Statistical Science
2011, Vol. 26, No. 3, 369–387
DOI: 10.1214/11-STS358

Covariance Estimation: The GLM and Regularization Perspectives

Mohsen Pourahmadi

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 two relatively complementary perspectives: (1) generalized linear models (GLM) or parsimony and use of covariates in low dimensions, and (2) regularization or sparsity for high-dimensional data. An emerging, unifying and powerful trend in both perspectives is that of reducing a covariance estimation problem to that of estimating a sequence of regression problems. We point out several instances of the regression-based formulation. A notable case is in sparse estimation of a precision matrix or a Gaussian graphical model leading to the fast graphical LASSO algorithm. Some advantages and limitations of the regression-based Cholesky decomposition relative to the classical spectral (eigenvalue) and variance-correlation decompositions are highlighted. The former provides an unconstrained and statistically interpretable reparameterization, and guarantees the positive-definiteness of the estimated covariance matrix. It reduces the unintuitive task of covariance estimation to that of modeling a sequence of regressions at the cost of imposing an *a priori order* among the variables. Elementwise regularization of the sample covariance matrix such as banding, tapering and thresholding has desirable asymptotic properties and the sparse estimated covariance matrix is positive definite with probability tending to one for large samples and dimensions.

Key words and phrases: 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 $Y = (y_1, \ldots, y_p)'$ 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, Jenkins and Reinsel, 1994), spatial data analysis (Cressie, 1993), variance components and longitudinal data analysis (Searle, Casella and McCulloch, 1992; Diggle et al., 2002), and in the modern and rapidly growing

Mohsen Pourahmadi is Professor, Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, USA e-mail: pourahm@stat.tamu.edu.

area of statistical and machine learning dealing with massive and high-dimensional data (Hastie, Tibshirani and Friedman, 2009). 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 these challenges become manageable, for example, 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 $p \gg n$. Here the use of a sample covariance matrix is problematic (Stein, 1956), particularly when its inverse is needed as, for example, in classification procedures (Anderson, 2003, Chapter 6), multivariate linear regression (Warton, 2008; Witten and Tibshirani, 2009), portfolio selection (Ledoit, Santa-Clara and Wolf, 2003) and Gaussian graphical models (Wong, Carter and Kohn, 2003; Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007). In these situations and others, it is desirable to find alternative covariance estimators that are more accurate and better-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 a poor estimator when $\frac{p}{n}$ is large (Johnstone, 2001). It distorts the eigenstructure of Σ, in the sense that the largest (smallest) sample eigenvalue will be biased upward (downward). Since then many improved estimators have been proposed by shrinking the eigenvalues of S toward 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. The Stein's family of shrinkage estimators leaving intact the eigenvectors of the sample covariance matrix are neither sparse nor parsimonious. However, lately the search for sparsity and parsimony has led to either shrinking the matrix S itself toward certain targets like diagonal and autoregressive structures as in Daniels and Kass (1999, 2001), or shrinking its eigenvectors as in Hoff (2009) and Johnstone and Lu (2009).

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. The regression-based approach of Meinshausen and Bühlmann (2006) provides a sparse estimate of the precision matrix or a Gaussian graphical model by fitting separate LASSO regression to each variable, using the others as predictors. This simple idea has inspired several direct and improved sparse estimators of Σ^{-1} using a penalized likelihood approach with a LASSO penalty on its off-diagonal terms (Yuan and Lin, 2007; Banerjee, El Ghaoui and d'Aspremont, 2008; Friedman, Hastie and Tibshirani, 2008; Rothman et al., 2008; Rocha, Zhao and Yu, 2008; Peng et al., 2009). Friedman, Hastie and Tibshirani (2008) graphical LASSO is the fastest available algorithm to date. Surprisingly, such a sparse covariance estimator is guaranteed to be positive definite (Banerjee, El Ghaoui and d'Aspremont, 2008).

A remarkable unifying *regression-based* theme has emerged from research on covariance estimation in the last decade or so. Some notable examples are as follows: (i) formulating principal component analysis (PCA) as regression optimization problems (Jong and Kotz, 1999; Zou, Hastie and Tibshirani, 2006), sparse loadings are then estimated by imposing the lasso constraint on the regression coefficients, (ii) regression-based derivation and interpretation of the modified Cholesky decomposition of a covariance matrix and its inverse (Pourahmadi, 1999, 2001, Section 3.5; Bilmes, 2000; Huang et al., 2006; Rothman, Levina and Zhu, 2010), (iii) the regression approach of Meinshausen and Bühlmann (2006) to the Gaussian graphical models, (iv) the graphical LASSO algorithm of Friedman, Hastie and Tibshirani (2008, 2010) and (v) the iteratively reweighted penalized likelihood of Fan, Feng and Wu (2009) where nonconcave penalties such as the smoothly clipped absolute deviation (SCAD) are imposed on the entries of the precision matrix. The problem of sparse estimation of the precision matrix is then recast as a sequence of penalized likelihood problems with a weighted LASSO penalty and solved using the graphical LASSO algorithm of Friedman, Hastie and Tibshirani (2008).

Among these approaches it seems only (ii) has the expressed goal of providing unconstrained and statistically interpretable regression parameters for the covariance (precision) matrix. Unfortunately, how-

ever, 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 give rise to covariance estimators which are sensitive to the order among the variables in Y. These approaches are suitable for 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, 2003, 2009) and generally those based on the Cholesky decomposition of the covariance matrix or its inverse (Pourahmadi, 1999, 2000; Rothman, Levina and Zhu, 2010) do impose an order among the components of Y and are not permutation-invariant. The idea of thresholding individual entries of S has been used in the estimation of large covariance matrices by Bickel and Levina (2008b), El Karoui (2008a, 2008b) and Rothman, Levina and Zhu (2009). Such estimators are permutation-invariant with desirable asymptotic properties.

