, add in the third variable that provides the closest fit for the common GARP of the regression of this variable on the previous two. ... Step (p-2): Conditional on first $(p-2)$ variables, choose the $(p-1)$st variable that is closest to common GARP for the regression of this variable on the $p-1$ already included. The result will likely be an ordering of the data for which the common GARP model fits well.

There are several other open computational problems in the context of simultaneous modelling of several covariance matrices. For the $(M''6)$ class of models, the model space is quite large and model search techniques are needed to move through the complex space of models when p or c is big. A fully Bayesian analysis using MC^3 approaches might be a good option here; some exploration of the models discussed here in a Bayesian setting can be found in Daniels (2005).

For the common correlation model (M3), finding fast and reliable algorithms remains a challenge. The constrained-step Newton algorithm mentioned in Section 3.1 has fast local convergence, but may take time to reach the vicinity of the maximizer if started from a distant point. Because of its high computational cost per step (compared to methods that do not use second derivatives), the Newton algorithm is particularly slow when used in conjunction with the EM algorithm, which can require many iterations to converge. (Only an approximate maximization is needed in EM, so the M-step can usually be performed with a single Newton step. But even one step can be costly in high dimensions.) Perhaps direct maximization of the observed-data likelihood would be preferable to using EM in this case.

Specialized algorithms for orthogonality-constrained optimization (like those of Edelman, Arias, and Smith (1998)) may provide better options for fitting the common PC models (M′3), (M′4), and (M′5). In fact, if the mean and variance (eigenvalue) parameters are unconstrained, such algorithms can be directly applied to (12) after further profiling over α. Efficient application in the

constrained case awaits generalization of these algorithms to simultaneous optimization over both the orthogonal matrix and unconstrained parameters.

Appendix

Derivation of (16)

To obtain the likelihood equations for the common GARPs, since the first $j-1$st and the jth entries of $\tilde{T}_j$ are zero and 1, respectively, and the rest are unconstrained, direct computation of the partial derivatives of $\ell(\cdot)$ with respect to $\tilde{T}_j$ in (15) could lead to complicated equations. Fortunately, due to the role of ϕ_{tj}'s as regression coefficients, we are able to rewrite $\ell(\cdot)$ as a quadratic form involving only the unconstrained entries of T and consequently reduce their estimation to that of solving a weighted least-squares problem.

To express $tr\Sigma_i^{-1}S_i$ in (8) as a quadratic form involving the nonredundant entries of T, recall that

$$n_i S_i = \sum_{\ell=1}^{n_i} (Y_{i\ell} - X_i\alpha)(Y_{i\ell} - X_i\alpha)',$$

consequently,

$$\begin{aligned} n_i tr\Sigma_i^{-1}S_i &= \sum_{\ell=1}^{n_i} trT'\nu_i^{-1}T(Y_{i\ell} - X_i\alpha)(Y_{i\ell} - X_i\alpha)' \\ &= \sum_{\ell=1}^{n_i} tr\nu_i^{-1}T(Y_{i\ell} - X_i\alpha)[T(Y_{i\ell} - X_i\alpha)]' \\ &= \sum_{\ell=1}^{n_i} (Ty_{i\ell})'\nu_i^{-1}(Ty_{i\ell}), \end{aligned} \tag{A1}$$

where $y_{i\ell} = Y_{i\ell} - X_i\alpha, \ell = 1, \cdots, n_i$. It is known (Pourahmadi, 2000) that the unit lower triangular matrix T transforms any mean-zero random vector with the covariance matrix Σ_i to its vector of successive prediction errors. More specifically,

$$Ty_{i\ell} = y_{i\ell} - \hat{y}_{i\ell}, \tag{A2}$$

where $\hat{y}_{i\ell} = (\hat{y}_{i\ell 1}, \cdots, \hat{y}_{i\ell p})$ and

$$\hat{y}_{i\ell t} = \sum_{j=1}^{t-1} \phi_{tj} y_{i\ell j}, t = 1, \cdots, p, \tag{A3}$$

