Biometrika (2010), **97**, 3, *pp*. 539–550
© 2010 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asq022
Advance Access publication 26 May 2010

A new approach to Cholesky-based covariance regularization in high dimensions

BY ADAM J. ROTHMAN, ELIZAVETA LEVINA AND JI ZHU

Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
ajrothma@umich.edu elevina@umich.edu jizhu@umich.edu

SUMMARY

In this paper we propose a new regression interpretation of the Cholesky factor of the covariance matrix, as opposed to the well-known regression interpretation of the Cholesky factor of the inverse covariance, which leads to a new class of regularized covariance estimators suitable for high-dimensional problems. Regularizing the Cholesky factor of the covariance via this regression interpretation always results in a positive definite estimator. In particular, one can obtain a positive definite banded estimator of the covariance matrix at the same computational cost as the popular banded estimator of Bickel & Levina (2008b), which is not guaranteed to be positive definite. We also establish theoretical connections between banding Cholesky factors of the covariance matrix and its inverse and constrained maximum likelihood estimation under the banding constraint, and compare the numerical performance of several methods in simulations and on a sonar data example.

Some key words: Cholesky decomposition; High-dimensional data; Large p small n; Lasso; Sparsity.

1. INTRODUCTION

Statistical inference for high-dimensional data has become increasingly necessary in recent years. Advances in computing have made high-dimensional data analysis possible in a number of important applications, including spectroscopy, functional magnetic resonance imaging, text retrieval, gene arrays, climate studies and imaging. Many multivariate data analysis techniques applied to high-dimensional data require an estimate of the covariance matrix or its inverse; however, traditional estimation by the sample covariance matrix is known to perform poorly when there are more variables than observations; see Johnstone (2001) and references therein for a detailed discussion. A number of alternative estimators have been proposed for high-dimensional problems, many of which exploit sparsity assumptions about the covariance matrix or its inverse.

The problems of estimating the covariance matrix and its inverse are usually considered separately in this context, since in high dimensions inversion is costly, noisy and it does not preserve sparsity. When the goal is to estimate the inverse covariance matrix, also known as the precision or concentration matrix, a popular method is to add the lasso penalty on the entries of the inverse covariance matrix to the normal likelihood (d'Aspremont et al., 2008; Yuan & Lin, 2007; Rothman et al., 2008; Friedman et al., 2008), extended to more general nonconvex penalties by Lam & Fan (2009) and to pseudolikelihood in a 2008 University of California at Berkeley technical report by Rocha, Zhao and Yu. A Bayesian approach for introducing sparsity in the inverse via a sparse prior was proposed by Wong et al. (2003).

Another large class of estimators relies on the assumption that variables have a natural ordering, and those far apart in the ordering have small partial correlations. There are many applications that fall in this class, such as longitudinal data and spectroscopy, and exploiting the natural ordering present in the data in such cases leads to improved performance. The inverse estimators in this case usually rely on the modified Cholesky decomposition of the inverse covariance matrix, which is described in § 2. This decomposition has a nice regression interpretation to which regularization can be applied more easily (Wu & Pourahmadi, 2003; Huang et al., 2006; Bickel & Levina, 2008b; Levina et al., 2008). A Bayesian approach using a sparse prior on the Cholesky factor of the inverse was proposed by Smith & Kohn (2002).

If the covariance matrix, rather than its inverse, is of interest, a simple way to improve on the sample covariance, both theoretically and in practice, is to threshold small elements to zero (Bickel & Levina, 2008a; El Karoui, 2008; Rothman et al., 2009), which does not require ordered variables. In the ordered variable case, assuming those far apart in the ordering are only weakly correlated, a better option is to band or taper the sample covariance matrix (Bickel & Levina, 2004, 2008b; Furrer & Bengtsson, 2007). These simple approaches are attractive for problems in very high dimensions since they have a small computational cost; however, estimators in this class are not generally guaranteed to be positive definite, although some forms of tapering can guarantee positive semidefinite estimates, including the one proposed by Furrer & Bengtsson (2007). Alternatively, a positive definite constrained maximum likelihood estimator can be computed under the constraint enforcing any given pattern of zeros (Chaudhuri et al., 2007), but this algorithm is only applicable when there are fewer variables p than observations n.

In this paper we show that the modified Cholesky factor of the covariance matrix, rather than its inverse, also has a natural regression interpretation, and therefore all Cholesky-based regularization methods can be applied to the covariance matrix itself instead of its inverse to obtain a sparse estimator with guaranteed positive definiteness. As with all Cholesky-based regularization methods, this approach exploits the assumption of naturally ordered variables where variables far apart in the ordering tend to have small correlations. The simplest estimator in this new class is banding the covariance Cholesky factor. Unlike banding the sample covariance matrix itself, it is guaranteed to be positive definite, but still has the same low computational complexity. We also derive some theoretical properties of banded estimators, connecting sparsity in a matrix to sparsity in its Cholesky factor and connecting banding Cholesky factors to constrained maximum likelihood.

2. MODIFIED CHOLESKY DECOMPOSITION OF THE COVARIANCE MATRIX

Throughout the paper we assume that the data $X_1, \ldots, X_n$ are independent and identically distributed p-variate random vectors with population covariance matrix Σ and, without loss of generality, mean zero. Let $\hat{\Sigma}$ denote the sample covariance matrix, $\hat{\Sigma} = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$. As a tool for regularizing the inverse covariance matrix, Pourahmadi (1999) and Wu & Pourahmadi (2003) suggested using the modified Cholesky factorization of Σ^{-1}. For a mean zero random vector $X = (X^{(1)}, \ldots, X^{(p)})^{\mathrm{T}}$ with covariance matrix Σ, this factorization arises from regressing each variable $X^{(j)}$ on $X^{(j-1)}, \ldots, X^{(1)}$, that is, fitting regressions

$$X^{(j)} = \sum_{q=1}^{j-1}(-t_{jq})X^{(q)} + \epsilon^{(j)} = \hat{X}^{(j)} + \epsilon^{(j)},$$

