Covariance Estimation: The GLM and Regularization Perspectives

Mohsen Pourahmadi
Department of Statistics
Texas A&M University
College Station, TX, USA
pourahm@stat.tamu.edu

Abstract

Finding an *unconstrained* and *statistically interpretable* reparameterization of a covariance matrix is still an open problem in statistics. Its solution is of central importance in covariance estimation, particularly in the recent high-dimensional data environment where enforcing the positive-definiteness constraint could be computationally expensive. We provide a survey of the progress made in modeling covariance matrices from the perspectives of generalized linear models (GLM) or parsimony and use of covariates in low dimensions, regularization (shrinkage, sparsity) for high-dimensional data, and the role of various matrix factorizations. A viable and emerging regression-based setup which is suitable for both the GLM and the regularization approaches is to link a covariance matrix, its inverse or their factors to certain regression models and then solve the relevant (penalized) least squares problems. We point out several instances of this regression-based setup in the literature. A notable case is in the Gaussian graphical models where linear regressions with LASSO penalty are used to estimate the neighborhood of one node at a time (Meinshausen and Bühlmann, 2006). Some advantages and a limitation of the regression-based Cholesky decomposition (Pourahmadi, 1999) relative to the classical spectral (eigenvalue) and variance-correlation decompositions are highlighted. It provides an unconstrained and statistically interpretable reparameterization, and guarantees the positive-definiteness of the estimated covariance at no additional computational cost, and reduces the unintuitive task of covariance estimation to that of modeling a sequence of regressions. However, its flexibility comes at the cost of imposing an *a priori order* among the variables in a random vector, so that the above problem is solved for time series, longitudinal, functional and spectroscopic data where the variables are naturally ordered, but for the general multivariate data where the variables are unordered it is still open.

Key Words: Bayesian estimation; Cholesky decomposition; Dependence and correlation; Graphical models; Longitudinal data; Parsimony; Penalized likelihood; Precision matrix; Sparsity; Spectral decomposition; Variance-Correlation decomposition

1 Introduction

The $p \times p$ covariance matrix Σ of a random vector $Y = (y_1, \cdots, y_p)'$ with as many as $\frac{p(p+1)}{2}$ constrained parameters plays a central role in virtually all of classical multivariate statistics

(Anderson, 2003), time series analysis (Box et al. 1994), spatial data analysis (Cressie, 1993), variance components and longitudinal data analysis (Searle et al. 1992; Diggle et al. 2002), and in the modern and rapidly growing area of statistical and machine learning dealing with massive and high-dimensional data (Hastie, Tibshirani and Friedman, 2009). More specifically, principal component analysis, factor analysis, classification and cluster analysis, inference about the means and regression coefficients, prediction and Kriging, and analysis of conditional independence in graphical models typically, require an estimate of a covariance matrix or its inverse. It is generally recognized that the two major challenges in covariance estimation are the positive-definiteness constraint and the high-dimensionality where the number of parameters grows quadratically in p. In this survey, we point out that the latter challenge is virtually eliminated by reducing covariance estimation to that of solving a series of penalized least-squares regression problems.

Nowadays, in microarray data, spectroscopy, finance, climate studies and abundance data in community ecology it is common to have situations where $n < p$, so that the use of sample covariance matrix is problematic (Stein, 1956), particularly when its inverse is needed as, for example, in the classification procedures (Anderson, 2003, Chap. 6), multivariate linear regression (Warton, 2008; Witten and Tibshirani, 2009), portfolio selection (Ledoit et al. 2003) and Gaussian graphical models (Wong et al. 2003 ; Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007). In these situations and others, the goal is to find alternative covariance estimators that are more accurate and well-conditioned than the sample covariance matrix.

It was noted rather early by Stein (1956, 1975) that the sample covariance matrix $S = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i'$, based on a sample of size n from a mean zero normal population with the covariance matrix Σ, though unbiased and positive-definite is not an accurate estimator when $\frac{p}{n}$ is large. In this situation, it distorts the eigenstructure of Σ, in the sense that the largest sample eigenvalue will be biased upward and the smallest sample eigenvalue will be biased downward. Since then many improved estimators have been proposed by shrinking

only the eigenvalues of S towards a central value (Haff, 1980, 1991; Lin and Perlman, 1985; Dey and Srinivasan, 1985; Yang and Berger, 1994; Ledoit and Wolf, 2004). These have been derived from a decision-theoretic perspective or by specifying an appropriate prior for the covariance matrix.

By now it is well-known that estimators like Stein's (1956, 1975) that focus on shrinking the eigenvalues, have smaller estimation risks and are usually more accurate than the sample covariance matrix (Lin and Perlman, 1985; Daniels and Kass, 1999; Ledoit and Wolf, 2004). However, when $p > n$, since the sample covariance matrix is singular, suitable ridge regularization (Hoerl and Kennard, 1970; Ledoit and Wolf) will lead to covariance estimators which are more accurate and well-conditioned (Bickel and Li, 2006; Warton, 2008). This class of estimators based on an optimal linear combination of the sample covariance matrix and the identity matrix also ends up shrinking only the eigenvalues of S.

The Stein's family of shrinkage estimators leaving intact the eigenvectors of the sample covariance matrix, are neither sparse nor parsimonious. In the search for sparsity and parsimony, one may directly shrink either the eigenvectors or the matrix S itself toward certain targets or structured covariance matrices like diagonal and autoregressive sturctures as in Daniels and Kass (1999; 2001), Hoff (2009), Fan, Fan and Lv (2008) and Johnstone and Lu (2009). Perhaps, the simplest, more direct and less computationally intensive ways of achieving sparsity are the techniques of banding and thresholding of the sample covariance matrix (Bickel and Levina, 2008a, b; Rothman, Levina and Zhu, 2009) which amounts to elementwise operations on S, and hence completely avoids the computationally expensive eigenvalue problem (Golub and Van Loan, 1989).

In many applications the need for the precision matrix Σ^{-1} is stronger than that for Σ itself. Though the former can be computed from the latter in $\mathcal{O}(p^3)$ operations, this could be computationally expensive and should be avoided when p is large. Recently, inspired by the approach of Meinshausen and Bühlmann (2006) for neighborhood selection in Gaussian

graphical models using p separate linear regressions to estimate the neighborhood of one node at a time, several direct sparse estimators of Σ^{-1} have been proposed using a penalized likelihood approach with a LASSO penaly on its off-diagonal terms (Yuan and Lin, 2007; Banerjee et al. 2008; d'Aspremont et al. 2008; Friedman et al. 2008; Rothman et al. 2008; Rocha et al. 2008; Peng et al. 2009). Friedman et al.'s (2008) graphical LASSO, which is the fastest available algorithm to date, relies on the equivalence of the Banerjee et al. (2008) blockwise interior point procedure, and recursively solving and updating a lasso least-squares regression problem using the coordinate descent algorithm for LASSO. Moreover, the sparse covariance estimator from the graphical LASSO is guaranteed to be positive-definite; this follows from another result due to Banerjee et al. (2008) showing that if the recursive procedure is initialized with a positive-definite matrix, then the subsequent iterates remain positive-definite and invertible. These methods share a common feature of treating the $\frac{p(p-1)}{2}$ off-diagonal entries of the precision matrix as if they were unconstrained regression coefficients. This view along with the use of fast LASSO least-squares regressions has greatly reduced the high-dimensionality challenge of the covariance estimation.

Interestingly, there are several *regression-based* approaches to covariance estimation in the literature where they rely first on the idea of regression to reparameterize a covariance (precision) matrix in a manner that its estimation can be recast later as linear least-squares regression problems. Some powerful examples are: (i) formulating principal component analysis (PCA) as a regression optimization problem (Jong and Kotz, 1999; Zou, Hastie and Tibshirani, 2006), sparse loadings are then obtained by imposing the lasso constraint on the regression coefficients, (ii) regression-based derivation (interpretation) of the modified Cholesky decomposition of a covariance matrix and its inverse (Pourahmadi, 1999; 2001, Sec. 3.5; Bilmes, 2000; Huang et al. 2006; Rothman, Levina and Zhu, 2010), (iii) the regression approach of Meinshausen and Bühlmann (2006), Rocha, Zhang and Yu (2008) and Peng, Wang, Zhou and Zhu (2009) merging all p regressions into a single least-squares

problem, (iv) the graphical LASSO algorithm of Friedman, Hastie and Tibshirani (2008, 2010); see also Friedman et al. (2009, Section 17.3), and (v) iteratively reweighted penalized likelihood of Fan, Feng and Wu (2010) where nonconcave penalties, such as the smoothly clipped absolute deviation (SCAD), are introduced on the entries of the precision matrix , then using the local linear approximation to the nonconcave penalty functions, the problem of sparse estimation of the precision matrix is recast as a sequence of penalized likelihood problems with a weighted LASSO penalty and solved using the graphical LASSO algorithm of Friedman et al. (2008).

Among these approaches it seems only (ii) using the Cholesky decomposition has the distinction of providing unconstrained regression parameters. Unfortunately, however, unlike the others which work for unordered variables and provide permutation-invariant covariance estimators, (ii) and a few other alternatives to the sample covariance matrix proposed in recent years work for ordered variables and give rise to covariance estimators which are sensitive to the order among the variables in Y. These approaches work well for the time series and longitudinal data which have a natural (time) order among the variables in Y, and assume that variables far apart in the ordering are less correlated. For example, regularizing a covariance matrix by tapering (Furrer and Bengtsson, 2007), banding (Bickel and Levina, 2004; 2008a; Wu and Pourahmadi, 2009) and generally those based on the Cholesky decomposition of the covariance matrix or its inverse (Pourahmadi, 1999, 2000; Rothman et al. 2010) do impose an order among the components of Y and are not permutation-invariant. Nevertheless, one could estimate the covariance matrix or its inverse using regression regularization tools like, covariance selection priors, AIC and LASSO penalties on the Cholesky factor (Smith and Kohn, 2002; Wu and Pourahmadi, 2003; Huang, Lin, Pourahmadi and Lin, 2006; Huang, Lin and Lin, 2007) and nested LASSO (Levina, Rothman and Zhu, 2007).

The recent surge of interest in regression-based approaches to *sparse* estimation of large covariance matrices of massive and high-dimensional data, bodes well with the long history

of interest in using regression ideas and covariates to achieve *parsimony* in estimating smaller covariance matrices in the traditional areas of statistics, biostatistics, econometrics and social sciences. For example, longitudinal data collected from expensive clinical trials and biological experiments, may have $n = 30$ subjects and $p \leq 10$. Accurate modeling and estimation of the covariance structure is important in these application areas (Cannon et al. 2001; Carroll, 2003; Qu and Lindsay, 2003; Ye and Pan, 2006; Fitzmaurice et al. 2009). While the process for modeling the mean is well-understood (McCullagh and Nelder, 1989; Diggle et al. 2002), the situation for modeling the covariance matrix is underdeveloped, where at one extreme it is modeled as $\sigma^2 I_p$ (independence) and at the other by an unstructured covariance matrix with $\frac{p(p+1)}{2}$ parameters (Zimmerman and Núñez-Antón, 2001, 2010). In these situations, it is highly desirable to bridge the gap between these two extremes by developing a bona fide GLM methodology and a data-based framework for modeling covariance matrices which include the three stages of model formulation, estimation and diagnostics, just like those for modeling the mean vector (McCullgh and Nelder, 1989). Attempts to develop such methods going beyond the traditional linear covariance models (Anderson, 1973), have been made in recent years by Chiu et al. (1996) and Pourahmadi (1999, 2000); Pan and MacKenzie (2003); Ye and Pan (2006); Wang and Lin (2008); Leng, Zhang and Pan (2010); Lin (2010) using the spectral and Cholesky decompositions of covariance matrices, respectively.

