Simultaneous Modelling of Covariance Matrices: GLM, Bayesian and Nonparametric Perspectives

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

Abstract

We provide a brief survey of the progress made in modelling covariance matrices from the perspective of generalized linear models (GLM) and the use of link functions (factorizations) that may lead to statistically meaningful and unconstrained reparameterization. We highlight the advantage of the Cholesky decomposition in dealing with the normal likelihood maximization and compare the findings with those obtained using the classical spectral (eigenvalue) and variance-correlation decompositions. These methods, which amount to decomposing covariance matrices into "dependence" and "variance" components, and then modelling them virtually separately using regression techniques, try to emulate the GLM principles to varying degrees, though the latter two are not satisfactory since their "dependence" components are severely constrained. Examples such as Flury's (1984, 1988) common principal components (CPC), Manly and Rayner's (1987) common correlations in multi-group situations, Bollerslev's (1990) constant-correlation, Engle's (2002) dynamic correlation and Vrontos et al.'s (2003) full-factor multivariate GARCH models are used to motivate various computational issues, when the number of parameters grows quadratically in the dimension of the response vector. Once a bona fide GLM framework for modelling covariance matrices is formulated (Chiu et al., 1996; Pourahmadi, 2000), its Bayesian (Daniels and Pourahmadi, 2002), nonparametric (Wu and Pourahmadi, 2003), generalized additive and other extensions can be derived in complete analogy with the respective extension of GLM for the mean vector (Nelder, 1998).

Key Words: Common principal components; Volatility matrices; Longitudinal data; Maximum likelihood estimation; Cholesky decomposition; Spectral decomposition; Variance-Correlation decomposition; GARCH models

1 Introduction

A single $p \times p$ covariance matrix with as many as $\frac{p(p+1)}{2}$ constrained parameters is known to play a central role in virtually all areas of classical multivariate statistics (Anderson, 2003). But their modern counterparts usually deal with several, say c groups, having covariance matrices $\Sigma_1, \cdots, \Sigma_c$ where both c and p are potentially large and often there is not enough data to estimate a separate Σ_i well for each group. Prominent examples include model-based principal component analysis (Flury, 1984, 1988); model-based cluster analysis and discriminant analysis (Murtagh and Raftery, 1984; Banfield and Raftery, 1993) and longitudinal data analysis (Diggle et al., 1994). The truly challenging case of multivariate volatility in finance (Bollerslev, Engle and Woodridge, 1988), for which the number of covariances to be estimated is the same as the number of observations, has bedeviled researchers for over two decades (Engle, 2002). Nowadays, it is well-known (Gourierroux, 1997; Tsay, 2002) that many tasks of financial management including portfolio selection, option pricing and risk management can be reduced to the prediction of a sequence of large $p \times p$ covariance matrices $\{\Sigma_t\}$ using the (conditionally) independent $N(0, \Sigma_t)$-distributed data $Y_t, t = 1, \cdots, n$, where Y_t is the shock (innovation) at time t of the multivariate time series of return of p assets in a portfolio. For example, in pension and mutual funds, a typical value of p could be around 500 while the standard volatility models can handle no more than 5 assets (Engle, 2002; Ledoit et al. 2003). The physical and biomedical sciences also give rise to spectroscopic and microarray data where usually the dimension of the response vector is larger than the sample size, leading to singular sample covariance matrices and hence the collapse of many standard multivariate techniques (Krzanowski et al. 1995).

Estimation of a covariance matrix is of fundamental importance in statistics and the sample covariance matrix which is positive-definite, unbiased and unstable for large p, is the most commonly used estimator. To stabilize this naive estimator, in the true tradition of mathematical statistics most decision-theoretic approaches seeking minimaxity or admissibility have led to estimators which are scalar multiples of the sample covariance matrix and known (Dempster, 1969) to distort the eigenstructure of Σ, unless p/n is small. A vast and growing literature on the problem of efficient estimation of the covariance matrix Σ of a normal distribution has focused on either 'correcting' this distortion in the eigenstructure and/or reducing the number of parameters to be estimated (Stein, 1975; Lin and Perlman, 1985; Yang and Berger, 1994; Daniels and Kass, 1999; Champion, 2003; Wong, Carter and

Kohn, 2003).

On the other hand, in the literature of applied statistics, particularly for repeated measure data, there is a growing tendency to pick a stationary covariance matrix, such as compound symmetry and autoregression of order one with few parameters from a growing menu provided by the popular software packages (Zimmerman and Núñez-Antón, 2001). The question of which covariance structure to pick is a challenge even for the experts. Of course, if the selected covariance structure is far from the truth, then the estimated covariance matrix could have considerable bias. To strike a balance between the variability of the sample covariance matrix and the bias of the estimated structured covariance matrix, it is prudent to rely on the data to formulate structures for the unknown underlying dependence in the data.

In this survey, we review the progress made in parsimonious modelling of covariance matrices that is compatible with the GLM principles, i.e. reducing the dimension of, and the constraint on, the parameter space of covariance matrices. We recall that the Nelder and Wedderburn's (1972) framework of generalized linear models (GLM) has successfully unified a vast collection of apparently disparate approaches developed over a span of two centuries, to model the mean or mean-like parameters (McCullagh and Nelder, 1989). Nowadays, GLM can handle normal, probit, logistic and Poisson regressions, log-linear models for contingency tables, variance component estimation from ANOVA mean squares (Searle et al. 1992), spectral estimation from periodogram ordinates (Cameron and Turner, 1987) and survival data among others. The success of GLM hinges on the concept of link functions to induce unconstrained reparameterization for the mean (first moment) of a distribution, and hence the ability to reduce the dimension of the parameter space via modelling the covariate effect additively by increasing the number of parameters gradually one at a time corresponding to inclusion of each covariate. The successful development of GLM has been the source of inspiration for a large class of models like nonparametric and generalized additive models (Hastie and Tibshirani, 1990), generalized linear mixed models (McCulloch and Searle, 2001), hierarchical generalized linear models (Lee and Nelder, 1996) and other extensions reviewed in Nelder (1998) and Bayesian GLM (Dey et al., 2001). We believe, in analogy with the above developments, once a bona fide GLM framework for modelling covariance matrices is formulated (Chiu et al., 1996, Pourahmadi, 2000), it will definitely facilitate further developments from the Bayesian, nonparametric and other perspectives.

In addition to the sample covariance matrix and estimated structured covariances, most successful approaches to covariance modelling can be divided into three distinct categories which in increasing order of adherence to the GLM principles, employ the variance-correlation, spectral (eigenvalue) and Cholesky decompositions. The variance-correlation decomposition (Styan, 1968; Liang and Zeger, 1986; Manly and Rayner, 1987; Barnard, McCulloch and Meng, 2000) is the simplest and the most direct method whose components are easy to interpret in terms of the original variables. On the other hand, the spectral decomposition (Hotelling, 1933; Flury, 1984, 1988; Boik, 2002; Fraley and Raftery, 2002) is the most familiar method whose components have interpretation as variances and coefficients of principal components which are linear combinations of the original variables. The factor-analytic decomposition is at the heart of the random-effect or mixed models (Laird and Ware, 1982; Searle et al., 1992; McCulloch and Searle, 2001) which are commonly used in handling covariance matrices in the literature of natural, biomedical and social sciences.