where $\epsilon^{(j)}$ denotes the error term in regression j ($j = 2, \ldots, p$), and $\epsilon^{(1)} = X^{(1)}$. Let $\epsilon = (\epsilon^{(1)}, \ldots, \epsilon^{(p)})^{\mathrm{T}}$, let $D = \mathrm{var}(\epsilon)$ be the diagonal matrix of error variances and let $T = (t_{jq})$

denote the lower-triangular matrix containing regression coefficients with the opposite sign, with ones on the diagonal. Then writing $\epsilon = X - \hat{X} = TX$ and using the fact that the errors are uncorrelated, $D = \mathrm{var}(\epsilon) = \mathrm{var}(TX) = T\Sigma T^{\mathrm{T}}$, and thus $\Sigma^{-1} = T^{\mathrm{T}} D^{-1} T$. This decomposition transforms inverse covariance matrix estimation into a regression problem, and hence regularization approaches for regression can be applied. If these regressions are not regularized, the resulting estimate is simply $\hat{\Sigma}^{-1}$. Banding the Cholesky factor of the inverse refers to regularizing by only including the immediate k predecessors in the regression, $X^{(j-k)}, \ldots, X^{(j-1)}$, for some fixed k (Wu & Pourahmadi, 2003; Bickel & Levina, 2008b).

The modified Cholesky factorization of Σ itself can be obtained from a latent variable regression model. Let $\Sigma = LDL^{\mathrm{T}}$ be the modified Cholesky decomposition of Σ, where D is diagonal and L is lower triangular with ones on the diagonal. Let ϵ be a normal vector with independent components, $\epsilon \sim N_p(0, D)$. Then if we let $X = L\epsilon$, we have

$$\Sigma = \mathrm{var}\,(L\epsilon) = LDL^{\mathrm{T}}. \tag{1}$$

Our main interest here is in the regression interpretation. The vector ϵ is unobserved, but because L is lower triangular, we can think of (1) as a sequence of regressions, where each variable $X^{(j)}$ is regressed on the previous regression errors $\epsilon^{(j-1)}, \ldots, \epsilon^{(1)}$. For $j = 2, \ldots, p$, we have

$$X^{(j)} = \sum_{q=1}^{j-1} l_{jq}\epsilon^{(q)} + \epsilon^{(j)} = \tilde{X}^{(j)} + \epsilon^{(j)}. \tag{2}$$

The decompositions above apply to the population matrices; Pourahmadi (2007) briefly mentions this decomposition for the population, but does not discuss any implications for estimation. Let $\mathcal{X}$ denote an n by p data matrix, where each column $x_j \in \mathbb{R}^n$ is centred by its sample mean. For the first variable, we set $e_1 = x_1$. For $j = 2, \ldots, p$, let $l_j = (l_{j1}, \ldots, l_{j,j-1})^{\mathrm{T}}$, $Z_j = (e_1, \ldots, e_{j-1})$, and compute coefficients and the residual, respectively, as

$$\hat{l}_j = \arg\min_{l_j} \|x_j - Z_j l_j\|^2, \quad e_j = x_j - Z_j \hat{l}_j. \tag{3}$$

The variances are estimated as $\hat{d}_{jj} = n^{-1}\|e_j\|^2$. Let Z denote the n by p matrix of residuals from carrying out the regressions in (2) sequentially. Here we assume that $p < n$ to ensure that all model matrices are of full column rank; § 3 discusses the rank deficient case when $p \geqslant n$. Performing the regressions in (3) amounts to, for each $j = 2, \ldots, p$, orthogonally projecting the response x_j onto the span of $e_1, \ldots, e_{j-1}$ to estimate $\hat{l}_j$. After the last projection, we have an orthogonal basis $(e_1, \ldots, e_p)$, and the estimates $\hat{L}$ and $\hat{D}$. This algorithm is a scaled version of Gram–Schmidt orthogonalization of the data matrix $\mathcal{X}$ for computing its QR decomposition, where the upper triangular matrix R is restricted to have positive diagonal entries. The orthonormal matrix Q is the matrix Z with its column vectors scaled to have unit length and $R^{\mathrm{T}} = \hat{L}(n\hat{D})^{1/2}$. If all regressions are fitted by least squares, the resulting estimate recovers the sample covariance matrix: $\hat{\Sigma} = n^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{X} = n^{-1}R^{\mathrm{T}}R = \hat{L}\hat{D}\hat{L}^{\mathrm{T}}$.

3. Regularized estimation of the Cholesky factor L

3·1. *Banding the Cholesky factor*

The simplest way to introduce sparsity in the Cholesky factor L is to estimate only the first k subdiagonals of L and set the rest to zero. This approach for the inverse was proposed by Wu & Pourahmadi (2003) and Bickel & Levina (2008b). In our case, each variable x_j is regressed on the k previous residuals $e_{j-k}, \ldots, e_{j-1}$, for all $j = 2, \ldots, p$. The index $j - k$ is understood

to mean $\max(1, j-k)$. Let $l_j^{(k)} = (l_{j,j-k}, \ldots, l_{j,j-1})^{\mathrm{T}}$ and $Z_j^{(k)} = (e_{j-k}, \ldots, e_{j-1})$. Then we compute

$$\hat{l}_j^{(k)} = \arg\min_{l_j^{(k)}} \left\| x_j - Z_j^{(k)} l_j^{(k)} \right\|^2, \quad e_j = x_j - Z_j^{(k)} \hat{l}_j^{(k)}. \tag{4}$$

In each regression, the design matrix $Z_j^{(k)}$ has orthogonal columns, which allows (4) to be solved with at most k univariate regressions. Hence the computational cost of banding the Cholesky factor in this manner is $O(kpn)$, the same order as banding the sample covariance matrix without the Cholesky decomposition. To ensure that design matrices are of full rank, the banding parameter k must be less than $\min(n-1, p)$. For sparse matrices, it is usually not necessary to search for values of $k \geqslant n$, since the optimal k is much smaller than n. We describe how to choose k in § 4. If we do need to perform regressions when $k \geqslant n-1$, we use a generalized inverse of $Z_j^{(k)\mathrm{T}} Z_j^{(k)}$ for fitting ordinary least squares, in which case the resulting estimator is positive semidefinite.