It should be noted that the recent surge of interest in regression-based approaches to *sparsity* in high-dimensional data bodes well with the long history of interest in *parsimony* and using covariates when modeling covariance matrices of low-dimensional data (Anderson, 1973). For example, longitudinal data collected from expensive clinical trials and biological experiments may have about $n = 30$ subjects and $p \leq 10$ measurements per subject. Parsimonious and accurate modeling of the covariance structure is important in these application areas (Cannon et al., 2001; Carroll, 2003; Fitzmaurice et al., 2009). However, the area of data-based covariance modeling is woefully underdeveloped. At present, a practitioner has the option of picking a structured covariance matrix from a long menu, where at one extreme the choice is $\sigma^2 I_p$ (independence) and at the other the unstructured covariance matrix with $\frac{p(p+1)}{2}$ parameters (Zimmerman and Núñez-Antón, 2001, 2010). Of course, it is desirable to bridge the gap between these two extremes and develop a bona fide GLM methodology and a data-based framework for modeling covariance matrices. Attempts to develop such methods going beyond the traditional linear covariance models (Anderson, 1973) have been made in recent years by Chiu, Leonard and Tsui (1996) and Pourahmadi (1999, 2000); Pan and MacKenzie (2003); Lin and Wang (2009); Leng, Zhang and Pan (2010); Lin (2011) using the spectral and Cholesky decompositions of covariance matrices, respectively.

Given the complex nature of the positive-definiteness constraint, in developing a GLM methodolgy it is plausible to factorize Σ into two or more 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. 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, orthogonal and lower triangular matrices, respectively. While the entries of the first two matrices are always constrained, those of the last are unconstrained. Interestingly, these three decompositions are subsumed (Zimmerman and Núñez-Antón, 2001, page 59) by a decomposition from the class of factor/mixed models (Anderson, 2003):

$$\Sigma = ZBZ' + W. \tag{1}$$

Here, the matrix Z is $p \times q$ with q standing for the number of latent factors, B and W are $q \times q$ and $p \times p$ unknown preferably diagonal matrices. The representation (1) is valid only when each of the p variables are well-approximated as linear combinations of the same latent factors plus an independent error. In principle, this may occur when q is large, and adding W to the reduced rank decomposition ensures the positive-definiteness of Σ. Technical difficulties with the use of (1) can be resolved to various extents by choosing the components of the quadruple (q, W, B, Z) close to the ideal values of $q = p$, $W = 0$, B diagonal and Z sparse or structured.

The outline of the paper is as follows. Section 2 covers some preliminaries on the GLM for covariance matrices, 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, tapering and thresholding), penalized likelihood estimation and improvement of the sample covariance matrix for high-dimensional data 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.

This survey emphasizes the importance of regression-based idea and hence the need for unconstrained reparameterization in both the GLM- and regularization-type approaches to covariance estimation for low- and high-dimensional data. 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 guide or a blueprint for further research in this active and growing area of current interest in statistics.

2. THE GLM AND 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. 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 develop a systematic, data-based statistical model-fitting procedure for covariance matrices. 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.

Thus, the GLM approach to covariance estimation hinges on finding link functions that induce unconstrained and statistically interpretable reparameterization. Not surprisingly, most 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 examples of unconstrained parameterizations of a covariance matrix are given which involve the variance-correlation, 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 both factors 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 factor is more important than the other (Lin and Perlman, 1985; Liang and Zeger, 1986; Barnard, McCulloch and Meng 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 and 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 a covariance structure by estimating the marginal variances via kernel smoothing and used specific parametric models for the correlation matrix such as the ARMA$(1,1)$.

2.2.2 *Decomposition of the precision matrix: Gaussian graphical models* Recall that the marginal (pairwise) dependence among the entries of a random vector is captured by the off-diagonal entries of Σ or the entries of the correlation matrix $R = (\rho_{ij})$. However, the conditional dependencies can be found in the off-diagonal entries of the precision matrix $\Sigma^{-1} = (\sigma^{ij})$. More precisely, for Y a mean zero normal random vector with a positive-definite covariance matrix, if the ijth component of the precision matrix is zero, then the variables y_i and y_j are conditionally independent, given the other variables.

Conditional independence structure in Y is often shown as a graphical model with the nodes corresponding to variables and the absence of edges indicating conditional independence (Anderson, 2003, Chapter 15).

In this section we 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) based on solving p separate LASSO 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 a variable given the rest. More precisely, standard regression calculations show that the *partial correlation* coefficient between y_i and y_j after removing the linear effect of the $p-2$ remaining variables is given by $\tilde{\rho}_{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$, and that $\tilde{d}_{ii}^2$, the *partial variance* of y_i after removing the linear effect of the remaining $p-1$ variables, 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. Similarly, its covariance and precision matrices will be partitioned conformally as

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

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 matrix and the vector of regression residuals be denoted by $Y_{2\cdot 1} = Y_2 - \Phi_{2|1}Y_1$. Recall that $\Phi_{2|1}$ and the corresponding prediction error covariance matrix can be found by requiring that the vector of residuals $Y_{2\cdot 1}$ be uncorrelated with Y_1. Thus,

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

and

$$\begin{aligned} \operatorname{Cov}(Y_{2\cdot 1}) &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\ &= \Sigma_{22\cdot 1} = (\Sigma^{22})^{-1}. \end{aligned} \tag{3}$$

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 precision matrix Σ^{-1}, as we discuss below.

