Regularized estimation of large covariance matrices

Peter J. Bickel and Elizaveta Levina
University of California, Berkeley and University of Michigan

September 14, 2006

Abstract

This paper considers estimating a covariance matrix of p variables from n observations by either banding the sample covariance matrix or estimating a banded version of the inverse of the covariance. We show that these estimates are consistent in the operator norm as long as $(\log p)^2/n \to 0$, and obtain explicit rates. The results are uniform over some fairly natural well-conditioned families of covariance matrices. We also introduce an analogue of the Gaussian white noise model and show that if the population covariance is embeddable in that model and well-conditioned then the banded approximations produce consistent estimates of the eigenvalues and associated eigenvectors of the covariance matrix. The results can be extended to smooth versions of banding and to non-Gaussian distributions with sufficiently short tails. A resampling approach is proposed for choosing the banding parameter in practice. This approach is illustrated numerically on both simulated and real data.

Keywords: covariance estimation, regularization, banded estimators, Cholesky decomposition.

1 Introduction

Estimation of population covariance matrices from samples of multivariate data has always been important for a number of reasons. Principal among these are: (1) estimation of principal components and eigenvalues in order to get an interpretable low-dimensional data representation (principal component analysis, or PCA); (2) construction of linear discriminant functions for classification of Gaussian data (linear discriminant analysis, or LDA); (3) establishing independence and conditional independence relations between components using exploratory data analysis and testing; and (4) setting confidence intervals on linear functions of the means of the components. Note that (1) requires estimation of the eigenstructure of the covariance matrix while (2) and (3) require estimation of the inverse.

The theory of multivariate analysis for normal variables has been well worked out – see Anderson (1958), the major monograph. However, it became apparent that exact expressions were cumbersome, even for small dimensions and sample sizes, and that multivariate data were rarely Gaussian. The remedy was asymptotic theory for large samples and fixed relatively small dimensions. In recent years, datasets that do not fit into this framework have become very common – the data are very high-dimensional and sample sizes can be very small relative to dimension. Examples include gene expression arrays, fMRI data, spectroscopic imaging, numerical weather forecasting, and many others.

It has long been known that the empirical covariance matrix for samples of size n from a p-variate Gaussian distribution, $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma_p)$ has unexpected features if both p and n are large. If $p/n \to c \in (0, 1)$ and the covariance matrix $\Sigma_p = I$ (the identity), then the empirical distribution of the eigenvalues of the sample covariance matrix $\hat{\Sigma}_p$ follows the Marĉenko-Pastur law (Marĉenko and Pastur, 1967), which is supported on $((1 - \sqrt{c})^2, (1 + \sqrt{c})^2)$. Thus, the larger p/n, the more spread out the eigenvalues, even asymptotically.

Further contributions to the theory of extremal eigenvalues of $\hat{\Sigma}_p$ have been made by Wachter (1978), Geman (1980) and Bai and Yin (1993), among others. In recent years, there have been great developments by Johnstone and his students in the theory of the largest eigenvalues (Johnstone, 2001; Paul, 2004) and associated eigenvectors (Johnstone and Lu, 2006), following similar work in mathematics on the spectra of random matrices by Tracy and Widom (1996). The implications of these results for inference, other than indicating the weak points of the sample covariance matrix, are not clear since the interest of statisticians is in the population covariance matrices.

Regularizing large empirical covariance matrices has already been proposed in some statistical applications – for example, as original motivation for ridge regression (Hoerl and Kennard, 1970) and in regularized discriminant analysis (Friedman, 1989). However, only recently has there been an upsurge of both practical and theoretical analyses of such procedures – see Ledoit and Wolf (2003), Wu and Pourahmadi (2003), Huang et al. (2006), and Furrer and Bengtsson (2006) among others. These authors study different ways of regularization. Ledoit and Wolf consider Steinian shrinkage toward the identity. Furrer and Bengtsson (2006) consider "tapering" the sample covariance matrix, i.e., gradually shrinking the off-diagonal elements towards zero. Wu and Pourahmadi use the Cholesky decomposition of the covariance matrix to perform what we shall call "banding the inverse covariance matrix" below, and Huang et al. impose L_1 penalties on the Cholesky factor to achieve extra parsimony. Other uses of L_1 penalty include applying it directly to the entries of the covariance matrix (Banerjee et al., 2006) and applying it to loadings in the context of PCA to achieve sparse representation (Zou et al., 2006). Johnstone and Lu (2006) consider a different regularization of PCA, which involves moving to a sparse basis and thresholding. Implicitly these approaches postulate different notions of sparsity. Wu and Pourahmadi's interest focuses, as does ours, on situations where we can expect that $|i - j|$ large implies near independence or conditional (given the intervening indexes) independence of X_i and X_j. At the very least our solutions are appropriate for applications such as climatology and spectroscopy, where there is a natural metric on the index set. The same is true for Huang et al.'s method of regularization, which is more flexible but also depends on the order of variables. Johnstone and Lu's method presupposes that the eigenvectors corresponding to the leading principal value are sparse in some basis. We give more discussion of these issues in Section 7.

Some of these papers derive expressions for their estimators $\hat{\Sigma}_p$ which can be used to study the rate of convergence to the population covariance Σ_p as n and p both tend to ∞. The asymptotic frameworks and convergence results, if at all considered, vary among these studies. Wu and Pourahmadi (2003) consider convergence in the sense of single matrix element estimates being close to their population values in probability, with $p_n \to \infty$ at a certain rate determined by the spline smoothers they used. Ledoit and Wolf (2003) show convergence of their estimator in "normalized" Frobenius norm $\|A\|_F^2/p$ if p/n is bounded, whereas Furrer and Bengtsson (2006) use the Frobenius norm itself, $\|A\|_F^2 = \text{tr}(AA^T)$, which we shall argue below is too big. Johnstone and Lu (2006) show convergence of the first principal component of their estimator when $p/n \to$ const.

We have previously studied (Bickel and Levina, 2004) the behavior of Fisher's discriminant function for classification as opposed to the so-called “naive Bayes” procedure which is constructed under the assumption of independence of the components. We showed that the latter rule continues to give reasonable results for well-conditioned Σ_p as long as $\frac{\log p}{n} \to 0$ while Fisher's rule becomes worthless if $p/n \to \infty$. We also showed that using k-diagonal estimators of the covariance achieves asymptotically optimal classification errors if Σ_p is Toeplitz and $k_n \to \infty$ at a certain rate. However, the performance of the banded estimators was only evaluated in the context of LDA.

In this paper we show how, by either banding the sample covariance matrix or estimating a banded version of the inverse population covariance matrix we can obtain estimates which are consistent at various rates in the operator norm as long as $\frac{(\log p)^2}{n} \to 0$ and Σ_p ranges over some fairly natural families. This implies that maximal and minimal eigenvalues of our estimates and Σ_p are close. We do this in Section 2, in which we introduce our procedures, and Section 3 where we give main results. In Section 4 we introduce an analogue of the Gaussian white noise model and show that if our matrices are embeddable in that model and well-conditioned then our banded approximations are such that the eigenstructures (individual eigenvalues and associated eigenvectors) of the estimate and population covariance are close. Another approximation result not dependent on existence of the limit model is presented as well. In Section 5 we discuss the choice of k. In Section 6 we give some numerical results. Section 7 concludes with discussion, and Section A is a technical appendix.

2 The model and two types of regularized covariance estimates.

We assume throughout that we observe $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, i.i.d. p-variate random variables with mean $\mathbf{0}$ and covariance matrix Σ_p, and write

$$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T.$$

In our treatment we will assume that the $\boldsymbol{X}_i$ are multivariate normal. We shall argue separately that if suffices for X_{1j}^2 to have sub-exponential tails for all j. That is,

$$P[X_{1j}^2 \geq t] \leq Ce^{-\gamma t} \tag{1}$$

for all $t > 0$, some C, $\gamma > 0$. We want to study the behavior of estimates of Σ_p as both p and $n \to \infty$. The features of Σ_p that are often of greatest interest are eigenvalues and eigenvectors as used for PCA, and Σ_p^{-1} appearing naturally for instance in LDA and the Kalman filter. It is well known that the usual MLE of Σ_p, the sample covariance matrix,

$$\hat{\Sigma}_p = \frac{1}{n} \sum_{i=1}^{n} \big(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\big)\big(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\big)^T \tag{2}$$

behaves optimally as one might expect if p is fixed, converging to Σ_p at rate $n^{-1/2}$. However, as discussed in the Introduction, if $p \to \infty$, $\hat{\Sigma}_p$ can behave very badly unless it is “regularized” in some fashion. Here we propose to consider two methods of regularization and will comment on others.

