

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

Adam J. Rothman, Elizaveta Levina, and Ji Zhu
 Technical Report # 480
 Department of Statistics
 University of Michigan

March 3, 2009

Abstract

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 proposed by Bickel and 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.

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, fMRI, 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 ($p > n$) – 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 various sparsity assumptions about the population 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 and not always accurate. When the goal is to estimate the inverse covariance matrix, also known as the concentration matrix, a popular method is to add the lasso (ℓ_1) penalty on the entries of the inverse covariance matrix to the normal likelihood (d’Aspremont et al., 2008; Yuan and Lin, 2007; Rothman et al., 2008; Friedman et al., 2008), which has been extended to more general penalties by Lam and Fan (2007). Other estimators of the inverse covariance exploit the assumption that variables have a natural ordering, and those far apart in the ordering have small partial correlations. These estimators usually

rely on the modified Cholesky decomposition of the inverse covariance matrix (see details in Section 2, since this decomposition has a nice regression interpretation and regression regularization can be applied; see Wu and Pourahmadi (2003), Huang et al. (2006), Bickel and Levina (2008b), and Levina et al. (2008).

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 and Levina, 2008a; El Karoui, 2008; Rothman et al., 2009). Under the assumption that variables are ordered and those far apart in the ordering are only weakly correlated, a better option is to band or taper the sample covariance matrix (Bickel and Levina, 2004; Furrer and Bengtsson, 2007; Bickel and Levina, 2008b; Cai et al., 2008). These simple approaches are attractive for problems in very high dimensions since they have a small computational cost; however, these estimators are not generally guaranteed to be positive definite, although some forms of tapering can guarantee positive semi-definite estimates. 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 than observations ($p < n$).

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 covariance matrix itself, it is guaranteed to be positive definite, but still has the same low computational complexity.

The rest of this paper is organized as follows: we discuss the modified Cholesky factorization of the covariance matrix and its regression interpretation in Section 2. Regularization techniques appropriate for the Cholesky factor of covariance are described in Section 3. In addition, we connect sparsity in the covariance matrix to sparsity in its Cholesky factor and use this to contrast the maximum likelihood properties of banding the Cholesky factor of covariance and banding the Cholesky factor of the inverse. In particular, we prove that Cholesky banding of the inverse is the constrained maximum likelihood estimator for normal data under the constraint that the inverse covariance matrix is banded. Numerical performance of regularized Cholesky-based estimators of the covariance is illustrated both on simulated data (Section 4) and on a spectroscopy data example (Section 5).

2 Modified Cholesky decomposition of the covariance matrix

Throughout the paper we assume that the data $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent and identically distributed p -variate random vectors with population covariance matrix Σ and, without loss of generality, mean $\mathbf{0}$. Let $\hat{\Sigma}$ denote the sample covariance matrix (the maximum likelihood version),

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T.$$

As a tool for regularizing the inverse covariance matrix, Pourahmadi (1999) suggested using the modified Cholesky factorization of Σ^{-1} . This factorization arises from regressing each variable X_j

on X_{j-1}, \dots, X_1 for $2 \leq j \leq p$. Fitting regressions

$$X_j = \sum_{q=1}^{j-1} (-t_{jq})X_q + \epsilon_j = \hat{X}_j + \epsilon_j ,$$

let ϵ_j denote the error term in regression j , $j \geq 2$, and let $\epsilon_1 = X_1$. Let $D = \text{var}(\epsilon)$ be the diagonal matrix of error variances and $T = [t_{jq}]$ the lower-triangular matrix containing regression coefficients (with the opposite sign), with ones on the diagonal. Then writing $\epsilon = \mathbf{X} - \hat{\mathbf{X}} = T\mathbf{X}$ and using the independence of errors we have,

$$D = \text{var}(\epsilon) = \text{var}(T\mathbf{X}) = T\Sigma T^T$$

and thus

$$\Sigma^{-1} = T^T D^{-1} T. \tag{1}$$

This decomposition transforms inverse covariance matrix estimation into a regression problem, and hence regularization approaches for regression can be applied. In practice, the coefficients are computed by regressing each variable X_j on its predecessors X_1, \dots, X_{j-1} (after centering all the variables). 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, $X_{\max(1, j-k)}, \dots, X_{j-1}$, for some fixed k (Wu and Pourahmadi, 2003; Bickel and Levina, 2008b).

The modified Cholesky factorization of Σ can be obtained by simply inverting (1). Let $L = T^{-1}$ and rewrite $\mathbf{X} = L\epsilon$. Then,

$$\Sigma = \text{var}(L\epsilon) = LDL^T . \tag{2}$$

Our main interest here is in the regression interpretation of this decomposition. By analogy to (1), we can interpret (2) as resulting from a new sequence of regressions, where each variable X_j is regressed on all the previous regression *errors* $\epsilon_{j-1}, \dots, \epsilon_1$ (rather than the variables themselves). For $j \geq 2$, we have the sequence of regressions,

$$X_j = \sum_{q=1}^{j-1} l_{jq}\epsilon_q + \epsilon_j = \tilde{X}_j + \epsilon_j . \tag{3}$$

The decompositions above apply to the population matrices. Let $X = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ be the $n \times p$ data matrix, where each column $\mathbf{x}_j \in \mathbb{R}^n$ is already centered by its sample mean. For the first variable, we set $\mathbf{e}_1 = \mathbf{x}_1$. For $2 \leq j \leq p$, let $\mathbf{l}_j = (l_{j1}, \dots, l_{j, j-1})^T$, $Z_j = [\mathbf{e}_1, \dots, \mathbf{e}_{j-1}]$, and compute coefficients and the residual, respectively, as

$$\begin{aligned} \hat{\mathbf{l}}_j &= \underset{\mathbf{l}_j}{\text{argmin}} \|\mathbf{x}_j - Z_j \mathbf{l}_j\|^2 , \\ \mathbf{e}_j &= \mathbf{x}_j - Z_j \hat{\mathbf{l}}_j . \end{aligned} \tag{4}$$