First, 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 $\tilde{\varepsilon}_i = y_i - \tilde{y}_i$, $\tilde{d}_i^2 = \operatorname{Var}(\tilde{\varepsilon}_i)$ be its prediction error and prediction error variance, respectively. Then,

$$y_i = \sum_{j \neq i} \beta_{ij} y_j + \tilde{\varepsilon}_i, \tag{4}$$

and it follows immediately from (2) and (3) that the regression coefficients of y_i on $Y_{-(i)}$, are given by

$$\beta_{i,j} = -\frac{\sigma^{ij}}{\sigma^{ii}}, \quad j \neq i, \tag{5}$$

and

$$\tilde{d}_i^2 = \operatorname{Var}(y_i | y_j, j \neq i) = \frac{1}{\sigma^{ii}}, \quad i = 1, \ldots, p. \tag{6}$$

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, note that $\beta_{j,j} = 0$ and $\beta_{i,j}$ is not symmetric in (i,j).

Writing (5) in matrix form gives another useful factorization of the precision matrix:

$$\Sigma^{-1} = \tilde{D}^2(I_p - \tilde{B}), \tag{7}$$

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 (7) 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 (4). This is the key conceptual tool behind the approach of Meinshausen and Bühlmann (2006). Note that the left-hand side of (7) is a symmetric matrix while the right-hand side is not necessarily so. Thus, one must impose the following symmetry constraint (Rocha, Zhao and Yu 2008; Friedman, Hastie and Tibshirani, 2010) for $j, k = 1, \ldots, p$:

$$\tilde{d}_k^2 \beta_{jk} = \tilde{d}_j^2 \beta_{kj}. \tag{8}$$

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 (3) 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:

$$\tilde{\rho}_{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}}\sigma^{ij}}, \tag{9}$$

as announced earlier. Moreover, from (5) and (9) it follows that

$$\beta_{ij} = \tilde{\rho}_{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}. \tag{10}$$

This representation which shows that Σ^{-1} and $\tilde{R}$ share the same sparsity pattern is the basis for the Peng, Zhou and Zhu (2009) SPACE algorithm which imposes a LASSO penalty on the off-diagonal entries of the matrix of partial correlations $\tilde{R}$; see also Friedman, Hastie and Tibshirani (2010).

2.2.3 *The spectral decomposition* The spectral decomposition of a covariance matrix given by

$$\Sigma = P\Lambda P' = \sum_{i=1}^{p} \lambda_i e_i e_i', \tag{11}$$

where Λ is a diagonal matrix of eigenvalues and P the orthogonal matrix of normalized eigenvectors with the e_i as 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 a new unconstrained reparameterization due to Leonard and Hsu (1992) and Chiu, Leonard and Tsui (1996). They observed that the logarithm of a covariance matrix Σ defined by

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

is an unconstrained symmetric matrix. 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, Liechty and Müller, 2004). From (11) and (12) it is evident that the entries of Σ and $\log\Sigma$ are similar functions of the entries of P and Λ, except that in (12) 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? This case is interesting as it points out to a sort of trade-off that exists between the requirements of unconstrained reparameterization of covariance matrices and statistical interpretability of the new parameters.

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 (Golub and Van Loan, 1989) is of the form

$$\Sigma = CC', \tag{13}$$

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 $D = \operatorname{diag}(c_{11}, \ldots, c_{pp})$ 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, (13) can be rewritten as

$$\Sigma = CD^{-1}DDD^{-1}C' = LD^2L', \tag{14}$$

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

$$T\Sigma T' = D^2, \quad \Sigma^{-1} = T'D^{-2}T, \tag{15}$$

where $T = L^{-1}$. Note that the second identity is, in fact, the modified Cholesky decomposition of the precision matrix Σ^{-1}, and the first identity in (15) looks a lot like the spectral decomposition, in that Σ is diagonalized by a lower triangular matrix. However, we show 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 (2009) have proposed an Isomap method for discovering an order among the variables based on their correlations. This could lead to block-diagonal or banded correlation structures which may help to fix a reasonable order before applying the Cholesky decomposition; see Section 5.

As in Section 2.2.2, we use the idea of regression 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 $y_{t-1}, \ldots, y_1$, and $\varepsilon_t = y_t - \hat{y}_t$ be its prediction error with variance $\sigma_t^2 = \mathrm{Var}(\varepsilon_t)$. Then, there are unique scalars ϕ_{tj} so that

$$y_t = \sum_{j=1}^{t-1} \phi_{tj} y_j + \varepsilon_t, \quad t = 1, \ldots, p. \tag{16}$$

Next, we show how to compute the regression coefficients ϕ_{tj} using the covariance matrix. For a fixed t, $2 \le t \le p$, set $\phi_t = (\phi_{t1}, \ldots, \phi_{t,t-1})'$ 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 (2) and (3) with $Y_1 = (y_1, \ldots, y_{t-1})', Y_2 = y_t$ it follows that

$$\phi_t = \Sigma_t^{-1} \tilde{\sigma}_t, \quad \sigma_t^2 = \sigma_{tt} - \tilde{\sigma}_t' \Sigma_t^{-1} \tilde{\sigma}_t. \tag{17}$$

Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)'$ be the vector of successive uncorrelated prediction errors with $\mathrm{Cov}(\varepsilon) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2) = D^2$. Then, (16) can be rewritten in matrix form as $\varepsilon = TY$, where T is the following unit lower triangular matrix:

$$T = \begin{pmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{n1} & -\phi_{n2} & \cdots & -\phi_{n,n-1} & 1 \end{pmatrix}. \tag{18}$$

Now, computing $\mathrm{Cov}(\varepsilon) = \mathrm{Cov}(TY) = T\Sigma T'$ gives the modified Cholesky decomposition (15).