Given the complex nature of a covariance matrix and the positive-definiteness constraint, to develop a GLM methodolgy it is plausible to factorize Σ into two components capturing the "variance" through a diagonal matrix and the "dependence" through a matrix with $\frac{p(p-1)}{2}$ functionally unrelated entries. A decomposition is ideal for the GLM purposes, if its "dependence" component is an unconstrained and statistically interpretable matrix, since then one may use covariates to achieve parsimony. The three most commonly used decompositions in increasing order of adherence to the GLM principles are the variance-correlation, spectral and Cholesky decompositions where their "dependence" components are correlation, orthog-

onal and lower triangular matrices, respectively. While the entries of the first two matrices are always constrained, those of the last are unconstrained. Consequently, computing the (penalized) maximum likelihood estimates of the parmaters of Cholesky decomposition involves unconstrained optimization compared to the orthogonally-constrained optimization algorithm of Flury and Gautschi (1986) for the spectral decomposition. Thus, finding at least an unconstrained reparameterization for Σ seems to be the first order of business for both the GLM and regularization approaches.

The outline of the paper is as follows. Section 2, covers some preliminaries on the GLM for covariance matrices, the roles of the three standard decompositions of a covariance matrix, a regression-based decomposition of the precision matrix useful in Gaussian graphical models, a review of covariance estimation from the GLM perspective and its evolution through linear/inverse, log and hybrid link functions. Steinian shrinkage, regularization (banding and thresholding), penalized likelihood estimation, and improvement of the sample covariance matrix from the shrinkage/sparsity perspective are discussed in Section 3. Some prior distributions on the parameters of the factors of the three decompositions and their roles in the Bayesian inference are reviewed in Section 4. Section 5 concludes the paper with some open problems for further research.

This survey emphasizes the importance of and the need for unconstrained reparameterization in both the GLM- and regularization-type approaches to covariance estimation for low- and high-dimensional data sets. As such it has a relatively narrow focus, important topics like robustness, use of random-effects models, nonparametric and semi-parametric methods in covariance estimation are not discussed. It is hoped to serve as a starting point or a blueprint for further research in this active and growing area of current interest in statistics.

2 The GLM & Matrix Decompositions

In this section, the importance of the GLM, the role of the three matrix decompositions in removing the positive-definiteness constraint on a covariance matrix, the connection between reparameterizing the precision matrix and the Gaussian graphical models, along with linear, log-linear and generalized linear models for covariance matrices are reviewed.

2.1 Positive-Definiteness and the GLM

A major stumbling block in covariance estimation, particularly when using covariates, is the notorious positive-definiteness constraint. Recall that constraints on the mean vector μ or the mean-like parameters of the distribution of a random vector Y has been handled quite successfully in the theory of generalized linear models (McCullagh and Nelder, 1989) using a *link function* $g(\cdot)$ and a *linear predictor* $g(\mu) = X\beta$. What are the analogues of these for a positive-definite covariance matrix?

Since a covariance matrix defined by $\Sigma = E(Y - \mu)(Y - \mu)'$, is a mean-like parameter, it is natural to exploit the idea of GLM to develope a systematic, data-based statistical model-fitting procedure for covariance matrices composed of model formulation, estimation and diagnostics. However, unlike the mean vector where a link function acts *elementwise*, *for covariance matrices elementwise transformations* are not enough as the positive-definiteness is a simultaneous constraint on *all* its entries. More global transformations engaging possibly all entries of a covariance matrix are needed to remove the constraint.

Conceptually, the GLM approach to covariance estimation is important due to its success and track record in unifying a vast collection of apparently disparate approaches, developed over the span of two centuries, to model the mean and mean-like parameters of various distributions (McCullagh and Nelder, 1989). It hinges on the concept of link functions to induce unconstrained and statistically interpretable reparameterization for the mean of a distribution, and to reduce the dimension of the parameter space using covariates. These

two GLM features or principles are precisely what are needed to tackle the two challenges in covariance estimation regardless of the size of p, and to possibly unify diverse approaches to covariance estimation.

Not surprisingly, some common and successful modeling approaches decompose a covariance matrix into its "variance" and "dependence" components, and write regression models using covariates for the logarithm of the "variances". However, writing such regression models for the entries of the "dependence" component is still a challenging problem because these are often constrained. In the next section, three examples of unconstrained parameterization of a covariance matrix are given which involve the spectral and Cholesky decompositions.

2.2 The Matrix Decompositions

In this section, we present the roles of the variance-correlation, spectral and Cholesky decompositions in potentially removing the positive-definiteness constraint on a covariance matrix, and paving the way for using covariates to reduce its high number of parameters.

2.2.1 The Variance-Correlation Decomposition

The simple decomposition

$$\Sigma = DRD,$$

where D is the diagonal matrix of standard deviations and $R = (\rho_{ij})$ is the correlation matrix of Y, has a strong practical appeal since these two factors of Σ are easily interpreted in terms of the original variables. It allows one to estimate D and R separately, which is important in situations where one component might be more important than the other (Lin and Perlman, 1985; Liang and Zeger, 1986; Barnard et al. 2000).

Note that while the logarithm of the diagonal entries of D are unconstrained, the correlation matrix R must be positive-definite with the additional constraint that all its diagonal entries are equal to 1. Thus, it is inconvenient to work with it in the framework of GLM to reduce its large number of parameters. In the literature of longitudinal data analysis

(Liang and Zeger, 1986; Diggle et al. 2002; Zimmerman and Núñez-Antón, 2010) and other application areas dealing with correlated data, in the interest of expediency, parsimony and ensuring positive-definiteness structured correlation matrices with a few parameters are preferred. Fan, Huang, and Li (2007) have studied a semiparametric model for the covariance structure by using the variance-correlation decomposition. They estimated the marginal variances via kernel smoothing and used specific parametric models for the correlation matrix such as AR(1) or ARMA(1, 1) to ensure positive-definiteness of the estimated covariance.

2.2.2 The Variance-Correlation Decomposition of the Precision Matrix: Gaussian Graphical Models

Recall that the marginal (pairwise) dependence or independence among the entries of a random vector are summarized by the off-diagonal entries of Σ or the entries of the correlation matrix $R = (\rho_{ij})$. Surprisingly, the conditional dependencies can be found in the off-diagonal entries of the precision matrix $\Sigma^{-1} = (\sigma^{ij})$. More precisely, take Y to be a mean zero normal random vector with a positive-definite covariance matrix, if the ijth component of the precision matrix is zero, then variables y_i and y_j are conditionally independent, given the other variables. A connection between graphs and statistical models is usually made by identifying the graph's nodes with random variables and translating the graph's edges into a parametrization that relates to the precision matrix of a multivariate normal distribution.

In this section, we focus on and give several regression interpretations of the entries of the variance-correlation decomposition of the precision matrix:

$$\Sigma^{-1} = (\sigma^{ij}) = \tilde{D}\tilde{R}\tilde{D}.$$

Most of these are motivated by the recent surge of activities in sparse estimation of Σ^{-1} in the context of Gaussian graphical models sparked by the approach in Meinshausen and Bühlmann (2006) using p LASSO linear least-squares regression problems. We show that the entries of $(\tilde{R}, \tilde{D})$ have direct statistical interpretations in terms of the partial correlations, and variance of predicting one variable given the rest. More precisely, using standard regression

calculations we show below that $\tilde{\rho}_{ij}$ is the *partial correlation* coefficient between y_i and y_j after removing the linear effect of the $p-2$ remaining variables, and is given by $-\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$, and that $\tilde{d}_{ii}^2$ is the *partial variance* of y_i after removing the linear effect of the remaining $p-1$ variables and is given by $\frac{1}{\sigma^{ii}}$.

For this and other regression-based techniques reviewed in this survey, it is instructive to partition a random vector Y into two components $(Y_1', Y_2')'$ of dimensions p_1 and p_2, respectively, and then partition its covariance and precision matrices conformally as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \; \Sigma^{-1} = \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix}.$$

Some useful relationships among the blocks of Σ and Σ^{-1} are obtained by considering the linear least-squares regression (prediction) of Y_2 based on Y_1. Let the $p_2 \times p_1$ matrix $\Phi_{2|1}$ be the regression coefficients and the vector of regression residuals be denoted by $Y_{2\cdot 1} = Y_2 - \Phi_{2|1}Y$. Recall that $\Phi_{2|1}$ is found so that the vector of residuals $Y_{2\cdot 1}$ is uncorrelated with Y_1. Then, it follows that (see the Appendix):

$$\Phi_{2|1} = \Sigma_{21}\Sigma_{11}^{-1}, \tag{1}$$

and

$$\mathrm{Cov}(Y_{2\cdot 1}) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \Sigma_{22\cdot 1}. \tag{2}$$

The following lemma re-expresses the covariance matrix of the vector of residuals $Y_{2\cdot 1}$ and the matrix of regression coefficients $\Phi_{2|1}$ in terms of the corresponding blocks of the precision matrix.

Lemma. With notation as above, we have

$$\Phi_{2|1} = \Sigma_{21}\Sigma_{11}^{-1} = -(\Sigma^{22})^{-1}\Sigma^{21}, \tag{3}$$

and

$$\mathrm{Cov}(Y_{2\cdot 1}) = \Sigma_{22\cdot 1} = (\Sigma^{22})^{-1}. \tag{4}$$

Certain special choices of Y_2 corresponding to $p_2 = 1, 2$, are helpful in connecting $\Phi_{2|1}, \Sigma_{22\cdot 1}$ directly to the entries of the concentration matrix Σ^{-1} as we discuss below.

For example, when $p_2 = 1$, $Y_2 = y_i$, for a fixed i, and $Y_1 = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_p)' = Y_{-(i)}$, then $\Sigma_{22\cdot 1}$ is a scalar, called the *partial variance* of y_i given the rest. Let $\tilde{y}_i$ be the linear least-squares predictor of y_i based on the rest $Y_{-(i)}$, and $\tilde{\varepsilon}_i = y_i - \tilde{y}_i$, $\tilde{d}_i^2 = Var(\tilde{\varepsilon}_i)$ be its prediction error and prediction error variance, respectively. Then,

$$y_i = \sum_{j \neq i} \beta_{ij} y_j + \tilde{\varepsilon}_i, \tag{5}$$

and it follows immediately from (3)-(4) that the regression coefficients of y_i on $Y_{-(i)}$, are given by:

$$\beta_{i,j} = -\frac{\sigma^{ij}}{\sigma^{ii}}, j \neq i, \tag{6}$$

and

$$\tilde{d}_i^2 = \mathrm{Var}(y_i | y_j, j \neq i) = \frac{1}{\sigma^{ii}}, \ i = 1, \cdots, p. \tag{7}$$

This shows that σ^{ij}, the (i, j) entry of the precision matrix is, up to a scalar, the regression coefficient of variable j in the multiple regression of variable i on the rest. As such each $\beta_{i,j}$ is an unconstrained real number, $\beta_{j,j} = 0$ and $\beta_{i,j}$ is not symmetric in (i, j).

Writing (6) in matrix form gives the following alternative and useful factorization of the precision matrix:

$$\Sigma^{-1} = \tilde{D}^2 (I_p - \tilde{B}), \tag{8}$$

where $\tilde{D}$ is a diagonal matrix with $\tilde{d}_j$ as its jth diagonal entry, and $\tilde{B}$ is a $p \times p$ matrix with zeros along its diagonal and $\beta_{j,k}$ in the (j, k)th position. Now, it is evident from (8) that the sparsity patterns of Σ^{-1} and $\tilde{B}$ are the same, and hence the former can be inferred from the latter using the regression setup (5) along with a LASSO penalty for each regression. This is essentially the key conceptual tool behind the approach of Meinshausen and Bühlmann (2006). Note that the left-hand side of (8) is a symmetric matrix while the right-side is not necessarily so. Thus, to increase computational and statistical efficiencies one must

impose the following symmetry constraint (Rocha et al. 2008; Friedman et al. 2010) for $j, k = 1, \cdots, p$:

$$d_k^2 \beta_{jk} = d_j^2 \beta_{kj}. \tag{9}$$

As another important example, take $p_2 = 2$, $Y_2 = (y_i, y_j)$, $i \neq j$ and $Y_1 = Y_{-(ij)}$ comprising the remaining $p-2$ variables. Then, it follows from (4) that the covariance matrix between y_i, y_j, after eliminating the linear effects of the other $p-2$ components, is given by

$$\Sigma_{22 \cdot 1} = \begin{pmatrix} \sigma^{ii} & \sigma^{ij} \\ \sigma^{ij} & \sigma^{jj} \end{pmatrix}^{-1} = \Delta^{-1} \begin{pmatrix} \sigma^{jj} & -\sigma^{ij} \\ -\sigma^{ij} & \sigma^{ii} \end{pmatrix},$$

where $\Delta = \sigma^{ii}\sigma^{jj} - (\sigma^{ij})^2$, the correlation coefficient in $\Sigma_{22 \cdot 1}$ is, indeed, the *partial correlation coefficient* between y_i and y_j:

$$\tilde{\rho}_{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}}\sigma^{ij}}, \tag{10}$$

as announced earlier. From (6) and (10), it follows that

$$\beta_{ij} = \tilde{\rho}_{ij} \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}. \tag{11}$$

This representation which shows that Σ^{-1} and $\tilde{R}$ share the same sparsity is the basis for the Peng et al.'s (2009) SPACE algorithm which imposes a LASSO penalty on the off-diagonal entries of the matrix of partial correlations $\tilde{R}$.