Method I: Banding the sample covariance matrix.

For any matrix $M = [m_{ij}]_{p \times p}$, and any $0 \leq k < p$, define,

$$B_k(M) = [m_{ij} \mathbf{1}(|i - j| \leq k)]$$

and estimate the covariance by $\hat{\Sigma}_{k,p} \equiv \hat{\Sigma}_k = B_k(\hat{\Sigma}_p)$. (In the rest of the paper, we sometimes suppress the dependence on p in $\hat{\Sigma}_{k,p}$ for the sake of compactness.) This kind of regularization is ideal in the situation where the indices have been arranged in such a way that in $\Sigma_p = [\sigma_{ij}]$

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

This assumption holds, for example, if Σ_p is the covariance matrix of $Y_1, \ldots, Y_p$, where Y_1, Y_2, ... is a finite inhomogeneous moving average process:

$$Y_t = \sum_{j=1}^{k} a_{t,t-j} \varepsilon_j \tag{3}$$

and ε_j are i.i.d. mean 0. Then the covariance matrix of $(Y_1, \ldots, Y_p)^T$ is k-banded. However, banding an arbitrary covariance matrix does not guarantee positive definiteness. Take for example $\sigma_{ij} = \rho + (1-\rho)\mathbf{1}(i=j)$, $p=3$, $k=1$, and $\rho > \frac{1}{\sqrt{2}}$. As we shall see, however, for k small compared to n and p, $B_k(\hat{\Sigma}_p)$ is positive definite with probability tending to 1 as $p, n \to \infty$.

We note that the lack of assured positive definiteness for this method can be eliminated altogether. Furrer and Bengtsson (2006) have pointed out that positive definiteness can be preserved by "tapering" the covariance matrix, that is, replacing $\hat{\Sigma}_p$ with $\hat{\Sigma}_p * R$, where $*$ denotes Schur (coordinate-wise) matrix multiplication, and $R = [r_{ij}]$ is a positive definite symmetric matrix, since the Schur product of positive definite matrices is also positive definite. Banding corresponds to $r_{ij} = \mathbf{1}(|i-j| \le k)$; we discuss other choices for R that guarantee positive definiteness below.

Method II: Banding the inverse.

This method is based on the Cholesky decomposition of the inverse which forms the basis of the estimators proposed by Wu and Pourahmadi (2003) and Huang et al. (2006). Here is our way of approaching this method. Suppose we have $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ defined on a probability space, with probability measure P, which is $\mathcal{N}_p(\mathbf{0}, \Sigma_p)$, $\Sigma_p \equiv [\sigma_{ij}]$. Let

$$\hat{X}_j = \sum_{t=1}^{j-1} a_{jt} X_t = \boldsymbol{Z}_j^T \boldsymbol{a}_j \tag{4}$$

be the $L_2(P)$ projection of X_j on the linear span of $X_1, \ldots, X_{j-1}$, with $\boldsymbol{Z}_j = (X_1, \ldots, X_{j-1})^T$ the vector of coordinates up to $j-1$, and $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{j,j-1})^T$ the coefficients. If $j=1$, let $\hat{X}_1 = 0$. Each vector $\boldsymbol{a}_j^T$ can be computed as

$$\boldsymbol{a}_j = (\mathrm{Var}(\boldsymbol{Z}_j))^{-1} \mathrm{cov}(X_j, \boldsymbol{Z}_j). \tag{5}$$

Note that we assumed mean $\mathbf{0}$ so here $\mathrm{cov}(\mathbf{U}, \mathbf{V}) \equiv E\mathbf{U}\mathbf{V}^T$, and we write $\mathrm{Var}(\mathbf{U})$ for $\mathrm{cov}(\mathbf{U}, \mathbf{U})$. Now, let

$$\varepsilon_j = \frac{X_j - \hat{X}_j}{d_j} \tag{6}$$

where

$$d_j^2 = \mathrm{Var} X_j - \mathrm{Var} \hat{X}_j = \sigma_{jj} - \sum_{l,m=1}^{j-1} a_{jl} a_{jm} \sigma_{lm}. \tag{7}$$

The geometry of $L_2(P)$ or standard regression theory show that $\varepsilon_1, \ldots, \varepsilon_p$ are independent. Under the normal assumption on the variables, they are also $\mathcal{N}(0,1)$. Let the lower triangular matrix A

$$A = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{p,p-1} & 0 \end{bmatrix}$$

contain the coefficients of the regressions (4). Let I be the identity and $D = \mathrm{diag}(d_j^2)$ a diagonal matrix with d_j^2 on the diagonal. Then (6) can be rewritten as

$$\boldsymbol{X} = (I - A)^{-1} D^{1/2} \boldsymbol{\varepsilon} \tag{8}$$

which implies

$$\begin{aligned} \Sigma_p &= (I-A)^{-1} D [(I-A)^{-1}]^T, \\ \Sigma_p^{-1} &= (I-A)^T D^{-1} (I-A), \end{aligned} \tag{9}$$

giving the modified Cholesky decompositions of Σ_p and Σ_p^{-1}.

Suppose now that $k < p$. It is natural to define an approximation to Σ_p by restricting the variables in regression (4) to $\boldsymbol{Z}_j^{(k)} = (X_{\max(j-k,1)}, \ldots, X_{j-1})$, that is, regressing each X_j on its closest k predecessors only. Note that for $j < k$, $\boldsymbol{Z}_j^{(k)} = \boldsymbol{Z}_j$. We can now similarly define $\hat{X}_1^{(k)} = 0$, and for $j > 1$,

$$\hat{X}_j^{(k)} = \sum_{t=\max(1,j-k)}^{j-1} a_{jt}^{(k)} X_t = (\boldsymbol{Z}_j^{(k)})^T \boldsymbol{a}_j^{(k)} \tag{10}$$

where $\varepsilon_1, \ldots, \varepsilon_p$ are again i.i.d. $\mathcal{N}(0,1)$. Replacing $\boldsymbol{Z}_j$ by $\boldsymbol{Z}_j^{(k)}$ in (5) gives the new coefficients

$$\boldsymbol{a}_j^{(k)} = (\mathrm{Var}(\boldsymbol{Z}_j^{(k)}))^{-1} \mathrm{cov}(X_j, \boldsymbol{Z}_j^{(k)}). \tag{11}$$

Let A_k be the k-banded lower triangular matrix containing the new vectors of coefficients $\boldsymbol{a}_j^{(k)}$ defined by (11), and let $D_k = \mathrm{diag}(d_{j,k}^2)$ be the diagonal matrix containing the corresponding residual variances

$$d_{j,k}^2 = \mathrm{Var} X_j - \mathrm{Var} \hat{X}_j^{(k)}. \tag{12}$$

Now we define

$$\begin{aligned} \Sigma_{k,p} &= (I-A_k)^{-1} D_k [(I-A_k)^{-1}]^T, \\ \Sigma_{k,p}^{-1} &= (I-A_k) D_k^{-1} (I-A_k)^T. \end{aligned} \tag{13}$$

Given a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$, the natural estimates of A_k and D_k, are obtained by performing the operations needed under $\hat{P}$, the empirical distribution, i.e., plugging in the ordinary least squares estimates of the coefficients in A_k and the corresponding residual variances in D_k. This means plugging in sample versions of covariances into (11) and (12), e.g.,

$$\widehat{\mathrm{Var}}(\boldsymbol{Z}_j^{(k)}) = \frac{1}{n} \sum_{i=1}^{n} (X_{i,j-k}, \ldots, X_{i,j-1})^T (X_{i,j-k}, \ldots, X_{i,j-1}).$$

Finally, so far we have been assuming that $\boldsymbol{X}_i$ have mean 0; in the general case, replace $\boldsymbol{X}_i$ by $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$ in the estimates of A_k and D_k, where $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$. We will refer to these sample estimates as $\tilde{A}_k = [\tilde{a}_{jt}^{(k)}]$, and $\tilde{D}_k = \text{diag}(\tilde{d}_{j,k}^2)$. Thus we obtain our final estimates of Σ_p^{-1} and Σ_p via the Cholesky decomposition:

$$\begin{aligned}\tilde{\Sigma}_{k,p}^{-1} \equiv \tilde{\Sigma}_k^{-1} &= (I - \tilde{A}_k)^T \tilde{D}_k^{-1}(I - \tilde{A}_k), \\ \tilde{\Sigma}_{k,p} \equiv \tilde{\Sigma}_k &= (I - \tilde{A}_k)^{-1}\tilde{D}_k[(I - \tilde{A}_k)^{-1}]^T.\end{aligned} \tag{14}$$

Note that since $\tilde{A}_k$ is a k-banded lower triangular matrix, $\tilde{\Sigma}_k^{-1}$ is k-banded nonnegative definite. Its inverse $\tilde{\Sigma}_k$ is in general not banded, and is different from $\hat{\Sigma}_k$. Similarly, $\tilde{\Sigma}_k^{-1}$ is not the same as $B_k(\hat{\Sigma}^{-1})$, which is in any case ill-defined when $p > n$.