Since the ϕ_{ij}'s in (17) 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 (16) it is evident that the regression or 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. Thus, one can bring the familiar regression analysis machinery to handle the unintuitive task of modeling covariance matrices (Smith and Kohn, 2002; Wu and Pourahmadi, 2003; Huang et al., 2006, Huang, Liu and Liu, 2007; Bickel and Levina, 2008a; Rothman, Levina and Zhu, 2009). An important consequence of (15) 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 (14) due to Chen and Dunson (2003), also obtained from (13), 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-effects models, and when the focus is on modeling the correlation matrix; see Bondell, Krishna and Ghosh (2010) and 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

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 the 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 the procedure as *covariance selection* models. It fits the framework of linear covariance models defined next.

Motivated by the simple and linear structure of covariance matrices of some time series and variance

component models, Anderson (1973) introduced the class of *linear covariance models* (LCM):

$$\Sigma^{\pm 1} = \alpha_1 U_1 + \cdots + \alpha_q U_q, \tag{19}$$

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 at least a set of coefficients where $\Sigma^{\pm 1}$ is positive definite. The model (19) is rather general, indeed, for $q = p^2$ any covariance matrix admits the representation

$$\Sigma = (\sigma_{ij}) = \sum_{i=1}^{p}\sum_{j=1}^{p} \sigma_{ij} U_{ij}, \tag{20}$$

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 (19), 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. 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 entries of the precision matrix, etc.; see (20).

A major drawback of (19) and (20) is the constraint on the coefficients which could make the estimation and other statistical problems difficult (Anderson, 1973). Szatrowski (1980) gives necessary and sufficient conditions for the existence of explicit maximum likelihood estimates, and the convergence of the iterative procedure in one iteration from any positive-definite starting point.

A good review of the MLE procedures for the model (19) and their applications to the problem of testing homogeneity of the covariance matrices of several dependent multivariate normals is 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 also in the spirit of (19), 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* A plausible way to remove the constraint on α_i's in (19) 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 a symmetric matrix with unconstrained entries taking values in $(-\infty, \infty)$.

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

$$\log \Sigma = \alpha_1 U_1 + \cdots + \alpha_q U_q, \tag{21}$$

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, the α_i's lack statistical interpretation (Brown, Le and Zidek, 1994; Liechty, Liechty and Müller, 2004). Fortunately, for Σ diagonal since $\log \Sigma = \operatorname{diag}(\log \sigma_{11}, \ldots, \log \sigma_{pp})$ is also diagonal, it can be seen that (21) amounts to log-linear models for heterogeneous variances which have a long history in econometrics and other areas; see Carroll and Ruppert (1988) and references therein.

Maximum likelihood estimation procedures for the parameters in (21) and their asymptotic properties are studied in Chiu, Leonard and Tsui (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, this does not seem to be the 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 (19) and (21) 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.

Recall that for an unstructured covariance matrix Σ, the nonredundant entries of its components $(T, \log D^2)$ in (15) 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 $t=1,\dots,p$; $j=1,\dots,t-1$,

$$\log\sigma_t^2 = z_t'\lambda, \quad \phi_{tj} = z_{tj}'\gamma. \tag{22}$$

Here, z_t, z_{tj} are $q\times 1$ and $d\times 1$ vectors of known covariates, $\lambda=(\lambda_1,\dots,\lambda_q)'$ and $\gamma=(\gamma_1,\dots,\gamma_d)'$ 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; Lin and Wang, 2009; Leng, Zhang and Pan, 2010) are in terms of powers of times and lags

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

A truly remarkable feature of (22) 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 or AIC to identify models such as (22) for the data; for more details see Pourahmadi (1999, 2001) and Pan and MacKenzie (2003). Liang and Zeger (1986) employ such parametrized models for covariance matrices in the context of the popular generalized estimating equations for longitudinal data.

Computing the MLE of the parameters is relatively simple due to the special form of the loglikelihood function for a sample $Y_1,\dots,Y_n$ from a normal population with mean zero and the common covariance Σ parameterized as in (22). Note that, except for a constant, we have

$$\begin{aligned} -2l(\lambda,\gamma) &= \sum_{i=1}^{n}(\log|\Sigma| + Y_i'\Sigma^{-1}Y_i) \\ &= n\log|D^2| + n\,\mathrm{tr}\,\Sigma^{-1}S \\ &= n\log|D^2| + n\,\mathrm{tr}\,D^{-2}TST' \\ &= n\log|D^2| \\ &\quad + n\,\mathrm{tr}\,D^{-2}(I_p-B)S(I_p-B)', \end{aligned} \tag{23}$$

where $S=\frac{1}{n}\sum_{i=1}^n Y_iY_i'$, $B=I_p-T$ and the last three equalities are obtained by replacing for Σ^{-1} from (15) and some basic matrix operations involving trace of a matrix. Since (23) is quadratic in B, 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; Huang, Liu and Liu, 2007). This simplicity in 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, Zhao and Yu (2008) SPLICE algorithm; see 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 asymptotic 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 Pourahmadi (2007a) 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; see Lin and Wang (2009) and Lin (2011).

3. REGULARIZATION OF THE SAMPLE COVARIANCE MATRIX

This section is devoted to high-dimensional data where the sample covariance matrix is known to be a poor estimator, and not even invertible when $p\gg n$. We review some alternative and improved estimators obtained by regularizing the sample covariance matrix in various ways. After presenting a few loss functions in Section 3.1, we review in Sections 3.2 and 3.3 shrinkage estimators obtained by minimizing certain risk functions. An early and inspiring example is the Stein's family of shrinkage estimators that shrinks the eigenvalues of the sample covariance matrix toward a central value. Penalized normal likelihood estimators with a LASSO penalty on the precision matrix are reviewed in Section 3.4 with a focus on the graphical LASSO algorithm. Regularization methods which act elementwise on the sample covariance matrix such as tapering, banding and thresholding are discussed in Section 3.5. Some conditions for consistency of such estimators are also reviewed.