2.2.3 The Spectral Decomposition

The spectral decomposition of a covariance matrix given by,

$$\Sigma = P \Lambda P' = \sum_{i=1}^{p} \lambda_i e_i e_i', \tag{12}$$

where Λ is a diagonal matrix of eigenvalues and P the orthogonal matrix of normalized eigenvectors with e_i its ith column, is familiar from the literature of principal component analysis (Anderson, 2003; Flury, 1988). The entries of Λ and P have interpretations as variances and coefficients of the principal components. The matrix P being orthogonal is

constrained, so that it is inconvenient to work with it in the framework of GLM or to use covariates to reduce its high number of parameters.

In spite of the severe constraint on the orthogonal matrix, the spectral decomposition is the source of two new reparameterizations that are unconstrained, but is believed to be hard to interpret statistically. The first, due to Leonard and Hsu (1992) and Chiu et al. (1996) exploits the fact that the logarithm of a covariance matrix Σ defined below is an unconstrained symmetric matrix:

$$\log \Sigma = P \log \Lambda P' = \sum_{i=1}^{p} (\log \lambda_i) e_i e_i'. \tag{13}$$

However, a drawback of this transformation (link function) seems to be the lack of statistical interpretability of the entries of $\log \Sigma$ (Brown, Le and Zidek, 1994; Liechty et al. 2004). From (12)-(13) it is evident that the entries of Σ and $\log \Sigma$ are similar functions of the entries of P and Λ, except that in (13) λ_i is replaced by $\log \lambda_i$. Can this "small" substitution be the reason for the "big" difference in the statistical interpretability of the entries of log of a covariance matrix and the matrix itself? Perhaps the apparent simplicity of (13) may have concealed the complicated and highly nonlinear relations that exist between the entries of Σ and $\log \Sigma$ as seen in the following simple example.

For $\log \Sigma = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$, using the eigenvalues and eigenvectors of the 2×2 matrix, it can be shown that

$$\sigma_{11} = \frac{1}{2\sqrt{\Delta}} \exp\left(\frac{\alpha+\gamma}{2}\right) \left\{\sqrt{\Delta} u - (\alpha - \gamma) u^-\right\},$$

where

$$\begin{aligned} \Delta &= (\alpha-\gamma)^2 + 4\beta^2, \\ u^{\pm} &= \exp\left(\frac{\sqrt{\Delta}}{2}\right) \pm \exp\left(-\frac{\sqrt{\Delta}}{2}\right). \end{aligned}$$

The second reparameterization relies on the logarithm of the eigenvalues and logit of the Givens angles (Golub and Van Loan, 1989) associated with pairs of columns of the orthogonal matrix P. Though the logit of the Givens angles are unconstrained (Daniels and Kass, 1999), they are hard to interpret statistically at this time.

2.2.4 The Cholesky Decompositions

The standard Cholesky decomposition of a positive-definite matrix encountered in some optimization techniques, software packages and matrix computation (Goulb and Van Loan, 1989), is of the form

$$\Sigma = CC', \tag{14}$$

where $C = (c_{ij})$ is a unique lower-triangular matrix with positive diagonal entries. Statistical interpretation of the entries of C is difficult in its present form (Pinheiro and Bates, 1996). However, reducing C to unit lower-triangular matrices through multiplication by the inverse of $D = diag(c_{11}, \cdots, c_{pp})$, makes the task of statistical interpretation of the diagonal entries of C, and the ensuing unit lower-triangular matrix much easier.

For example, using basic matrix multiplication, (14) can be rewritten as

$$\Sigma = CD^{-1}DDD^{-1}C' = LD^2L', \tag{15}$$

where $L = CD^{-1}$ is obtained from C by dividing the entries of its ith column by c_{ii}. This is usually called the modified Cholesky decomposition of Σ, it can also be written in the forms

$$T\Sigma T' = D^2, \Sigma^{-1} = T'D^{-2}T, \tag{16}$$

where with $T = L^{-1}$. Note that the second identity is, in fact, the modified Cholesky decomposition of the precision matrix Σ^{-1}, and the first identity in (16) looks a lot like the spectral decomposition, in that Σ is diagonalized by a lower triangular matrix. However, we show next that unlike the constrained entries of the orthogonal matrix of the spectral decomposition, the nonredundant entries of $T = L^{-1}$ are unconstrained and statistically meaningful. Furthermore, the argument makes it clear that the parameters in the factors of the Cholesky decomposition are dependent on the *order* in which the variables appear in the random vector Y. Wagaman and Levina (2008) have proposed an Isomap method for discovering meaningful orderings of variables based on their correlations. This method may

result in block-diagonal or banded correlation structure, resulting in an Isoband estimator, or may help to fix a reasonable ordering before applying the Cholesky decomposition, see Section 5.

As in Section 2.2.2, we use the idea of regression or the Gram-Schmidt orthogonalization to orthogonalize the random variables sequentially, and to show that T and D can be constructed directly by regressing a variable y_t on its predecessors. In what follows, it is assumed that Y is a random vector with mean zero and a positive-definite covariance matrix Σ. Let $\hat{y}_t$ be the linear least-squares predictor of y_t based on its predecessors $y_{t-1}, \ldots, y_1$, $\varepsilon_t = y_t - \hat{y}_t$ be its prediction error with variance $\sigma_t^2 = Var(\varepsilon_t)$. Then, there are unique scalars ϕ_{tj} so that

$$y_t = \sum_{j=1}^{t-1} \phi_{tj} y_j + \varepsilon_t, \ t = 1, \cdots, p. \tag{17}$$

Next, we show how to compute the regression coefficients ϕ_{tj} using the covariance matrix. For a fixed $t, 2 \leq t \leq p$, set $\phi_t = (\phi_{t1}, \cdots, \phi_{t,t-1})'$ and let Σ_t be the $(t-1) \times (t-1)$ leading principal minor of Σ and $\tilde{\sigma}_t$ be the column vector composed of the first $t-1$ entries of the tth column of Σ. Then, from (1)-(2) it follows that

$$\phi_t = \Sigma_t^{-1} \tilde{\sigma}_t, \sigma_t^2 = \sigma_{tt} - \tilde{\sigma}_t' \Sigma_t^{-1} \tilde{\sigma}_t. \tag{18}$$

Let $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_p)'$ be the vector of successive uncorrelated prediction errors with $Cov(\varepsilon) = diag(\sigma_1^2, \cdots, \sigma_p^2) = D^2$. Then, (17) can be rewritten in matrix form as $\varepsilon = TY$, where T is the following unit lower triangular matrix :

$$T = \begin{pmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{n1} & -\phi_{n2} & \cdots & -\phi_{n,n-1} & 1 \end{pmatrix}. \tag{19}$$

Now, computing $Cov(\varepsilon) = Cov(TY) = T\Sigma T'$, gives the modified Cholesky decomposition (16).

Since the ϕ_{ij}'s in (18) are simply the regression coefficients computed from an unstructured covariance matrix, these coefficients along with $\log \sigma_t^2$ are unconstrained (Pourahmadi, 1999, 2000). From (17) it is evident that, the orthogonalization process reduces the task of modeling a covariance matrix to that of a sequence of p varying-coefficient and varying-order regression models so that one can bring the whole regression analysis machinery to the service of the unintuitive task of modeling covariance matrices (Smith and Kohn, 2002; Wu and Pourahmadi, 2003; Huang et al. 2006, 2007; Bickel and Levina, 2008a; Rothman et al. 2009). An important consequence of (16) is that for any estimate $(\hat{T}, \hat{D}^2)$ of the Cholesky factors, the estimated precision matrix $\hat{\Sigma}^{-1} = \hat{T}'\hat{D}^{-2}\hat{T}$ is guaranteed to be positive-definite.

An alternative form of the Cholesky decomposition (15) due to Chen and Dunson (2003) which can be obtained from (14) is

$$\Sigma = D\tilde{L}\tilde{L}'D,$$

where $\tilde{L} = D^{-1}C$ is obtained from C by dividing the entries of its ith row by c_{ii}. This form has proved useful for joint variable selection for fixed and random effects in the linear mixed-efects models, and when the focus is on modeling the correlation matrix, see Bondell, Krishna and Ghosh (2010), Pourahmadi (2007a).

Some early and implicit examples of the use of the Cholesky decomposition in the literature of statistics include Bartlett'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. Some of the more explicit uses are in Kalman (1960) for filtering of state-space models and the Gaussian graphical models (Wermuth, 1980). For other uses of Cholesky decomposition in multivariate quality control and related areas see Pourahmadi (2007b).

2.3 GLM for Covariance Matrices

Research on estimation of covariance matrices has followed paths of developments very much similar to those for the estimation of the mean vector or regression analysis, and modeling dependence in time series analysis (Klein, 1997). It has gone or is currently going through the phases of linear, log-linear and generalized linear models (McCullagh and Nelder, 1989), ridge regularization (Hoerl and Kennard, 1970), LASSO (Tibshirani, 1996) and aspects of penalized normal likelihood, see Fan and Lv (2010) for a review of the current state of regression analysis for high-dimensional data.

2.3.1 Linear Covariance Models

The origin of linear models for covariance matrices can be traced to the work of Yule (1927) and Gabriel (1962) and the implicit parameterization of a multivariate normal distribution in terms of entries of either Σ or its inverse. However, Dempster (1972) was the first to recognize the entries of $\Sigma^{-1} = (\sigma^{ij})$ as the canonical parameters of the exponential family of normal distributions. He proposed to select or estimate a covariance matrix efficiently and sparsely by identifying zeros in its inverse, and referred to these as *covariance selection* models which fit the framework of linear covariance models defined next.

Motivated by the simple and linear structure of covariance matrices of some time series and variance components, Anderson (1973) introduced the class of *linear covariance models* (LCM):

$$\Sigma^{\pm 1} = \alpha_1 U_1 + \cdots + \alpha_q U_q, \tag{20}$$

where U_i's are some known symmetric basis matrices and α_i's are unknown parameters, they must be restricted so that the matrix is positive-definite. It is usually assumed that there is a least a set of coefficients where the $\Sigma^{\pm 1}$ is positive-definite. The model (20) is general enough to include any covariance matrix. Indeed, for $q = p^2$ any covariance matrix admits

the representation:

$$\Sigma = (\sigma_{ij}) = \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma_{ij} U_{ij}, \tag{21}$$

where U_{ij} is a $p \times p$ matrix with one on the (i, j)th position and zero elsewhere.

Replacing Σ by S in the left-hand side of (20), it can be viewed as a collection of $\frac{p(p+1)}{2}$ linear regression models. The same regression models viewpoint holds with the precision matrix on the left-hand side. Similar to linear regression models, the class of linear covariance models is omnipresent when dealing with covariance matrices. It includes virtually any estimation method that acts elementwise on a covariance matrix such as tapering, banding, thresholding, covariance selections models, penalized likelihood with LASSO penalty on the off-diagonal entries of the precision matrix, etc. see (21).