3 Main Results

Our results can be made uniform on sets of covariance matrices which we now define. All our sets will be subsets of the set which we shall refer to as *well conditioned covariance matrices*, Σ_p, such that, for all p,

$$0 < \varepsilon \leq \lambda_{\min}(\Sigma_p) \leq \lambda_{\max}(\Sigma_p) \leq 1/\varepsilon < \infty.$$

Here, $\lambda_{\max}(\Sigma_p)$, $\lambda_{\min}(\Sigma_p)$ are the maximum and minimum eigenvalues of Σ_p, and ε is independent of p.

As noted in Bickel and Levina (2004), examples of such matrices include covariance matrices of $(U_1, \ldots, U_p)^T$ where $\{U_i, i \leq 1\}$ is a stationary ergodic process with spectral density f, $0 < \varepsilon \leq f \leq \frac{1}{\varepsilon}$ and, more generally, of $X_i = U_i + V_i$, $i = 1, \ldots,$ where $\{U_i\}$ is a stationary process as above and $\{V_i\}$ is a noise process independent of $\{U_i\}$. This model includes the "spike model" of Paul (2004) since a matrix of bounded rank is Hilbert-Schmidt.

In what follows we will use several vector and matrix norms which we now define. For a vector $\boldsymbol{x} = (x_1, \ldots, x_p)^T$, let

$$\|\boldsymbol{x}\| = \big(\sum_{j=1}^p x_j^2\big)^{1/2}, \; \|\boldsymbol{x}\|_1 = \sum_{j=1}^p |x_j|, \|\boldsymbol{x}\|_\infty = \max_j |x_j|.$$

For a matrix $M = [m_{ij}]$, the corresponding operator norms from l_2 to l_2, l_1 to l_1, and l_∞ to l_∞ are, respectively,

$$\begin{aligned}&\|M\| \equiv \sup\{\|M\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\} = \lambda_{\max}^{1/2}(M^T M), \\ &\|M\|_{(1,1)} \equiv \sup\{\|M\boldsymbol{x}\|_1 : \|\boldsymbol{x}\|_1 = 1\} = \max_j \sum_i |m_{ij}|, \\ &\|M\|_{(\infty,\infty)} \equiv \sup\{\|M\boldsymbol{x}\|_\infty : \|\boldsymbol{x}\|_\infty = 1\} = \max_i \sum_j |m_{ij}|.\end{aligned} \tag{15}$$

We will also write $\|M\|_\infty \equiv \max_{i,j} |m_{ij}|$.

As is well known, if M is symmetric,

$$\|M\| = \max\{|\lambda_1|, \ldots, |\lambda_p|\} \tag{16}$$

since all eigenvalues are real, and if M is invertible,

$$\|M^{-1}\| = [\min\{|\lambda_1|, \dots, |\lambda_p|\}]^{-1} . \tag{17}$$

For symmetric matrices, $\|M\|_{(1,1)} = \|M\|_{(\infty,\infty)}$. The l_1 to l_1 norm arises naturally through the inequality (see e.g. Golub and Van Loan (1989))

$$\|M\| \leq \left[\|M\|_{(1,1)}\|M\|_{(\infty,\infty)}\right]^{\frac{1}{2}} = \|M\|_{(1,1)} \text{ for } M \text{ symmetric.} \tag{18}$$

We define classes of positive definite symmetric matrices $\Sigma \equiv [\sigma_{ij}]$ as follows.

$$\begin{aligned}\mathcal{U}(\varepsilon_0, C, \alpha) = \big\{\Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Ck^{-\alpha} \text{ for all } k \geq 0, \\ \text{and } 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0\big\}.\end{aligned} \tag{19}$$

Contained in $\mathcal{U}$ for suitable ε_0, α, C is the class,

$$\begin{aligned}&\mathcal{L}(\varepsilon_0, m, C) = \{\Sigma : \ \sigma_{ij} = \sigma(i-j) \text{ (Toeplitz)} \\ &\text{with spectral density } f_\Sigma, \ 0 < \varepsilon_0 \leq \|f_\Sigma\|_\infty \leq \varepsilon_0^{-1}, \ \|f_\Sigma^{(m)}\|_\infty \leq C\} ,\end{aligned}$$

where $f^{(m)}$ denotes the m-th derivative of f. By Grenander and Szegö (1984), if Σ is symmetric, Toeplitz, $\Sigma \equiv [\sigma(i-j)]$, with $\sigma(-k) = \sigma(k)$, and Σ has an absolutely continuous spectral distribution with Radon-Nikodym derivative, which is continuous on $(-1, 1)$,

$$f_\Sigma(u) \equiv \sum_{t=0}^{\infty} \sigma(t) \cos(2\pi t u)$$

then,

$$\|\Sigma\| = \sup_u |f_\Sigma(u)|, \tag{20}$$

$$\|\Sigma^{-1}\| = [\inf_u |f_\Sigma(u)|]^{-1}. \tag{21}$$

Since $\|f_\Sigma^{(m)}\|_\infty \leq C$ implies that

$$|\sigma(t)| \leq Ct^{-m} \tag{22}$$

which in turn implies $\sum_{t>k} \sigma(t) \leq C(m-1)^{-1}k^{-m+1}$, we conclude from (20), (21) and (22) that,

$$\mathcal{L}(\varepsilon_0, m, C) \subset \mathcal{U}(\varepsilon_0, m-1, C) . \tag{23}$$

A second uniformity class of nonstationary covariance matrices is defined by

$$\mathcal{K}(m, C) = \left\{\Sigma : \sigma_{ii} \leq Ci^{-m}, \text{ all } i\right\} .$$

The bound C independent of dimension identifies any limit as being of "trace class" as operator for $m > 1$.

Although $\mathcal{K}$ is not a well conditioned class,

$$\mathcal{T}(\varepsilon_0, m_1, m_2, C_1, C_2) \equiv \Big\{\Sigma : \Sigma = L + K, \ L \in \mathcal{L}(\varepsilon_0, m_1, C_1), \ K \in \mathcal{K}(m_2, C_2)\Big\} \subset \mathcal{U}(\varepsilon, \alpha, C), \tag{24}$$

where $\alpha = \min\{m_1 - 1, m_2/2 - 1\}$, $C \le (C_1/(m_1 - 1) + C_2/(m_2/2 - 1)$, $\varepsilon^{-1} \le \varepsilon_0^{-1} + C_2$. To check claim (24), note that

$$\varepsilon_0 \le \lambda_{\min}(L) \le \lambda_{\min}(L + K) \le \lambda_{\max}(L + K) \le \|L\| + \|K\| \le \varepsilon_0^{-1} + C_2,$$

and

$$\max_{j \ge k} \sum_{i:|i-j|>k} |K_{ij}| \le \max_{j \ge k} \sum_{i:|i-j|>k} |K_{ii}|^{1/2} |K_{jj}|^{1/2} \le C_2 (m_2/2 - 1)^{-1} k^{-m_2/2+1}$$

$$\max_{j<k} \sum_{i:|i-j|>k} |K_{ii}|^{1/2} |K_{jj}|^{1/2} \le C_2^{1/2} \sum_{i=k+2}^{p} |K_{ii}|^{1/2} \le C_2 (m_2/2 - 1)(k + 2)^{-m_2/2+1}$$

We will use the $\mathcal{T}$ and $\mathcal{L}$ classes for Σ_p and Σ_p^{-1} for convenience.