The variances are estimated as

$$\hat{d}_{jj} = \frac{1}{n} \|\mathbf{e}_j\|^2 .$$

Let $Z = [\mathbf{e}_1, \dots, \mathbf{e}_p]$ denote the $n \times p$ matrix of residuals from carrying out the regressions in (3) sequentially. Here we assume that $p < n$ to ensure that all model matrices are of full column

rank; Section 3 discusses the rank deficient case when $p \geq n$. Performing the regressions in (4) amounts to, for each $j \geq 2$, orthogonally projecting the response \mathbf{x}_j onto the span of $\mathbf{e}_1, \dots, \mathbf{e}_{j-1}$ to estimate $\hat{\mathbf{l}}_j$. After the last projection we have an orthogonal basis $\{\mathbf{e}_1, \dots, \mathbf{e}_p\}$, and the estimates \hat{L} and \hat{D} . This algorithm is nothing but a scaled version of Gram-Schmidt orthogonalization of the data matrix 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 $R^T = \hat{L}(n\hat{D})^{\frac{1}{2}}$. If all regressions are fitted without any regularization, simply by least squares (as described above), the resulting estimate recovers the sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n} X^T X = \frac{1}{n} R^T R = \hat{L} \hat{D} \hat{L}^T .$$

3 Regularized estimation of the Cholesky factor L

It is clear that in order to improve on the sample covariance, the regressions in (4) need to be regularized. In this section we describe several estimators that introduce sparsity in covariance Cholesky factor L . We also connect sparsity patterns in positive definite matrices with sparsity patterns in their Cholesky factors and use this to analyze the connection between banding Cholesky factors and constrained maximum likelihood estimation.

3.1 Banding the Cholesky factor

The simplest way to introduce sparsity in the Cholesky factor L is to estimate only the first k sub-diagonals of L and set the rest to zero. This approach for banding the Cholesky factor of the inverse was proposed by Wu and Pourahmadi (2003) and Bickel and Levina (2008b). In practice, it means that each variable \mathbf{x}_j is regressed on the k previous residuals $[\mathbf{e}_{j-k}, \dots, \mathbf{e}_{j-1}]$, for all $j \geq 2$. Note that the index $j - k$ everywhere is understood to mean $\max(1, j - k)$. Let $\mathbf{l}_j^{(k)} = (l_{j,j-k}, \dots, l_{j,j-1})^T$ and $Z_j^{(k)} = [\mathbf{e}_{j-k}, \dots, \mathbf{e}_{j-1}]$. Then we compute,

$$\begin{aligned} \hat{\mathbf{l}}_j^{(k)} &= \underset{\mathbf{l}_j^{(k)}}{\operatorname{argmin}} \|\mathbf{x}_j - Z_j^{(k)} \mathbf{l}_j^{(k)}\|^2 , \\ \mathbf{e}_j &= \mathbf{x}_j - Z_j^{(k)} \hat{\mathbf{l}}_j^{(k)} . \end{aligned} \tag{5}$$

In each regression, the design matrix $Z_j^{(k)}$ has orthogonal columns, which allows (5) 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 $\min(n - 1, p)$. Also note that while each design matrix $Z_j^{(k)}$ has orthogonal columns, all of the residual vectors $\mathbf{e}_1, \dots, \mathbf{e}_p$ are not necessarily mutually orthogonal; \mathbf{e}_j and $\mathbf{e}_{j'}$ are only guaranteed to be orthogonal if $|j - j'| \leq k$.

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, the answer depends on whether the banding is applied to the Cholesky factor of the inverse or of the covariance matrix itself: the former estimator coincides with constrained maximum likelihood estimator, and the latter does not. In order to show this, we first establish some relationships between zero patterns in positive definite matrices and their Cholesky factors.

Proposition 1. *Given a positive definite matrix Σ with modified Cholesky decomposition $\Sigma = LDL^T$, where L is lower triangular, for any row i and $c(i) < i$, $\sigma_{i1} = \dots = \sigma_{i,c(i)} = 0$ if and only if $l_{i1} = \dots = l_{i,c(i)} = 0$.*

Proof. Using the expression

$$\sigma_{ij} = \sum_{m=1}^j l_{im}l_{jm}d_{mm},$$

it is obvious that $l_{i1} = \dots = l_{i,c(i)} = 0$ implies $\sigma_{i1} = \dots = \sigma_{i,c(i)} = 0$.

Now assume $\sigma_{i1} = \dots = \sigma_{i,c(i)} = 0$ for some i . The sequential column-wise formula for computing the modified Cholesky factorization (Watkins, 1991), is given by, for $i > j$,

$$\begin{aligned} d_{ii} &= \sigma_{ii} - \sum_{m=1}^{i-1} l_{im}^2 d_{mm} , \\ l_{ij} &= \frac{1}{d_{jj}} \left(\sigma_{ij} - \sum_{m=1}^{j-1} l_{im}l_{jm}d_{mm} \right) . \end{aligned} \quad (6)$$

This formula allows one to compute L one column at a time, starting from the first column. We proceed by induction: for the first column of L , $l_{i1} = \sigma_{i1}/\sigma_{11}$, hence $l_{i1} = 0$. Assume that for some column $u < c(i)$ we have $l_{i1} = \dots = l_{iu} = 0$, then using (6),

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

which implies $l_{i,u+1} = 0$. □

Proposition 1 states that a Cholesky factor with banded rows of arbitrary band length (by band length k_i of row i we mean that k_i is the smallest integer such that $l_{ij} = 0$ for all $j < i - k_i$) corresponds to a covariance matrix with banded rows of the same band lengths. In particular, the Cholesky factor L is k -banded if and only if the covariance matrix itself is k -banded. An analogous result holds for the inverse covariance matrix Ω , with rows replaced by columns.