A major drawback of (20)–(21) is the constraint on the coefficients which could make the estimation and other statistical problems difficult (Anderson, 1973; Szatfrowski, 1980). For a sample of size n from a normal distribution, the score equations for the maximum likelihood estimates can be solved by an iterative method; in each step a set of linear equations is solved (Anderson, 1973). He showed that if consistent estimates of $\alpha_1, \cdots, \alpha_q$ are used as initial values to obtain the coefficients of the linear equations, then its solution or the MLE is asymptotically efficient for n large. Szatrowski (1980) gives necessary and sufficient conditions for the existence of explicit maximum likelihood estimates, and the convergence of the iterative procedure, proposed by Anderson (1973), in one iteration from any positive-definite starting point. In fact, Szatrowski (1980) showed that, using a linear covariance model for Σ, the MLE of the coefficient vector has an explicit representation, i.e., is a vector of known linear combinations of elements of the sample covariance matrix, if and only if Σ^{-1} has the same LCM pattern. This happens, for example, when Σ has a compound symmetry (exchangeable) structure.

A good review of the MLE procedures for the model (20) and their applications to the problem of testing homogeneity of the covariance matrices of several dependent multivariate

normals are presented in Jiang, Sarkar and Hsuan (1999). They derive a likelihood ratio test, and show how to compute the MLE of Σ, in both the restricted (null) and unrestricted (alternative) parameter spaces using SAS PROC MIXED software. They also provide the code and the implementation is explained using several examples.

The notion of covariance regression introduced by (Hoff and Niu, 2009) is in the spirit of (20), but unlike the LCM the covariance matrix is quadratic in the covariates, and positive-definiteness is guaranteed through the special construction.

2.3.2 Log-Linear Covariance Models

In analogy with the use of log-linear models to handle variance heterogeneity in regression analysis where the variability depends on some predictors, a plausible way to remove the constraint on α_i's in (20) is to work with the logarithm of a covariance matrix. The key fact needed here is that for a general covariance matrix with the spectral decomposition $\Sigma = P\Lambda P'$, its *matricial logarithm* defined by $\log \Sigma = P \log \Lambda P'$ is simply a symmetric matrix with unconstrained entries taking values in $(-\infty, \infty)$.

This idea has been pursued by Leonard and Hsu (1992) and Chiu et al. (1996) who introduced the *log-linear covariance* models for Σ as

$$\log \Sigma = \alpha_1 U_1 + \cdots + \alpha_q U_q, \tag{22}$$

where U_i's are known matrices as before and the α_i's are now unconstrained. However, since $\log \Sigma$ is a highly nonlinear operation on Σ, see the example in Section 2.2.3, the α_i's lack statistical interpretation (Brown et al. 1994; Liechty et al. 2004). Fortunately, for Σ diagonal since $\log \Sigma = \text{diag}(\log \sigma_{11}, \ldots, \log \sigma_{pp})$, then (22) amounts to log-linear models for heterogeneous variances which has a long history in econometrics and other areas, see Carroll and Ruppert (1988) and references therein.

Maximum likelihood estimation procedures to estimate the parameters in (22) and their asymptotic properties are studied in Chiu et al. (1996) along with the analysis of two real

data sets. Given the flexibility of the log-linear models, one would expect them to be used widely in practice, however, a literature search shows this not to be case. An interesting application to spatial autoregressive (SAR) models and some of its computational advantages are discussed in LeSage and Pace (2007).

2.3.3 GLM via the Cholesky Decomposition

In this section, the constraint and lack of interpretation of α_i's in (20) and (22) are resolved simultaneously by relying on the Cholesky decomposition of a covariance matrix described in Section 2.2.4. A bona fide GLM for the precision matrix in terms of covariates is introduced and its maximum likelihood estimation (MLE) is discussed. An important consequence of the approach based on the modified Cholesky decomposition is that for any estimate of the Cholesky factors, the estimated precision matrix $\hat{\Sigma}^{-1} = \hat{T}'\hat{D}^{-2}\hat{T}$ is guaranteed to be positive-definite, see (16).

Recall that for an unstructured covariance matrix Σ, the nonredundant entries of its components $(T, \log D^2)$ in (16) obtained from Σ^{-1} are unconstrained. Thus, following the GLM's tradition one may write parameteric models for them using covariates (Pourahmadi, 1999; Pan and MacKenzie, 2003; Zimmerman, and Núñez-Antón, 2010). We consider the following parametric models for ϕ_{tj} and $\log \sigma_t^2$,for $t = 1, \ldots, p$; $j = 1, \ldots, t-1$:

$$\log \sigma_t^2 = z_t'\lambda, \ \phi_{tj} = z_{tj}'\gamma, \tag{23}$$

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 (Pourahmadi, 1999). The most common covariates used in the analysis of several real longitudinal data sets (Pourahmadi, 1999; Pourahmadi and Daniels, 2002; Pan and MacKenzie, 2003; Ye and Pan, 2006; Lin and Wang, 2009; Leng et al. 2010) are in terms of powers of times and lags :

$$z_t = (1, t, t^2, \cdots, t^{d-1})',$$

$$z_{tj} = (1, t-j, (t-j)^2, \cdots, (t-j)^{p-1})'.$$

A truly remarkable feature of (23) is its flexibility in reducing the potentially high-dimensional and constrained parameters of Σ or the precision matrix to $q+d$ unconstrained parameters λ and γ. Furthermore, one can rely on graphical tools such as the regressogram, a nonstationary analogue of the time series correlogram or AIC, to identify models such as (23) for the data; for more details on these, see Pourahmadi (1999, 2001) and Pan and MacKenzie (2003). Ye and Pan (2006) employ such parametrized models for covariance matrices in the context of the popular (Liang and Zeger, 1986) generalized estimating equations for longitudinal data.

To study the MLE of the parameters, recall that minus twice the loglikelihood function for a sample $Y_1, \cdots, Y_n$ from a normal population with mean zero and the common covariance Σ, except for a constant, is given by

$$\begin{aligned} -2l &= \sum_{i=1}^{n} (\log|\Sigma| + Y_i^{'}\Sigma^{-1}Y_i) \\ &= n\log|D^2| + ntr\Sigma^{-1}S \\ &= n\log|D^2| + ntrD^{-2}TST' \\ &= n\log|D^2| + ntrD^{-2}(I_p - B)S(I_p - B)', \end{aligned} \tag{24}$$

where $S = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i'$, $B = I_p - T$ and the last three equalities are obtained by replacing for Σ^{-1} from (16) and basic matrix operations involving trace of a matrix. Note that (24) is quadratic in B, thus for a given D^2 the MLE of B or ϕ_{tj}'s has a closed form, the same is true of the MLE of D^2 for a given B (Pourahmadi, 2000; Huang et al. 2006, 2007). This observation for computing the MLE of the saturated (unstructured) model for (T, D) is important when comparing the computational aspects of Cholesky-based estimation of the precision matrix with the Rocha et al.'s (2008) SPLICE algorithm in Section 3.4.

An algorithm for computing the MLE of the parameters (γ, λ) using the iterative Newton-Raphson algorithm with Fisher scoring is given in Pourahmadi (2000) along with the asymp-

totic properties of the estimators. An unexpected finding is the asymptotic orthogonality of the MLE of the parameters λ and γ, in the sense that their Fisher information matrix is block-diagonal, see Ye and Pan (2006), Pourahmadi (2007) and references therein. When the assumption of normality is questionable like when the data exhibit thick tails, then a multivariate t-distribution might be a reasonable alternative. Lin and Wang (2009) and Lin (2010) have extended the above theory of MLE to this situation.

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. The idea of smoothing along its subdiagonals is motivated by the similarity of the regressions in (17) to the varying-coefficients autoregressions (Kitagawa and Gersch, 1985; 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 nearly 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), for a related problem in multivariate time series see Dai and Guo (2004) and Stoffer and Rosen (2007). Huang, Liu and Liu (2007) have proposed a more direct and efficient approach using splines to smooth the subdiagonals of T. Leng et al. (2010) estimate a covariance matrix by writing linear models for T and semi-parametric models for D^2.

3 Regularization of the Sample Covariance Matrix

Knowing that the sample covariance matrix behaves poorly in high dimensions and is not even invertible when $n < p$, it is natural to look for alternative and improved estimators by regularizing the sample covariance matrix in various ways. In Sections 3.2 and 3.3, we review shrinkage estimators obtained by minimizing risk functions. A good example is the Stein's family of shrinkage estimators that shrinks the eigenvalues in the spectral decomposition. Regularization methods which act elementwise on the sample covariance matrix such as tapering, banding and thresholding are discussed in Section 3.5. Penalized normal likelihood estimators with LASSO penalty on the precision matrix are reviewed in Section 3.4 .

3.1 The Loss and Risk Functions

Regularized estimators are usually obtained by minimizing suitable risk/objective functions. The two common loss functions used when $n > p$ are

$$L_1(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\ \Sigma^{-1}) - \log|\hat{\Sigma}\Sigma^{-1}| - p,$$

$$L_2(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1} - I)^2,$$

where $\hat{\Sigma} = \hat{\Sigma}(S)$ is an arbitrary estimator, with the corresponding risk functions:

$$R_i(\hat{\Sigma}, \Sigma) = E_\Sigma L_i(\hat{\Sigma}, \Sigma), i = 1, 2.$$

An estimator $\hat{\Sigma}$ is considered better than S if its risk function is smaller than that of S. The loss function L_1 was advocated by Stein (1956) and is usually called the entropy loss or the Kullback-Liebler divergence of two multivariate normal densities corresponding to the two covariance matrices. The second, called a quadratic loss function is essentially the Euclidean or the Frobenius norm of its matrix argument which involves squaring the difference between aspects of the estimator and the target. Consequently, it penalizes overestimates more than underestimates, and "smaller" estimates are more favored under L_2 than under L_1. For example, among all scalar multiples aS, $a > 0$, it is known (Haff, 1980) that S is optimal under L_1, while the smaller estimator $\frac{nS}{n+p+1}$ is optimal under L_2.

Following the lead of Muirhead and Leung (1987), Ledoit and Wolf (2004) have used a slight modification of the Frobenius norm as the loss function:

$$L_3(\hat{\Sigma}, \Sigma) = p^{-1}\|\hat{\Sigma} - \Sigma\|^2 = p^{-1}tr(\hat{\Sigma} - \Sigma)^2,$$

with the corresponding risk function. Note that though dividing by the dimension p is not standard, it has the advantage that norm of the identity matrix is simply one, regardless of the size of p. Also, the loss L_3 does not involve matrix inversion which is ideal with regard to computational cost for the "small n, large p" case. The heuristics behind this loss function are the same those for L_2, it has the additional and attractive feature that the optimal covariance estimator under L_3 turns out to be the penalized normal likelihood estimator with $tr\Sigma^{-1}$ as the penaly (Warton, 2008; Yuan and Huang, 2009). Since the penalty function becomes large when Σ gets closer to singularity, such a penalty forces the covariance estimators to be nonsingular and well-conditioned.

3.2 Shrinking The Spectrum and The Correlation Matrix

In this section we prsent one of the earliest improvements of S obtained by shrinking only its eigenvalues. Having observed that the sample covariance matrix systematically distorts the eigenstructure of Σ, particularly when $\frac{p}{n}$ is large, Stein (1956, 1975) initiated the task of improving it. He considered orthogonally invariant estimators of the form

$$\hat{\Sigma} = \hat{\Sigma}(S) = P\Phi(\lambda)P',$$

where $\lambda = (\lambda_1, \cdots, \lambda_p)', \lambda_1 > \cdots > \lambda_p > 0$ are the ordered eigenvalues of S, and P is the orthogonal matrix whose jth column is the normalized eigenvector of S corresponding to λ_j, and $\Phi(\lambda) = diag(\varphi_1, \cdots, \varphi_p)$ is a diagonal matrix where $\varphi_j = \varphi_j(\lambda)$ estimates the jth largest eigenvalue of Σ. For example, the choice of $\varphi_j = \lambda_j$ corresponds to the usual unbiased estimator S, where it is known that λ_1 and λ_p have upward and downward biases, respectively. Stein's method chooses $\Phi(\lambda)$ so as to counteract the biases of the eigenvalues

of S by shrinking them toward some central values. For the L_1 risk, his modified estimators of the eigenvalues of Σ are $\varphi_j = \frac{n\lambda_j}{\alpha_j}$, where

$$\alpha_j = \alpha_j(\lambda) = n - p + 1 + 2\ \lambda_j \sum_{i \neq j} \frac{1}{\lambda_j - \lambda_i}.$$

Note that the φ_j's will differ the most from λ_j when some or all of the λ_j's are nearly equal and $\frac{n}{p}$ is not small. Since some of the φ_j's could be negative and may not even satisfy the order restriction, Stein has suggested an isotonizing procedure to obtain modified estimators satisfying the above constraints, for more details on this procedure see Lin and Perlman (1985).