Theorem 1. *Suppose that $\boldsymbol{X}$ is Gaussian and $\mathcal{U}(\varepsilon_0, \alpha, C)$ is the class of covariance matrices defined above. Then, if $k_n \asymp (n^{-1/2} \log p)^{-\frac{1}{\alpha+1}}$,*

$$\|\hat{\Sigma}_{k_n,p} - \Sigma_p\| = O_P\left(\left(n^{-1/2} \log p\right)^{\frac{\alpha}{\alpha+1}}\right) = \|\hat{\Sigma}_{k_n,p}^{-1} - \Sigma_p^{-1}\| \tag{25}$$

uniformly on $\Sigma \in \mathcal{U}$.

Immediately, we obtain,

Corollary 1. *If $\alpha = \min\left\{m_1 - 1, \frac{m_2}{2} - 1\right\}$, $m_1 > 1, m_2 > 2$, then (25) holds uniformly for $\Sigma \in \mathcal{T}(\varepsilon_0, m_1, m_2, C_1, C_2)$.*

Proof of Theorem 1: It is easy to see that (18) and the difinitions (15) imply

$$\|B_k(\hat{\Sigma}) - B_k(\Sigma)\| = O_P\big(k \|B_k(\hat{\Sigma}) - B_k(\Sigma)\|_\infty\big). \tag{26}$$

Let $\hat{\Sigma}^0 = \frac{1}{n} \Sigma_{i=1}^n \boldsymbol{X}_i^T \boldsymbol{X}_i$ and w.l.o.g. $E\boldsymbol{X}_1 = \boldsymbol{0}$. By a simplification of Lemmas 3 and 4 of Bickel and Levina (2004) (see Lemma 3 in the Appendix) and the union sum inequality,

$$P\big[\|B_k(\hat{\Sigma}^0) - B_k(\Sigma)\|_\infty \ge t\big] \le (2k + 1)p\ C(\varepsilon_0) \exp\{-nc(t, \varepsilon_0)\} \tag{27}$$

where $c(t, \varepsilon_0) = \min\{a(\varepsilon_0)t, b(\varepsilon_0)t^2\}$ and $a, b > 0$. By choosing $t = Mn^{-1/2} \log(pk)$ for M arbitrary we conclude that, uniformly on $\mathcal{U}$,

$$\|B_k(\hat{\Sigma}^0) - B_k(\Sigma_p)\|_\infty = O_P\big(n^{-1/2} \log(pk)\big) = O_P\big(n^{-1/2} \log(p)\big) \tag{28}$$

since $k < p$. On the other hand, by (19),

$$\|B_k(\Sigma_p) - \Sigma_p\|_\infty \le Ck^{-\alpha} \tag{29}$$

for $\Sigma_p \in \mathcal{U}(\varepsilon_0, \alpha, C)$.

Combining (28) and (29) the result follows for $B_k(\hat{\Sigma}^0)$. But, if $\bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_p)^T$,

$$\|B_k(\hat{\Sigma}^0) - B_k(\hat{\Sigma})\| \le \|B_k(\bar{\boldsymbol{X}}^T \bar{\boldsymbol{X}})\| \le (2k+1) \max_{1 \le j \le p} |\bar{X}_j|^2 = O_P\left(\frac{k \log p}{n}\right) = O_P\left(\left(n^{-1/2} \log p\right)^{\frac{\alpha}{\alpha+1}}\right)$$

Since

$$\big\|\big[B_{k_n}(\hat{\Sigma})\big]^{-1} - \Sigma_p^{-1}\big\| = \Omega_P\big(\|B_{k_n}(\hat{\Sigma}) - \Sigma_p\|\big)\ ,$$

uniformly on $\mathcal{U}$, the result follows. □

Extensions

I. The Gaussian assumption may be replaced by the following. Suppose $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ are i.i.d., $X_{1j} \sim F_j$, where F_j is the c.d.f. of X_{1j}, and $G_j(t) = F_j\big(\sqrt{t}\big) - F_j\big(-\sqrt{t}\big)$ is the c.d.f. of X_{1j}^2. Then for Theorem 1 to hold it suffices to assume that

$$\bar{G}_j(t) \equiv 1 - G_j(t) \le Ce^{-\gamma t} \tag{30}$$

for all $t \ge 0$, $1 \le j \le p$, $\gamma > 0$ and C fixed. This follows from

$$\begin{aligned} &\textstyle\int_0^\infty \exp(\lambda t) dG_j(t) = -\int_0^\infty \exp(\lambda t) d\bar{G}_j(t) \\ &= 1 + \lambda \int\limits_0^\infty \bar{G}(t)\exp(\lambda t)dt \le 1 + C\tfrac{\lambda}{\gamma-\lambda} \ . \end{aligned} \tag{31}$$

for $\lambda < \gamma$. Arguing as in Lemma 3 of Bickel and Levina (2004), (27) follows for $t \ge \frac{a(\varepsilon_0)}{b(\varepsilon_0)}$.

II. If we only assume $E|X_{ij}|^\beta \le C$, $\beta > 2$, for all j, we can replace (27) by

$$P\Big[\|B_k\big(\hat{\Sigma}^0\big) - B_k\big(\Sigma_p\big)\|_\infty \ge t\Big] \le Cn^{-\frac{\beta}{4}}(2k+1)pt^{-\frac{\beta}{2}} \ . \tag{32}$$

Then (26), (28) and (29) imply that if $k_n \asymp (n^{-1/2}p^{2/\beta})^{-\gamma(\alpha)}$ where $\gamma(\alpha) = \big(1 + \alpha + 2/\beta\big)^{-1}$, then,

$$\|B_{k_n}\big(\hat{\Sigma}^0\big) - \Sigma_p\| = O_P\left((n^{-1/2}p^{2/\beta})^{\alpha\gamma(\alpha)}\right) \ . \tag{33}$$

Again, the passage from $\hat{\Sigma}^0$ to $\hat{\Sigma}$ is straightforward.

Remarks

1) Theorem 1 implies that $\|B_{k_n}\big(\hat{\Sigma}\big) - \Sigma_p\| \xrightarrow{P} 0$ if $\frac{(\log p)^2}{n} \to 0$, uniformly on $\mathcal{U}$. It is not hard to see that if $\Sigma_p = S + K$ where S is Toeplitz, $\varepsilon_0 \le f_S \le \varepsilon_0^{-1}$ and K is trace class in the sense of Section 4, $\Sigma_i K(i,i) < \infty$, then, if $\frac{(\log p)^2}{n} \to 0$, there exist $k_n \uparrow \infty$ such that, for the given $\{\Sigma_p\}$

$$\big\|B_{k_n}\big(\hat{\Sigma}\big) - \Sigma_p\big\| + \big\|\big[B_{k_n}\big(\hat{\Sigma}\big)\big]^{-1} - \Sigma_p^{-1}\big\| \xrightarrow{P} 0 \ . \tag{34}$$

2) The same claim can be made under (30). On the other hand, under only the moment bound of II with $Ee^{\lambda X_{ij}^2} = \infty$, $\lambda > 0$ we may only conclude that (34) holds if

$$\frac{p^{\frac{4}{\beta}}}{n} \to 0 \ . \tag{35}$$

Related results of Furrer and Bengtsson (2006) necessarily have rates of the type (35) not because of tail conditions on the variables, but because they consider the Frobenius norm.

3) The rate $\frac{(\log p)^2}{n} \to 0$ appears in these cases rather than the rate $\frac{\log p}{n} \to 0$ as in Bickel and Levina (2004) because we are not assuming stationarity. In particular, our initial estimate of Σ_p is just $\hat{\Sigma}$, while in the stationary case we estimated $S(k)$ by $\frac{1}{p-k}\Sigma\big\{\hat{\sigma}_{ij} : j = i + k\big\}$ which enabled us to use the $\exp\{-b(\varepsilon_0)t^2\}$ part of the bound of Lemma 3 of Bickel and Levina (2004).

There is an important generalization of Theorem 1. Let A be a countable set of labels of cardinality $|A|$. We can think of a matrix as $[m_{ab}]_{a\in A,\ b\in A}$.

Let $\rho : A \times A \to R^+,\ \rho(a,a) = 0$ for all a, be a function we can think of as distance of the point (a,b) from the diagonal. As an example think of a and b as identified with points in R^m and $\rho(a,b) = |a-b|$ where $|\cdot|$ is a norm on R^m. It is then clear how to generalize the notion of banding by defining, if $M = [m_{ab}]_{a,b\in A}$,

$$B_k(M) = \big[m_{ab} 1(\rho(a,b) \le k)\big] \ .$$

We can even go further, following the example of Furrer and Bengtsson (2006), and use a smoother method of regularization than banding.