Although each design matrix $Z_j^{(k)}$ has orthogonal columns, all of the residual vectors $e_1, \ldots, e_p$ are not necessarily mutually orthogonal; e_j and $e_{j'}$ are only guaranteed to be orthogonal if $|j - j'| \leqslant k$.

3·2. *Connection to constrained maximum likelihood*

Given that a Cholesky-based banded estimator is always positive definite, it is natural to ask whether it coincides with the maximum likelihood estimator under the banded constraint. Here we show that, somewhat surprisingly, banding the Cholesky factor of the inverse coincides with constrained maximum likelihood, and banding the Cholesky factor of the covariance matrix itself does not. First, we establish some relationships between zero patterns in positive definite matrices and their Cholesky factors. The proofs for the following propositions are given in the Appendix.

PROPOSITION 1. *Let Σ and Ω be positive definite matrices with modified Cholesky decompositions $\Sigma = LDL^{\mathrm{T}}$ and $\Omega = T^{\mathrm{T}} D^{-1} T$, where L and T are both lower triangular. Then for any row i and $c(i) < i$, $\sigma_{i1} = \cdots = \sigma_{i,c(i)} = 0$ if and only if $l_{i1} = \cdots = l_{i,c(i)} = 0$; and for any column j and $r(j) > j$, $\omega_{p,j} = \cdots = \omega_{r(j),j} = 0$ if and only if $t_{p,j} = \cdots = t_{r(j),j} = 0$.*

Proposition 1 is a simple matrix property, but we are not aware of a source to cite, so we give a proof in the Appendix for completeness. Proposition 1 implies that a covariance Cholesky factor with banded rows of arbitrary band lengths, not necessarily all the same, corresponds to a covariance matrix with banded rows of the same band lengths. On the other hand, the modified Cholesky factor of the inverse covariance matrix T with arbitrary column band lengths corresponds to an inverse covariance matrix Ω with the same column band lengths. In particular, the Cholesky factor of either the covariance matrix or the inverse is k-banded if and only if the corresponding matrix itself is k-banded.

PROPOSITION 2. *Banding the modified Cholesky factor T of the inverse covariance matrix Ω maximizes the normal likelihood subject to the banded constraint, $\omega_{ij} = 0$ for $|i-j| > k$.*

PROPOSITION 3. *Banding the modified Cholesky factor L of the covariance matrix Σ does not maximize the normal likelihood under the constraint that $\sigma_{ij} = 0$ for $|i-j| > k$.*

Intuitively, the constrained maximum likelihood result holds for the inverse only because the inverse is the canonical parameter of the normal likelihood. The constrained maximum likelihood estimator of the covariance matrix can be computed by the algorithm proposed by Chaudhuri et al. (2007), but this algorithm only works for $p < n$. We are not aware of suitable

constrained maximum likelihood estimation algorithms for $p > n$, which makes banding the Cholesky factor a more attractive option for computing a positive definite estimator for large p. In § 4, we briefly compare the numerical performance of banding the Cholesky factor of covariance to the constrained maximum likelihood estimator when $p < n$, and find that the two estimators are in practice very close, even though they differ theoretically.

3·3. *The penalized regression approach*

Once we have the regression interpretation (2), all penalty-based approaches proposed for regularizing the inverse become equally applicable to the covariance matrix itself. In general, we can estimate the Cholesky factor by

$$\hat{l}_j = \arg\min_{l_j} \{\|x_j - Z_j l_j\|^2 + P_\lambda(l_j)\}.$$

Penalty functions P_λ that encourage sparsity in the coefficient vector l_j are of particular interest. Huang et al. (2006) applied the lasso penalty in the inverse covariance Cholesky estimation problem, and here we can analogously use

$$P_\lambda^L(l_j) = \lambda \sum_{t=1}^{j-1} |l_{jt}|.$$

The lasso penalty function can result in zeros in arbitrary locations in the Cholesky factor, which may or may not lead to any zeros in the resulting covariance matrix. To impose additional structure, Levina et al. (2008) proposed the nested lasso penalty, which in our context is given by

$$P_\lambda^{NL}(l_j) = \lambda \left(|l_{j,j-1}| + \frac{|l_{j,j-2}|}{|l_{j,j-1}|} + \frac{|l_{j,j-3}|}{|l_{j,j-2}|} + \cdots + \frac{|l_{j,1}|}{|l_{j,2}|} \right), \tag{5}$$

where 0/0 is defined as 0. This penalty imposes the restriction that $l_{jt} = 0$ if $l_{j,t+1} = 0$. By Proposition 1, this means that all the zeros estimated in the Cholesky factor of covariance $\hat{L}$ will be preserved in $\hat{\Sigma}$. This is not the case in the inverse Cholesky decomposition for which this penalty was originally proposed by Levina et al. (2008), although some zeros are preserved in that case as well. In practice, Levina et al. (2008) recommend using a slightly modified version of (5), where the first term is divided by the univariate regression coefficient from regressing x_j on e_{j-1} alone, to address a potential difference of scales, which is the version we used in simulations. Both lasso and nested lasso have much higher computational cost than banding, and are not appropriate for very large p; however, the additional flexibility of the sparsity structure of nested lasso's variable bandwidths may work well in some cases.

4. Numerical results

4·1. *Simulation settings*

Our simulation study compares the performance of all the covariance estimators discussed in § 3, banding the sample covariance matrix directly (Bickel & Levina, 2008b), which is not positive definite, and, as a benchmark, the shrinkage estimator of Ledoit & Wolf (2003), which does not depend on the order of variables. The Ledoit–Wolf estimator is a linear combination of the identity matrix and the sample covariance matrix, with coefficients optimal in a certain sense; it does not introduce any sparsity.