Simulation studies in Lin and Perlman (1985) show that the above shrinkage estimator has significant improvement in risk over the sample covariance matrix, and performs the best when the eigenvalues of the population covariance matrix are nearly equal or form clusters within each of which the eigenvalues are nearly equal, see Ledoit and Wolf (2004) for more simulation-based comparisons of a few other alternatives to the sample covariance.

Lin and Perlman (1985) have applied the James-Stein shrinkage estimators to the sample correlation in order to improve the sample covariance matrix for large p. They shrink the Fisher z-transform of the individual correlation coefficients (and the logarithm of the variances) toward a common target value. Their simulation study shows that, the greatest potential improvement can be expected when the correlations and/or the the standard deviations are nearly equal, or when these can be partitioned into clusters within each of which the values are nearly the same. Liechty et al. (2004) provide a Bayesian method for correlation estimation that exploits the prior knowledge that one may have on the clustering of the correlations, see also Daniels and Kass (1999) on using hierarchical priors on the Fisher z-transform of the entries of a correlation matrix.

Stein (1956), James and Stein (1961) have considered shrinking the Cholesky factor of the sample covariance matrix: $S = CC'$, where C is a lower triangular matrix with positive diagonal entries, see Anderson (2003, Sec. 7.8) for more discussions and details.

3.3 Ledoit-Wolf Shrinkage Estimator

As in regression analysis to ensure nonsingularity of the estimated covariance matrix in the "n small, p large" case the idea of ridge regularization (Hoerl and Kennard, 1970) seems promising. Using this idea, Ledoit and Wolf (2004) present a shrinkage estimator that is asymptotically the optimal convex linear combination of the sample covariance matrix and the identity matrix with respect to L_3.

Alternatively, one can motivate such a ridge regularization by recalling that the sample covariance matrix S is unbiased for Σ, but unstable with considerable risk when $n < p$, and a structured covariance matrix estimator, like the identity matrix, has very little estimation error, but can be severely biased when the structure is misspecified. A natural compromise between these two extremes is a linear combination of them, giving a simple shrinkage or ridge candidate of the form

$$\hat{\Sigma} = \alpha_1 I + \alpha_2 S.$$

Now, one may choose α_1 and α_2 to optimize certain criterion (Ledoit and Wolf, 2004; Warton, 2008).

Using the Frobenius norm or minimizing the risk corresponding to the loss function L_3, Ledoit and Wolf (2004) showed that the optimal choices of α_1 and α_2 depend only on the following four scalar functions of the true (but unknown) covariance matrix Σ :

$$\mu =< \Sigma, I >, \alpha^2 = \|\Sigma - \mu I\|^2, \beta^2 = E\|S - \Sigma\|^2, \delta^2 = E\|S - \mu I\|^2,$$

where $< A, B >= p^{-1} tr(AB')$. Consistent estimators for these scalars are provided by Ledoit and Wolf (2004), so that substitution in $\hat{\Sigma}$ results in a positive-definite estimator of Σ. Through extensive simulation studies they establish the superiority of this estimator to the sample covariance matrix and the empirical Bayes estimator (Haff, 1980) among others. Applications of this procedure to the estimation of the spectral density matrix of multivariate stationary time series are discussed in Böhm and von Sachs (2008).

Warton (2008) taking $\alpha_2 = 1$, showed that such ridge estimators can be obtained using the penalized normal likelihood where the penalty term is proportional to $tr\Sigma^{-1}$. Evidently, such a penalty ensures that the estimator is a nonsingular matrix. He suggests using the cross-validation of the likelihood function for estimation of the ridge and the penalty parameters, and extends the approach to the ridge estimation of the correlation matrix. His method of estimation leads to the definition of a suitable test statistics for the parameters in multivariate linear regression in high-dimensional situations. The power properties of the test statistic are studied and compared with the principal components and generalized inverse test statistics used in dealing with high dimensionality.

3.4 The Penalized Likelihood Approach

In this section, we review various methods for solving Dempster's (1972) covariance selection problem, that is, inducing sparsity in the precision matrix, which is of great interest in the literature of Gaussian graphical models (Tibshirani et al. 2009, Chap. 17).

Motivated by the success of the LASSO estimators in the context of linear regression with a large number of covariates (Tibshirani, 1996), and in view of (20) and (21) it is plausible to induce sparsity in the precision matrix estimation by adding to the normal loglikelihood (24) a penalty on the off-diagonal entries of the precision matrix Σ^{-1} or its Cholesky factor (Huang et al. 2006) :

$$-2l + \sum_{i<j} p_{\lambda_{ij}}(\sigma^{ij}), \tag{25}$$

where σ^{ij} is the (i,j)th entry of the precision matrix and λ_{ij} is the corresponding tuning parameter. Note that the LASSO and ridge penalties correspond to choosing $p_\lambda(|x|) = \lambda|x|^p$ for $p = 1, 2$, respectively. Such an approach will inherit many desirable computational and statistical properties of LASSO and its many improved variants, provided that a regression-based interpretation can be found for the entries of the precision matrix or its factors. Of particular interests are LASSO's abilities to select models and estimate parameters si-

multaneously, and the recent improved computational algorithms for LASSO such as the homotopy/LARS–LASSO (Efron et al. 2004; Rocha et al. 2008), see Fan and Lv (2010, Sec. 3.5) for other improved algorithms.

Some early attempts at inducing sparsity in the precision matrix are, Bilmes (2000), Smith and Kohn (2002), Wu and Pourahmadi (2003) and Levina et al.(2007) who, for a fixed order of the variables in Y, use a parametrization of the precision matrix in terms of the modified Cholesky decomposition (16) . Covariance selection priors and AIC were used to promote sparsity in T. Huang et al. (2006) propose a covariance selection estimate by adding to the normal loglikelihood the LASSO penalty on the off-diagonal entries of T, cross-validation is used to select a common regularization parameter. Using a local quadratic approximation to the penalty, it was shown that the method is computationally tractable, see also Huang et al. (2007) and Levina et al. (2007) for some improvents. Bickel and Levina (2008a) provide conditions ensuring consistency in the operator norm (spectral norm) for precision matrix estimates based on banded Cholesky factors.

Chang and Tsay (2010) extend the Huang et al. (2006) setup using an equi-angular penalty which imposes different penalty on each regression or row of T, where the penalties are inversely proportional to the prediction variance σ_t^2 of the tth regression. They use extensive simulations to compare the performance of their method with others including the sample covariance matrix, the banding (Bickel and Levina, 2008a), and the L_1-penalized normal loglikelihood (Huang et al. 2006). Contrary to the banding method, the method of Huang et al. and the equi-angular method work reasonably well for six covariance matrices with the equi-angular method outperforming the others. Since the modified Cholesky decomposition is not permutation-invariant, they also use a random permutation of the variables before estimation to study the sensitivity to permutation of each method. They conclude that permuting the variables introduces some difficulties for each estimation method, except the sample covariance matrix, but the equi-angular method remains the best with the band-

ing method having the worst sensitivity to permutation. They also compare these methods by applying them to a portfolio selection problem with $p = 80$ series of actual daily stock returns. In the context of space-time data, Zhu and Liu (2009) rely on on the Cholesky decomposition of the precision matrix, based on several ordering schemes using the spatial locations of the observations. Their theoretical and simulation studies show that the regression-based penalized normal likelihood method performs competitively.

Two disadvantage of imposing the sparsity on the factor T are that its sparsity does not necessarily imply sparsity of the precision matrix; and the sparsity structure in T could be sensitive to the order of the random variables within Y. Some alternative methods which tackle these issues penalize the precision matrix directly . For example, d'Aspremont et al. (2008), Yuan and Lin (2007), and Friedman et al. (2008) consider an estimate defined by the normal loglikelihood penalized by the L_1-norm of the entries of Σ^{-1}. These methods produce sparse, permutation-invariant estimators of the precision matrix, though some are computationally expensive. Yuan and Lin (2007) use the max-det algorithm to compute the estimator while imposing the positive-definiteness constraint, this seems to have limited their numerical results to $p \leq 10$ (Rothman et al. 2008, p. 496). A faster semi-definite programming algorithm based on Nesterov's method for interior point optimization was used by d'Aspremont et al. (2008). Rothman et al. (2008) analyze the properties of the solution of the same problem using the Cholesky decomposition to avoid the computational cost of semi-definite programming. To date, the fastest available algorithm is the graphical lasso (glasso), proposed by Friedman et al. (2008) which relies on the equivalence of the d'Aspremont et al. (2008) blockwise interior point procedure and recursively solving and updating a lasso least-squares regression problem using the coordinate descent algorithm for LASSO.

The sparse pseudo-likelihood inverse covariance estimation (SPLICE) algorithm of Rocha et al. (2008) and the SPACE (Sparse PArtial Correlation Estimation) algorithm of Peng et

al. (2009) also impose sparsity constraints directly on the precision matrix, but with slightly different regression-based reparameterizations of Σ^{-1} , see (8) and (10). They are designed to improve several shortcomings of the approach of Meinshausen and Bühlmann (2006) including its lack of symmetry for neighborhood selection in Gaussian graphical models. While Meinshausen and Bühlmann (2006) use p separate linear regressions to estimate the neighborhood of one node at a time, Rocha et al. and Peng et al. propose merging all p linear regressions into a single least squares problem where the observations associated to each regression are weighted differently according to their conditional variances.

To appreciate the need for using approximate or pseudo-likelihood it is instructive to note that unlike the sequence of prediction errors in (17), the $\tilde{\varepsilon}_j$'s from Section 2.2.2 are correlated so that $\tilde{D}^2$ is not really the covariance matrix of the vector of regression errors $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \cdots, \tilde{\varepsilon}_p)'$. The use of its true and full covariance matrix in the normal loglikelihood would increase the computational cost at the estimation stage. This problem is circumvented in Rocha et al. (2008) and Friedman et al. (2010) by using a pseudo-likelihood function which in the normal case amounts to pretending that the $Cov(\tilde{\varepsilon})$ is $\tilde{D}^2$. To this pseudo-loglikelihood function, they add the symmetry constraints (9) and a weighted LASSO penalty on the off-diagonal entries to promote sparsity. A drawback of the SPLICE and SPACE algorithms is that they do not enforce the positive-definiteness constraint, hence the resulting covariance estimator are not guaranteed to be positive-definite.

The *sparsistency* and rates of convergence for sparse covariance and precision matrix estimation using the penalized likelihood with nonconvex penalty functions have been studied in Lam and Fan (2009). By sparsistency they refer to the property that all zero entries are actually estimated as zero with probability tending to one. In a given situation, sparsity might be present in the covariance matrix, its inverse or Cholesky fator. They develop a unified framework to study these three sparsity problems with a general penalty function and show that the rates of convergence for these problems under the Frobenius norm are

of order $(\frac{s \log p}{n})^{1/2}$, where $s = s_n$ is the number of nonzero elements, $p = p_n$ is the size of the covariance matrix and n is the sample size. This reveals that the contribution of high-dimensionality is merely of a logarithmic factor.

3.5 Elementwise Shrinkage

In this section, we review a few alternative estimators like *banding*, *tapering* and *thresholding* which are based on the elementwise shrinkage of the sample covariance matrix. These covariance estimators require minimal amount of computation, except in the cross-validation for selecting the tuning parameter which is computationally comparable to that for the penalized likelihood method. However, due to their emphasis on elementwise transformations such estimators are not guaranteed to be positive-definite.