Proposition 2. *For a positive definite matrix Ω with modified Cholesky decomposition $T^T D^{-1} T = \Omega$, where T is lower triangular, for any column j and $r(j) > j$, $\omega_{p,j} = \dots = \omega_{r(j),j} = 0$ if and only if $t_{p,j} = \dots = t_{r(j),j} = 0$.*

The proof of Proposition 2 is similar to that of Proposition 1 and is omitted. Proposition 2 states that the modified Cholesky factor of the inverse T with arbitrary column band lengths corresponds to an inverse covariance matrix Ω with the same column band lengths, and thus an inverse covariance matrix is k -banded if and only if its Cholesky factor is k -banded.

With these propositions we can investigate maximum likelihood properties of Cholesky and inverse Cholesky banding.

Proposition 3. *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 $|i - j| > k$.*

Proof. Let $\Omega_{(k)}$ be a symmetric positive definite matrix with k non-zero main sub-diagonals, i.e., $\omega_{(k)ij} = 0$ for $|i - j| > k$. The negative normal log-likelihood of $\mathbf{x}_1, \dots, \mathbf{x}_n \sim N(\mathbf{0}, \Omega_{(k)}^{-1})$, up to a constant, is given by,

$$f(\Omega_{(k)}) = \text{trace}(\hat{\Sigma}\Omega_{(k)}) - \log |\Omega_{(k)}|,$$

where f is a function of the non-zero unique parameters in $\Omega_{(k)}$. The k -banded constrained maximum likelihood estimator $\hat{\Omega}_{(k)}$ satisfies $\nabla f(\hat{\Omega}_{(k)}) = 0$. Let $T_{(k)}^T D_{(k)}^{-1} T_{(k)} = \Omega_{(k)}$ be the modified Cholesky decomposition of $\Omega_{(k)}$. By Proposition 2, $t_{(k)ij} = 0$ for $|i - j| > k$. Let $g(T_{(k)}, D_{(k)}) \equiv f(T_{(k)}^T D_{(k)}^{-1} T_{(k)})$, where g is a function of non-zero unique parameters in $(T_{(k)}, D_{(k)})$.

We continue by establishing that if $\nabla g(\hat{T}_{(k)}, \hat{D}_{(k)}) = 0$ then $\hat{T}_{(k)}^T \hat{D}_{(k)}^{-1} \hat{T}_{(k)} = \hat{\Omega}_{(k)}$. Let $h(T_{(k)}, D_{(k)}) = T_{(k)}^T D_{(k)}^{-1} T_{(k)}$. 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

$$\nabla h(T_{(k)}, D_{(k)})[u] = T_{(k)}^T D_{(k)}^{-1} A_T + A_T^T D_{(k)}^{-1} T_{(k)} - T_{(k)}^T D_{(k)}^{-2} A_D T_{(k)}, \quad (7)$$

where A_T is written as a $p \times p$ matrix with non-zero entries in the same positions as the non-zero 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, we have that

$$\nabla g(T_{(k)}, D_{(k)})[u] = \nabla f(T_{(k)}^T D_{(k)}^{-1} T_{(k)})[u] \cdot \nabla h(T_{(k)}, D_{(k)})[u].$$

Since f is convex with unique minimizer $\hat{\Omega}_{(k)}$ it follows that $\nabla f(T_{(k)}^T D_{(k)}^{-1} T_{(k)})[u] = 0$ if and only if $T_{(k)}^T D_{(k)}^{-1} T_{(k)} = \hat{\Omega}_{(k)}$ unless $u = 0$. Hence we have that $\nabla g(T_{(k)}, D_{(k)})[u] = 0$ iff $\nabla f(T_{(k)}^T D_{(k)}^{-1} T_{(k)})[u] = 0$ and $\hat{T}_{(k)}^T \hat{D}_{(k)}^{-1} \hat{T}_{(k)} = \hat{\Omega}_{(k)}$.

Minimizing $g(T_{(k)}, D_{(k)})$, which can be expressed as,

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

is equivalent to minimizing,

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

for each row $1 \leq j \leq p$. For row j , the solution to $\nabla g_j(\hat{t}_{(k)j,j-k}, \dots, \hat{t}_{(k)j,j-1}, \hat{d}_{(k)jj}) = 0$, gives exactly the ordinary least squares regression coefficients (with the opposite sign) from regressing \mathbf{x}_j on $\mathbf{x}_{j-k}, \dots, \mathbf{x}_{j-1}$, and the sample variance of the n residuals from this fit. Thus the solution coincides with the output of the inverse Cholesky banding algorithm. \square

Next, we show that banding the Cholesky factor of the covariance matrix itself does not give the constrained maximum likelihood estimator. This is due to the inverse being the natural canonical parameter of the multivariate normal distribution.

Proposition 4. *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 $|i - j| > k$.*

Proof. We show that the first-order necessary condition for optimality is not met using $p = 3$ variables. Let the function g be the negative normal log-likelihood parameterized by the inverse Cholesky factor $T = L^{-1}$ and D , which is given up to a constant by,

$$g(T, D) \equiv \ell(T^T D^{-1} T) = \sum_{j=1}^p \left(n \log d_{jj} + \sum_{i=1}^n \frac{1}{d_{jj}} \left(x_{ij} + \sum_{v=1}^{j-1} t_{jv} x_{iv} \right)^2 \right) \quad (8)$$

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 inverse Cholesky factor T in terms of the entries in the Cholesky factor L is given by,

$$T = \begin{pmatrix} 1 & 0 & 0 \\ -l_{21} & 1 & 0 \\ -l_{31} + l_{32}l_{21} & -l_{32} & 1 \end{pmatrix}$$

Minimizing the negative log-likelihood subject to $l_{31} = 0$ is equivalent to minimizing the unconstrained function

$$b(l_{21}, l_{32}, D) = n \sum_{j=1}^3 \log d_{jj} + \frac{1}{d_{11}} \|\mathbf{x}_1\|^2 + \frac{1}{d_{22}} \|\mathbf{x}_2 - l_{21}\mathbf{x}_1\|^2 + \frac{1}{d_{33}} \|\mathbf{x}_3 + l_{32}l_{21}\mathbf{x}_1 - l_{32}\mathbf{x}_2\|^2.$$