3.1 Some Loss and Risk Functions

Regularized estimators are usually obtained by minimizing suitable norms, risks or objective functions. For covariance matrix estimation the Frobenius and operator (spectral) norms are quite nat-

ural and have proved useful in establishing theoretical properties of covariance estimators. For example, consistency in operator norm guarantees the consistency of the eigenstructure used in principal component analysis (Johnstone and Lu, 2009) and other related methods in multivariate statistics; see Section 3.5.

The two commonly used loss functions when $n > p$ are

$$L_1(\hat{\Sigma}, \Sigma) = \operatorname{tr}(\hat{\Sigma}\Sigma^{-1}) - \log|\hat{\Sigma}\Sigma^{-1}| - p,$$
$$L_2(\hat{\Sigma}, \Sigma) = \operatorname{tr}(\hat{\Sigma}\Sigma^{-1} - I)^2,$$

where $\hat{\Sigma} = \hat{\Sigma}(S)$ is an arbitrary estimator. The corresponding risk functions are

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

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 of the sample covariance matrix, 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

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

Note that though dividing by the dimension p is not standard, it has the advantage that norm of the identity matrix is 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 as those for L_2. However, it has an additional and attractive feature that the optimal covariance estimator under L_3 turns out to be the penalized normal likelihood estimator with $\operatorname{tr}\Sigma^{-1}$ 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 better conditioned.

3.2 Shrinking the Spectrum and the Correlation Matrix

In this section we present 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 $\lambda = (\lambda_1, \ldots, \lambda_p)', \lambda_1 > \cdots > \lambda_p > 0$, 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 $\Phi(\lambda) = \operatorname{diag}(\varphi_1, \ldots, \varphi_p)$ 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).

Lin and Perlman (1985) have applied the James–Stein shrinkage estimators (James and Stein, 1961) to the sample correlation in order to improve it 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.

3.3 Ledoit–Wolf Shrinkage Estimator

To ensure nonsingularity of the estimated covariance matrix in the "n small, p large" case, 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.

One can motivate such an estimator by recalling that the sample covariance matrix S is unbiased for Σ, but unstable with considerable risk when $p \gg n$. By contrast, 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).

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-dimensional aspects of the true (but unknown) covariance matrix Σ:

$$\mu = \operatorname{tr}(\Sigma)/p, \quad \alpha^2 = \|\Sigma - \mu I\|^2,$$
$$\beta^2 = E\|S - \Sigma\|^2, \quad \delta^2 = E\|S - \mu I\|^2.$$

Consistent estimators of these low-dimensional parameters 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.

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 $\operatorname{tr}\Sigma^{-1}$. 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 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 regularization methods based on penalizing the normal likelihood. These methods differ mostly on the LASSO penalty imposed on certain segments of the precision matrix. For example, Huang et al. (2006), Banerjee, El Ghaoui and d'Aspremont (2008), Friedman, Hastie and Tibshirani (2008), Rothman et al. (2008) and Warton (2008), respectively, impose penalty on the Cholesky factor, all the entries, off-diagonal entries and the diagonal entries of the precision matrix. These can be viewed as methods for solving Dempster's (1972) covariance selection problem of inducing sparsity in the precision matrix. However, Warton's (2008) penalty leads to the Ledoit–Wolf estimator where neither Σ nor its inverse is sparse.

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 (19) and (20), it is plausible to induce sparsity in the precision matrix estimate by adding to the normal loglikelihood (23) a penalty on the entries of the precision matrix Σ^{-1} or its Cholesky factor (Huang et al., 2006)

$$-2l + \sum_{i<j} p_{\lambda_{ij}}(\sigma^{ij}), \tag{24}$$

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 penalty corresponds to $p_\lambda(|x|) = \lambda|x|$. Such an approach will inherit many desirable computational and statistical properties of LASSO and its many improved variants (Efron et al., 2004; Rocha, Zhao and Yu, 2008; Fan and Lv, 2010, Section 3.5).

Some early attempts at inducing sparsity in the precision matrix are Bilmes (2000), Smith and Kohn (2002), Wu and Pourahmadi (2003) and Levina, Rothman and Zhu (2008) who, for a fixed order of the variables in Y, use a parametrization of the precision matrix in terms of the modified Cholesky decomposition (15). Covariance selection priors and AIC were used to promote sparsity in T. Huang et al. (2006) proposed a covariance selection estimator by adding to the normal loglikelihood the LASSO penalty on the off-diagonal entries of T, and cross-validation was used to select a common regularization parameter; see also Huang, Liu and Liu (2007) and Levina, Rothman and Zhu (2008) for some improvements. Bickel and Levina (2008a) provide conditions ensuring consistency in the operator norm for the 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 row of T and the penalties are inversely proportional to the prediction variance σ_t^2 of the tth regression. Extensive simulations were used to compare the performance of their method with others, including the sample covariance matrix, 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 worked 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 banding 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.

Two disadvantages of imposing 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, Banerjee, El Ghaoui and d'Aspremont (2008), Yuan and Lin (2007) and Friedman, Hastie and Tibshirani (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) used 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, page 496).