Suppose $R = [r_{ab}]_{a,b\in A}$ is symmetric positive definite with $r_{ab} = g\big(\rho(a,b)\big)$, $g : R^+ \to R^+$. Then, if M is also symmetric nonnegative definite,

$$R * M \equiv [m_{ab} r_{ab}]_{a,b\in A}$$

where $*$ denotes Schur multiplication, is positive definite (unless $M = 0$). Suppose further that $g(0) = 1$ and g is decreasing to 0. Then $R * M$ is a regularization of M. Note that $g(t) = 1(t \le k)$, $\rho(i,j) = |i-j|$ gives banding. However, $\big[g(|i-j|)\big]$ is not nonnegative definite.

In general, let $R_\sigma = [r_\sigma(a,b)]$, where

$$r_\sigma(a,b) = g\left(\frac{\rho(a,b)}{\sigma}\right) \ , \quad \sigma \ge 0.$$

Assumption A. g is continuous, $g(0) = 1$, g is non-increasing, $g(\infty) = 0$.

Examples of positive definite symmetric R_σ are,

$$r_\sigma(i,j) = \left(1 - \frac{|i-j|}{\sigma}\right)_+$$

or

$$r_\sigma(i,j) = e^{-\frac{|i-j|}{\sigma}} \ .$$

With this notation define,

$$R_\sigma(M) \equiv \big[m_{ab} g_\sigma\big(\rho(a,b)\big)\big]$$

with $R_0(M) = M$. Clearly, as $\sigma \to \infty$, $R_\sigma(M) \to M$.

Our generalization is the following. Denote the range of $g_\sigma\big(\rho(a,b)\big)$ by $\{g_\sigma(\rho_1), \ldots, g_\sigma(\rho_L)\}$ where $\{0 < \rho_1 < \ldots < \rho_L\}$ is the range of $\rho(a,b)$, $a \in A$, $b \in A$. Note that L depends on $|A| = p$.

Theorem 2. *Let $\Delta(\sigma^\varepsilon) = \Sigma_{l=1}^L g_\sigma(\rho_l)$. Note that Δ depends on $|\mathcal{A}| = p$ and the range of ρ. Suppose Assumption A holds. Then if*

$$\Delta \asymp (n^{-1/2} \log p)^{-\frac{1}{\alpha+1}}$$

the conclusion of Theorem 1 holds for $R_\sigma(\hat\Sigma)$.

The proof of Theorem 2 closely follows the proof of Theorem 1 with (26) replaced by Lemma 1 in the Appendix. Both the result and the lemma are of independent interest.

The remarks after Theorem 1 generalize equally. Note that Theorem 1 is a special case of Theorem 2 with $\mathcal{A} = \{1, 2, \ldots, p\}$, $\rho(a, b) = |a - b|$ and $g(u) = \mathbf{1}(u \leq 1)$.

Theorems 1 and 2 give the scope of what can be accomplished by banding the sample covariance matrix. "Banding the inverse" yields similar results.

If $\Sigma^{-1} = T(\Sigma)^T D^{-1}(\Sigma) T(\Sigma)$ with $T(\Sigma)$ lower triangular, $T(\Sigma) \equiv [t_{ij}(\Sigma)]$, let

$$\begin{aligned} U^{-1}(\varepsilon_0, C, \alpha) = \big\{\Sigma : 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \varepsilon_0^{-1}, \\ \max_i \sum_{j < i-k} |t_{ij}(\Sigma)| \leq Ck^{-\alpha} \quad \text{for all } k \leq p-1 \ \big\} \end{aligned}$$

Theorem 3. *Uniformly for $\Sigma \in U^{-1}(\varepsilon_0, C, \alpha)$, if $k_n \asymp (n^{-1/2} \log p)^{-\frac{1}{\alpha+1}}$ and $n^{-1/2} \log p = o_P(1)$,*

$$\|\tilde{\Sigma}^{-1}_{k_n,p} - \Sigma^{-1}_p\| = O_P\left(\left(n^{-1/2} \log p \right)^{\frac{\alpha}{\alpha+1}} \right) = \|\tilde{\Sigma}_{k_n,p} - \Sigma_p\| \ .$$

The proof is given in the Appendix. Note that the condition $n^{-1/2} \log p = o_P(1)$ is needed solely for the purpose of omitting a cumbersome and uninformative term from the rate (see Lemma 2 in the Appendix for details).

It is *a priori* not clear what $\Sigma \in \mathcal{U}^{-1}$ means in terms of Σ. The following corollary to Theorem 3 gives a partial answer.

Corollary 2. *For $m \geq 2$, uniformly on $\mathcal{L}(\varepsilon_0, m, C)$, if $k_n \asymp (n^{-1/2} \log p)^{-\frac{1}{m}}$,*

$$\begin{aligned} &\|\tilde{\Sigma}^{-1}_{k_n,p} - \Sigma^{-1}_p\| = O_P\left((n^{-1/2} \log p)^{\frac{m-1}{m}} \right) \\ &= \|\tilde{\Sigma}_{k_n,p} - \Sigma\| \ . \end{aligned}$$

Essentially, banding the inverse works just as well as banding for suitably ergodic stationary processes. The proof of Corollary 2 is given in the Appendix. The reason that the argument of Theorem 1 can not be invoked simply for Theorem 3 is that, as we noted before, $\tilde{\Sigma}^{-1}$ is not the same as $B_k(\hat{\Sigma}^{-1})$, which is not well defined if $p > n$.

4 An analogue of the Gaussian white noise model and eigenstructure approximations

In estimation of the means $\boldsymbol{\mu}_p$ of p-vectors of i.i.d. variables, the Gaussian white noise model (Donoho et al., 1995) is the appropriate infinite dimensional model into which all objects of interest are embedded. In estimation of matrices, a natural analogue is the space $\mathcal{B}(l_2, l_2)$, which we write as $\mathcal{B}$, of bounded linear operators from l_2 to l_2. These can be represented as matrices $[m_{ij}]_{i \geq 1, j \geq 1}$ such that, $\sum\limits_i [\sum\limits_j m_{ij} x_j]^2 < \infty$ for all $\boldsymbol{x} = (x_1, x_2, \ldots) \in l_2$. It is well known, see Böttcher (1996) for example, that if M is such an operator, then,

$$\|M\|^2 = \sup\{(M\boldsymbol{x}, M\boldsymbol{x}) : \ |\boldsymbol{x}| = 1\} = \sup \mathcal{S}(M^* M)$$

where $M^* M$ is a self adjoint member of $\mathcal{B}$ with nonnegative spectrum $\mathcal{S}$. Recall that the spectrum $\mathcal{S}(A)$ of a self adjoint operator is $\mathcal{R}^c(A)$ where, $\mathcal{R}(A) \equiv \{\lambda \in R : \ A - \lambda J \in \mathcal{B}\}$ where J is the identity. To familiarize ourselves with this space we cite some results from functional analysis and some properties of $\Sigma \in \mathcal{B}$ where

$$\Sigma = \big[\mathrm{cov}\big(X(i), X(j)\big)\big]_{i,j \geq 1} \tag{36}$$

is the matrix of covariances of a Gaussian stochastic process $\{X(t) : t = 1, 2, \ldots\}$.

Functional analytic properties of self adjoint $T \in \mathcal{B}$

1. Properties (16) and (17) continue to hold, save that max and min eigenvalues are replaced by supremum and infimum of $\mathcal{S}(T)$.

2. The spectral theorem holds (Riesz and Sz-Nagy, 1955). If T is as above, there exists a unique projection valued measure $E(\cdot)$ on $\mathcal{S}$ with $E(\mathcal{S}) = J$, $E(\emptyset) = 0$, and $E(\cup_{j=1}^{\infty} A_j) = \sum_j E(A_j)$ if the A_j are mutually disjoint, such that,

$$T = \int_{\mathcal{S}(T)} \lambda E(d\lambda) \ .$$

3. Suppose that T is a Toeplitz matrix

$$T = [\rho(i-j)]_{i \geq 1, j \geq 1}$$

$\rho(k) = \rho(-k)$ for all k. Then $T \in \mathcal{B}$ iff T has a spectral density,

$$f_T(u) = \sum_{k=-\infty}^{\infty} \rho(k) \exp\{2\pi\sqrt{-1}ku\}$$

which is bounded on $[-1, 1]$ and then

$$\|T\| = \operatorname*{ess\,sup}_u |f_T(u)| \tag{37}$$

and $T^{-1} \in \mathcal{B}$ iff

$$\operatorname{ess\,inf} |f_T(u)| > 0$$

and then

$$\|T^{-1}\| = \{\operatorname{ess\,inf} |f_T(u)|\}^{-1}$$

(Grenander and Szegö (1984) for example.)