Taking the partial derivative of b with respect to l_{21} ,

$$\frac{\partial}{\partial l_{21}} b(l_{21}, l_{32}, D) = \frac{1}{d_{22}} (2l_{21}\mathbf{x}_1^T \mathbf{x}_1 - 2\mathbf{x}_1^T \mathbf{x}_2) + \frac{1}{d_{33}} (2l_{32}\mathbf{x}_1^T \mathbf{x}_3 - 2l_{32}^2 \mathbf{x}_1^T \mathbf{x}_2 + 2l_{21}l_{32}^2 \mathbf{x}_1^T \mathbf{x}_1),$$

and evaluating at the Cholesky banding solution gives,

$$\frac{\partial}{\partial l_{21}} b(\hat{l}_{21}, \hat{l}_{32}, \hat{D}) = \frac{2\hat{l}_{32}\mathbf{x}_1^T \mathbf{x}_3}{\hat{d}_{33}}.$$

Since $\frac{\partial}{\partial l_{21}} b(\hat{l}_{21}, \hat{l}_{32}, \hat{D}) \neq 0$ 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. \square

The constrained maximum likelihood estimator 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 Section 4, we briefly compare the numerical performance of banding the Cholesky factor to the constrained maximum likelihood estimator when $p < n$, and find that the two estimators are in practice very close.

3.3 The penalized regression approach

Instead of banding the Cholesky factor, more sophisticated regularization approaches can be applied to regressions involved in the computation. In general, for $2 \leq j \leq p$ we can estimate the Cholesky factor by,

$$\hat{\mathbf{l}}_j = \underset{\mathbf{l}_j}{\operatorname{argmin}} \{ \|\mathbf{x}_j - Z_j \mathbf{l}_j\|^2 + P_\lambda(\mathbf{l}_j) \}. \quad (9)$$

Penalty functions P_λ that encourage sparsity in the coefficient vector \mathbf{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

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

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,

$$P_\lambda^{NL}(\mathbf{l}_j) = \lambda \left(|l_{j,j-1}| + \frac{|l_{j,j-2}|}{|l_{j,j-1}|} + \frac{|l_{j,j-3}|}{|l_{j,j-2}|} + \dots + \frac{|l_{j,1}|}{|l_{j,2}|} \right), \quad (10)$$

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 \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 (not all) zeros are preserved in that case as well.

In practice, Levina et al. (2008) recommend using a slightly modified version of (10) where the first term is divided by the univariate regression coefficient from regressing \mathbf{x}_j on \mathbf{e}_{j-1} alone, to address a potential difference of scales, which is the version we used in simulations. Note that 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 may work well in some cases.

4 Numerical results

In this section we present a simulation study which compares the performance of all the covariance estimators discussed in Section 3, banding the sample covariance matrix directly (Bickel and Levina, 2008b), and, as a benchmark, the shrinkage estimator of Ledoit and Wolf (2003). The main difference between banding the sample covariance directly and regularizing the Cholesky factor is the guaranteed positive definiteness of the latter. The Ledoit-Wolf estimator is a linear combination of the identity matrix and the sample covariance matrix, where linear coefficients are estimates of asymptotically optimal coefficients under Frobenius loss; it does not introduce any sparsity.

4.1 Simulation Settings

We consider two standard covariance structures for ordered variables,

1. $\Sigma_1: \sigma_{ij} = (7/10)^{|i-j|}$;

2. $\Sigma_2 : \sigma_{ij} = \mathbf{1}(i = j) + (4/10)\mathbf{1}(|i - j| = 1) + (2/10)\mathbf{1}(|i - j| = 2) + (2/10)\mathbf{1}(|i - j| = 3) + (1/10)\mathbf{1}(|i - j| = 4)$.

The AR(1) model Σ_1 has a dense Cholesky factor while the MA(4) 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 and Levina (2008b), and Σ_2 by Yuan and Lin (2007).

We generate $n = 100$ training observations and another 100 independent validation observations from $N_p(\mathbf{0}, \Sigma)$. The dimensions considered were $p = 30, 100, 200, 500$, and 1000. Note that lasso and nested lasso were not run for $p = 500$ and 1000 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. Alternatively, one could select tuning parameters using the random-splitting scheme of Bickel and Levina (2008b), which we use in the data example in Section 5. The whole process was repeated 50 times.

To compare estimators, we used the operator norm loss, also known as the matrix 2-norm ($\|M\|^2 = \lambda_{\max}(MM^T)$), of the difference between the covariance estimator and the truth,

$$\Delta(\hat{\Sigma}, \Sigma) = E\|\hat{\Sigma} - \Sigma\|.$$

We also compute the true positive rate (TPR) and true negative rate (TNR), defined as

$$\text{TPR}(\hat{\Sigma}, \Sigma) = \frac{\#\{(i, j) : \hat{\sigma}_{ij} \neq 0 \text{ and } \sigma_{ij} \neq 0\}}{\#\{(i, j) : \sigma_{ij} \neq 0\}}, \quad (11)$$

$$\text{TNR}(\hat{\Sigma}, \Sigma) = \frac{\#\{(i, j) : \hat{\sigma}_{ij} = 0 \text{ and } \sigma_{ij} = 0\}}{\#\{(i, j) : \sigma_{ij} = 0\}}. \quad (12)$$

Note that the sample covariance has a true positive rate of 1, and a diagonal estimator has a true negative rate of 1. Additionally we measure eigenspace agreement between the estimate and the truth using the measure, for $q = 1, \dots, p$

$$K(q) = \sum_{i=1}^q \sum_{j=1}^q (\hat{e}_{(i)}^T e_{(j)})^2, \quad (13)$$

introduced by Krzanowski (1979), where $\hat{e}_{(i)}$ denotes the estimated eigenvector corresponding to the i -th largest estimated eigenvalue, and $e_{(i)}$ the true eigenvector corresponding to the i -th largest true eigenvalue. Note that $K(q) = q$ indicates perfect agreement of the eigenspaces spanned by the first q eigenvectors.