In the literature of finance, a totally different approach based on difference equations has emerged for capturing the volatility of a financial asset or portfolio which is of fundamental importance for many of the common tasks of financial management. In the last two decades, the key to successful modelling of volatility has been the empirical fact that the first differences of many asset prices are uncorrelated, but their squares display serial correlation. The latter has been used as proxy for the time-varying variances $\{\sigma_t^2\}$ of the returns $\{y_t\}$ in Engle's (1982) class of autoregressive conditional heteroscedastic (ARCH) and Bollerslev's (1986) generalized ARCH (GARCH) models. These are, indeed, analogues of the familiar

ARMA models for the variances. One of the simplest models of this class, the GARCH (1,1):

$$\sigma_t^2 = \alpha_0 + \alpha y_{t-1}^2 + \beta\ \sigma_{t-1}^2, \tag{1}$$

where $\alpha_0 > 0, \alpha, \beta \geq 0$, has been very effective in capturing the volatility of univariate financial time series (Tsay, 2002). However, there is a great need for their multivariate counterparts (Engle, 2002). Motivated by the success of univariate models, some of the early variants of multivariate GARCH models (Bollerslev, Engle and Woodridge, 1988; Engle and Kroner, 1995) simply amounted to writing multivariate analogues of (1) with matrix coefficients either for the vectorized sequence of covariance matrices $\{vec(\Sigma_t)\}$ or the sequence $\{\Sigma_t\}$ itself, with the number of parameters proportional to p^4 and p^2, respectively. Of course, simplification occurs when the coefficient matrices are taken to be diagonal. In the vectorized case, a diagonal coefficient matrix means each variance/covariance term $\sigma_{ij,t}$ in Σ_t follows a univariate GARCH model with the lagged variance/covariance terms and squares and cross-products of the data:

$$\sigma_{ij,t} = \alpha_{0,ij} + \alpha_{ij} y_{i,t-1} y_{j,t-1} + \beta_{ij} \sigma_{ij,t-1},$$

but complicated restrictions on the coefficient parameters are needed to guarantee the positive-definiteness of the estimated Σ_t's. These restrictions are often too difficult to satisfy in the course of iterative optimization of the likelihood function. To remedy some of these complications, few alternatives such as Bollerslev's (1990) constant-correlation models, Engle's (2002) dynamic correlation models and Alexander's (2001) orthogonal GARCH models have been proposed. We relate these to the three factorizations of a covariance matrix and show that they have the flexibility of univariate GARCH models in dealing with the "variance" component coupled with parsimonious models for the "dependence" component.

Given the current surge of interest in proper handling of dependence in modern statistics, it is desirable to have a statistically meaningful unconstrained reparameterization and a systematic covariate-based procedure for parsimonious modelling of several covariance matrices that could guarantee their positive-definiteness. In this paper, we point out the unique feature of the Cholesky decomposition in solving this problem. In particular, in many application areas where the spectral decomposition and variance-correlation decomposition have proved useful, the Cholesky decomposition may do even better by providing statistically meaningful and unconstrained parameterizations and guaranteeing the positive definiteness of the estimated covariance matrices (Pourahmadi, 1999, 2000). An important determining factor for the possibility of a switch is the trade-off between the computational difficulty and the scientific (practical) relevance of the new parameters. Note that the entries of the correlation and orthogonal matrices are always constrained while those appearing in the unit lower triangular matrix of the Cholesky decomposition referred to as the generalized autoregressive parameters (GARP) are unconstrained. Consequently, computing the maximum likelihood estimates (MLE) of the Cholesky decomposition involves unconstrained optimization compared to the orthogonally-constrained optimization algorithm of Flury and Gautschi (1986) for the spectral decomposition, see also Boik (2002, 2003).

The outline of the paper is as follows. An overview of the earlier methods of modelling a single covariance matrix in the linear mixed-effects and time series literatures, multivariate statistics and finance leading to linear covariance models, logarithmic covariance models and GLM is presented in Section 2. The three decompositions of covariance matrices, the ensuing parameterizations, their roles in introducing parsimonious covariance structures and applications to finance and multivariate GARCH models are introduced in Section 3. Algorithms for computing the normal theory MLE of the parameters under the common GARP, CPC and common correlation matrices are presented in Section 4. The roles of the three decompositions in the Bayesian inference and nonparametric estimation of covariance matrices are discussed in Sections 5 and 6, respectively. Section 7 concludes the paper. This survey emphasizes the need to GLM-type approach for modelling covariances, echoed by the author (Pourahmadi, 2002) at the First International Workshop in Correlated Data Modeling, Trieste, 1999. It is hoped to serve as a blueprint for further research in this important area of statistics.

2 An Overview of GLM for a Single Covariance Matrix

Modeling a covariance matrix is an important and difficult problem of great interest in the statistical science. The sample covariance matrix used in virtually all multivariate techniques (Anderson, 2003) is generally positive-definite and unbiased, but neither parsimonious nor stable due to its high-dimensionality. More specifically, given $Y_1, \cdots, Y_n$, a sample of size n from a p-variate $N(\mu, \Sigma)$, the sample covariance matrix

$$S = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})',$$

serves as an estimator of the $\frac{p(p+1)}{2}$ parameters of the unstructured covariance matrix Σ. Since the number of parameters of $\Sigma = (\sigma_{ij})$ grows quadratically with the dimension and the parameters are to satisfy the notorious positive-definiteness condition

$$c'\Sigma c = \sum_{i,j=1}^{p} c_i c_j \sigma_{ij} \geq 0, \text{ for all } c \epsilon R^p,$$

the problem of parsimonious modeling of a covariance matrix while heeding the above constraint is a truly challenging problem in the current state of statistical science and areas of applications.

While systematic and data-based modeling of covariance matrices is hampered by the positive-definiteness constraint and high-dimensionality, similar though simpler obstacles in modeling the mean vector μ (first moments) of the distribution of a random vector $Y = (y_1, \ldots, y_p)$ has been handled quite successfully in the framework of regression analysis, leading to the powerful theory of generalized linear models (GLM). The success of GLM in handling variety of continuous and discrete data is mainly due to relying on a *link function* $g(\cdot)$ and a *linear predictor* $g(\mu) = X\beta$ to induce unconstrained parameterization and reduce the parameter space dimension simultaneously. Since the covariance matrix of a random vector Y, defined by $\Sigma = E(Y-\mu)(Y-\mu)'$, is a mean-like parameter, one would like to exploit the idea of GLM along with the experience and progress in fitting the mixed-effects and time series models in developing a systematic, data-based procedure for covariance matrices.

We shall focus on the correlated data of continuous type. Table 1 depicts the ideal, rectangular shape of such data where n units (human and/or animal subjects, stocks or financial instruments, firms, households, schools, ...) are measured either repeatedly on one variable (repeated measure, longitudinal data, panel data, ...) or once on several variables.

Table 1: Ideal Shape of Correlated Data

		Occasion					
		1	2	$\cdots$	t	$\cdots$	p
	1	y_{11}	y_{12}	$\cdots$	y_{1t}	$\cdots$	y_{1p}
	2	y_{21}	y_{22}	$\cdots$	y_{2t}	$\cdots$	y_{2p}
Unit	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	i	y_{i1}	y_{i2}	$\cdots$	y_{it}	$\cdots$	y_{ip}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	n	y_{n1}	y_{n2}	$\cdots$	y_{nt}	$\cdots$	y_{np}

Special cases of Table 1, in increasing order of difficulty as indicated by the degrees of dependence of the rows of the table, are the following familiar data structures:

- **Univariate Time Series Data**: $n = 1$, p large;
- **Multivariate Data**: $n > 1, p$ small to moderate and rows are independent; in particular this includes longitudinal data, panel data, and cluster data.
- **Multiple Time Series**: $n > 1, p$ large and rows are dependent;
- **Spatial Data**: n and p are (hopefully) large and rows are dependent.

2.1 Modeling Covariances via Decompositions

The areas of time series analysis (Klein, 1997) and variance components (Searle, Casella and McCulloch, 1992, Chap. 2) are among the oldest in dealing with modeling covariance matrices using covariates implicitly and explicitly, respectively. In a sense, they provide the much needed core methods and ideas. In fact, time series techniques based on spectral and Cholesky decompositions provide suitable tools for handling the awkward positive-definiteness constraint on a stationary covariance matrix (function). However, unlike modeling the mean vector where a link function acts component-wise on the vector μ, *link functions for covariance matrices cannot act componentwise* since positive-definiteness is a simultaneous constraint on all entries of a matrix. Not surprisingly, because of the complicated structure of a general covariance matrix, the most successful modeling approaches need to rely on decomposing a covariance matrix into its "variance" and "dependence" components. The idea of regression and its diagnostic techniques work well for the logarithm of the "variances", but their analogues need to be developed for the more complicated "dependence" component. The three major methods for producing such pairs, i.e. the variance-correlation, spectral (eigenvalue) and the Cholesky decompositions of several covariances are reviewed in Section 2. However, the latter being less familiar is described next for a single covariance matrix.