3.5.1 Banding the Sample Covariance Matrix

Many entries of the sample covariance matrix $S = (s_{ij})$ could be small or unstable in the "n small, p large" case, where, in fact, one is estimating far too many parameters based on a few observations. The most extreme case of this occurs in time series analysis where one has to work with only a single (long) realization ($n = 1$). Here the requirement of stationarity helps to reduce the number of distinct entries of the $p \times p$ covariance matrix Σ from $p(p+1)/2$ to just p, which is still large. The moving average (MA) and autoregressive (AR) models in time series analysis which further reduce the number of parameters, are the prototypes of banding a covariance matrix or its inverse, i.e. replacing all entries outside a band around the main diagonal by zeros (Wu and Pourahmadi, 2009; Bickel and Gel, 2010; McMurry and Politis, 2010).

Given the sample covariance matrix $S = (s_{ij})$ and any integer k, $0 \leq k < p$, its k-banded (Bickel and Levina, 2008a) version defined by

$$B_k(S) = [s_{ij}\mathbf{1}(|i-j| \leq k)],$$

can serve as an estimator for Σ. This kind of regularization is ideal when the indices have been arranged so that

$$|i-j| > k \Longrightarrow \sigma_{ij} = 0.$$

This occurs, for example, if Σ is the covariance matrix of $Y = (y_1, \cdots, y_p)'$ where $y_1, y_2, \cdots$, is a finite inhomogenous moving average process

$$y_t = \sum_{j=1}^{k} \theta_{t,t-j}\varepsilon_j,$$

and ε_j's are i.i.d. with mean 0 and finite variances.

Note that a k-banded matrix $B_k(S)$ is not necessarily positive-definite. Tapering the covariance matrix is frequently used in time series and spatial models, it has been used recently to improve the performance of covariance matrix estimates used by classifiers based on linear discriminant analysis (Bickel and Levina, 2004) and in Kalman filter ensembles (Furrer and Bengtsson, 2007). Banding is a special case of tapering, that is, replacing S by $S * R$, where $(*)$ denotes the Schur (coordinate-wise) matrix multiplication and $R = (r_{ij})$ is a positive-definite symmetric matrix (Furrer and Bengtsson, 2007). It is known that the Schur product of two positive-definite matrices is also positive-definite. Note that banding corresponds to using $R = r_{ij} = (\mathbf{1}(|i-j| \leq k)$ which is not a positive definite matrix. The idea of banding has also been used on the lower triangular matrix of the Cholesky decomposition of Σ^{-1} by Wu and Pourahmadi (2003) and Huang et al.(2006), Bickel and Levina (2008a). While Furrer and Bengtsson (2007) have used tapering as a regularization technique for the ensemble Kalman filter, Kaufman, Schervish and Nychka (2008) use it for purely computational purposes in the likelihood-based estimation of the parameters of a structured covariance function for large spatial data sets.

Asymptotic analysis of banding is possible when n, p and k are large. Bickel and Levina (2008a, Theorems 1 and 2) have shown that, for normal data the banded estimator is consistent in the operator norm (spectral norm), uniformly over a class of approximately "bandable" matrices, as long as $\frac{\log p}{n} \to 0$. They obtain explicit rate of convergence which

depends on how fast $k \to \infty$, see also Cai et al. (2010). The consistency in operator norm guarantees the consistency of principal component analysis (Johnstone and Lu, 2009) and other related methods in multivariate statistics when n is small and p is large. Cai et al. (2010) propose a tapering procedure for the covariance matrix estimation and derive the optimal rate of convergence for estimation under the operator norm. They also carry out a simulation study to compare the finite sample performance of their proposed estimator with that of the banding estimator introduced in Bickel and Levina (2008a). The simulation shows that their proposed estimator has good numerical performance, and nearly uniformly outperforms the banding estimator.

3.5.2 Thresholding the Sample Covariance Matrix

When both n and p are large, it is plausible that many elements of the population covariance matrix are equal to 0, and hence Σ is sparse. How does one develop an estimator other than S to cope with this situation? The concept of thresholding originally developed in nonparametric function estimation has been used in the estimation of large covariance matrices by Bickel and Levina (2008b), El Karoui (2008 a, b) and Rothman et al. (2009).

For a sample covariance matrix $S = (s_{ij})$ the thresholding operator T_s for $s \geq 0$ is defined by

$$T_s(S) = [s_{ij}\mathbf{1}(|s_{ij}| \geq s)]\,,$$

so that thresholding S at s amounts to replacing by zero all entries with absolute value less than s. Its biggest advantage is its simplicity as it carries no major computational burden compared to its competitors like the penalized likelihood with the LASSO penalty (Huang et al. 2006; Rothman et al. 2008; Rocha et al. 2008). A potential disadvantage is the loss of positive-definiteness as in banding. However, just as in banding, Bickel and Levina (2008b) have established the consistency of the threshold estimator in the operator norm, uniformly over the class of matrices that satisfy a notion of sparsity, provided that $\frac{\log p}{n} \to 0$, with an explicit rate of convergence. An immediate consequence of the consistency result is that a

threshold estimator will be positive definite with probability tending to one for large samples and dimensions.

4 Bayesian Modeling of Covariances

Traditionally, in Bayesian approaches to inference for Σ the Jefferys' improper prior and the conjugate inverse Wishart (IW) priors are used. For some reviews of the earlier work in this direction, see Lin and Perlman (1985) and Brown et al. (1994). However, the success of Bayesian computation and Markov Chain Monte Carlo (MCMC) in the late 1980's did open up the possibility of using more flexible and elaborate non-conjugate priors for covariance matrices, see Yang and Berger (1994), Daniels and Kass (1999), Wong et al. (2003), Liechty, Liechty and Müller (2004) and Hoff (2009). Some of these priors were constructed and inspired by certain useful and desirable features of IW, such as the generalized inverse Wishart (GIW) and marginally uniform priors introduced by Brown et al. (1994), Daniels and Pourahmadi (2002), Pourahmadi and Daniels (2002), Smith and Kohn (2002), Barnard et al. (2000), Wong et al. (2003) and Liechty et al. (2004), which rely on the Cholesky and the variance-correlation decompositions, respectively. We present a very brief review of the progress in Bayesian covariance estimation in a somewhat chronological order starting with priors put on the components of the spectral decomposition.

4.1 Priors on the Spectral Decomposition

Starting with the remarkable work of Stein (1956, 1975) efforts to improve estimation of a covariance matrix, have been confined mostly to shrinking the eigenvalues of the sample covariance matrix toward a common value (Dey and Srinivasan, 1985; Lin and Perlman, 1985; Haff 1991; Yang and Berger 1994; Daniels and Kass, 1999; Hoff, 2009). Such covariance estimators have been shown to have lower risk than the sample covariance matrix. Intuitively, shrinking the eigenvectors is expected to further improve or reduce the estimation

risk (Daniels and Kass, 1999, 2001; Johnstone and Lu, 2009).

There are three broad classes of priors that are based on unconstrained parameterizations of a covariance matrix using its spectral decomposition. These have the goal of shrinking some functions of the off-diagonal entries of Σ or the corresponding correlation matrix toward a common value like zero. Consequently, estimation of the $\frac{p(p-1)}{2}$ dependence parameters is reduced to that of estimating a few parameters.

Perhaps, the first breakthrough with the GLM principles in mind is the log matrix prior due to Leonard and Hsu (1992) which is based on the matricial logarithm defined in Section 2.2.3. Thus, formally a multivariate normal prior with a large number of hyperparameters is introduced. They show the flexibility of this class of priors for the covariance matrix of a multivariate normal distribution, yielding much more general hierarchical and empirical Bayes smoothing and inference, when compared with a conjugate analysis involving an IW prior. The prior is not conditionally conjugate, and according to Brown et al. (1994), its major drawback is the lack of statistical interpretability of the entries of $\log \Sigma$ and their complicated relations to those of Σ as seen in Section 2.2.3. Consequently, prior elicitation from experts and substantive knowledge cannot be used effectively in arriving at priors for the entries of $\log \Sigma$ and their hyperparameters, see Lietchy et al. (2004, p. 2) for a discussion on the lack of intuition and relationship between log-eigenvalues and correlations.

The reference (noninformative) prior for a covariance matrix in Yang and Berger (1994) is of the form,

$$p(\Sigma) = c[|\Sigma| \prod_{i<j} (\lambda_i - \lambda_j)]^{-1},$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ are the ordered eigenvalues of Σ and c is a constant. Yang and Berger(1994, p. 1199) note that compared to the Jeffreys prior, the reference prior puts considerably more mass near the region of equality of the eigenvalues. Therefore, it is intuitively plausible that the reference prior would produce a covariance estimator with better eigenstructure shrinkage. Furthermore, they point out that the reference priors for

Σ^{-1}, and the eigenvalues of the covariance matrix are the same as $p(\Sigma)$. Expression for the Bayes estimator of the covariance matrix using this prior involve computation of high-dimensional posterior expectations, the computation is done using the hit-and-run sampler in a Markov chain Monte Carlo setup. An alternative noninformative reference prior for Σ (and the precision matrix) which allows for closed-form posterior estimation is given in Rajaratnam et al. (2008).

It is known (Daniels, 2005) that the Yang and Berger's (1994) reference prior implies a uniform prior on the orthogonal matrix P and flat improper priors on the logoarithm of the eigenvalues of the covariance matrix. The shrinkage priors of Daniels and Kass (1999) also rely on the spectral decomposition of the covariance matrix, but are designed to shrink the eigenvectors by reparametrizing the orthogonal matrix in terms of $\frac{p(p-1)}{2}$ Givens angles (Golub and Van Loan, 1989) θ between pairs of the columns of the orthogonal matrix P. Since θ is restricted to lie in the interval $(-\pi/2, \pi/2)$, a logit transform will make it unconstrained so as to conform to the GLM principles. They put a a mean-zero normal prior on the logit tranformation of the Givens angles. The statistical relevance and interpretation of the Givens angles are not well-understood at this time. The local parametrization of orthogonal matrices in Boik (2002) could shed some light on the problem of interpretation of the new parameters. The idea of introducing matrix Bingham distributions as priors on the group of orthogonal matrices (Hoff, 2009) could also be useful in shrinking the eigenvectors of the sample covariance matrix.

Using simulation experiments, Yang and Berger (1994) compared the performance of their reference prior Bayes covariance estimator to the covariance estimators of Stein (1975) and Haff (1991) and found it to be quite competitive based on the risks corresponding to the loss functions $L_i, i = 1, 2$. Daniels and Kass (1999) also using simulations compared the performance of their shrinkage estimator to several other Bayes estimators of covariance matrices, using only the risk corresponding to the L_1 loss function. It turns out that the

Bayes estimators from the Yang and Berger's (1994) reference prior does as well as those from Givens-angle prior for some nondiagonal and ill-conditioned matrices, but suffers when the true matrix is diagonal and poorly conditioned.

4.2 The Generalized Inverse Wishart Priors

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); Liu (1993). Though the ensuing parameters have nice statistical interpretation, they are not permutation-invariant. It is shown by Brown et al. (1994) that a regression dissection of the inverse Wishart (IW) distribution reveals some of its noteworthy features making it possible to define flexible generalized inverted Wishart (GIW) priors for general covariance matrices.

These priors are constructed by first partitioning a multivariate normal random vector Y with mean zero and covariance matrix Σ, into $k \leq p$ subvectors: $Y = (Z_1, \cdots, Z_k)'$, and writing its joint density as the product of a sequence of conditionals:

$$f(y) = f(z_1)f(z_2|z_1)\cdots f(z_k|z_{k-1}, \cdots, z_1).$$

Now, in each conditional distribution one places normal prior distributions on the regression coefficients and inverse Wishart on the prediction variances. The hyperparameters can be structured so as to maintain the conjugacey of the resulting priors. It is known (Daniels and Pourahmadi, 2002; Rajaratnam, Massam and Carvalho, 2008) that such priors offer considerable flexibility as there are many parameters to control the variability in contrast to the one parameter for IW.

These ideas and techniques have been further refined in Gaithwaite and Al-Awadi (2001) in prior distribution elicitation from experts, and extended to longitudinal and panel data setup in Daniels and Pourahmadi (2002) and Smith and Kohn (2002) . The GIW prior was further refined in Daniels and Pourahmadi (2002) using the finest partition of Y, i.e. using

$k = p$. In this case all restrictions on the hyperparameters are removed from the normal and inverse Wishart (gamma) distributions and the prior remains conditionally conjugate, in the sense that the full-conditional of the regression coefficients is normal given the prediction variances, and the full-conditional of prediction variances is inverse gamma given the regression coefficients. For a review of certain advantages of this approach in the context of longitudinal data and some examples of analysis of such data, see Daniels (2005) and Daniels and Hogan (2008).