We consider the following three covariance structures: Σ_1 has entries $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0{\cdot}7$, Σ_2 has entries $\sigma_{ij} = I(i = j) + 0{\cdot}4\, I(|i - j| = 1) + 0{\cdot}2 I(2 \leqslant |i - j| \leqslant 3) + 0{\cdot}1 I(|i - j| = 4)$

and Σ_3 has entries $\sigma_{ij} = I(i = j) + 0{\cdot}5I(i \neq j)$. The first-order autoregressive model Σ_1 has a dense Cholesky factor, but its entries decay as one moves away from the diagonal. We only report results for $\rho = 0{\cdot}7$, but the same pattern is observed over the whole range of ρ. The fourth-order moving average model Σ_2 is a banded matrix with $k = 4$, and therefore its Cholesky factor is also 4-banded. The model Σ_1 was considered by Bickel & Levina (2008b), and Σ_2 by Yuan & Lin (2007). Model Σ_3 is a full matrix, where introducing sparsity cannot improve estimation, and thus we expect the regularization methods to perform similarly to the covariance matrix.

We generate $n = 100$ training observations and another 100 independent validation observations from $N_p(0, \Sigma)$, with $p = 30$, 100, 200, 500 and 1000. Lasso and nested lasso were not run for $p \geqslant 500$ due to their high computational cost. Tuning parameters were selected by minimizing the Frobenius norm, $\|M\|_F^2 = \sum_{i,j} m_{ij}^2$, of the difference between the regularized estimate computed with the training observations and the sample covariance computed with the validation observations. The results are not sensitive to the choice of loss; we have also tested matrix 1-norm and matrix 2-norm losses and obtained very similar results, and we selected the Frobenius norm because it had a very slight edge in simulations and because there are general theoretical results justifying crossvalidation via Frobenius norm (Bickel & Levina, 2008a). The whole process was repeated 200 times.

To compare estimators, we used the operator norm, also known as the matrix 2-norm, $\|M\|^2 = \lambda_{\max}(MM^{\mathrm{T}})$, of the difference between the covariance estimator and the truth, $\Delta(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|$. This loss is commonly used to assess covariance estimators because convergence in this norm implies convergence of all eigenvectors and eigenvalues. Other losses such as Frobenius norm, matrix 1-norm and entropy loss are omitted to save space; they produce very similar results.

We also compute the true positive rate and true negative rate, defined respectively as

$$\mathrm{TPR}(\hat{\Sigma}, \Sigma) = \frac{\#\{(i, j) : \hat{\sigma}_{ij} \neq 0 \quad \text{and} \quad \sigma_{ij} \neq 0\}}{\#\{(i, j) : \sigma_{ij} \neq 0\}},$$

$$\mathrm{TNR}(\hat{\Sigma}, \Sigma) = \frac{\#\{(i, j) : \hat{\sigma}_{ij} = 0 \quad \text{and} \quad \sigma_{ij} = 0\}}{\#\{(i, j) : \sigma_{ij} = 0\}}.$$

The sample covariance has $\mathrm{TPR}(\hat{\Sigma}, \Sigma) = 1$, and a diagonal estimator has $\mathrm{TNR}(\hat{\Sigma}, \Sigma) = 1$.

4·2. *Results*

The averages and standard errors over 200 replications of the operator norm loss for the three models are given in Table 1. For models Σ_1 and Σ_2, where the true Cholesky factor is either banded or has entries decaying fast as one goes away from the diagonal, banding the Cholesky factor provides the best performance in every case. In particular, it outperforms banding the sample covariance directly, particularly in high dimensions, presumably due to its ability to enforce positive definiteness. Both banding methods outperform the Ledoit–Wolf estimator, which is not sparse at all, and lasso applied to the Cholesky factor, which cannot create a banded structure and loses sparsity in the matrix itself. The nested lasso does have the ability to create a banded structure in the Cholesky factor, but its extra flexibility, not needed for these models, leads to noisier estimates in this case. As expected, the margin by which sparse regularized estimators outperform nonsparse estimators, the sample and Ledoit–Wolf, is larger for the sparse population covariance Σ_2. For the full matrix Σ_3, introducing sparsity cannot help, and thus all sparse estimators, excluding nested lasso, perform similarly to the sample covariance. The Ledoit–Wolf estimator is very close to the sample covariance because one eigenvalue of Σ_3 is very large relative to others, which makes the coefficient of the sample covariance term very close to 1. Nested lasso has a large risk for $p = 200$ because it can only estimate up to $n - 1$ nonzeros in any

Table 1. *Averages and standard errors of the operator norm loss for the sample covariance, Ledoit–Wolf's estimator, the banded sample covariance, and regularization of the Cholesky factor of the covariance by banding, lasso, and nested lasso*

Model	p	Sample	Ledoit–Wolf	Sample Band.	Chol. Band.	Lasso	Nested lasso
	30	1·82 (0·03)	1·70 (0·02)	1·31 (0·02)	1·30 (0·02)	1·73 (0·02)	1·47 (0·02)
	100	4·10 (0·04)	3·10 (0·01)	1·61 (0·02)	1·61 (0·02)	3·53 (0·01)	1·83 (0·02)
Σ_1	200	6·59 (0·04)	3·83 (0·01)	1·77 (0·02)	1·76 (0·01)	3·91 (0·01)	1·97 (0·01)
	500	12·47 (0·04)	4·43 (0·00)	1·96 (0·02)	1·91 (0·01)	–	–
	1000	20·64 (0·04)	4·64 (0·00)	2·08 (0·02)	2·01 (0·01)	–	–
	30	1·44 (0·02)	1·14 (0·01)	0·76 (0·01)	0·74 (0·01)	1·24 (0·01)	0·87 (0·01)
	100	3·27 (0·02)	1·63 (0·00)	0·92 (0·01)	0·89 (0·01)	1·63 (0·00)	1·03 (0·01)
Σ_2	200	5·33 (0·02)	1·77 (0·00)	1·00 (0·01)	0·95 (0·01)	1·72 (0·00)	1·08 (0·01)
	500	10·37 (0·03)	1·84 (0·00)	1·09 (0·01)	1·06 (0·01)	–	–
	1000	17·58 (0·03)	1·85 (0·00)	1·17 (0·01)	1·14 (0·01)	–	–
	30	2·62 (0·07)	2·64 (0·07)	2·63 (0·07)	2·69 (0·07)	2·62 (0·07)	2·68 (0·07)
	100	8·83 (0·22)	8·86 (0·23)	8·83 (0·22)	8·88 (0·23)	8·82 (0·22)	8·86 (0·23)
Σ_3	200	17·63 (0·43)	17·85 (0·44)	17·63 (0·43)	17·73 (0·43)	17·62 (0·43)	68·11 (0·67)
	500	44·58 (1·25)	44·77 (1·29)	44·58 (1·25)	44·62 (1·25)	–	–
	1000	86·62 (2·43)	87·13 (2·47)	86·62 (2·43)	86·69 (2·43)	–	–

Sample Band., banded sample covariance; Chol. Band., regularization of the Cholesky factor of the covariance by banding.