Properties of covariance matrices of Gaussian processes $X(\cdot)$

1. It is easy to see that the operators Σ for all ergodic AR processes, $X(t) = \rho X(t-1) + \varepsilon(t)$ where $\varepsilon(t)$ are i.i.d. $\mathcal{N}(0, 1)$ and $|\rho| < 1$ are in $\mathcal{B}$, and $\Sigma^{-1} \in \mathcal{B}$. This is, in fact, true of all ergodic ARMA processes. On the other hand, $X(t) \equiv \sum_{j=1}^{t} \varepsilon(j)$ has

$$\Sigma = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots \\ 1 & 2 & 2 & 2 & \cdots \\ 1 & 2 & 3 & 3 & \cdots \\ & & \cdots & & \end{bmatrix}$$

which is evidently not a member of $\mathcal{B}$.

2. The property $\Sigma \in \mathcal{B}$, $\Sigma^{-1} \in \mathcal{B}$ which we shall refer to as being well conditioned, has strong implications. By a theorem of Kolmogorov and Rozanov (see Ibragimov and Linnik (1971)), if Σ is Toeplitz, this property holds iff the corresponding stationary Gaussian process is strongly mixing.

We now consider sequences of covariance matrices Σ_p such that Σ_p is the upper $p \times p$ matrix of the operator $\Sigma \in \mathcal{B}$. That is, Σ is the covariance matrix of $\{X(t) : \ t = 1, 2, \ldots\}$ and Σ_p that of $\big(X(1), \ldots, X(p)\big)$.

By Böttcher (1996) if Σ is well conditioned then,

$$\Sigma_p(x) \to \Sigma(x)$$

as $p \to \infty$ for all $x \in l_2$. We now combine Theorem 6.1 p. 120 and Theorem 5.1, p. 474 of Kato (1966) to indicate in what ways the spectra and eigenstructures (spectral measures) of $B_{k_n}\big(\hat{\Sigma}_p\big)$ are close to those of Σ_p.

Suppose that the conditions of Remark 1) following Theorem 1 hold. That is, Σ_p corresponds to $\Sigma = S + K$ where $S \in \mathcal{B}$ is a Toeplitz operator with spectral density f_S such that, $0 < \varepsilon_0 \le f_S \le \varepsilon_0^{-1}$ and K is trace class, $\sum_u K(u,u) < \infty$ which implies $K \in \mathcal{B}$.

Let M be a symmetric matrix and $\mathcal{O}$ be an open set containing $\mathcal{S}(M) \equiv \{\lambda_1, \ldots, \lambda_p\}$ where $\lambda_1(M) \ge \lambda_2(M) \ge \ldots \ge \lambda_p(M)$ are the ordered eigenvalues of M and let $E(M)(\cdot)$ be the spectral measure of M which assigns to each eigenvalue the projection operator corresponding to its eigenspace. Abusing notation, let $E_p \equiv E(\Sigma_p)$, $\hat{E}_p \equiv E\big(\hat{\Sigma}_{k,p}\big)$, $\mathcal{S} \equiv \mathcal{S}(\Sigma_p)$. Then, $E_p(\mathcal{O}) = E_p(\mathcal{S}) = J$, the identity.

Theorem 4. *Under the above conditions on Σ_p,*

$$|\hat{E}_p(\theta)(x) - x| \xrightarrow{P} 0 \tag{38}$$

for all $x \in l_2$. Further, if $\mathcal{I}$ is any interval whose endpoints do not belong to $\mathcal{S}$ then,

$$|\hat{E}_p(\mathcal{I} \cap \mathcal{S})(x) - E_p(\mathcal{I})(x)| \xrightarrow{P} 0 \ .$$

Similar remarks apply to $\tilde{\Sigma}_{k,p}$. This result gives no information about rates. It can be refined (Theorem 5.2, p.475 of Kato (1966)) but still yields very coarse information. One basic problem is that Σ typically has at least in part continuous spectrum and another is that the errors involve the irrelevant bias $|(\Sigma_p - \Sigma)(x)|$. Here is a more appropriate formulation whose consequences for principal component analysis are clear. Let

$$\mathcal{G}(\varepsilon, \alpha, C, \Delta, m) = \{\Sigma_p \in \mathcal{U}(\varepsilon, \alpha, C) : \lambda_j(\Sigma_p) - \lambda_{j-1}(\Sigma_p) \ge \Delta, \ 1 \le j \le m\} \tag{39}$$

Thus the top m eigenvalues are consecutively separated by at least Δ and all eigenvalues λ_j with $j \ge m+1$ are separated from the top m by at least Δ. Furthermore, the dimension of the sum of the eigenspaces of the top m eigenvalues is bounded by l independent of n and p. We can then state

Theorem 5. *Uniformly on $\mathcal{G}$ as above, for k as in Theorem 1, $\boldsymbol{X}$ Gaussian,*

$$|\lambda_j\big(\hat{\Sigma}_{k,p}\big) - \lambda_j(\Sigma_p)| = O_P\Big(\Big[n^{-1/2}\log p\Big]\big(\log n + \frac{\alpha}{2}\log p\big)\Big) = \|E_j\big(\hat{\Sigma}_{k,p}\big) - E_j(\Sigma_p)\| \tag{40}$$

for $1 \le j \le m$.

That is, the top m eigenvalues and principal components of Σ_p, if the eigenvalues are all simple, are well approximated by those of $\hat{\Sigma}_{k,p}$. If we make an additional assumption on Σ_p,

$$\frac{\sum_{j=m+1}^{p} \lambda_j(\Sigma_p)}{\sum_{j=1}^{p} \lambda_j(\Sigma_p)} \leq \delta \ , \tag{41}$$

we can further conclude that the top m principal components of $\hat{\Sigma}_{k,p}$ capture $100(1-\delta)\%$ of the variance of $\boldsymbol{X}$. To verify (41) we need that,

$$\frac{\operatorname{tr}\bigl(\hat{\Sigma}_p - \Sigma_p\bigr)}{\operatorname{tr}(\Sigma_p)} = o_P(1) \tag{42}$$

This holds if, for instance, $\operatorname{tr}(\Sigma_p) = \Omega_p(p)$ which is certainly the case for all $\Sigma_p \in \mathcal{T}$. Then, Theorem 5 follows from Theorem 6.1, p.120 of Kato (1966), for instance. For simplicity, we give a self-contained proof.

Proof of Theorem 5. We employ a famous formula of Kato (1949) and Sz.-Nagy (1946). If $R(\lambda, M) \equiv (M - \lambda J)^{-1}$ for $\lambda \in \mathcal{S}^c$, the resolvent set of M and λ_0 is an isolated eigenvalue, $|\lambda - \lambda_0| \geq \Delta$ for all $\lambda \in \mathcal{S}$, $\lambda \neq \lambda_0$, then (Formula (1.16), p.67, Kato (1966))

$$E_0(x) = \frac{1}{2\pi i} \int_{\Gamma} R(\lambda, M) d\lambda \tag{43}$$

where E_0 is the projection operator on the eigenspace corresponding to λ_0 and Γ is a closed simple contour in the complex plane about λ_0 containing no other member of $\mathcal{S}$. The formula is valid not just for symmetric M but we only employ it there. We argue by induction on m. For $m = 1$, $|\lambda_1(M) - \lambda_1(N)| \leq \|M - N\|$ for M, N symmetric by the Courant-Fischer Theorem. Thus, if $\|\hat{\Sigma}_{k,p} - \Sigma_p\| \leq \frac{\Delta}{2}$ (say) we can find Γ containing $\lambda_1\bigl(\hat{\Sigma}_{k,p}\bigr)$ and $\lambda_1(\Sigma_p)$ and no other eigenvalues of either matrix with all points on Γ at distance at least $\Delta/4$ from both $\lambda_1\bigl(\hat{\Sigma}_{k,p}\bigr)$ and $\lambda_1(\Sigma_p)$. Applying (43) we conclude that,

$$\|\hat{E}_1 - E_1\| \leq \max_{\Gamma} \bigl\{ \|R(\lambda, \Sigma_p)\| \|R(\lambda, \hat{\Sigma}_{k,p})\|^{-2} \bigr\} \|\hat{\Sigma}_{k,p} - \Sigma\| \ .$$