with the convention that $\sum_{j=1}^{0} = 0$. Substituting from (A2)-(A3) into (A1) leads to

$$\begin{aligned} n_i tr\Sigma_i^{-1} S_i &= \sum_{\ell=1}^{n_i}\sum_{t-1}^{p} \nu_{it}^{-2} \left(y_{i\ell t} - \hat{y}_{i\ell t}\right)^2 \\ &= \sum_{\ell=1}^{n_i}\sum_{t=1}^{p} \nu_{it}^{-2} \left(y_{i\ell t} - \sum_{j=1}^{t-1} \phi_{tj} y_{i\ell j}\right)^2 \\ &= \sum_{\ell=1}^{n_i}\sum_{t=1}^{p} \nu_{it}^{-2} \left(y_{i\ell t} - \phi'_{(t)} y_{i\ell(t)}\right)^2 \\ &= \sum_{\ell=1}^{n_i} Z'_{i\ell} \nu_i^{-1} Z_{i\ell}, \end{aligned} \tag{A4}$$

where

$$\phi_{(t)} = (\phi_{t1}, \cdots, \phi_{t,t-1}), y_{i\ell(t)} = (y_{i\ell 1}, \cdots, y_{i\ell,t-1})$$

and

$$\begin{aligned} Z_{i\ell} &= \left(y_{i\ell 1} - \phi'_{(1)} y_{i\ell(1)}, \cdots, y_{i\ell p} - \phi'_{(p)} y_{i\ell(p)}\right) \\ &= y_{i\ell} - \left(\phi'_{(1)} y_{i\ell(1)}, \cdots, \phi'_{(p)} y_{i\ell(p)}\right) \\ &= y_{i\ell} - \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ y_{i\ell 1} & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & y_{i\ell 1} & y_{i\ell 2} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & 0 & y_{i\ell 1} & \cdots & y_{i\ell,p-1} \end{pmatrix} \begin{bmatrix} \phi_{21} \\ \phi_{31} \\ \phi_{32} \\ \vdots \\ \phi_{p,p-1} \end{bmatrix} \\ &= y_{i\ell} - \mathbf{Y}_{i\ell}\Phi, \end{aligned} \tag{A5}$$

with the obvious definitions for the $p \times \frac{p(p-1)}{2}$ matrix $\mathbf{Y}_{i\ell}$ and the $\frac{p(p-1)}{2}$-dimensional column vector Φ. Thus, from (8), (A1)-(A5) we have

$$\ell(\Sigma_1, \cdots, \Sigma_c, a) = \sum_{i=1}^{c} \left[-\frac{n_i}{2} \log|\nu_i| - \frac{n_i}{2} \sum_{\ell=1}^{n_i} (y_{i\ell} - \mathbf{Y}_{i\ell}\Phi)' \nu_i^{-1} (y_{i\ell} - \mathbf{Y}_{i\ell}\Phi) \right], \tag{A6}$$

and

$$\frac{\partial \ell}{\partial \Phi} = \sum_{i=1}^{c}\sum_{\ell=1}^{n_i} n_i \mathbf{Y}'_{i\ell} \nu_i^{-1} \left\{ y_{i\ell} - \mathbf{Y}_{i\ell}\Phi \right\}$$

gives an estimator for Φ with a familiar formula:

$$\hat{\Phi} = \left(\sum_{i=1}^{c}\sum_{\ell=1}^{n_i} n_i \mathbf{Y}'_{i\ell} \nu_i^{-1} \mathbf{Y}_{i\ell} \right)^{-1} \left(\sum_{i=1}^{c}\sum_{\ell=1}^{n_i} n_i \mathbf{Y}'_{i\ell} \nu_i^{-1} y_{i\ell} \right). \tag{A7}$$

Acknowledgments

The authors would like to thank David MacLean (Memorial Hospital of Rhode Island and Pfizer) for providing the data. The first two authors' research were partially supported by NSF grant DMS-0307055 and NIH grant CA85295 , respectively.