Table 2. *Averages and standard errors of true positive/true negative percentages for* Σ_2*, based on* 200 *replications*

p	Sample banding	Cholesky banding	Lasso	Nested lasso
30	87·47 (0·84)/100·00 (0·00)	90·19 (0·83)/99·69 (0·12)	99·71 (0·05)/3·90 (0·27)	94·07 (0·32)/89·06 (0·46)
100	88·31 (0·87)/100·00 (0·00)	93·72 (0·76)/99·99 (0·01)	90·38 (0·14)/37·35 (0·21)	93·82 (0·20)/97·25 (0·08)
200	87·22 (0·88)/100·00 (0·00)	93·92 (0·76)/100·00 (0·00)	90·59 (0·10)/34·93 (0·14)	94·11 (0·14)/98·69 (0·03)
500	85·42 (0·87)/100·00 (0·00)	96·51 (0·61)/100·00 (0·00)	–	–
1000	85·77 (0·88)/100·00 (0·00)	98·13 (0·47)/100·00 (0·00)	–	–

row of the Cholesky factor in a band extending from the diagonal. The covariance Cholesky factor of Σ_3 is $l_{ij} = (j+1)^{-1} I(i > j) + I(i = j)$ and thus the true row coefficients are the smallest in a band extending from the diagonal. The lasso is also only able to estimate up to $n - 1$ nonzeros in any row of the Cholesky factor; however, these nonzeros could be estimated in any location, and estimating most of the first $n - 1$ coefficients in a row as nonzero is enough to come close to the sample covariance in this model.

The banded maximum likelihood estimator was also computed using the algorithm of Chaudhuri et al. (2007) for $p = 30$, since the algorithm is only applicable when $p < n$. Its loss values are 1·32 (0·02) for Σ_1, 0·74 (0·01) for Σ_2 and 2·77 (0·07) for Σ_3, which are essentially the same as those for Cholesky banding for $p = 30$.

For the sparse matrix Σ_2, we also report true positive and true negative rates of estimating zeros in Table 2. Both Cholesky banding and sample covariance banding have nearly perfect true negative rates, but banding the Cholesky factor has a better true positive rate than for banding the sample, which means that banding the sample tends to set more subdiagonals to zero than necessary. The lasso method has a low true negative rate because zeros in the Cholesky factor are not preserved in the matrix, and the nested lasso does reasonably well on both but not as well as Cholesky banding.

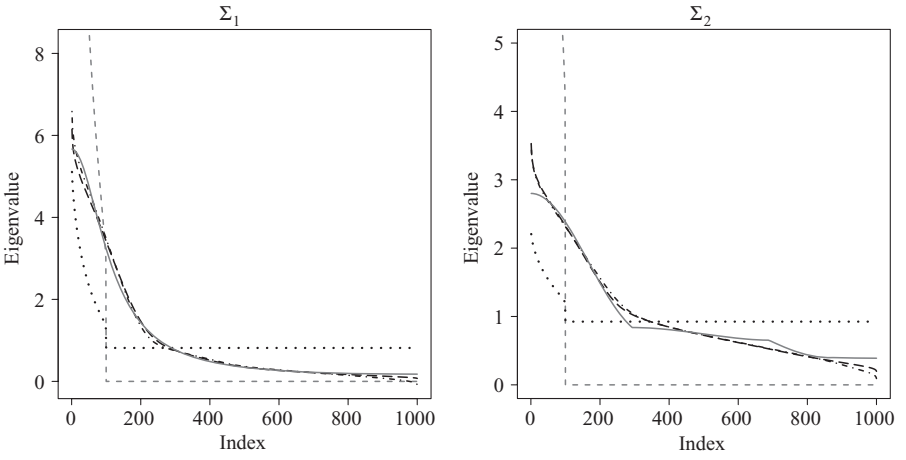

Fig. 1. Scree plots for the sample covariance (grey dashes), Ledoit–Wolf (dots), banding the sample covariance (dash-dot), Cholesky banding (black dashes), and the truth (solid) for $p = 1000$, averaged over 200 replications.

Table 3. *Percentage of positive definite banded sample realizations*

p	30	100	200	500	1000
Σ_1	61·5	16·5	3·0	0·0	0·0
Σ_2	100·0	100·0	99·5	98·0	98·0
Σ_3	98·0	4·0	0·0	0·0	0·0

In Fig. 1 we plot the average estimated eigenvalues in descending order for sample banding, Cholesky banding, the sample covariance and the Ledoit–Wolf estimator, as well as the true eigenvalues, for both models with $p = 1000$. Since $n = 100$, the sample covariance matrix only has 99 nonzero eigenvalues. Cholesky banding and sample banding perform similarly for both models, with Cholesky banding having a slight edge for the small eigenvalues. The banding methods outperform both the sample covariance and the Ledoit–Wolf estimator by a considerable amount, especially for larger eigenvalues. This is expected since the banding methods performed best under the operator norm loss, and the truth is banded or almost banded. For Σ_3, the plots are indistinguishable and are omitted to save space.

Since sample covariance banding does not necessarily produce a positive definite estimator, we also report the percentage of estimates that are positive definite in Table 3. It is clear that larger p and denser truth make it harder to keep positive definiteness.