To date, the fastest available algorithm is the graphical lasso (glasso), proposed by Friedman, Hastie and Tibshirani (2008). It relies on the equivalence of the Banerjee, El Ghaoui and d'Aspremont (2008) blockwise interior point procedure and recursively solving and updating a series of LASSO regression problems using the coordinate descent algorithm for LASSO. Fortunately, the sparse covariance estimator from the graphical LASSO is guaranteed to be positive definite. This important property follows from a result due to Banerjee, El Ghaoui and d'Aspremont (2008) showing that if the iterative procedure is initialized with a positive-definite matrix, then the subsequent iterates remain positive definite.

The sparse pseudo-likelihood inverse covariance estimation (SPLICE) algorithm of Rocha, Zhao and Yu (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 (7) and (9). 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 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 (16), 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 $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_p)'$. 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, Zhao and Yu (2008) and Friedman, Hastie and Tibshirani (2010) by using a pseudo-likelihood function which in the normal case amounts to pretending that the $\mathrm{Cov}(\tilde{\varepsilon})$ is $\tilde{D}^2$. To this pseudo-loglikelihood function, they add the symmetry constraints (8) 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 estimators 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). Sparsistency refers 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 factor. 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 the 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 a 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 and tapering 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. The most extreme case of this occurs in time series analysis where one has to work with only a single (long) realization ($n = 1$). The requirement of stationarity reduces 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 which further reduce the number of parameters are the prototypes of banding a covariance/precision matrix (Bickel and Levina, 2004; Wu and Pourahmadi, 2009; McMurry and Politis, 2010).

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

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

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

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

This occurs, for example, if $y_1, y_2, \ldots,$ form 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.

Banding is a special case of tapering which replaces 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. Banding corresponds to using $R = r_{ij} = (\mathbf{1}(|i-j| \le k)$, 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), Huang et al. (2006) and 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, Zhang and Zhou (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, Zhang and Zhou (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 (2008a, 2008b) and Rothman, Levina and Zhu (2009).

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

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

Thus, 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; Friedman, Hastie and Tibshirani, 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 $\frac{\log p}{n} \to 0$. 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

Heuristically, there is an implicit equivalence between regularization and Bayesian estimation in statistics. This can be seen by suitable exponentiation of the penalty term in (24) and viewing it as a prior on the parameter space, or conversely by viewing a prior as a means of imposing constraints on the parameters.

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, Le and Zidek (1994). However, the success of Bayesian computation and Markov Chain Monte Carlo (MCMC) in the late 1980s did open up the possibility of using more flexible and elaborate nonconjugate priors for covariance matrices; see Leonard and Hsu (1992), Yang and Berger (1994), Daniels and Kass (1999) and Hoff (2009). We present a 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, Le and Zidek (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 Liechty, Liechty and Müller [(2004), page 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

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

where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ are the ordered eigenvalues of Σ and c is a constant. Yang and Berger [(1994), page 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 involves 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, Massam and Carvalho (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 do as well as those from the 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); see Liu (1993). It is shown by Brown, Le and Zidek (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: $Y = (Z_1, \ldots, Z_k)'$, and writing its joint density as the product of a sequence of conditionals:

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

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 Garthwaite 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, that is, 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 uses of variance-correlation decomposition in Bayesian covariance estimation seems to be due to Barnard, McCulloch and Meng (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. The latter is capable of inducing uniform $(-1, 1)$ priors on the entries ρ_{ij} of the correlation matrix R; 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

$$f_p(\Sigma|\nu) = c|\Sigma|^{-(1/2)(\nu+p+1)} \exp(-\tfrac{1}{2} \operatorname{tr} \Sigma^{-1}).$$

It turns out that

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

where 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 ρ_{ij}, $i \neq j$, is obtained as

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

The latter can be viewed as a Beta $(\frac{\nu-p+1}{2}, \frac{\nu-p+1}{2})$ 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 $f(\rho_{ij}|\nu)$, that is, 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, Liechty and Müller (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.

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, Jenkins and Reinsel, 1994; Pourahmadi, 2001, Chapter 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, Jenkins and Reinsel, 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, we focus on reparameterizing the correlation matrix $R = (\rho_{ij})$ in terms of entries of a simpler symmetric matrix $\Pi = (\pi_{ij})$, where $\pi_{ii} = 1$ and for $i < j$, π_{ij} is the *partial autocorrelation* between y_i and y_j adjusted for the *intervening* (not the remaining) variables. More precisely, $\pi_{i,i+1} = \rho_{i,i+1}, i = 1, \ldots, p - 1$, are the lag-1 correlations and for $j - i \geq 2$, $\pi_{ij} = \rho_{ij|i+1,\ldots,j-1}$ in the notation of Anderson (2003), page 41. Note that, unlike R, and the matrix of full partial correlations (ρ^{ij}) constructed from Σ^{-1} in Section 2.2.2, Π has a much 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 $\tilde{\Pi}$ where its off-diagonal entries take values in the entire real line $(-\infty, +\infty)$.

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), page 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,

$$|\Sigma| = \left(\prod_{i=1}^{p} \sigma_{ii}\right) \prod_{i=2}^{p} \prod_{j=1}^{i-j} (1 - \pi_{ij}^2), \tag{25}$$

has been rediscovered 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), equation (25).

The identity (25) 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. It is used in Daniels and Pourahmadi (2009) to introduce priors for the Bayesian analysis of correlation matrices. These papers employ inde-

pendent 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

We have reviewed progress in covariance estimation for low- and high-dimensional data, 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. Fortunately, mostly due to the computational-algorithmic advances centered around LASSO, the high-dimensionality challenge in covariance estimation has been become manageable, 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 problem of removing the positive-definiteness constraint 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.

ACKNOWLEDGMENTS

Research supported in part by the NSF Grants DMS-05-05696 and DMS-09-06252. Comments from the Associate Editor and two referees have greatly improved the presentation, focus and scope of the paper.