The Cholesky decomposition is similar to the spectral decomposition in that Σ is diagonalized by a lower triangular matrix T:

$$T\Sigma T' = D, \tag{2}$$

where the nonredundant entries of T are unconstrained and more meaningful statistically than those of the orthogonal matrix of the spectral decomposition. The matrix T is constructed from the regression coefficients when y_t is regressed on its predecessors $y_{t-1}, \ldots, y_1$.

More precisely, with

$$y_t = \sum_{j=1}^{t-1} \phi_{tj} y_j + \varepsilon_t, \tag{3}$$

the $(t, j)^{\text{th}}$ entry of T is $-\phi_{tj}$, the negatives of the regression coefficients and the $(t, t)^{\text{th}}$ entry of D is $\sigma_t^2 = \text{var}(\varepsilon_t)$, the innovation variance. A schematic view of the components of a covariance matrix obtained through successive regressions (Gram-Schmidt orthogonalization procedure) is given in Table 2. Since ϕ_{ij}'s are regression coefficients, it is evident that for any unstructured covariance matrix these and $\log \sigma_t^2$ are unconstrained, in the sequel they are referred to as the **generalized autoregressive parameters** (GARP) and **innovation variances** (IV) of Y or Σ (Pourahmadi, 1999, 2000). Interestingly, this regression approach reveals the equivalence of modeling a covariance matrix to that of dealing with a sequence of $p-1$ varying-coefficient and varying-order regression models. Consequently, one can bring the entire regression machinery to the service of the unintuitive task of modeling covariance matrices. Stated differently, the framework above is similar to that of using increasing order autoregressive models in approximating the covariance matrix or the spectrum of a stationary time series.

Table 2: Regression coefficients and residual variances of successive regressions.

y_1	y_2	y_3	$\cdots$	y_{p-1}	y_p
1					
ϕ_{21}	1				
ϕ_{31}	ϕ_{32}	1			
$\vdots$	$\vdots$		$\ddots$		
$\vdots$	$\vdots$			$\ddots$	
ϕ_{p1}	ϕ_{p2}	$\cdots$	$\cdots$	$\phi_{p,p-1}$	1
σ_1^2	σ_2^2	$\cdots$	$\cdots$	σ_{p-1}^2	σ_p^2

The sequence of regressions in (3) written in the matrix form

$$\begin{bmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{32} & -\phi_{31} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{p1} & -\phi_{p2} & \cdots & \phi_{p,p-1} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}, \tag{4}$$

immediately leads to the modified Cholesky decomposition (2). It also can be used to clarify the close relation between the decomposition (2) and the time series ARMA models in that the latter is means to diagonalize a Toeplitz covariance matrix, for details see Pourahmadi (2001, Sec. 4.2.5).

Some notable examples of the implicit use of the Cholesky decomposition in the literature of statistics include Szegö's orthogonal polynomials on the unit circle which is closely related to fundamental problems in time series analysis and prediction theory of stationary process and signal processing such as the Durbin-Levinson algorithm and fast inversion of Toeplitz matrices (Szegö and Grenander, 1958; Pourahmadi, 2001, Chap. 7); Bartelett's (1933) decomposition of a sample covariance matrix; Wright's (1934) path analysis; Roy's (1958) step-down procedures and Wold's (1960) causal chain models which assume the existence of an *a priori* order among the p variables of interest, and relies on the regression coefficients and residual variances of one variable regressed on its predecessors. Other more explicit examples are the introduction of Kalman filtering of state-space models in the 1960's (Jones, 1993) and the Gaussian graphical models (Wermuth, 1980; Whittaker, 1990; Roverato, 2000). The use of Cholesky decomposition in multivariate quality control to find important variables (Mason and Young, 2002), the Mahalanobis-Taguchi-Gram-Schmidt (MTGS) method to build a multidimensional scaling system to measure the level of unusualness (abnormality) of vector of observations (Taguchi and Jugulum, 2002) are among the most recent and promising applications. Hawkins (2003) provides an excellent discussion revealing the close link between the MTGS method and the more familiar diagnostic techniques of multivariate statistics.

2.2 A History of GLM for a Single Covariance Matrix

The origin of GLM for covariance matrices can be traced to the work of Edgeworth (1892), Yule (1927) and Gabriel (1962), who implicitly parameterized a multivariate normal distribution in terms of entries of the concentration matrix Σ^{-1}. This was made more explicit by Dempster (1972) who recognized the entries of $\Sigma^{-1} = (\sigma^{ij})$ as the canonical parameters of the exponential family of normal distributions with mean zero and unknown covariance matrix Σ:

$$\log f(y; \Sigma^{-1}) = tr\Sigma^{-1}(-yy'/2) + \log|\Sigma|^{-1/2} - p\log\sqrt{2\pi}. \tag{5}$$

These and subsequent developments are reviewed next in direct analogy with those in GLM (McCullagh and Nelder, 1989, Chap. 2). Therefore, particular attention is paid to the linear covariance models and their evolution to log-linear covariance models.

Linear Covariance Models:

Motivated by the simple structure of many time series and variance component models, Anderson (1973), defined the class of *linear covariance models* (LCM) by

$$\Sigma = \alpha_1 U_1 + \cdots + \alpha_q U_q, \tag{6}$$

where U_i's are some known symmetric matrices and α_i's are unknown parameters restricted so that the matrix is positive-definite. The model (6) is general enough to include all linear mixed-effect, certain time series and graphical models. Indeed, for q large any covariance matrix admits representation (6). It suffices to note that every covariance matrix (and hence its inverse) can be written as

$$\Sigma = (\sigma_{ij}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij}U_{ij}, \tag{7}$$

where U_{ij} is a $p \times p$ matrix with one on the (i,j)th position and zero elsewhere. A major drawback of (6)–(7) is the constraint on the coefficients which makes the estimation and other statistical problems difficult (Anderson, 1973; Szafrowski, 1980).

A common method of reducing the high-dimensionality of a covariance structure is to set to zero certain entries of Σ or its inverse. From (6)–(7), it can be seen that this amounts to removing the corresponding matrix U_{ij} from the list of covariates in the LCM. Of course, an extreme instance of this occurs when Σ is diagonal corresponding to independent and heterogeneous observations. Other intermediate cases correspond to the familiar and important classes of stationary *moving average* (MA) and *autoregressive* (AR) models introduced in the literature of time series analysis by Slutsky around 1926 and Yule (1927), see also Table 3. More specifically, for the MA(q) and AR(p) models we have,

$$\sigma_{ij} = 0, \quad \text{for } |i-j| > q, \tag{8}$$

and

$$\sigma^{ij} = 0, \quad \text{for } |i-j| > p, \tag{9}$$

namely, either Σ or Σ^{-1} is a *banded* matrix. Nonstationary analogues of such models with more irregular patterns of zeros in Σ^{-1} than (9) introduced by Gabriel (1962); Dempster (1972); Macchiavelli and Arnold (1994), are related to the Gaussian graphical models (Whittaker, 1990). We review some interesting special cases in more detail next.

Antedependence Models:

A set of random variables $y_1, y_2, \ldots, y_p$ (indexed by time) whose joint distribution is multivariate normal is said to be *pth-order antedependent* (Gabriel, 1962), AD(p) for short, if y_t and y_{t+s+1} are independent given the intervening values $y_{t+1}, \ldots, y_{t+s}$ for $t = 1, \ldots, p-s-1$ and all $s \geq p$. It can be shown that $Y = (y_1, \ldots, y_p)$ is AD(p), if and only if, its covariance matrix satisfies (9). The equivalent classes of variable-order AD (Macchiavelli and Arnold, 1994) and varying-order, varying-coefficient autoregressive models (Kitagawa and Gersch, 1985; Kitagawa and Gersch, 1996) in which the coefficients and order of antedependence depend on time are clearly related to AD models. The latter two references show that these models are much easier to define in terms of their dynamics than properties of their inverse covariance matrices.