5. SONAR DATA EXAMPLE

In this section we illustrate the effects of Cholesky banding and sample covariance banding on sonar data from the UCI machine learning data repository, available at http://www.ics.uci.edu/~mlearn/MLRepository.html. This dataset has 111 spectra from metal cylinders and 97 spectra from rocks, where each spectrum has 60 frequency band energy measurements. These spectra were measured at multiple angles for the same objects, but following previous analyses of the dataset, we assume independence of the spectra.

The top panel of Fig. 2 shows heatmaps of the absolute values of the sample correlation matrices for metal and rock, where we standardize the variables first to facilitate comparison for metal and rock spectra, which are on different scales. Both matrices show a general pattern of correlations decaying away from the diagonal, which makes banding a reasonable option.

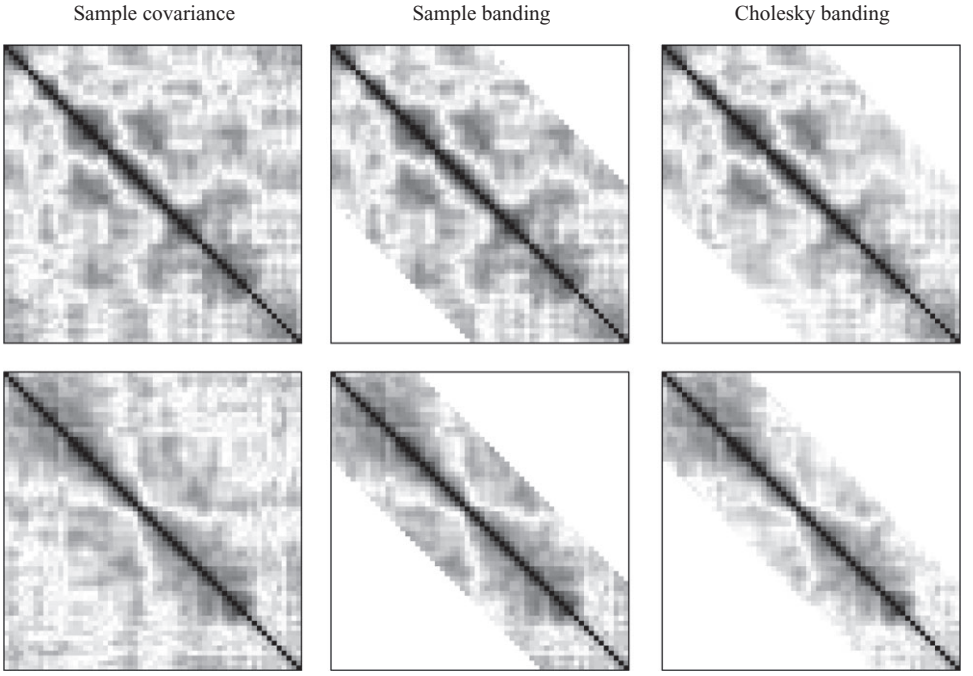

Fig. 2. Heatmaps of the absolute values of entries in the correlation matrix estimates, where a correlation of magnitude 0 is white and a correlation of magnitude 1 is black. The top row is for metal spectra and the bottom row is for rock spectra.

The banding parameter k for both banding methods was selected using the random-splitting scheme of Bickel & Levina (2008b), $\hat{k} = \arg\min_k N^{-1} \sum_{v=1}^{N} \|\hat{\Sigma}_{(k)}^{(v)} - \tilde{\Sigma}^{(v)}\|_F$, where $\hat{\Sigma}_{(k)}^{(v)}$ is the banded estimator with k bands computed on the training data, and $\tilde{\Sigma}^{(v)}$ is the sample covariance of the validation data. To obtain these training and validation sets, the data were split at random $N = 100$ times, with one-third of the sample used for training. For metal, Cholesky banding and sample banding both chose $\hat{k} = 31$ subdiagonals; for rock, Cholesky banding chose $\hat{k} = 17$ and sample banding chose $\hat{k} = 18$. Since these values are so close, for easier visual comparison we show both with $\hat{k} = 17$ for the rock spectra. The heatmaps of the absolute values of correlations from the banded estimators are shown in Fig. 2. We see that Cholesky banding shrinks the nonzero correlations whereas the sample banding does not, which is the property that allows Cholesky banding to achieve positive definiteness.

We also show eigenvalue plots for these estimators in Fig. 3(a) and (b). We see that the sample covariance has the most spread out eigenvalues, and the eigenvalues from Cholesky banding have the least spread, as we would expect.

We also compared the performance of the various estimators if they are used in quadratic discriminant analysis to discriminate between rock and metal. An observation x is classified as rock $j = 0$ or metal $j = 1$ using the rule, $G(x) = \arg\max_j \{\log|\hat{\Omega}_j|/2 - (x - \hat{\mu}_j)^{\mathrm{T}}\hat{\Omega}_j(x - \hat{\mu}_j)/2 + \log\hat{\pi}_j\}$, where $\hat{\pi}_j$ is the proportion of class j observations, $\hat{\mu}_j$ is the class j sample mean, and $\hat{\Omega}_j$ is the inverse covariance estimate for class j, all computed on the training data. More details can be found in Mardia et al. (1979). In addition to banding the Cholesky factor of covariance and of the inverse, we also added a diagonal estimator of the covariance matrix, which corresponds to the naive Bayes classifier. Banding the sample covariance was omitted because it is not invertible. Leave-one-out crossvalidation was used to estimate the testing error,