REFERENCES

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1** 135–141. MR0331612

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

Barnard, J., McCulloch, R. and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. MR1804544

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

Bickel, P. J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes,' and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040

Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

Bilmes, J. A. (2000). Factored sparse inverse covariance matrices. In *IEEE International Conference on Accoustics, Speech and Signal Processing (Istanbul, Turkey)* **2** II1009–II1012.

Boik, R. J. (2002). Spectral models for covariance matrices. *Biometrika* **89** 159–182. MR1888370

Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66** 1069–1077.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ. MR1312604

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, eds.) 77–92. Wiley, Chichester. MR1309689

Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

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. Med.* **20** 1461–1467.

Carroll, R. J. (2003). Variances are not always nuisance parameters. *Biometrics* **59** 211–220. MR1987387

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, New York. MR1014890

Chang, C. and Tsay, R. S. (2010). Estimation of covariance matrix via the sparse Cholesky factor with lasso. *J. Statist. Plann. Inference* **140** 3858–3873. MR2674171

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59** 762–769. MR2025100

Chiu, T. Y. M., Leonard, T. and Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *J. Amer. Statist. Assoc.* **91** 198–210. MR1394074

Cressie, N. A. C. (1993). *Statistics for Spatial Data,* rev ed. Wiley, New York. MR21239641

DANIELS, M. J. (2005). A class of shrinkage priors for the dependence structure in longitudinal data. *J. Statist. Plann. Inference* **127** 119–130. MR2103028

DANIELS, M. J. and HOGAN, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Monographs on Statistics and Applied Probability* **109**. Chapman & Hall/CRC, Boca Raton, FL. MR2459796

DANIELS, M. J. and KASS, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Amer. Statist. Assoc.* **94** 1254–1263. MR1731487

DANIELS, M. J. and KASS, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57** 1173–1184. MR1950425

DANIELS, M. J. and POURAHMADI, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89** 553–566. MR1929162

DANIELS, M. J. and POURAHMADI, M. (2009). Modeling covariance matrices via partial autocorrelations. *J. Multivariate Anal.* **100** 2352–2363. MR2560376

DÉGERINE, S. and LAMBERT-LACROIX, S. (2003). Partial autocorrelation function of a nonstationary time series *J. Multivariate Anal.* **89** 135–147.

DEMPSTER, A. (1972). Covariance selection models. *Biometrics* **28** 157–175.

DEY, D. K. and SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13** 1581–1591. MR0811511

DIGGLE, P., LIANG, K. Y., ZEGER, S. L. and HEAGERTY, P. J. (2002). *Analysis of Longitudinal Data,* 2nd ed. Clarendon Press, Oxford.

EAVES, D. and CHANG, T. (1992). Priors for ordered conditional variance and vector partial correlation. *J. Multivariate Anal.* **41** 43–55. MR1156680

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407–499. MR2060166

EL KAROUI, N. (2008a). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

EL KAROUI, N. (2008b). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Statist.* **3** 521–541.

FAN, J., HUANG, T. and LI, R. (2007). Analysis of longitudinal data with semiparametric estimation of convariance function. *J. Amer. Statist. Assoc.* **102** 632–641. MR2370857

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G. and MOLENBERGHS, G., eds. (2009). *Longitudinal Data Analysis.* CRC Press, Boca Raton, FL. MR1500110

FLURY, B. (1988). *Common Principal Components and Related Multivariate Models.* Wiley, New York. MR0986245

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 Univ.

FURRER, R. and BENGTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.* **98** 227–255. MR2301751

GABRIEL, K. R. (1962). Ante-dependence analysis of an ordered set of variables. *Ann. Math. Statist.* **33** 201–212. MR0145611

GARTHWAITE, P. H. and AL-AWADHI, S. A. (2001). Non-conjugate prior distribution assessment for multivariate normal sampling. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 95–110. MR1811993

GOLUB, G. H. and VAN LOAN, C. F. (1989). *Matrix Computations*, 2nd ed. *Johns Hopkins Series in the Mathematical Sciences* **3**. Johns Hopkins Univ. Press, Baltimore, MD. MR1002570

HAFF, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8** 586–597. MR0568722

HAFF, L. R. (1991). The variational form of certain Bayes estimators. *Ann. Statist.* **19** 1163–1190. MR1126320

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR2722294

HOFF, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *J. Roy. Statist. Soc. Ser. B* **71** 971–992.

HOFF, P. D. and NIU, X. (2009). A covariance regression model. Technical report, Univ. Washington.

HUANG, J. Z., LIU, L. and LIU, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *J. Comput. Graph. Statist.* **16** 189–209. MR2345752

HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. MR2277742

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. Probab.* **I** 361–379. Univ. California Press, Berkeley. MR0133191

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. *J. Statist. Plann. Inference* **81** 95–111.

JOE, H. (2006). Generating random correlation matrices based on partial correlations. *J. Multivariate Anal.* **97** 2177–2189. MR2301633

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

JONES, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22** 389–395. MR0585635

JONES, M. C. (1987). Randomly choosing parameters from the stationarity and invertibility region of autoregressive-moving average models. *J. Roy. Statist. Soc. Ser. C* **36** 134–138. MR0897452

JONG, J.-C. and KOTZ, S. (1999). On a relation between principal components and regression analysis. *Amer. Statist.* **53** 349–351. MR1728916

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. *J. Amer. Statist. Assoc.* **103** 145–155.

KUROWICKA, D. and COOKE, R. (2003). A parameterization of positive definite matrices in terms of partial correlation vines. *Linear Algebra Appl.* **372** 225–251. MR1999149

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LEDOIT, O., SANTA-CLARA, P. and WOLF, M. (2003). Flexible multivariate GARCH modeling with an application to international stock markets. *Rev. Econom. Statist.* **85** 735–747.