4.3 Priors on Correlation Matrices

One of the first use of variance-correlation decomposition in Bayesian covariance estimation seems to be due to Barnard et al. (2000), who using $p(\Sigma) = p(D, R) = p(D)p(R|D)$ introduced independent priors for the standard deviations in D and the correlations in R. Specifically, they used log normal priors on variances independently of a prior on the whole matrix R capable of inducing uniform $(-1, 1)$ priors on its entries ρ_{ij}, see Liu and Daniels (2006).

This is done by first deriving the marginal distribution of R when Σ has a standard IW distribution, $W_p^{-1}(I, \nu), \nu \geq p$, with the density

$$f_p(\Sigma|\nu) = c|\Sigma|^{-\frac{1}{2}(\nu+p+1)} \exp\left(-\frac{1}{2} tr\Sigma^{-1}\right).$$

It turns out that

$$f_p(R|\nu) = c|R|^{\frac{1}{2}(\nu-1)(p-1)-1} \prod_{i=1}^{p} |R_{ii}|^{-\nu/2},$$

when R_{ii} is the principal submatrix of R, obtained by deleting its ith row and column. Then, using the marginalization property of the IW (i.e. a principal submatrix of an IW is still an IW), the marginal distribution of each $\rho_{ij}(i \neq j)$ is obtained as

$$f(\rho_{ij}|\nu) = c(1 - \rho_{ij}^2)^{\frac{\nu-p-1}{2}}, \ |\rho_{ij}| \leq 1.$$

The latter can be viewed as a Beta $\left(\frac{\nu-p+1}{2}, \frac{\nu-p+1}{2}\right)$ on $(-1, 1)$, which is uniform when $\nu = p + 1$. Moreover, by choosing $p \leq \nu < p + 1$ or $\nu > p + 1$, one can control the

tail of $f(\rho_{ij}|\nu)$, i.e. making it heavier or lighter than the uniform. Thus, the above family of priors for R is indexed by a single "tuning" parameter ν.

Liechty et al. (2004) note that few existing probability models and parameterizations for covariance matrices allow for easy interpretation and prior elicitation. They propose priors in which correlations are grouped based on similarities among the correlations or based on groups of variables. A good example of this situation is in financial time series where it is often known that returns of stocks in the same industries are more closely related than others. Alternatively, one could rely on the idea of reparametrizing the correlation matrix R using the partial autocorrelations, discussed next, and then introduce various independent beta priors on the latter.

4.4 Reparameterization via Partial Autocorrelations

In this section, we present yet another unconstrained and statistically interpretable reparameterization of Σ, but now using the notion of partial autocorrelation function (PACF) from time series analysis (Box et al. 1994; Pourahmadi, 2001, Chap. 7). As expected, this approach just like the Cholesky decomposition requires an *a priori* order among the random variables in Y. It is motivated by and tries to mimic the phenomenal success of the PACF of a stationary time series in model formulation (Box et al. 1994) and removing the positive-definiteness constraint on the autocorrelation function (Ramsey, 1974). We note that reparameterizing the stationarity-invertibility domain of ARMA models by Jones (1980) had a profound impact on algorithms for computing the MLE of the ARMA coefficients and guaranteeing that the estimates are in the feasible region.

Starting with the variance-correlation decomposition of a general $\Sigma = DRD$, we focus on reparameterizing its correlation matrix $R = (\rho_{ij})$ in terms of entries of a simpler symmetric matrix $\Pi = (\pi_{ij})$ where $\pi_{ii} = 1$ and for $i < j, \pi_{ij}$ is the *partial autocorrelation* between y_i and y_j adjusted for the *intervening* (not the remaining) variables. More precisely, $\pi_{i,i+1} = \rho_{i,i+1}, i = 1, \cdots, p-1$ are the lag-1 correlations and for $j - i \geq 2, \pi_{ij} = \rho_{ij|i+1,\cdots,j-1}$ in

the notation of Anderson (2003, p.41). Note that unlike R, and the matrix of full partial correlations (ρ^{ij}) constructed from Σ^{-1}, Π has a simpler structure in that its entries are free to vary in the interval $(-1, 1)$. If needed, using the Fisher z-transform Π can be mapped to the matrix $\widetilde{\Pi}$ where its off-diagonal entries take values in the entire real line $(-\infty, +\infty)$. Of course, the process of going from a constrained R to Π or $\widetilde{\Pi}$ is reminiscent of finding a link function in the theory of GLM (McCullagh and Nelder, 1989).

Compared to the long history of using the PACF in time series analysis (Quenouille, 1949), research on establishing a one-to-one correspondence between a general covariance matrix and (D, Π) has a rather short history. An early work in the Bayesian context is due to Eaves and Chang (1992), followed by Zimmerman (2000) and Pourahmadi (1999, 2001, p.102) for longitudinal data, Dégerine and Lambert-Lacroix (2003) for the nonstationary time series, and Kurowicka and Cooke (2003) and Joe (2006) for a general random vector. The fundamental determinantal identity:

$$|\Sigma| = \left(\prod_{i=1}^{p} \sigma_{ii}\right) \prod_{i=2}^{p} \prod_{j=1}^{i-j} \left(1 - \pi_{ij}^2\right), \tag{26}$$

has been redicovered recently by Dégerine and Lambert-Lacroix (2003), Kurowicka and Cooke (2003) and Joe (2006), but its origin can be traced to a notable and somewhat neglected paper of Yule (1907, equ. 25).

The identity (26) plays a central role in Joe's (2006) method of generating random correlation matrices whose distributions are *independent of the order of variables* in Y, and in Daniels and Pourahmadi's (2009) introduction of priors for the Bayesian analysis of correlation matrices. These papers employ independent linearly transformed Beta priors on $(-1, 1)$ for the partial autocorrelations π_{ij}. However, Jones (1987) seems to be the first to use such Beta priors in simulating data from "'typical"' ARMA models.

5 Conclusions and Future Research

We have reviewed progress in covariance estimation for low- and high-dimensional data sets, from the narrow perspectives of the GLM and regularization or parsimony and sparsity. Recent appearance of many regression-based techniques and the use of LASSO-type penalties show that covariance estimation can benefit greatly from mimicking/using the conceptual and computational tools of regression analysis, and the GLM's insistence on using unconstrained parameters. Fortunately, mostly due to the computational-algorithmic advances centered around LASSO, the high-dimensionality challenge in covariance estimation has been virtually eliminated, however, the positive-definiteness challenge still remains. Its removal could not only further reduce the computational cost due to high-dimensionality, but is also crucial for parsimony and writing simple, interpretable models using covariates. Among the three matrix decompositions, the spectral and Cholesky decompositions are the most helpful in removing the positive-definiteness constraint. These along with some recent covariance estimation algorithms enforcing the positive-definiteness suggest that there are trade-offs among the requirements of unconstrained parameterization, statistical interpretability, and the computational cost.

In summary, the key problem of removing the positive-definiteness constraint which is central to developing a GLM setup for covariance modeling remains open, in the sense that, as yet, no **unconstrained and statistically interpretable** reparameterization exists for a general covariance matrix without imposing an order on the variables. In addition to this, the following topics/problems related to covariance estimation are worthy of further study:

- **Positive-Definiteness Plus Aother Constraint:** There are several situations in statistics where it is desirable to have sparse or parsimonious estimators of certain structured covariance matrices. For example:

    1. Some off-diagonal entries are zero, as in graphical models with known structures,

2. Correlation matrices, where the diagonal entries are constant and equal to one,

3. Stationary (Toeplitz) covariance matrices, where all (sub-) diagonals are constant.

An iterative conditional fitting procedure for solving the first problem is proposed by Chaudhuri et al. (2007). Friedman et al. (2009, Chap.17) provide a regression-based approach very much similar to the graphical LASSO. A permutation-invariant solution to the second problem is given in Rothman et al. (2008). Assuming an order on the variables, a general correlation matrix can be reparameterized using the PACF as in Section 4.4, where in principle covariates can be used to to write GLM for correlation matrices. Some preliminary results are given in Daniels and Pourahmadi (2009). The PACF also works for stationary covariance matrices (Jones, 1980), so long as an order is imposed on the variables. The problem is open for unordered variables.

- **Ordering The Variables:** Variables in many application areas do not have a natural order as in time series or longitudinal studies. A method for discovering meaningful orderings among variables based on their correlations using the Isomap is proposed in Wagaman and Levina (2009). If the variables can be rearranged so that the resulting covariance matrix is approximately "bandable", then this order can be used to proceed with the Cholesky decomposition, construct a sparse covariance estimator that is block-diagonal and/or banded. In the context of spatial data analysis, Zhu and Liu (2009) also propose an ordering strategy that uses the location information to minimize the bandwidth of the Cholesky factor of the precision matrix. The problem of finding good permutations of a sparse symmetric matrix to induce extra nonzero elements in its Cholesky factor has been studied extensively in the numerical analysis literature for the purpose of minimizing storage and speeding up computation; see Section 3 in Zhu and Liu for an extended review of this literature. The need for ordering a set of variables or parameters is also present in other areas of statistics, for example, in the construction of reference priors (Berger and Yang, 1994; Brown et al. 1994;

Liechty et al. 2004) and spatial data where the idea of ordering groups of variables or parameters (regression dissection) is advocated. In the literature of statistics, this problem has not received the rigorous and systematic attention that it deserves. To achieve parsimony in covariance estimation, how can one order the variables so that the ensuing covariance matrix can be modeled, say, as in (23) using a minimal number of parameters/covariates? How sensitive is a particular inference to the choice of an order among the variables?

Acknowledgments

This research was partially supported by the NSF grants DMS-0505696 and DMS-0906252. Comments from an AE and two referees have greatly improved the presentation, focus and the scope of the paper.

Appendix: Proof of the Lemma

Recall that the matrix of regression coefficients $\Phi_{2|1}$ must be found so that the vector of residuals $Y_{2\cdot 1}$ is uncorrelated with the data Y_1:

$$\begin{aligned} 0 = \mathrm{Cov}(Y_{2\cdot 1}, Y_1) &= \mathrm{Cov}(Y_2, Y_1) - \Phi_{2|1}\mathrm{Cov}(Y_1, Y_1) \\ &= \Sigma_{21} - \Phi_{2|1}\Sigma_{11}, \end{aligned}$$

or

$$\Phi_{2|1} = \Sigma_{21}\Sigma_{11}^{-1}. \tag{27}$$

The covariance matrix of $Y_{2\cdot 1}$, the regression residuals, denoted by $\Sigma_{22\cdot 1}$, is

$$\begin{aligned} \Sigma_{22\cdot 1} &= \mathrm{Cov}(Y_{2\cdot 1}) = \mathrm{Cov}(Y_2 - \Phi_{2|1}Y_1,\ Y_2 - \Phi_{2|1}Y_1) \\ &= \mathrm{Cov}(Y_2 - \Phi_{2|1}Y_1,\ Y_2) \\ &= \Sigma_{22} - \Phi_{2|1}\Sigma_{12} \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \end{aligned} \tag{28}$$

Since the vector of residuals $Y_{2\cdot 1}$ is a linear transformation of Y_1, Y_2, it is evident that

$$\begin{pmatrix} Y_1 \\ Y_{2\cdot 1} \end{pmatrix} = T \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \tag{29}$$

with T a block lower-triangular matrix with the identity blocks down the diagonals and $-\Phi_{2|1}$ in the $(2,1)$ block. The covariance matrix of the left-hand side in (29) is by construction a block-diagonal matrix. Thus, with

$$D = \text{diag}(\Sigma_{11}, \Sigma_{22\cdot 1}), \tag{30}$$

computing the covariance of both sides of (29) we obtain

$$T\Sigma T' = D, \tag{31}$$

and the precision matrix Σ^{-1} has a similar decomposition:

$$\Sigma^{-1} = T'D^{-1}T. \tag{32}$$

Important consequences of multiplying out the partitioned matrices in the right-hand side of this identity and matching with the $(2,2)$ and $(1,2)$ blocks of Σ^{-1} are precisely the results given in the lemma.