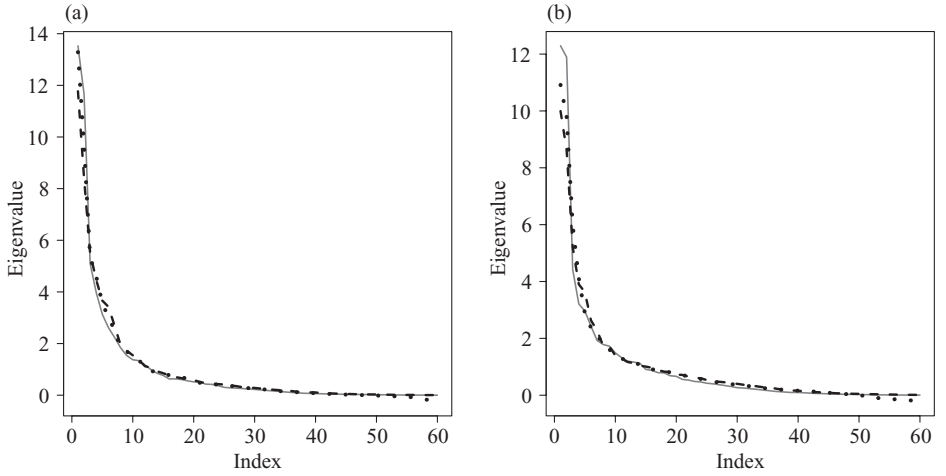

Fig. 3. Scree plots of the sample covariance (solid), sample banding (dots), and Cholesky banding (dashes) for the metal spectra in panel (a) and for the rock spectra in panel (b).

and the banding parameters were selected with 10 random splits with one-third of the data used for training, using Frobenius loss for covariance Cholesky banding and the validation likelihood for the inverse covariance Cholesky banding. The test errors were 24·0% for the sample covariance, 32·7% for naive Bayes, 20·2% for covariance Cholesky banding and 14·9% for inverse Cholesky banding. Both banding methods are substantially better than either estimating the whole dependency structure by the sample covariance or not estimating it at all with naive Bayes. We conjecture that the inverse Cholesky banding does better because it introduces sparsity directly in the inverse.

6. DISCUSSION

In terms of convergence rates, one would expect a convergence result analogous to the one for inverse Cholesky banding established by Bickel & Levina (2008b) to hold here as well, but this case presents substantial extra technical difficulties in analysis, because the errors used as predictors in the regressions required to compute the Cholesky factor are unobservable and have to be estimated by residuals. Nonetheless, we expect the method to be equally useful based on its good practical performance.

The regression representation of the covariance matrix and its inverse have obvious parallels with time series models for moving average and autoregressive processes, respectively. However, we do not fit a parametric model here, and do not assume stationarity, which would correspond to imposing a Toeplitz structure on the matrix, and thus fitting and model selection methods are very different from time series. As a rule of thumb in practice, if it is not clear from the problem whether it is preferable to regularize the covariance or the inverse, we would recommend fitting both and choosing the sparser estimate.

ACKNOWLEDGEMENT

We thank Richard Davis for pointing out the use of regression on residuals in time series, and Bala Rajaratnam for helpful discussions on sparse Cholesky factors. A. J. R. was supported in part by the Yahoo! PhD Student Fellowship, and E. L. and J. Z. were supported in part by grants from the National Science Foundation, U.S.A.

APPENDIX

Proof of Proposition 1. We prove the first claim only since the proof of the second one is very similar. From $\sigma_{ij} = \sum_{m=1}^{j} l_{im} l_{jm} d_{mm}$, it is obvious that $l_{i1} = \cdots = l_{i,c(i)} = 0$ implies $\sigma_{i1} = \cdots = \sigma_{i,c(i)} = 0$.

Now assume $\sigma_{i1} = \cdots = \sigma_{i,c(i)} = 0$ for some i. The formula for computing the modified Cholesky factorization L one column at a time, starting from the first column, is given by, for $i > j$ (Watkins, 1991),

$$d_{ii} = \sigma_{ii} - \sum_{m=1}^{i-1} l_{im}^2 d_{mm}, \quad l_{ij} = \frac{1}{d_{jj}} \left(\sigma_{ij} - \sum_{m=1}^{j-1} l_{im} l_{jm} d_{mm} \right). \tag{A1}$$

We proceed by induction: for the first column of L, $l_{i1} = \sigma_{i1}/\sigma_{11}$, hence $l_{i1} = 0$. Assuming that for a column $u < c(i)$ we have $l_{i1} = \cdots = l_{iu} = 0$, using (A1) gives,

$$l_{i,u+1} = \frac{1}{d_{u+1,u+1}} \left(\sigma_{i,u+1} - \sum_{m=1}^{u} l_{im} l_{u+1,m} d_{u+1,u+1} \right) = \frac{\sigma_{i,u+1}}{d_{u+1,u+1}},$$

which implies $l_{i,u+1} = 0$. □

Proof of Proposition 2. Let $\Omega_{(k)}$ be a symmetric positive definite matrix with k nonzero main subdiagonals, $\omega_{(k)ij} = 0$ for $|i - j| > k$. The negative normal loglikelihood up to a constant, as a function of the nonzero unique parameters in $\Omega_{(k)}$ is, $f(\Omega_{(k)}) = \operatorname{tr}(\hat{\Sigma}\Omega_{(k)}) - \log|\Omega_{(k)}|$. The k-banded constrained maximum likelihood estimator $\hat{\Omega}_{(k)}$ satisfies $\nabla f(\hat{\Omega}_{(k)}) = 0$. Let $T_{(k)}^{\mathrm{T}} D_{(k)}^{-1} T_{(k)} = \Omega_{(k)}$ be the modified Cholesky decomposition of $\Omega_{(k)}$. By Proposition 1, $t_{(k)ij} = 0$ for $|i - j| > k$. Let $g(T_{(k)}, D_{(k)}) \equiv f(T_{(k)}^{\mathrm{T}} D_{(k)}^{-1} T_{(k)})$, where g is a function of nonzero unique parameters in $(T_{(k)}, D_{(k)})$.