Table 3: A Chronology of Linear Covariance Model (LCM)

	$\Sigma = (\sigma_{ij})$	$\Sigma^{-1} = (\sigma^{ij})$
Edgeworth (1892)		Parameterized $N(0, \Sigma)$ in terms of entries of the **concentration** matrix.
Slutsky (1926)	**Banded: Stationary** MA(q)	
Yule (1927)		**Banded: Stationary** AR(p); $y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t$.
Gabriel (1962)		**Banded: Nonstationary** AR(p), **Ante-dependence** (AD) structure. $y_t = \phi_{t1} y_{t-1} + \cdots + \phi_{tp} y_{t-p} + \varepsilon_t$,
Dempster (1972)		**Sparse:** Certain $\sigma^{ij} = 0$. Σ^{-1} , the **canonical param.** of MVN. **Graphical Models.** **Matrix completion** problem.
Anderson (1973)	**Linear**	**Linear**

Covariance Selection Models:

Dempster (1972) suggested to model the covariance structure parsimoniously by identifying zeros in its inverse. These so-called *covariance selection* models have interpretation in terms of conditional independence of variables as in Gabriel's (1962) AD(p) models. But, unlike AD(p) models, setting to zero entries of a covariance matrix in an irregular manner may destroy its positive-definiteness. This problem has given rise to the emerging area of *matrix completion* in the literature of linear algebra.

Log-Linear Covariance Models:

The constraint on α_i's in (6) were removed by Leonard and Hsu (1992) and Chiu et al. (1996) who introduced the *log-linear covariance* models upon noting that for a general covariance matrix with the spectral decomposition $\Sigma = P\Lambda P'$, see (15), its *matricial logarithm* denoted by $\log\Sigma$ and defined by $\log\Sigma = P\log\Lambda P'$ is simply a symmetric matrix with unconstrained entries taking values in $(-\infty,\infty)$. Thus, one may write a log-linear model for Σ as

$$\log\Sigma = \alpha_1 U_1 + \cdots + \alpha_q U_q, \tag{10}$$

where U_i's are as before and the α_i's are now unconstrained . However, since $\log\Sigma$ is a highly nonlinear operation on Σ, the α_i's lack statistical interpretation (Brown et al. 1994). Fortunately, for Σ diagonal since $\log\Sigma = \mathrm{diag}(\log\sigma_{11},\ldots,\log\sigma_{pp})$, (10) amounts to log-linear modeling of heterogeneous variances which has a long history in econometrics and other areas, see Carroll and Ruppert (1988), Verbyla (1993) and references therein.

GLM for Covariances:

The constraint and lack of interpretation of α_i's in (6) and (10) are the source of potential problems in the inference. However, both of these can be resolved simultaneously by relying on the decomposition (2), and modeling the nonredundant entries of T linearly as in (6) and those of $\log D$ as in (10). Note that for an unstructured covariance matrix Σ, the nonredundant entries of its components $(T, \log D)$ in (2) or equivalently the entries of

$$g(\Sigma) = 2I - T - T' + \log D, \tag{11}$$

are unconstrained and hence can be modeled parametrically, semi-parametrically, nonparametrically and Bayesianlly using covariates. Thus, the $g(\cdot)$ above is, indeed, a link function and following the GLM's tradition we may begin by formulating parametric models for ϕ_{tj} and $\log\sigma_t^2$ for $t = 1,\ldots,p$; $j = 1,\ldots,t-1$ by

$$\log\sigma_t^2 = z_t'\lambda,\ \phi_{tj} = z_{tj}'\gamma, \tag{12}$$

where z_t, z_{tj} are $q\times 1$ and $d\times 1$ vectors of known covariates, $\lambda = (\lambda_1,\ldots,\lambda_q)'$ and $\gamma = (\gamma_1,\ldots,\gamma_d)'$ are parameters related to the innovation variances and dependence in Y, respectively. A truly remarkable feature of (12) is its flexibility in reducing the potentially high-dimensional and constrained parameters of Σ to $q + d$ unconstrained parameters λ and γ. Furthermore, one can rely on regressogram, a nonstationary analogue of the time series correlogram or AIC, to identify models such as (12) for the data; for more details, see Pourahmadi (1999, 2001) and Pan and MacKenzie (2003).

In analogy with the theory of GLM (McCullagh and Nelder, 1989, p. 30), the developments outlined above can be summarized in terms of their link functions. Note that the *identity link* functions leads to the linear covariance models (Anderson, 1973) with constraint coefficients as in the early history of classical linear models. The *inverse link* used by Edgeworth (1892); Gabriel (1962); Anderson (1973) and Dempster (1972) has a similar drawback. The Table 3 provides a historical outline of the developments with time series models, mixed models and covariance selection models paving the way for introducing parsimonious covariance structures without heeding the positive-definiteness requirement).

The *log link* (10) as used in Leonard and Hsu (1992); Pinheiro and Bates (1996); Chiu, Leonard and Tsui (1996) leads to unconstrained parameterization of a covariance matrix, but the new parameters are not statistically meaningful. However, the *hybrid link* (11), constructed from the modified Cholesky decomposition of Σ^{-1} (Pourahmadi, 1999; 2000) is somewhere between the identity and log links and combines the ideas of Edgeworth (1892); Gabriel (1962); Anderson (1973); Dempster (1972); Leonard and Hsu (1992); Macchiavelli and Arnold (1994); Chiu et al. (1996) and Zimmerman and Núñez-Antón (1997). It leads to unconstrained and statistically meaningful reparameterization of the covariance matrix so that the ensuing GLM overcomes most of the shortcomings of the linear and log-linear models.

3 Decompositions and Parameterizations

The most methodic and successful approaches to covariance modelling decompose a complicated and high-dimensional matrix into its "variance" and "dependence" components. Three common decompositions of covariance matrices, their adherence to the GLM principles and roles in providing a hierarchy of covariance models with the number of parameters ranging in $1, 2, \cdots, \frac{cp(p+1)}{2}$ are described next. The first two, namely variance-correlation and spectral decompositions have constrained "dependence" components and are not flexible in using covariates.

3.1 The Variance-Correlation Decomposition

Perhaps, the most familiar and simplest decomposition of the covariance matrices $\{\Sigma_i\}$ is

$$\Sigma_i = D_i R_i D_i, \tag{13}$$

where $D_i = \text{diag}\left(\sqrt{\sigma_{i11}}, \cdots, \sqrt{\sigma_{ipp}}\right)$ stands for a diagonal matrix whose diagonal entries are the square-roots of those of Σ_i, and R_i is the corresponding correlation matrix. It has a strong practical appeal since the standard deviations are in the original scale of the responses, and one can manage to separate out the estimation of $\{D_i\}$ and $\{R_i\}$ as either sequence might be fixed or one might be more important than the others in some applications, see Styan (1968); Liang and Zeger (1986); Manly and Rayner (1987); Bollerslev (1990); Barnard et al. (2000), Engle's (2002) dynamic correlation model is actually motivated by the fact that variances (volatilities) of individual assets are more important than their time-varying correlations. Note that while the diagonal entries of D_i are nonnegative, their logarithms are unconstrained, but the correlation matrix R_i is positive-definite with severely constrained entries.