4.2 Results

The averages and standard errors over 50 replications of the operator norm loss for both models are given in Table 1. One can see that banding the Cholesky factor provides the best performance in every case. It outperforms banding the sample covariance directly, particularly in high dimensions, and both banding methods outperform the Ledoit-Wolf estimator as well as both regularized regression methods.

The banded maximum likelihood estimator was also computed using the algorithm of Chaudhuri et al. (2007) for $p = 30$ (the algorithm is only applicable when $p < n$). Its loss values are 1.27(0.04) for

Table 1: Operator Norm Loss, average(SE) over 50 replications

p	Sample	Ledoit-Wolf	Sample Banding	Cholesky Banding	Lasso	Nested Lasso
Σ_1						
30	1.75(0.04)	1.67(0.04)	1.27(0.04)	1.27(0.03)	1.68(0.05)	1.45(0.04)
100	4.14(0.07)	3.06(0.03)	1.58(0.03)	1.56(0.03)	3.50(0.03)	1.78(0.03)
200	6.55(0.07)	3.79(0.02)	1.75(0.03)	1.74(0.03)	3.90(0.01)	1.93(0.03)
500	12.57(0.08)	4.42(0.01)	1.95(0.03)	1.91(0.02)	–	–
1000	20.65(0.09)	4.64(0.00)	2.08(0.03)	2.00(0.02)	–	–
Σ_2						
30	1.44(0.03)	1.13(0.02)	0.77(0.02)	0.75(0.02)	1.23(0.02)	0.88(0.02)
100	3.34(0.04)	1.64(0.01)	0.92(0.02)	0.89(0.02)	1.63(0.01)	1.00(0.01)
200	5.36(0.04)	1.78(0.00)	0.99(0.02)	0.93(0.02)	1.71(0.00)	1.08(0.01)
500	10.36(0.05)	1.84(0.00)	1.09(0.02)	1.05(0.02)	–	–
1000	17.60(0.07)	1.85(0.00)	1.19(0.02)	1.14(0.02)	–	–

Σ_1 and 0.76(0.02) for Σ_2 , which are essentially the same as those for Cholesky banding for $p = 30$. As expected, the margin by which sparse regularized estimators outperform non-sparse estimators (the sample and Ledoit-Wolf) is larger for the sparse population covariance Σ_2 .

For the sparse matrix Σ_2 , we also report true positive and true negative rates of estimating zeros in Table 2. These rates also depend on the tuning parameter (k for banding and λ for the lasso and nested lasso). Both Cholesky banding and sample covariance banding have perfect true negative rates, meaning that all of the realizations had at most 4 non-zero sub-diagonals. We see a better true positive rate for banding the Cholesky factor than for banding the sample covariance matrix, which means that banding the sample tends to set more diagonals to zero than necessary. This is partly because the entries on the fourth sub-diagonal of Σ_2 are quite small. The lasso method has a low true negative rate, which is expected since zeros in the Cholesky factor are not preserved, and the nested lasso does reasonably well on both but not as well as Cholesky banding.

Table 2: True Positive/True Negative Ratea for Σ_2 (%), average(SE) over 50 replications

p	Banding	Cholesky Banding	Lasso	Nested Lasso
30	88.18(1.69) / 100(0)	91.00(1.78) / 100(0)	99.71(0.08) / 3.86(0.40)	93.89(0.75) / 88.9(0.88)
100	88.68(1.75) / 100(0)	94.09(1.50) / 100(0)	90.69(0.27) / 36.45(0.40)	94.44(0.36) / 97.07(0.16)
200	88.59(1.77) / 100(0)	95.04(1.42) / 100(0)	90.76(0.18) / 34.63(0.28)	94.01(0.32) / 98.72(0.05)
500	88.04(1.78) / 100(0)	96.01(1.31) / 100(0)	–	–
1000	87.02(1.78) / 100(0)	96.51(1.24) / 100(0)	–	–

In Figure 1 we plot the averaged 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 and $p = 1000$. Since $n = 100$, the sample covariance matrix only has 99 non-zero 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,

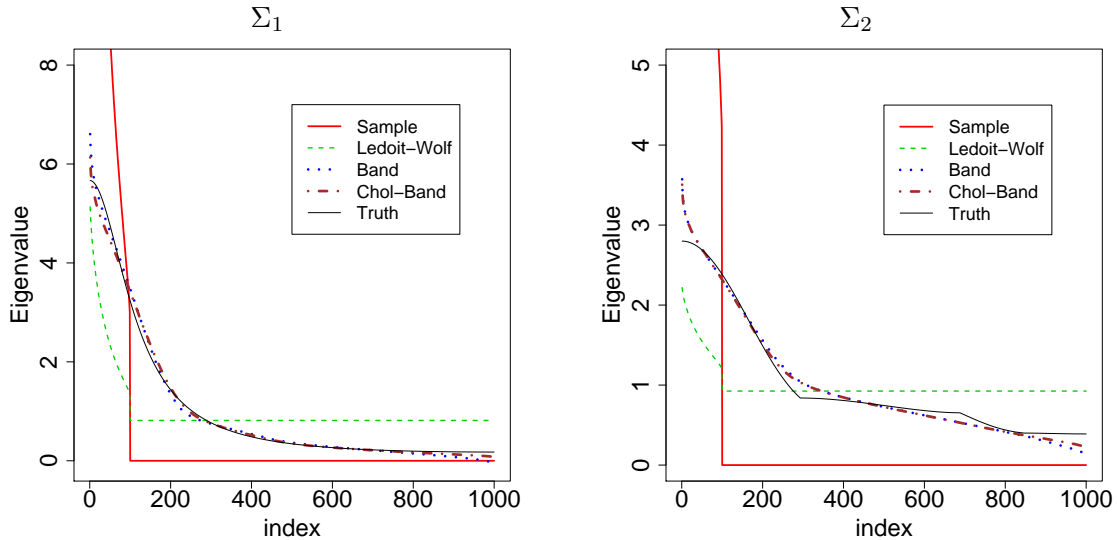


Figure 1: Scree plots (averaged over 50 replications) for $p = 1000$.

especially for larger true eigenvalues.