References

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.

Banerjee, O., Ghaoui, L.E. and dAspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, **9**, 485516.

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.

Bickel, P.J. and Gel, Y. R. (2010). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. Technical Report, University of Waterloo.

Bickel, P.J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes". and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989-1010.

Bickel, P.J. and Levina, L. (2008a). Regularized estimation of large covariance matrices.Ann. of Statist. **36**,199-227.

Bickel, P.J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. of Statist.* **36**, 2577-2604.

Bickel, P. J. and Li, B. (2006). Regularization in statistics (with discussion). *Test*, **15**, 271-344.

Bilmes, J. A. (2000). Factored sparse inverse covariance matrices. In IEEE International Conference on Accoustics, Speech and Signal Processing.

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

Böhm, H. and von Sachs, R. (2008). Structural shrinkage of nonparametric spectral estimators for multivariate time series. *Electronic Journal of Statistics*, **2**,696-721.

Bondell, H.D., Krishna, A. and Ghosh, S.K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, To appear.

Box, G.E.P., Jenkins, G.M. and Reinsel, G. (1994). *Time Series Analysis-Forecasting and Control*, Revised Third Edition,Prentice Hall,NJ.

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.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*. To Appear.

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.

Chang, C. and Tsay, R.S. (2010). Estimation of covariance matrix via the sparse Cholesky factor with Lasso. *Journal of Statistical Planning and Inference*. To appear.

Chaudhuri, S., Drton, M. and Richardson, T.S. (2007). Estimation of a covariance matrix with zeros. *Biometrika*, **94**, 199-216.

Chen, Z. and Dunson,D.(2003). Random effects selection in linear mixed models. *Biometrics*, **59**, 762-769.

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

Cressie, N.A.C. (2003). *Statistics for Spatial Data*, Revised Edition, Wiley, New York.

dAspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection.*SIAM Journal on Matrix Analysis and its Applications* ,**30**, 5666.

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

Dai, M. and Guo, W. (2004). Multivariate spectral analysis using Cholesky decomposition. *Biometrika*, **91**, 629-643.

Daniels, M.J. (2005). A class of shrinkage priors for the dependence structure in longitudinal data. *J. of Statist. Planning and Inference*, **127**, 119-130.

Daniels, M.J. and Hogan, J. (2008). *Missing Data In Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis* , Chapman & Hall/CRC.

Daniels, M.J. and 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 Kass, R. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, **57**, 1173-1184.

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

Daniels,M.J. and Pourahmadi,M. (2009). Modeling covariance matrices via partial autocorrelations. *J of Multivariate Analysis*, **100**, 2352-2363.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, **30**,5666.

Dégerine, S., Lambert-Lacroix, S. (2003). Partial autocorrelation function of a nonstationary time series. *J. of Multivariate Analysis*, 89, 135-147.

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

Dey, D.K., Srinivasan, C. (1985). Estimation of a covariance matrix under Steins loss. *Ann. Statist.* ,**13**, 15811591.

Diggle, P., Liang, K.Y. , Zeger, S.L. and Heagerty,P.J. (2002). *Analysis of Longitudinal Data,Second Edition*, Oxford, Clarendon Press.

Eaves, D. and Chang, T. (1992). Priors for ordered conditional variances and vector partial correlation. *J. of Multivariate Analysis*, 41, 43-55.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **35**, 407499.

El Karoui, N. (2008a). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.*, **36**, 27172756.

El Karoui, N. (2008b). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.*, **36**, 27572790.

Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, **102**, 632641.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, **147**, 186197.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics*, **3**, 521541.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high-dimensional feature space. *Statistica Sinica*, **20**, 101-148.

Fitzmaurice, G., Davidian, M., Verbke, G. and Molenberghs, G. (eds.) (2009). *Longitudinal Data Analysis*, Handbooks of Modern Statistical Methods, Chapman & Hall/CRC.

Flury, B. (1988). *Common Principal Components and Related Multivariate Models* ,Wiley, NewYork.

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.

Friedman, J., Hastie, T. and Tibshirani, R.(2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical Report, Stanford University.

Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, **98**, 227-255.

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.

Golub, G.H. and Van Loan, C.F. (1989). *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Second Edition.

Haff, L.R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.*, **8** ,586597.

Haff, L. R. (1991). The variational form of certain Bayes estimators. *Annals of Statistics*, **19**, 1163-1190.

Hastie, T.J. and Tibshirani, R. J.(1990). *Generalized Additive Models*, Chapman &HALL/CRC.

Hastie, T., Tibshirani, R. and Frieman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second Edition, Springer, New York.

Hoerl, A. E., and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 5567.

Hoff, P.D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *J. of Royal Statistical Society, Ser. B*, **71**, 971-992.

Hoff, P.D. and Niu, X. (2009). A covariance regression model. Technical Report, University of Washington.

Huang, J.Z., Liu, N. and Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalized normal likelihood. *Biometrika*, **93**, 85-98.

Huang, J.Z., Liu, L. and Liu, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *J. of Statistical Computation and Graphics*, **16**, 189-209.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Sympossium on Mathematical and Statistical Probability*, (Jerzy Neyman, ed.), Vol. I, 361-379, University of California, Berkeley.

Jiang, G., Sarkar, S. K. and Hsuan, F. (1999). A likelihood ratio test and its modifications for the homogeneity of the covariance matrices of dependent multivariate normals. *Journal of Statistical Planning and Inference*, **81**, 95-111.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *J. of Multivariate Analysis*, **97**,2177-2189.

Jong, J.C. and Kotz, S. (1999). On a relation between principal components and regression analysis. *The Amer. Statist.*, **53**, 349-351.

Johnstone, I.M. and Lu, A.Y. (2009).

On consistency and sparsity for principal components analysis in high dimensions. *J. of Amer. Statist. Assoc.*, **104**, 682-693.

Jones, M.C. (1987). Randomly choosing parameters from the stationarity and invertibility region of autoregressive-moving average models. *Appl. Statist.*, *36*, 134-138.

Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, *22*, 389-395.

Kalman, A.E. (1960). A new approach to linear filtering and prediction problems, *Trans. Amer. Soc. Mech. Eng., J. Basic Engineering*, **82**, 35-45.

Kaufman, C.G., Schervish, M.J. and Nychka, W. (2008). Covariance tapering for likelihood-based estimation in large data sets. *Journal of the Amer. Statist. Assoc.*, **103**, 145-155.

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.

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

Kurowicka, D. and Cooke, R. (2003). A parameterization of positive definite matrices in terms of partial correlation vines. *Linear Algebra and its Applications*, **372**, 225-251.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics*, **37**, 4254-4278.

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.

Leng, C., Zhang, W. and Pan, J.(2010). Semiparametric meancovariance regression analysis for longitudinal data. *J. of Amer. Statist. Assoc.*, **105**, 181-193.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, **88**, 365-411.

Leonard, T. and Hsu, J.S.J. (1992). Bayesian inference for a covariance matrix. *Annals of Statistics*, **20**,1669-1696.

LeSage, J.P. and Pace, R.K. (2007). A matrix exponential spatial specification. *Journal of Econometrics*, **140**, 198-214.

Levina, E., Rothman, A.J. and Zhu, J.(2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Annals of Applied Statistics*, **2**, 245-263.

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

Liechty,J.C., Liechty, M.W. and Müller, P. (2004). Bayesian correlation estimation. *Biometrika*, **91**, 1-14.

Lin, T.I. (2010). A Bayesian approach to joint modelling of location and scale parameters of the t distribution for longitudinal data. *Journal of Statistical Planning and Inference*, to appear.

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, 411-429. Amsterdam: North-Holland.

Tsung-I. Lin, T.-I., Wang, Y. J. (2009). A robust approach to joint modeling of mean and scale covariance for longitudinal data. *Journal of Statistical Planning and Inference* , **139**, 3013–3026.

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.

Liu, X. and Daniels, M.J. (2006). A new algorithm for simulating a correlation matrix based on parameter expansion and re-parameterization. *Journal of Computational and Graphical Statistics*, **15**, 897-914.

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

McMurry, T.L. and Politis, D.N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. Technical Report, UC San Diego.

Meinshausen, N. and and Bühlman, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. of Statist.*, **34**, 1436-1462.

Muirhead, R.J. and Leung, P.L. (1987). Estimation of parameter matrices and eigenvalues in MANOVA and canonical correlation analysis. *Ann. Statist.*, **15**, 16511666.

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

Peng, J., Wang, P., Zhou, N., Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. of Amer. Statist. Assoc.*, **104**, 735-746.

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. (2007a). Cholesky decompositions and estimation of a multivariate normal covariance matrix: Parameter orthogonality. *Biometrika*, **94**,1006-1013.

Pourahmadi, M. (2007b). Simultaneous modeling of covariance matrices : GLM, Bayesian and nonparametric perspective. *Correlated Data Modelling 2004*, D. Gregori et al. (eds), FrancoAngeli, Milan, Italy.

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

Pourahmadi M, Daniels M. and Park T. (2007). Simultaneous modelling of the Cholesky decomposition of several covariance matrices with applications. J. of Multivariate Analysis, **98**,568-587.

Qu, A. and Lindsay, B. (2003). Building adaptive estimating equations when inverse of covariance estimation is difficult. *Journal of the Royal Statistical Society, Series B*, **65**, 127-142.

Quenouille, M.H. (1949). Approximate tests of correlation in time series. *J. of Roy. Statist. Soc. B*, **11**, 68-84.

Rajaratnam, B., Massam, H., and Corvallo, C. (2008). Flexible covariance estimation in graphical gaussian models. *Ann. Statist.*, **36**, 28182849.

Ramsey, F.L. (1974). Characterization of the partial autocorrelation function. *Ann. of Statist.*, **2**, 1296-1301.

Rocha, G. V., Zhao, P., and Yu, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). Technical Report 759, Department of Statistics, UC Berke- ley.

Rosen, O. and Stoffer, D. (2007). Automatic estimation of multivariate spectra via smoothing splines. *Biometrika*, **94**, 335-345.

Rothman, A.J., Bickel, P.J., Levina, E. and Zhu, J. (2008). *Sparse Permutation Invariant Covariance Estimation. Electronic Journal of Statistics*, **2**, 494-515.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* (Theory and Methods), **104**, 177186.

Rothman, A.J., Levina, E. and Zhu, J. (2010). A new approach to Cholesky-based estimation of high-dimensional covariance matrices. *Biometrika*, to appear.

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.

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

Stein, C.(1956). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Sympossium on Mathematical and Statistical Probability*, (Jerzy Neyman, ed.), Vol. I, 197-206, University of California, Berkeley.

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

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.

Wagaman, A. S. and Levina, E. (2009). Discovering sparse covariance structures with the Isomap. *J. Comp. Graph. Statist.*, **18**, 551-572.

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

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *Journal of the Royal Statistical Society, Series B*, **71**, 615636.

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.

Wu, W.B. and Pourahmadi, M. (2009). Banding sample covariance matrices of stationary processes. *Statistica Sininca*, **19**, 1755-1768.

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

Yuan, M. and Huang, J. Z. (2009). Regularised parameter estimation of high dimensional t distribution. *Journal of Statistical Planning and Inference*, **139**, 2284-2292.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19-35.

Yule, G.U. (1907). On the theory of correlation for any number of variables, treated by a new system of notation. *Roy. Soc. Proc.*, **79**, 85-96.

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.

Zimmerman, D.L. (2000). Viewing the correlation structure of longitudinal data through a PRISM. *The American Statistician*, **54**, 310-318.

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

Zimmerman, D.L. and Núñez-Antón, V. (2010). *Antedependence Models for Longitudinal Data*, CRC Press, New York.

Zhu, Z. and Liu, Y. (2009). Estimating spatial covariance using penalised likelihood with weighted L1 penalty. *J. of Nonparametric Statistics*, **21**, 925-942.

Zou, H., Hastie, T. and Tibshirani. R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286.