One could reduce the number of parameters in $\{\Sigma_i\}$ by imposing various relationships or hierarchies among components of the decomposition in (13). The first hierarchy among Σ_i's introduced by Manly and Rayner (1987) has four coarse levels:

(M1) *equality*, $\Sigma_1 = \cdots = \Sigma_c$ with $d_1 = p(p+1)/2$ parameters;

(M2) *proportionality*, $\Sigma_i = \rho_i \Sigma_1, i = 2, \cdots, c$ with $d_1 + c - 1$ 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 4, (M2)-(M3) are of particular interest in the recent literature of finance. In fact, starting with the variance - correlation decomposition (13), Bollerslev's (1990) *constant-correlation models* assume that the correlation matrices $\{R_t\}$ are constant, that is $R_t \equiv R$ with $p(p-1)/2$ parameters. The MLE of R turns out to be the sample correlation matrix of the suitably standardized vector of returns. Recognizing that constancy of correlations over time is not often satisfied, Engle (2002) and Tse and Tsui (2002) have recently introduced a dynamic model for $\{R_t\}$ with scalar coefficients. In the first-order case it takes the form

$$R_t = (1-\alpha-\beta)\bar{R} + \alpha R_{t-1} + \beta\Psi_{t-1}, t = 1, \dots, n, \tag{14}$$

where $\bar{R}$ is the sample correlation matrix of $Y_1, \cdots, Y_n$ and Ψ_{t-1} is a positive-definite correlation matrix whose elements are functions of the lagged observations, the two parameters α, β are nonnegative with $\alpha + \beta \leq 1$, so that R_t as weighted average of positive-definite matrices with nonnegative coefficients is guaranteed to be positive-definite. Though such models are highly parsimonious relative to the full multivariate GARCH models, they are neither realistic nor flexible, in that all pairwise correlations satisfy the same simple dynamic. Nevertheless, simplicity of such models has made it possible to analyze a 100-dimensional vector of stock returns (Engle and Sheppard 2001). Furthermore, the fact that the variances in (13) are in the original scale of the returns make the results easier to interpret compared to the next two methods.

3.2 The Spectral Decomposition

This widely used decomposition has been central to the development of principal component analysis (Hotelling, 1933), factor analysis and variety of other applied techniques in multivariate statistics (Anderson, 2003). The spectral decomposition of the covariance matrices

$\{\Sigma_i\}$ is given by

$$\Sigma_i = P_i \Lambda_i P_i', i = 1, \cdots, c, \tag{15}$$

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. Flury's (1984, 1988, Chap. 7) hierarchy replaces (M3) above 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.

The MLE of the parameters under (M$'$3) is reviewed in Section 4, it is of interest in the current literature of finance. Since correlations among the asset returns are the main reason for complexity of multivariate GARCH models (Alexander, 2001), they can be removed using the spectral decompositions (15) of their volatility matrices. Indeed, the orthogonal matrices P_t is known to transform the vector of returns Y_t to their uncorrelated principal components, standard univariate GARCH(1,1) model is developed for each principal component separately and then transformed back using the matrix P_t to the volatility of the original vector of returns. For common principal components, we have $P_t \equiv P$ which is the analogue of Bollerslev's (1990) constant-correlation GARCH models, the methodology for this new multivariate GARCH model along with empirical studies is developed in Alexander (2001) using the orthogonal matrix of eigenvectors of the sample covariance matrix of the whole data. However, the more flexible cases of (M$'$4) and (M$'$5), which allow more time-variation in the "dependence" component, has not been pursued in the finance literature as yet. For arbitrary time-varying $\{\Sigma_t\}$, the problem is even more challenging because orthogonality of P_t's makes it difficult to write an analogue of (14). A possible way around this problem is to reparameterize the $p \times p$ orthogonal matrix by its $\frac{p(p-1)}{2}$ Givens angles (Daniels and Kass, 1999) $\theta_t = (\theta_{21t}, \theta_{31,t}, \cdots, \theta_{p,p-1,t})$ and then write an analogue of (14) or a first-order difference equation for $\{\theta_t\}$. It should be noted that the idea of using spectral decomposition of conditional covariances to simplify multivariate GARCH models has a longer history going back, at least, to the work of Engle and co-authors in 1987 (see Gouriéroux, 1997, pp. 109-111 and p. 114).

3.3 The Cholesky Decomposition

Though the form of this decomposition is similar to the spectral decomposition, it is much more flexible because the nonredundant entries of the "dependence" component are unconstrained, see Sec. 2.1. Following (2), the Cholesky decompositions of several Σ_i's are given by

$$T_i \Sigma_i T_i' = \boldsymbol{\nu}_i, \tag{16}$$

where 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 $\boldsymbol{\nu}_i = \text{ diag } (\nu_{i1}^2, \cdots, \nu_{ip}^2)$ are the corresponding *innovation (residual) variances*.

Analogues of (M′3)-(M′5) for the decomposition (16) 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 common 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.

Next, we show that the framework of factor GARCH models (Diebold and Nerlove, 1989) employing latent processes is related to the Cholesky decomposition. Recall that for a positive integer k, a k-factor model for the returns is usually written as

$$Y_t = Lf_t + e_t, \tag{17}$$

where $f_t = (f_{1t}, \cdots, f_{kt})$ is a vector of common factors with mean zero and diagonal $k \times k$ covariance matrix $D_t =$ diag $(\sigma_{1t}^2, \cdots, \sigma_{kt}^2), L$ is a $p \times k$ matrix of factor loadings and e_t is a vector of specific (idiosyncratic) errors independent of $\{f_t\}$, with mean zero and a diagonal covariance matrix W_t. Univariate GARCH (1,1) models are used for the time-varying common factors and specific variances to reduce their high number of parameters.

Since for any $k \times k$ orthogonal matrix P we have $Lf_t = LPP'f_t$, the matrix of factor loadings L and the common factors f_t are not unique, but can be identified up to a rotation matrix P. The nonuniqueness of the pair (L, f_t) is the source of some controversies and opportunities so far as interpretation of $\{f_t\}$ is concerned. However, lately in the literature of finance (Geweke and Zhou, 1996; Aguilar and West, 2000) a unique k-factor model has been introduced by restricting L to be full-rank k with a "hierarchical" factor structure

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ l_{k1} & l_{k2} & \cdots & \cdots & 1 \\ \ell_{k+1,1} & \ell_{k+1,2} & \cdots & \cdots & \ell_{k+1,k} \\ \vdots & \vdots & & & \vdots \\ \ell_{p,1} & \ell_{p,2} & \cdots & \cdots & \ell_{p,k} \end{pmatrix}. \tag{18}$$

It is evident from (17) - (18) that such choice of L corresponds to an *a priori* ordering of the components of Y_t in the sense that the first series $\{y_{1t}\}$ is essentially, the first latent process $\{f_{1t}\}$ save an additive noise, the second series $\{y_{2t}\}$ is a linear combination of the first two latent process plus a noise and so on. This approach suggests ordering the components of

Y_t according to some rules, here we order them based on their sample variances from the smallest to the largest. Instances of using this mode of "ordering" is implicit in several areas including: (i) in the capital asset pricing models (CAPM) where for $p = 2$ stocks, the more volatile stock return is regressed on a less volatile market index to introduce the *beta* or the market risk; (ii) in multivariate statistics, the PCs are defined according to variances of linear combinations of the original variables; (iii) in introducing priors about covariance matrices in multivariate regression (Zellner, 1979). The choice of $k = p$ in (17)-(18), known as the *full factor models*, leads to the Cholesky decomposition (16) with $T_t = L_t^{-1}$. Vrontos, Dellaportas and Politis (2003) have advocated its use in place of the traditional multivariate GARCH models and provide details of ML and Bayesian estimation of the model parameters and averaging out the impact of "order" among the variables along with an empirical study. A related example of the use of Cholesky decomposition in finance is given in Smith and Kohn (2002).

4 The Normal Likelihood Procedures

The maximization of the normal likelihood under the three decompositions of covariance matrices, namely variance-correlation, spectral and modified Cholesky are reviewed in this section. Perhaps the most steady research on this topic is conducted in the context of principal components analysis (Anderson, 2003, Chap. 11) and variants of Flury's (1984, 1988) CPC. The introduction of Boik (2002) provides an excellent review of this research. However, the orthonormality of the eigenvectors puts a great deal of strain on further reducing the number of parameters by using covariates and the estimation as one has to deal with orthogonally-constrained optimization. In sharp contrast, when using the Cholesky decomposition these tasks are relatively easy and, in fact, closed-formula for the MLE of common GARPs can be derived. For the sake of reference, we present a review of MLE in general and then present MLE for CPC, common correlation matrices and common GARPs. A complete analysis of a dataset from a clinical trial using the above estimation procedure is reported in Pourahmadi et al. (2004).