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. We see that for the dense matrix Σ_1 , sample banding has 0 out of 50 positive definite realizations for $p \geq 200$; for the sparse matrix Σ_2 , sample banding has 50 out of 50 positive definite realizations for $p \leq 200$, 49 for for $p = 500$ and 48 for $p = 1000$; it is clear that, for both models, the larger p , the harder it is to keep positive definiteness.

Table 3: Percentage of banded sample covariance realizations that are positive definite (based on 50 replications)

Model	p				
	30	100	200	500	1000
Σ_1	66	8	0	0	0
Σ_2	100	100	100	98	96

Finally, Figure 2 shows a plot of the averaged eigenspace agreement measure $K(q)$ versus q , for $p = 1000$ variables, along with the line $K(q) = q$ representing perfect eigenspace agreement. We see that both Cholesky banding and ordinary banding perform roughly the same under this measure; both outperform the sample covariance matrix and the Ledoit-Wolf estimator, which have the same eigenvectors, since the Ledoit-Wolf estimator is a linear combination of the sample covariance and the identity.

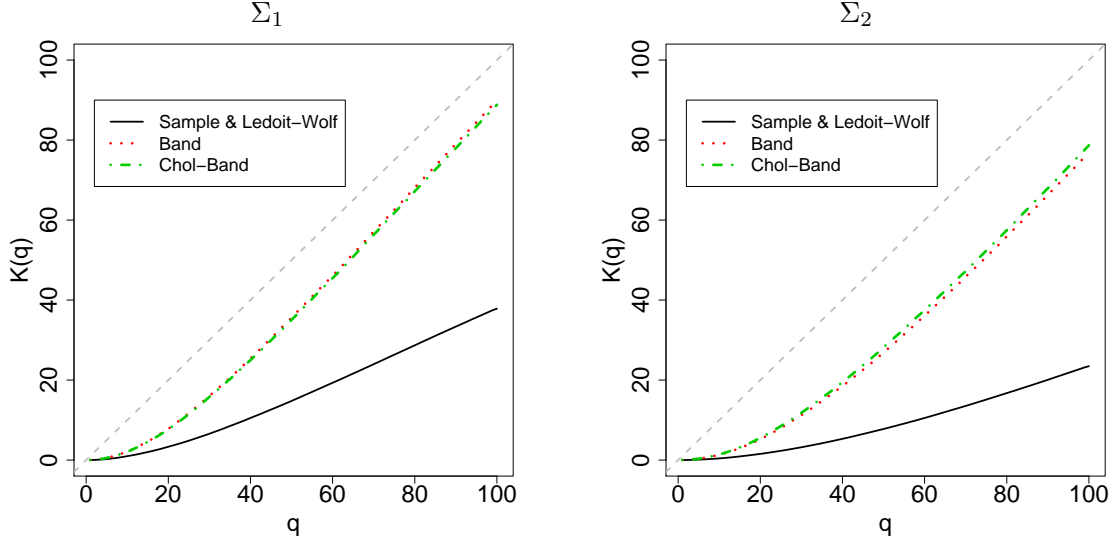


Figure 2: $K(q)$ versus q (averaged over 50 replications) for $p = 1000$. $K(q) = q$ corresponds to perfect agreement.

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 (Asuncion and Newman, 2007). 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 Figure 3 shows heatmaps of the absolute values of the sample correlation matrices for metal and rock (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 as one moves away from the diagonal, which makes banding a reasonable option.

The banding parameter k for both banding methods was selected using the random-splitting scheme of Bickel and Levina (2008b),

$$\hat{k} = \operatorname{argmin}_k \frac{1}{N} \sum_{v=1}^N \|\hat{\Sigma}_{(k)}^{(v)} - \tilde{\Sigma}^{(v)}\|_F,$$

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 was split at random $N = 100$ times, with $1/3$ of the sample used for training. For metal, Cholesky banding and sample banding both chose $\hat{k} = 31$ sub-diagonals; 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 Cholesky banding and sample banding both computed with $\hat{k} = 17$ for the rock spectra. The heatmaps of the absolute values of the banded estimators are shown in Figure 3. We see that Cholesky banding shrinks the non-zero correlations whereas the sample banding does not, which is the property that allows Cholesky banding to achieve positive definiteness.