By hypothesis, $\|R(\lambda, \Sigma_p)\| \leq |\lambda - \lambda_p(\Sigma_p)|^{-1}$ and similarly for $\|R(\lambda, \hat{\Sigma}_{k,p})\|$. Therefore,

$$\|\hat{E}_1 - E_1\| \leq 16\Delta^{-2} \|\hat{\Sigma}_{k,p} - \Sigma\| \ . \tag{44}$$

and the claims (40) and (41) are established for $m = 1$. We describe the induction step from $m = 1$ to $m = 2$ which is repeated with slightly more cumbersome notation for all m (omitted). Consider a unit vector,

$$\begin{aligned} x &= \sum_{j=2}^{p} E_j x \perp E_1 x \\ &= (\hat{E}_1 - E_1)x + (J - \hat{E}_1)x \ . \end{aligned} \tag{45}$$

Then,

$$\begin{aligned} &\bigl|(x, \hat{\Sigma}_{k,p} x) - \bigl((J - \hat{E}_1)\hat{\Sigma}_{k,p}(J - \hat{E}_1)x, x\bigr)\bigr| \\ &\leq \|\hat{\Sigma}_{k,p}\| \bigl(2\|\hat{E}_1 - E_1\| + \|\hat{E}_1 - E_1\|^2\bigr) \end{aligned} \tag{46}$$

Therefore,

$$\begin{aligned}\lambda_2(\hat{\Sigma}_{k,p}) &= \max\big\{\big(x, (J-\hat{E}_1)\hat{\Sigma}_{k,p}(J-\hat{E}_1)x\big) :\ |x|=1\big\} \\ &\le O(\|\hat{E}_1 - E_1\|) + \lambda_2(\Sigma_p)\ .\end{aligned}$$

Inverting the roles of $\hat{\Sigma}_{k,p}$ and Σ_p we obtain,

$$|\lambda_2(\hat{\Sigma}_{k,p}) - \lambda_2(\Sigma_p)| = O_p(\|\hat{\Sigma}_{k,p} - \Sigma_p\|)\ .$$

Now repeating the argument we gave for (44) we obtain,

$$\|\hat{E}_2 - E_2\| = O_p(\|\hat{\Sigma}_{k,p} - \Sigma_p\|)\ . \tag{47}$$

The theorem follows from the induction and Theorem 1. □
Note that if we track the effect of Δ and m, we in fact have,

$$\|\hat{E}_j - E_j\| = O_p(j\Delta^{-2}\|\hat{\Sigma}_{k,p} - \Sigma_p\|),\ 1 \le j \le m.$$

Also note that the dimension of $\sum_{j=1}^m E_j$ is immaterial.

5 Choice of the banding parameter

The results in Section 3 give us the rate of $k = k_n$ that guarantees convergence of the banded estimator $\hat{\Sigma}_k$, but they do not offer much practical guidance for selecting k for a given dataset. The standard way to select a tuning parameter is to minimize the risk

$$R(k) = E\|\hat{\Sigma}_k - \Sigma\|_{(1,1)}, \tag{48}$$

with the "oracle" k given by

$$k_0 = \arg\min_k R(k). \tag{49}$$

The choice of matrix norm in (48) is somewhat arbitrary. In practice, we found the choice of k is not sensitive to the choice of norm; the l_1 to l_1 matrix norm does just slightly better than others in simulations, and is also faster to compute.

We propose a resampling scheme to estimate the risk and thus k_0: divide the original sample into two samples at random and use the sample covariance matrix of one sample as the "target" to choose the best k for the other sample. Let n_1, $n_2 = n - n_1$ be the two sample sizes for the random split, and let $\hat{\Sigma}_1^{\nu}$, $\hat{\Sigma}_2^{(\nu)}$ be the two sample covariance matrices from the ν-th split, for $\nu = 1, \ldots, N$. Alternatively, N random splits could be replaced by K-fold cross-validation. Then the risk (48) can be estimated by

$$\hat{R}(k) = \frac{1}{N}\sum_{\nu=1}^{N} \|B_k(\hat{\Sigma}_1^{(\nu)}) - \hat{\Sigma}_2^{(\nu)}\|_{(1,1)} \tag{50}$$

and k is selected as

$$\hat{k} = \arg\min_k \hat{R}(k). \tag{51}$$

Generally we found little sensitivity to the choice of n_1 and n_2, and used $n_1 = n/3$ throughout this paper. If n is sufficiently large, another good choice (see, e.g., Bickel et al. (2006)) is $n_1 = \log n$.

The oracle k_0 provides the best choice in terms of expected loss, whereas $\hat{k}$ tries to adapt to the data at hand. Another, and more challenging, comparison is that of $\hat{k}$ to the best band choice for the sample in question:

$$k_1 = \arg\min_k \|\hat{\Sigma}_k - \Sigma\|_{(1,1)}. \tag{52}$$

Here k_1 is a random quantity, and its loss is always smaller than that of k_0. The results in Section 6 show that $\hat{k}$ generally agrees very well with both k_0 and k_1, which are quite close for normal data. For heavier-tailed data, one would expect more variability; in that case, the agreement between $\hat{k}$ and k_1 is more important that that between $\hat{k}$ and k_0.

It may be surprising that using the sample covariance $\hat{\Sigma}_2$ as the target in (50) works at all, since it is known to be a very noisy estimate of Σ. It is, however, an unbiased estimate, and we found that even though (50) tends to overestimate the actual value of the risk, it gives very good results for choosing k.

Criterion (50) can be used to select k for the Cholesky-based $\tilde{\Sigma}_k$ as well. An obvious modification – replacing the covariance matrices with their inverses in (50) – avoids additional computational cost and instability associated with computing inverses. One has to keep in mind, however, that while $\hat{\Sigma}_k$ is always well-defined, $\tilde{\Sigma}_k$ is only well-defined for $k < n$, since otherwise regressions become singular. Hence, if $p > n$, k can only be chosen from the range $0, \ldots, n-1$, not $0, \ldots, p-1$.

6 Numerical results

In this section, we investigate the performance of the proposed banded estimator of the covariance $\hat{\Sigma}_k$ and the resampling scheme for the choice of k, by simulation and on a real dataset. The Cholesky-based $\tilde{\Sigma}_k$ and its variants have been numerically investigated by extensive simulations by Wu and Pourahmadi (2003) and Huang et al. (2006), and shown to outperform the sample covariance matrix. Because of that, we omit $\tilde{\Sigma}_k$ from simulations, and only include it in the real data example.

6.1 Simulations

We start from investigating the banded estimator by simulating data from $\mathcal{N}(0, \Sigma_p)$ with several different covariance structures Σ_p. For all simulations, we report results for $n = 100$, and $p = 10$, 100, and 200. Qualitatively, these represent three different cases: $p \ll n$, $p \sim n$, and $p > n$. We have also conducted selected simulations with $p = 1000$, $n = 100$, which qualitatively corresponds to the case $p \gg n$; all the patterns observed with $p > n$ remain the same, only more pronounced. The number of random splits used in (50) was $N = 50$, and the number of replications was 100.

Example 1: Moving average covariance structure

We take Σ_p to be the covariance of the MA(1) process, with

$$\sigma_{ij} = \rho^{|i-j|} \cdot \mathbf{1}\{|i-j| \le 1\},\ 1 \le i, j \le p,$$

The true Σ_p is banded, and the oracle $k_0 = 1$ for all p. For this example we take $\rho = 0.5$. Figure 1 shows plots of the true risk $R(k)$ and the estimated risk $\hat{R}(k)$ from (50). While the risk values

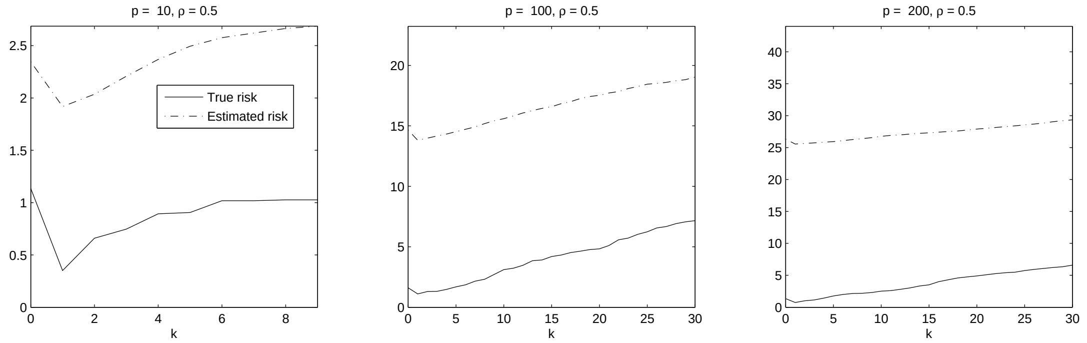


Figure 1: MA(1) covariance: True (averaged over 100 realizations) and estimated risk (single realization) as a function of k, plotted for $k \leq 30$. Both risks are increasing after $k = 1$ for all p.