We continue by establishing that if $\nabla g(\hat{T}_{(k)}, \hat{D}_{(k)}) = 0$, then $\hat{T}_{(k)}^{\mathrm{T}} \hat{D}_{(k)}^{-1} \hat{T}_{(k)} = \hat{\Omega}_{(k)}$. Let $h(T_{(k)}, D_{(k)}) = T_{(k)}^{\mathrm{T}} D_{(k)}^{-1} T_{(k)}$. Denote the differential of h in the direction $u = (A_T, A_D)$ evaluated at $(T_{(k)}, D_{(k)})$, by $\nabla h(T_{(k)}, D_{(k)})[u]$. Then

$$\nabla h\big(T_{(k)}, D_{(k)}\big)[u] = T_{(k)}^{\mathrm{T}} D_{(k)}^{-1} A_T + A_T^{\mathrm{T}} D_{(k)}^{-1} T_{(k)} - T_{(k)}^{\mathrm{T}} D_{(k)}^{-2} A_D T_{(k)},$$

where A_T is written as a $p \times p$ matrix with nonzero entries in the same positions as the nonzero lower triangular entries in $T_{(k)}$, and A_D is written as a $p \times p$ diagonal matrix. Since the diagonal entries of $T_{(k)}$ are all equal to 1 and the diagonal entries of $D_{(k)}$ are positive, one can show by induction that $\nabla h(T_{(k)}, D_{(k)})[u] = 0$ implies $u = 0$. By the chain rule, $\nabla g(T_{(k)}, D_{(k)})[u] = \nabla f(T_{(k)}^{\mathrm{T}} D_{(k)}^{-1} T_{(k)})[u] \cdot \nabla h(T_{(k)}, D_{(k)})[u]$. Since f is convex with global minimizer $\hat{\Omega}_{(k)}$, it follows that $\nabla f(T_{(k)}^{\mathrm{T}} D_{(k)}^{-1} T_{(k)})[u] = 0$ if and only if $T_{(k)}^{\mathrm{T}} D_{(k)}^{-1} T_{(k)} = \hat{\Omega}_{(k)}$ unless $u = 0$. Hence we have that $\nabla g(T_{(k)}, D_{(k)})[u] = 0$ if and only if $\nabla f(T_{(k)}^{\mathrm{T}} D_{(k)}^{-1} T_{(k)})[u] = 0$ and $\hat{T}_{(k)}^{\mathrm{T}} \hat{D}_{(k)}^{-1} \hat{T}_{(k)} = \hat{\Omega}_{(k)}$.

Minimizing

$$g\big(T_{(k)}, D_{(k)}\big) = \sum_{j=1}^{p} \left\{ n \log d_{(k)jj} + \sum_{i=1}^{n} \frac{1}{d_{(k)jj}} \left(x_{ij} + \sum_{v=j-k}^{j-1} t_{(k)jv} x_{iv} \right)^2 \right\},$$

where $\hat{\Sigma} = n^{-1} \mathcal{X}^{\mathrm{T}} \mathcal{X}$, is equivalent to minimizing

$$g_j\big(t_{(k)j,j-k}, \ldots, t_{(k)j,j-1}, d_{(k)jj}\big) = n \log d_{(k)jj} + \sum_{i=1}^{n} \frac{1}{d_{(k)jj}} \left\{ x_{ij} - \sum_{v=j-k}^{j-1} \big(-t_{(k)jv}\big) x_{iv} \right\}^2$$

for each row $j = 1, \ldots, p$. For row j, the solution to $\nabla g_j(\hat{t}_{(k)j,j-k}, \ldots, \hat{t}_{(k)j,j-1}, \hat{d}_{(k)jj}) = 0$ gives the exactly similar ordinary least squares regression coefficients with the opposite sign from regressing x_j on $x_{j-k}, \ldots, x_{j-1}$, and the sample variance of the n residuals from this fit. □

Proof of Proposition 3. We show this by counterexample for $p = 3$. Let the function g be the negative normal loglikelihood parameterized by the inverse Cholesky factor $T = L^{-1}$ and D. Consider a 3×3

covariance matrix Σ with the banding constraint $\sigma_{31} = \sigma_{13} = 0$. This constraint is equivalent to $l_{31} = 0$ by Proposition 1. The unique parameters in the inverse Cholesky factor T in terms of the entries in the Cholesky factor L are: $t_{21} = -l_{21}$, $t_{31} = -l_{31} + l_{32}l_{21}$ and $t_{32} = -l_{32}$. Minimizing the negative loglikelihood subject to $l_{31} = 0$ is equivalent to minimizing the unconstrained function

$$b(l_{21}, l_{32}, D) = n \sum_{j=1}^{3} \log d_{jj} + \frac{1}{d_{11}} \|x_1\|^2 + \frac{1}{d_{22}} \|x_2 - l_{21}x_1\|^2 + \frac{1}{d_{33}} \|x_3 + l_{32}l_{21}x_1 - l_{32}x_2\|^2.$$

Since $\partial b(\hat{l}_{21}, \hat{l}_{32}, \hat{D})/\partial l_{21} = 2\hat{l}_{32}x_1^{\mathrm{T}}x_3\hat{d}_{33}^{-1} \neq 0$ with probability 1, the Cholesky banding solution does not satisfy the first-order necessary condition for being an optimum of an unconstrained differentiable function b, and hence Cholesky banding does not maximize the constrained normal likelihood. □

REFERENCES

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

BICKEL, P. J. & LEVINA, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–2604.

BICKEL, P. J. & LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

CHAUDHURI, S., DRTON, M. & RICHARDSON, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika* **94**, 199–216.

D'ASPREMONT, A., BANERJEE, O. & EL GHAOUI, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30**, 56–66.

EL KAROUI, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* **36**, 2717–56.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.

FURRER, R. & BENGTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Mult. Anal.* **98**, 227–55.

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

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

LAM, C. & FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37**, 4254–78.

LEDOIT, O. & WOLF, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *J. Mult. Anal.* **88**, 365–411.

LEVINA, E., ROTHMAN, A. J. & ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Statist.* **2**, 245–63.

MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1979). *Multivariate Analysis*. New York: Academic Press.

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

POURAHMADI, M. (2007). Cholesky decompositions and estimation of a covariance matrix: orthogonality of varianceà-correlation parameters. *Biometrika* **94**, 1006–13.

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

ROTHMAN, A. J., LEVINA, E. & ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Am. Statist. Assoc.* **104**, 177–86.

SMITH, M. & KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Am. Statist. Assoc.* **97**, 1141–53.

WATKINS, D. S. (1991). *Fundamentals of Matrix Computations*. New York: John Wiley & Sons.

WONG, F., CARTER, C. & KOHN, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90**, 809–30.

WU, W. B. & POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831–44.

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

[*Received February* 2009. *Revised January* 2010]