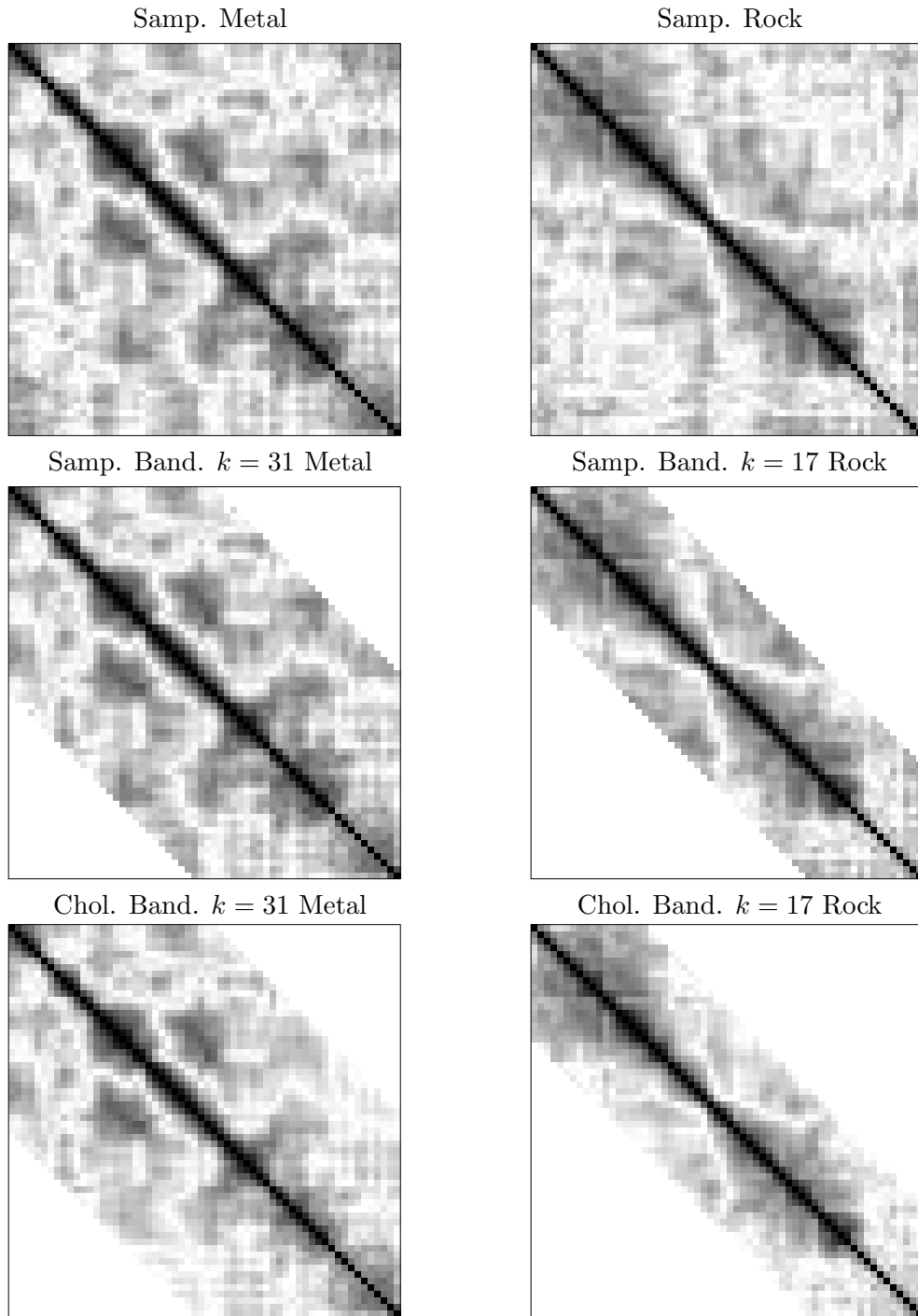


Figure 3: Heatmaps of the absolute values of the correlation estimates. White is magnitude 0 and black is magnitude 1.

We also show eigenvalue plots for these estimators in Figure 4(a) and (b), and the eigenspace agreement measure between the banded estimators and the sample covariance in Figure 4(c) and