Throughout 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. Flury (1984), for example, did not consider modeling of the mean vectors, and instead used a marginal likelihood for $\Sigma_1, \cdots, \Sigma_c$, based on the marginal distribution of the usual sample covariance matrices (under the assumption that $E[Y_{ij}] = \mu_i$), which differs from the full likelihood only in that the power of $|\Sigma_i|$ is $-(n_i - 1)/2$. Thus, in our setup the log-likelihood is

$$l(\Sigma_1, \cdots, \Sigma_c, \alpha) = \sum_{i=1}^{c} \left[-\frac{n_i}{2}\log|\Sigma_i| - \frac{n_i}{2}tr(\Sigma_i^{-1}S_i(\alpha))\right], \tag{19}$$

up to an additive constant that can be neglected. Since proper and parsimonious modeling of both the mean vectors and covariance matrices is becoming increasingly important in longitudinal data analysis (Cannon et al. 2001; Carroll, 2003) and other areas of application, we include α in our estimation algorithms. For a single covariance matrix a complete theory of MLE for its eigenvectors and eigenvalues has been available for a while (Anderson, 2003, Chap. 11). Its analogue for several covariance matrices reviewed below was developed later by Flury (1986).

Assuming 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:

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

This can be formulated as an (unconstrained) optimization problem by using Lagrange multipliers. Following the derivation of Flury (1984) and adding in a third equation for the regression parameters, we obtain the likelihood equations:

$$\begin{aligned} \alpha &= \left(\sum_{i=1}^{c} n_i X_i' \Sigma_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{c} n_i X_i' \Sigma_i^{-1} \bar{Y}_i \right), & & \\ \lambda_{ij} &= \beta_j' S_i(\alpha) \beta_j, & i = 1, \cdots, c, j = 1, \cdots, p, & \qquad (20) \\ \beta_\ell' &\left(\sum_{i=1}^{c} n_i \frac{\lambda_{i\ell} - \lambda_{ij}}{\lambda_{i\ell} \lambda_{ij}} S_i(\alpha) \right) \beta_j = 0. & \ell, j = 1, \cdots, p, \ell \neq j & \end{aligned}$$

An iterative procedure for solving the last two equations in (20) 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 and α:

$$l(\beta_1, \ldots, \beta_p, \alpha) = -\frac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{p} n_i \log \beta_j' S_i(\alpha) \beta_j. \qquad (21)$$

Recently, special Newton-style algorithms have been developed for optimization problems involving orthogonal matrices (notably, those introduced by Edelman, Arias, and Smith (1998)). An alternative approach to computing the MLE of β_j would be to maximize (21) as a function of $P = (\beta_1, \ldots, \beta_p)$ using such an algorithm. Simultaneously maximizing over α, however, would require an extension of current methods.

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). For constant-correlation GARCH models Bollerslev (1990) pursues an alternative computational approach in the context of multivariate time series models. Parsimonious representation of a single correlation matrix via its spectral decomposition has recently been proposed in Boik (2003), which is an adaptation of his earlier models for a covariance matrix (Boik, 2002).

For a single covariance matrix the theory of MLE for GARPs and innovation variances are developed in Pourahmadi (1999, 2000). Their analogues and ramifications for several covariance matrices are presented next. In analogy with the estimation of CPC and common correlation, we compute the MLE of common *generalized autoregressive parameters* (GARP) when (16) 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 $\boldsymbol{\nu}_i = diag(\nu_{i1}^2, \cdots, \nu_{ip}^2)$ is a diagonal matrix of *innovation variances* (IV) changing across the c populations. First, we allow the nonredundant entries of T and $\boldsymbol{\nu}_i$'s to remain unstructured. For normal populations, the likelihood equations for α and ν_{ij}^2's are similar to those in (20), 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.

¿From (16), and $\boldsymbol{\nu}_i$ being diagonal, we obtain

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

and

$$\begin{aligned} tr\Sigma_i^{-1} S_i &= tr(T'\boldsymbol{\nu}_i^{-1} T S_i) = tr(\boldsymbol{\nu}_i^{-1} T S_i T') \\ &= \sum_{j=1}^{p} \tilde{T}_j' S_i \tilde{T}_j / \nu_{ij}^2. \end{aligned} \tag{22}$$

Therefore,

$$\ell(\Sigma_1, \cdots, \Sigma_c, \alpha) = \sum_{i=1}^{c} \sum_{j=1}^{p} \left(-\frac{n_i}{2} \log \nu_{ij}^2 - \frac{n_i}{2} \tilde{T}_j' S_i \tilde{T}_j / \nu_{ij}^2 \right), \tag{23}$$

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 (20) for α, and

$$\hat{\nu}_{ij}^2 = \tilde{T}_j' S_i \tilde{T}_j, \quad i = 1, \cdots, c, j = 1, \cdots, p,$$

$$\hat{\Phi} = \left[\sum_{i=1}^{c} n_i \sum_{\ell=1}^{n_i} \mathbf{Y}_{i\ell}' \boldsymbol{\nu}_i^{-1} \mathbf{Y}_{i\ell} \right]^{-1} \left[\sum_{i=1}^{c} n_i \sum_{\ell=1}^{n_i} \mathbf{Y}_{i\ell}' \boldsymbol{\nu}_i^{-1} y_{i\ell} \right] \tag{24}$$

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

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

is of size $p \times \frac{p(p-1)}{2}$. Furthermore, it follows from (22) and the first equation in (24) that

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

Using the likelihood equations (24) one can devise an iterative three-step method for computing the MLE of α, IV's ν_{ij}^2's and GARPs ϕ_{ij}'s. 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 (24) one obtains an estimate of IV's, and iterate until a convergence criterion is met. Although the last formula in (24) seems to require inversion of a matrix of order $\frac{p(p-1)}{2}$, its block diagonal structure can be exploited to save computation time.

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

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 full matrix to be inverted is of order $p-1$. For more detailed information on the estimation of common GARPs see Pourahmadi et al. (2004).

5 Bayesian Modelling of Covariances

Traditionally, in developing Bayesian approaches to inference for Σ an inverse Wishart (IW) for the covariance matrix is used, for an excellent review see Brown et al. (1994) and Wong et al. (2003). However, the success of Bayesian computation and Markov chain Monte Carlo (MCMC) in the late 1980's has opened up the possibility of using more flexible and elaborate non-conjugate priors for covariance matrices. Some of the most notable contributions and their connections with the three decompositions are outlined here.

Perhaps the first breakthrough with the GLM principles in mind is due to Leonard and Hsu (1992). They employed the matrix logarithmic transformation of Σ, computed using its spectral decomposition, i.e. $\log \Sigma = P \log \Lambda P'$ where $\log \Lambda = diag(\log \lambda_1, \cdots, \log \lambda p)$, see (15). The matrix $\log \Sigma$ turns out to be symmetric with unconstrained entries, so that formally a multivariate normal prior with a large number of hyperparameters is appropriate. However, according to Brown et al. (1994) a major drawback of this approach lies in the lack of statistical interpretability of logarithms of the eigenvalues of Σ and hence the entries of $\log \Sigma$. Consequently, the substantive knowledge cannot be used in handling the priors and their hyperparameters.

The alternative reference prior of Yang and Berger (1994) and the hierarchical priors of Daniels and Kass (1999) also rely on the spectral decomposition of the covariance matrix. They reparametrize the orthogonal matrix in terms of logit of the Given angles, so that they are unconstrained and hence conform to the GLM principles. Here too, the statistical relevance and interpretation of these angles are not well understood at this time, therefore the choice of suitable priors is not guided by substantive knowledge, particularly when the sizes of the matrices are large. The local parametrization of orthogonal matrices in Boik (2002) could shed some light on the problem of interpretation of the new parameters in both approaches.

The use of Cholesky decomposition of a covariance matrix or the regression dissection of the associated random vector has a long history and can be traced at least to the work of Bartlett (1933). Though the ensuing parameters have nice statistical interpretation, there remains the choice of "order" among the variables or a coordinate system that could not affect the conclusions. It is shown by Dawid (1988) and Brown et al. (1994) that a regression dissection of the inverse Wishart (IW) distribution reveals some of its noteworthy features which makes it possible to define flexible generalized inverted Wishart (GIW) priors for general covariance matrices. These ideas and techniques are further refined by Gaithwaite and Al-Awadi (2001) in prior distribution elicitation from experts, and extended to longitudinal and panel data setup by Daniels and Pourahmadi (2002) and Smith and Kohn (2002) where the "order" problem is not an issue due to the natural time-order for such data.