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339

LENG, C., ZHANG, W. and PAN, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *J. Amer. Statist. Assoc.* **105** 181–193. MR2656048

LEONARD, T. and HSU, J. S. J. (1992). Bayesian inference for a covariance matrix. *Ann. Statist.* **20** 1669–1696. MR1193308

LESAGE, J. P. and PACE, R. K. (2007). A matrix exponential spatial specification. *J. Econometrics* **140** 190–214. MR2395921

LEUNG, P. L. and MUIRHEAD, R. J. (1987). Estimation of parameter matrices and eigenvalues in MANOVA and canonical correlation analysis. *Ann. Statist.* **15** 1651–1666. MR0913580

LEVINA, E., ROTHMAN, A. and ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Statist.* **2** 245–263. MR2415602

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

LIECHTY, J. C., LIECHTY, M. W. and MÜLLER, P. (2004). Bayesian correlation estimation. *Biometrika* **91** 1–14. MR2050456

LIN, T. I. (2011). A Bayesian inference in joint modelling of location and scale parameters of the t distribution for longitudinal data. *J. Statist. Plann. Inference* **141** 1543–1553.

LIN, S. P. and PERLMAN, M. D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. In *Multivariate Analysis VI (Pittsburgh, PA, 1983)* 411–429. North-Holland, Amsterdam. MR0822310

LIN, T.-I. and WANG, Y.-J. (2009). A robust approach to joint modeling of mean and scale covariance for longitudinal data. *J. Statist. Plann. Inference* **139** 3013–3026. MR2535179

LIU, C. (1993). Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. *J. Multivariate Anal.* **46** 198–206. MR1240420

LIU, X. and DANIELS, M. J. (2006). A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization. *J. Comput. Graph. Statist.* **15** 897–914. MR2297634

MCMURRY, T. L. and POLITIS, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *J. Time Series Anal.* **31** 471–482. MR2732601

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

PAN, J. and MACKENZIE, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika* **90** 239–244. MR1966564

PENG, J., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

PINHEIRO, J. D. and BATES, D. M. (1996). Unconstrained parameterizations for variance–covariance matrices. *Stat. Comput.* **6** 289–366.

POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786

POURAHMADI, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87** 425–435. MR1782488

POURAHMADI, M. (2001). *Foundations of Time Series Analysis and Prediction Theory*. Wiley, New York. MR1849562

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. In *Correlated Data Modelling 2004* (D. Gregori et al., eds.) 41–64. FrancoAngeli, Milan, Italy.

POURAHMADI, M. and DANIELS, M. J. (2002). Dynamic conditionally linear mixed models for longitudinal data. *Biometrics* **58** 225–231. MR1891383

QUENOUILLE, M. H. (1949). Approximate tests of correlation in time-series. *J. Roy. Statist. Soc. Ser. B* **11** 68–84. MR0032176

RAJARATNAM, B., MASSAM, H. and CARVALHO, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.* **36** 2818–2849. MR2485014

RAMSEY, F. L. (1974). Characterization of the partial autocorrelation function. *Ann. Statist.* **2** 1296–1301. MR0359219

ROCHA, G. V., ZHAO, P. and YU, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). Technical Report 759, Dept. Statistics, Univ. California, Berkeley.

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97** 539–550. MR2672482

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

ROY, J. (1958). Step-down procedure in multivariate analysis. *Ann. Math. Statist.* **29** 1177–1187. MR0100938

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York. MR1190470

SMITH, M. and KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.* **97** 1141–1153. MR1951266

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Probab.* **I** 197–206. Univ. California Press, Berkeley. MR0084922

STEIN, C. (1975). Estimation of a covariance matrix. In *Rietz Lecture. 39th Annual Meeting IMS. Atlanta, Georgia*.

SZATROWSKI, T. H. (1980). Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances. *Ann. Statist.* **8** 802–810. MR0572623

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

WAGAMAN, A. S. and LEVINA, E. (2009). Discovering sparse covariance structures with the Isomap. *J. Comput. Graph. Statist.* **18** 551–572.

WARTON, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J. Amer. Statist. Assoc.* **103** 340–349. MR2394637

WERMUTH, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.* **75** 963–972. MR0600984

WITTEN, D. M. and TIBSHIRANI, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *J. Roy. Statist. Soc. Ser. B* **71** 615–636.

WOLD, H. O. A. (1960). A generalization of causal chain models. *Econometrica* **28** 443–463. MR0120036

WONG, F., CARTER, C. K. and KOHN, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90** 809–830. MR2024759

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

WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844. MR2024760

WU, W. B. and POURAHMADI, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statist. Sinica* **19** 1755–1768. MR2589209

YANG, R.-Y. and BERGER, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22** 1195–1211. MR1311972

YUAN, M. and HUANG, J. Z. (2009). Regularized parameter estimation of high dimensional t distribution. *J. Statist. Plann. Inference* **139** 2284–2292. MR2507990

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

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. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **226** 267–298.

ZIMMERMAN, D. L. (2000). Viewing the correlation structure of longitudinal data through a PRISM. *Amer. Statist.* **54** 310–318. MR1815571

ZIMMERMAN, D. L. and NÚÑEZ-ANTÓN, V. (2001). Parametric modelling of growth curve data: An overview (with discussion). *Test* **10** 1–73. MR1856193

ZIMMERMAN, D. L. and NÚÑEZ-ANTÓN, V. A. (2010). *Antedependence Models for Longitudinal Data. Monographs on Statistics and Applied Probability* **112**. CRC Press, Boca Raton, FL. MR2640722

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527