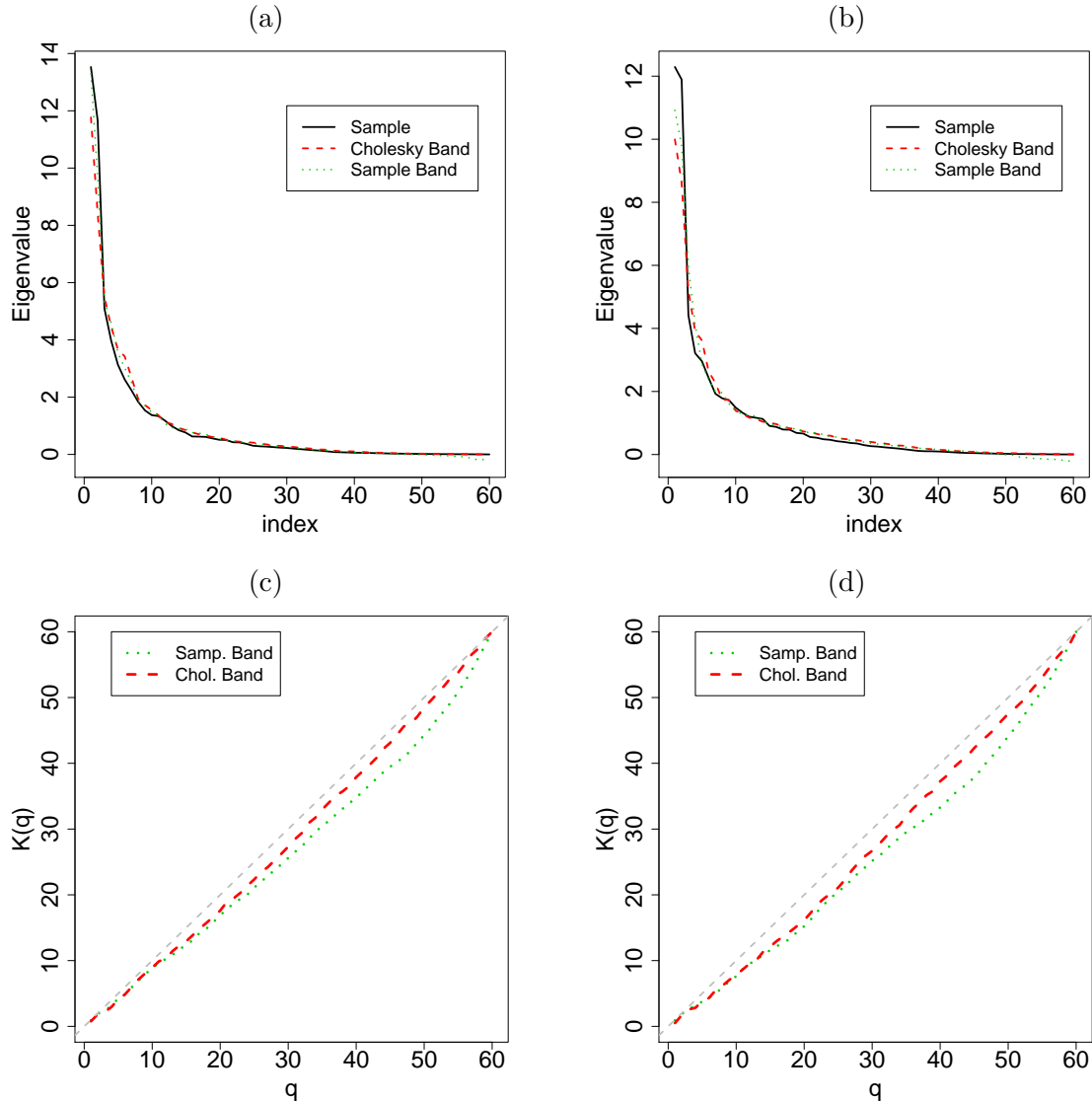


Figure 4: (a) Scree plots of the banded estimators, using the metal spectra; (b) Scree plots using the rock spectra; (c) Eigenspace agreement with the sample covariance matrix using the metal spectra. (d) Eigenspace agreement with the sample covariance matrix using the rock spectra. Note that in (c) and (d), $K(q) = q$, drawn as the gray dashed line, corresponds to perfect agreement, see (13) for the definition of $K(q)$.

(d), using the agreement measure (13). 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. For eigenvectors, there are no major differences between the estimators, a result consistent with simulations.

We also compared the performance of the various estimators if they are used in quadratic discriminant analysis (QDA) to discriminate between rock and metal. An observation \mathbf{x} is classified

as rock ($k = 0$) or metal ($k = 1$) using the QDA rule,

$$G(\mathbf{x}) = \operatorname{argmax}_k \left\{ \frac{1}{2} \log |\hat{\Omega}_k| - \frac{1}{2} (\mathbf{x} - \hat{\boldsymbol{\mu}}_k)^T \hat{\Omega}_k (\mathbf{x} - \hat{\boldsymbol{\mu}}_k) + \log \hat{\pi}_k \right\},$$

where $\hat{\pi}_k$ is the proportion of class k observations in the training sample, $\hat{\boldsymbol{\mu}}_k$ is the training class k sample mean vector, and $\hat{\Omega}_k$ is the inverse covariance estimate computed with the class k training observations. A full description of QDA 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). Leave-one-out cross validation was used to estimate the testing error, and the banding parameters were selected with 10 random splits with 1/3 of the data used for training, using Frobenius loss for covariance Cholesky banding and the validation likelihood for the inverse covariance Cholesky banding. Banding the sample covariance was omitted because its lack of positive definiteness led to inversion problems. The test errors (%) were 24.0(3.0) for the sample covariance, 32.7(3.3) for naive Bayes, 20.2(2.8) for covariance Cholesky banding, and 14.9(2.5) 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 (naive Bayes), and the inverse Cholesky banding does better in this case because it introduces sparsity directly in the inverse covariance.

6 Summary and discussion

In this paper we proposed a new regression interpretation of the Cholesky factor of the covariance matrix, which was previously only available for the Cholesky factor of the inverse. Banding of this Cholesky factor gives a banded positive definite estimator of the covariance, unlike banding the sample covariance matrix, and was shown to perform better numerically. An attractive property of the banded Cholesky estimator is its low computational cost, the same as that of banding the sample covariance matrix itself, and thus there is no computational penalty to pay for enforcing positive definiteness. More complicated regularization obtained from penalties such as the lasso or the nested lasso can be applied using the same regression interpretation, but at an additional computational cost. The proposed estimators perform well numerically under a variety of measures.

We also connected sparsity in banded Cholesky factors with sparsity in the covariance matrix and the inverse covariance matrix, which allows us to show that inverse Cholesky banding is equivalent to constrained maximum likelihood under the banded constraint. Banding the Cholesky factor of the covariance itself is not equivalent to constrained maximum likelihood, but we found empirically they perform similarly. In terms of convergence rates, one would expect a convergence result analogous to the one for inverse Cholesky banding established by Bickel and Levina (2008b) to hold here as well, but this case presents substantial extra technical difficulties in analysis, due to the fact that 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.

Acknowledgments

We thank Richard Davis (Columbia) for pointing out the use of regression on residuals in time series, and Bala Rajaratnam (Stanford) for helpful discussions on sparse Cholesky factors. A.J.

Rothman’s research is supported in part by the Yahoo! Ph.D. Fellowship. E. Levina’s research is supported in part by grants from the NSF (DMS-0505424, DMS-0805798). J. Zhu’s research is supported in part by grants from the NSF (DMS-0705532 and DMS-0748389).

References

- Asuncion, A. and Newman, D. (2007). UCI machine learning repository. <http://www.ics.uci.edu/~mlern/MLRepository.html>.
- Bickel, P. J. and Levina, E. (2004). Some theory for Fisher’s linear discriminant function, “naive Bayes”, and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.
- Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604.
- Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.
- Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2008). Optimal rates of convergence for covariance matrix estimation. Manuscript.
- Chaudhuri, S., Drton, M., and Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216.
- d’Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 30(1):56–66.
- El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.*, 36(6):2717–2756.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.
- Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255.
- Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.
- Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327.
- Krzanowski, W. (1979). Between-groups comparison of principal components. *J. Amer. Statist. Assoc.*, 74(367):703–707.
- Lam, C. and Fan, J. (2007). Sparsistency and rates of convergence in large covariance matrices estimation. Manuscript.
- Ledoit, O. and Wolf, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.

- Levina, E., Rothman, A. J., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Annals of Applied Statistics*, 2(1):245–263.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, New York.
- Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86:677–690.
- Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.
- Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc. (Theory and Methods)*, 104. To appear.
- Watkins, D. S. (1991). *Fundamentals of matrix computations*. John Wiley & Sons, Inc., New York, NY, USA.
- Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90:831–844.
- Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.