The first use of variance-correlation separation in Bayesian covariance modelling seems to be due to Barnard et al. (2000) who introduced independent priors for the standard deviations and correlations. Two prior models used for the correlation matrices are, (i) the jointly uniform prior, where the correlation matrix is assumed a priori uniformly distributed over all possible correlation matrices, and (ii) marginally uniform prior, in which the marginal prior for each entry of the correlation matrix is a modified beta distribution over $[-1, 1]$. This approach has been adopted lately by Wong et al. (2003) and Liechty, Liechty and Müller (2003) in the context of finance.

6 Nonparametric Estimation of Covariances

The nonparametric estimation of the spectral density of a stationary time series is perhaps the earliest and most successful example of nonparametric estimation of a (Toeplitz) covariance structure. The Fourier transform (spectral decomposition) of the covariance plays a central role in this development. Schuster's reliance on *periodogram* in the search for *hidden periodicity*, in the late nineteenth century, and Yule's (1927) use of *autoregression* in the search for *flexible periodicity* were not necessarily inspired by the idea of orthogonalizing the observations of a time series. However, in the 1930's publication and power of the spectral representation theorem and the Wold decomposition of stationary processes elevated them as the prototypes of two distinct techniques for eliminating dependence among time series observations. For an excellent account on predecessors of periodogram and autoregression, see Klein (1997, Chaps. 9,10) or Pourahmadi (2001, Sec. 2.4). Nowadays, it is taken for granted that in the spectral-domain and time-domain analysis of a stationary time series, the spectral decomposition and the Cholesky decomposition of the covariance matrix (function) are the main tools, respectively.

For general covariance matrices, such transformations are crucial to loosen the tight grip of the positive-definiteness constraint, and nonparametric methods can suggest the form of the underlying covariance structure. More generally, as in nonparametric regression, nonparametric estimators of covariance structures could be used either as a guide to the formulation of parsimonious parametric models or as the basis for formal inference without imposing parametric restrictions. To date, most nonparametric estimators of covariance matrices are developed either for stationary processes, see Glasbey (1988); Hall et al. (1994); Hall and Patil (1994) or without heeding the positive-definiteness, see Shapiro and Botha (1991); Sampson and Guttorp (1992); Nott and Dunsmuir (2002) for spatial data and Diggle and Verbyla (1998) and references therein for longitudinal data. In the latter, the repeated measurements over time for subjects are viewed as samples from a Gaussian process with an unknown covariance structure. Diggle and Verbyla (1998) have proposed nonparametric estimators for the covariance structures without assuming stationarity using a kernel weighted local linear regression smoothing of sample variogram ordinates and of squared residuals from the mean. Unfortunately, such an estimator is not guaranteed to be positive-definite.

The use of spectral decomposition in nonparametric estimation of covariance of functional data, arising from experiments where the basic observed responses are curves, were initiated by Rice and Silverman (1991) and has been pursued vigorously by others (Silverman, 1996; Boente and Fraiman, 2000; Ramsay and Silverman, 1997). The covariance structure is estimated using the so-called functional principal component analysis, this amounts to smoothing the principal components of functional data using penalized least squares of the normalized eigenvectors subject to the orthogonality constraint. A kernel-based principal component analysis for functional data is proposed by Boente and Fraiman (2000) which allows derivation of the asymptotic distribution of the smooth principal components. Maintaining orthogonality of the smooth principal components remains a major computational challenge in both approaches.

In sharp contrast, the fact that the lower triangular matrix T in the Cholesky decomposition of a covariance matrix Σ is unconstrained makes it ideal for nonparametric estimation. Wu and Pourahmadi (2003) have used local polynomial estimators to smooth the subdiagonals of T. For the moment, denoting such estimators of T and D in (2) by $\hat{T}$ and $\hat{D}$, an estimaor of Σ given by $\hat{\Sigma} = \hat{T}^{-1}\hat{D}\hat{T}^{-1\prime}$ is guaranteed to be positive-definite. Although one could smooth rows and columns of T, or the whole T viewed as a bivariate function, the idea of smoothing along its subdiagonals is motivated by the similarity of the regressions in (3) to the varying-coefficients autoregressions (Kitagawa and Gersch, 1985, 1996; Dahlhaus, 1997):

$$\sum_{j=0}^{m} f_{j,p}(t/p) y_{t-j} = \sigma_p(t/p)\varepsilon_t, t = 0, 1, 2, \cdots,$$

where $f_{0,p}(\cdot) = 1, f_{j,p}(\cdot),\ 1 \leq j \leq m$, and $\sigma_p(\cdot)$ are continuous functions on $[0, 1]$ and $\{\varepsilon_t\}$

is a sequence of independent random variables each with mean zero and variance one. This analogy and comparison with the matrix T for stationary autoregressions having constant entries along subdiagonals suggest taking the subdiagonals of T to be realizations of some smooth univariate functions:

$$\phi_{t,t-j} = f_{j,p}(t/p)\ ,\ \sigma_t = \sigma_p(t/p).$$

The details of smoothing and selection of the order m of the autoregression and a simulation study comparing performance of the sample covariance matrix to smoothed estimators are given in Wu and Pourahmadi (2003).

Due to the closer connection between entries of T and the family of regression (3), it is conceivable that some of the entries of T could be zero or close to it. Smith and Kohn (2002) have used a prior that allows for zero entries in T and have obtained a parsimonious model for Σ without assuming a parametric structure. Similar results are reported in Huang, Liu and Pourahmadi (2004) using penalized likelihood with L_1-penalty to estimate T for Gaussian data.

7 What Have We Learned?

The three decompositions and unconstrained reparametrizations of their "dependence" components hold the key to successful research in parsimonious modelling of covariance matrices in the spirit of GLM. The trade-off between the unconstrained reparametrizations and statistical interpretability of the new parameters plays the key role in choosing the suitable decomposition in a given situation. In the most familiar variance-correlation decomposition, the correlation matrix is interpreted easily, but is highly constrained; in the Cholesky decomposition, the lower triangular matrix or the GARPs are unconstrained and statistically interpretable, but requires an "order" among the underlying variables. The orthogonal matrix of the spectral decomposition is less constrained than a correlation matrix, but is "interpretable" as the coefficients of the principal components. The reparametrization of an orthogonal matrix in terms of the logits of the Givens angles is known to be unconstrained. What is the statistical meaning of the Givens angles? This question is worth pursuing. The entries of the matrix-logarithm of a covariance matrix constructed from the logarithm of its eigenvalues is unconstrained, their statistical interpretability has received a harsh verdict (Brown et al. 1994) that needs reconsideration and further research.

Acknowledgments

This research was partially supported by the NSF grant DMS-0307055.

References

Aguilar, O. and West, M. (2000), Bayesian dynamic factor models and portfolio allocation, *Journal of Business and Economic Statistics*, **18**, 338-357

Alexander, C. (2001). *Market Models: A Guide to Financial Data Analysis.* Wiley, New York.

Anderson, T.W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.*, **1**, 135-141.

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.

Bartlett, M.S. (1933). On the theory of statistical regression. *Proc. Roy. Soc. Edinburgh* **53**, 260-283.

Boente, G., Fraiman, R. (2000). Kernel-based functional principal components. *Statistics and Probability Letters*, **48**, 335-345.

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.

Brown P.J., Le, N.D. and Zidek, J.V. (1994). Inference for a covariance matrix. In *Aspects of Uncertainty* (P.R. Freeman and A.F.M. Smith, ed.). Wiley, Chichester, U.K., 77-90.

Cameron, M.A. and Turner, T.R. (1987). Fitting models to spectra using regression packages. *Applied Statistics*, **36**, 47-57.