References

Anderson, T.W. (2003). *An Introduction to Multivariate Statistics*, 3rd ed., Wiley, New York.

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803-821.

Barnard, J., McCulloch, R. and Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica*, **10**, 1281–1312.

Boik, R.J. (2002). Spectral models for covariance matrices. *Biometrika*, **89**, 159-182.

Boik, R.J. (2003). Principal component models for correlation matrices. *Biometrika*, **90**, 679-70.

Bollerslev, T., Engle, R. and Woodridge, J.M. (1988). A capital asset pricing model with time varying coefficients. *J. of Political Economy.* **96**, 116-131.

Bollerslev, T. (1990). Modeling the coherence in short run nominal exchange rates: A multivariate generalized arch model, *The Review of Economics and Statistics*, **72**, 498-505.

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

Daniels, M.J. (2005) Bayesian modelling of several covariance matrices and some results on propriety of the posterior for linear regression with correlated and/or heterogeneous errors. Tentatively accepted to *Journal of Multivariate Analysis.*

Daniels, M.J. and Hogan, J.W. (2000) Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*, 56, 1241-1248.

Daniels M, Kass R. (1999) Nonconjugate Bayesian estimation of covariance matrices in hierarchical models. *Journal of the American Statistical Association*, **94**, 1254-1263.

Daniels, M.J. and Pourahmadi, M. (2002). Dynamic models and Bayesian analysis of covariance matrices in longitudinal data. *Biometrika*, **89**, 553-566.

Diggle, P., and Kenward, M. (1994) Informative dropout in longitudinal data analysis. *Applied Statistics*, 43, 49-73.

Diggle, P., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd edition, Oxford, Clarendon Press.

Edelman, A., Arias, T. and Smith, S. (1998) The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20, 303-353.

Engle, R.F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroscedasticity models. *J. of Business and Economic Statistics*, **20**, 339-350.

Fletcher, R. (1987) *Practical Methods of Optimization* (2nd Ed.), Chichester, Wiley.

Flury, B. (1984). Common principal components analysis. *J. Am. Statist. Assoc.*, **79**, 892-898.

Flury, B. (1986). Asymptotic theory for common principal components analysis. *Ann. Statist.*, **14**, 418-430.

Flury, B. (1988). *Common Principal Components and Related Multivariate Models.* New York: John Wiley, New York.

Flury, B. and Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM J. Statist. Comp.* **7**, 169-184.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discrimination analysis and density estimation. *J. of Amer. Statist. Assoc.*, **97**, 611-631.

Kendall, M.G. and Stuart, A. (1967) *The Advanced Theory of Statistics* (Vol. 2, 2nd Ed.), Hafner, New York.

Kiel, D.P., Puhl, J., Rosen, C.J., Berg, K., Murphy, J.B., and MacLean, D.B. (1998). Lack of association between insulin-like growth factor I and body composition, muscle strength, physical performance or self reported mobility among older persons with functional limitations. *Journal of the American Geriatrics Society*, **46**, 822-828.

Krzanowski, W.J. (1984). Principal component analysis in the presence of group structure. *Appl. Statist.*, **33**, 164-168.

Little, R.J.A. (1994) A class of pattern mixture models for normal incomplete data. *Biometrika*, 81, 471-483.

Little, R.J.A. and Rubin, D.B. (2002) Statistical analysis with missing data. John Wiley & Sons, New York.

Liu, C. and Rubin, D.B. (1998). Ellipsoidally symmetric extension of the general location model for mixed categorical and continuous data. *Biometrika*, **85**, 673–688.

Manly, B.F.J. and Rayner, J.C.W. (1987). The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika*, **74**, 841-847.

Murtagh, F., Raftery, A.E. (1984). Fitting straight lines to point patters, *Pattern Recognition*, **17**, 479-483.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J.R. Statist. Soc. A*, **135**, 370-784.

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

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

Pourahmadi M, Daniels M.J. (2002) Dynamic conditionally linear mixed models. *Biometrics*, 58:225-231.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.