Cannon, M.J., Warner, L., Taddei, J.A. and Kleinbaum, D.G. (2001). What can go wrong when you assume that correlated data are independent: an illustration from the evaluation of a childhood health intervention in Brazil. *Statist. in Med.* **20**, 1461-1467.

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

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, London.

Champion, C.J. (2003). Empirical Bayesian estimation of normal variance and covariances. *J. of Multivariate Analysis*, **87**, 60-79.

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

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **36**, 1-37.

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.

Davison, A.C. and Sardy, S. (2000). The partial scatterplot matrix. *J. of Computational and Graphical Statistics*, **9**, 750-758.

Dawid, A.P. (1988). The infinite regress and its conjugate analysis. In *Bayesian Stastics 3*, Ed. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, 95-110. Oxford, Claredon Press.

Day, D.K., Ghosh, S.K. and Mallick, B.K. (2000). *Generalize Linear Models: A Bayesian Perspective*. Marcel Dekker, Inc., New York.

Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis.* Reading, MA, Addison-Wesley.

Dempster, A. (1972). Covariance selection models. *Biometrics*, **28**, 157-175.

Diebold, F.X. and Nerlove, M. (1989). The dynamics of exchange rate volatility: A multivariate latent factor ARCH Model. *J. of Applied Econometrics*, **4**, 1-21.

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

Diggle, P., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*, 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.

Edgeworth, F.Y. (1892). Correlated averages. *Philosophical Magazine, 5th Series*, **34**, 190-204.

Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, 50, 987-1008.

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.

Engle, R.F. and Kroner, K.F. (1995) Multivariate simultaneous generalized ARCH. *Econometric Theory*, 11, 122-150.

Engle, R.F. and Sheppard, K. (2001) Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. preprint.

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.

Gabriel, K. R. (1962), Ante-dependence analysis of an ordered set of variables, *The Annals of Mathematical Statistics*, 33 , 201-212.

Gaithwaite, P.H. and Al-Awadhi, S.A. (2001). Nonconjugate prior distribution assessment for multivariate normal sampling. *J. of Royal Statist B*, **63**, 95-110.

Geweke, J.F. and Zhou, G. (1996) Measuring the price error of arbitrage pricing theory. *The Review of Financial Studies*, 9, 557-587.

Gourieroux, C. (1997), *ARCH models and financial applications*, Springer-Verlag Inc (Berlin; New York)

Hawkins, D.M. (2003). Discussion of "A Review and Analysis of the Mahalanobis-Taguchi System". *Technometrics*, **45**, 25-29.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417-441, 498-520.

Huang, J.Z., Liu, N. and Pourahmadi, M. (2004). Parsimonious covariance matrix estimation for longitudinal data via penalized likelihood. In preparation.

James, G.M. and Sugar, C.A. (2003). Clustering for sparsely sampled functional data, *J. of the Amer. Statist. Assoc.*, **98**, 397-408.

Jones, R.H. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach.* Chapman and Hall, London.

Kitagawa, G. and Gersch, W. (1985). A smoothness priors time varying AR coefficient modeling of nonstationary time series. *IEEE Trans. on Automatic Control*, AC-30, 48-56.

Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series.* Springer-Verlag, New York.

Klein, J. L. (1997. *Statistical Visions in Time: A History of Time Series Analysis, 1662-1938*. Cambridge University Press.

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

Krzanowski, W.J., Jonathan, P., McCarthy, W.V. and Thomas, M.R. (1995). Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Appl. Statist.*, **44**, 101-115.

Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Ledoit, O., Santa-Clara, P. and Wolf, M. (2003). Flexible multivariate GARCH modeling with an application to international stock markets. *The Review of Economics and Statistics*, **85**, 735-747.

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

Lin, S.P. and Perlman, M.D. (1985). *A Monte Carlo comparison of four estimators of a covariance matrix. In Multivariate Analysis*, 6, Ed. P.R. Krishnaiah, pp. 411-429. Amsterdam: North-Holland.

Liu, C. (1993). Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. *J. of Multiv. Analysis*, **46**, 198-206.

Macchiavelli, R. and Arnold, S. (1994). Variable order ante-dependence models. *Communications in Statistics*, **23**, 2683-2699.

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

Mason, R.L. and Young, J.C. (2002), Multivariate statistical process control with industrial applications, SIAM [Society for Industrial and Applied Mathematics] (Philadelphia).

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

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models.* Wiley, New York.

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.

Nelder, J. (1998). A large class of models derived from generalized linear models, *Statistics in Medicine*, **17**, 2747-2753.

Nott, D.J. and Dunsmuir, T.M. (2002). Estimation of nonstationary spatial covariance structure. *Biometrika*, **89**, 819-829.

Pan, J. and MacKenzie, G. (2003). On modelling mean-covariance structure in longitudinal studies. *Biometrika*, **90**, 239-244.

Pinheiro, J.D. and Bates, D.M. (1996). Unconstrained parameterizations for variance-covariance matrices. *Stat. Comp.*, **6**, 289-366.

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

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

Pourahmadi, M. (2002). Joint modeling of mean, dispersion and correlation: generalized Gaussian estimating equations. Correlated Data Modeling, 27-34, Proceedings, Trieste, 1999, Franco Angeli, Milano.

Pourahmadi M, Daniels M. and Park T. (2004). Simultaneous modelling of the Cholesky decomposition of several covariance matrices with applications. Preprint.

Ramsey, J.O., Silverman, B.W. (1997). *Functional Data Analysis.* Springer, New York.

Rice, J., Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53**, 233-243.

Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*, **87**, 99-112.

Roy, J. (1958) Step-down procedure in multivariate-analysis. *Annals of Mathematical Statistics*, 29, 1177-1187.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.

Silverman, B.W. (1996). Smoothed functional principal component analysis by choice of norm. *Ann. Statist.* **24**, 1-24.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. of Amer. Statist. Assoc.*, **97**, 1141-1153.

Stein, C. (1975) Estimation of a covariance matrix. In *Reitz lecture*, Atlanta, Georgia, 1975. 39th annual meeting IMS.

Styan, G.P.H. (1968). Inference in multivariate normal populations with structure. Part I: inference when correlations are known. Technical report, Department of Statistics, University of Minnesota, USA.

Szatrowski, T.H. (1980). Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances. *Ann. of Statist.*, **8**, 802-810.

Szegö, G., Grenander, U. (1958). *Toeplitz Forms and Their Applications.* University of California Press. Berkeley, CA.

Taguchi, G. and Jugulum, R. (2002). *The Mahalanobis-Taguchi Strategy*, Wiley, New York.

Tsay, R. (2002). *Analysis of Financial Time Series*, Wiley, New York.

Tse, Y.K. and Tsui, A.K. (2002) A multivariate generalized conditional heteroscedasticity model with time varying correlations. *Journal of Business and Economics Statistics*, 20, 351-362.

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

Vrontos, I.D., Dellaportas, P., and Politis, D.N. (2003) A multivariate latent GARCH model. *Econometric Theory*, to appear.

Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *J. of Amer. Statist. Assoc.* **75**, 963-972.

Wold, H.O.A. (1960). A generalization of causal chain models. *Econometrica*, **28**, 443-463.

Wong, F., Carter, C.K. and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, **90**, 809-830.

Wright, S. (1934). The method of path coefficients. *The Ann. of Math. Statist.*, **5**, 161-215.

Wu, W.B., Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **90**, 831-844.

Yang, R. and Berger, J.O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.*, **22**, 1195-1211.

Yule, G.U. (1927). On a model of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Phil. Trans. A*, **226**, 267-298.

Zellner, A. (1979) An error-components procedures (ECP) for introducing prior information about covariance matrices and analysis of multivariate regression models. *International Economic Review*, 20, 201-214.

Zimmerman, D.L. and Núñez-Antón, V. (1997). Structured antedependence models for longitudinal data. In *Modeling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions* (T.G. Gregiore et al, eds.). Springer Lecture Notes in Statistics, 122, Spring-Verlag, New York, 63-76.

Zimmerman, D.L. and Núñez-Antón, V. (2001). Parametric modelling of growth curve data: An overview. *Test*, **10**, 1-73.