		Mean(SD)			Loss			
p	k_0	k_1	$\hat{k}$	$k_1 - \hat{k}$	$\hat{\Sigma}_{\hat{k}}$	$\hat{\Sigma}_{k_0}$	$\hat{\Sigma}_{k_1}$	$\hat{\Sigma}$
10	1	1(0)	1(0)	0(0)	0.5	0.5	0.5	1.2
100	1	1(0)	1(0)	0(0)	0.8	0.8	0.8	10.6
200	1	1(0)	1(0)	0(0)	0.9	0.9	0.9	20.6

Table 1: MA(1): Oracle and estimated k and the corresponding loss values.

themselves are overestimated by (50) due to the extra noise introduced by $\hat{\Sigma}_2$, the agreement of the minima is very good, and that is all that matters for selecting k.

Table 1 shows the oracle values of k_0 and k_1, the estimated $\hat{k}$, and the losses corresponding to all these along with the loss of the sample covariance $\hat{\Sigma}$. When the true model is banded, the estimation procedure always picks the right banding parameter $k = 1$, and performs exactly as well as the oracle. The covariance matrix, as expected, does worse.

Example 2: Autoregressive covariance structure

Let Σ_p be the covariance of an AR(1) process,

$$\sigma_{ij} = \rho^{|i-j|},\ 1 \leq i, j \leq p.$$

For this simulation example, we take $\rho = 0.1$, 0.5, and 0.9. The covariance matrix is not sparse, but the entries decay exponentially as one moves away from the diagonal. Results in Figure 2 and Table 2 show that the smaller ρ is, the smaller the optimal k. Results in Table 2 also show the variability in $\hat{k}$ increases when the truth is far from banded (larger ρ), which can be expected from the flat risk curves in Figure 2. Variability of k_1 increases as well, and $k_1 - \hat{k}$ is not significantly different from 0. In terms of the loss, the estimate again comes very close to the oracle.

Example 3: Long-range dependence.

This example is designed to challenge the banded estimator, since conditions (19) will not hold for covariance matrix of a process exhibiting long-range dependence. Fractional Gaussian noise (FGN),

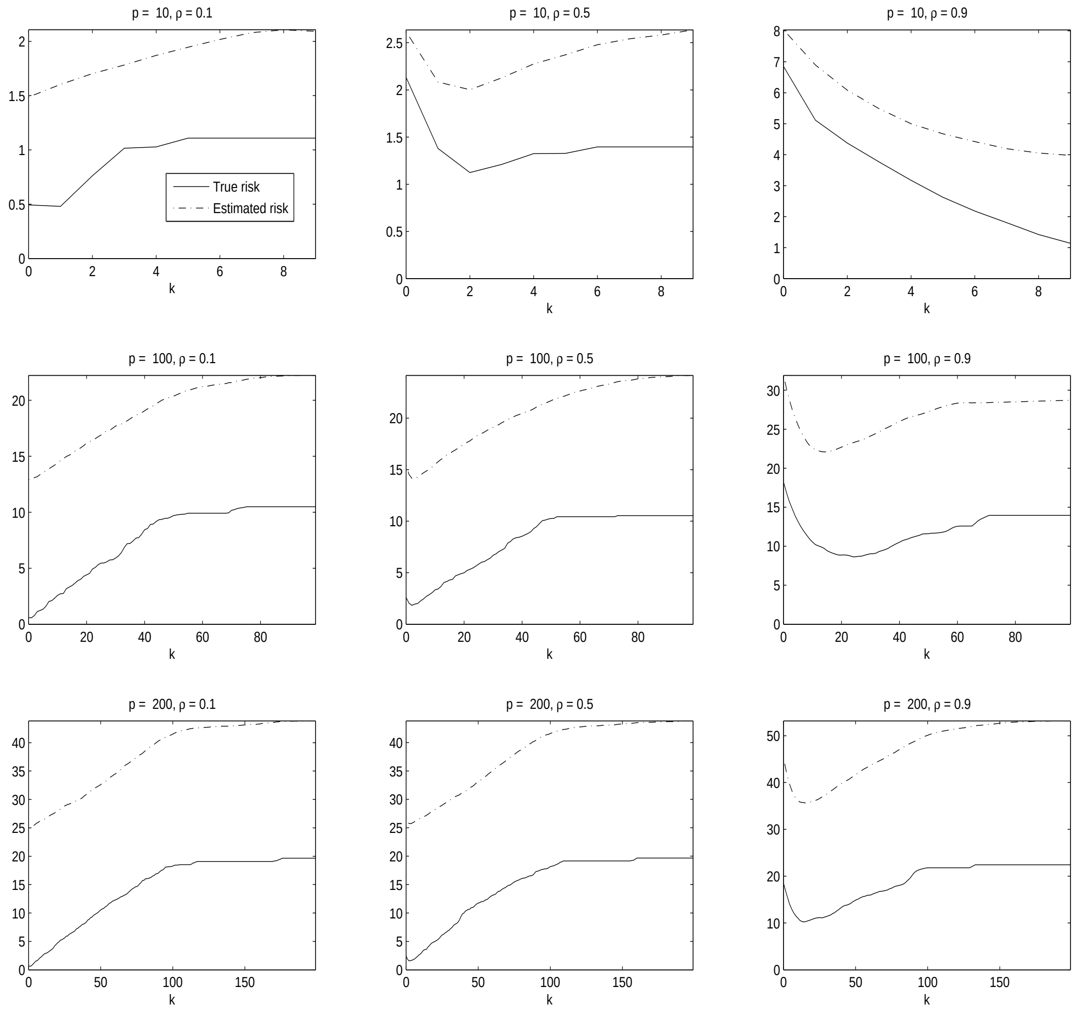


Figure 2: AR(1) covariance: True (averaged over 100 realizations) and estimated risk (single realization) as a function of k.

			Mean(SD)			Loss			
p	ρ	k_0	k_1	$\hat{k}$	$k_1 - \hat{k}$	$\hat{\Sigma}_{\hat{k}}$	$\hat{\Sigma}_{k_0}$	$\hat{\Sigma}_{k_1}$	$\hat{\Sigma}$
10	0.1	1	0.5(0.5)	0.0(0.2)	0.5(0.6)	0.5	0.5	0.4	1.1
10	0.5	3	3.3(0.8)	2.0(0.6)	1.3(1.1)	1.1	1.0	1.0	1.3
10	0.9	9	8.6(0.7)	8.9(0.3)	-0.4(0.7)	1.5	1.5	1.5	1.5
100	0.1	0	0.2(0.4)	0.1(0.3)	0.1(0.6)	0.6	0.6	0.6	10.2
100	0.5	3	2.7(0.7)	2.3(0.5)	0.4(1.0)	1.6	1.6	1.5	10.6
100	0.9	20	21.3(4.5)	15.9(2.6)	5.5(5.8)	9.2	8.8	8.5	13.5
200	0.1	1	0.2(0.4)	0.2(0.4)	-0.0(0.6)	0.7	0.6	0.6	20.4
200	0.5	3	2.4(0.7)	2.7(0.5)	-0.2(1.0)	1.8	1.7	1.7	20.8
200	0.9	20	20.2(4.5)	16.6(2.4)	3.6(5.6)	9.9	9.7	9.5	24.5

Table 2: AR(1): Oracle and estimated k and the corresponding loss values.

			Mean(SD)			L_1 Loss			
p	H	k_0	k_1	$\hat{k}$	$k_1 - \hat{k}$	$\hat{\Sigma}_{\hat{k}}$	$\hat{\Sigma}_{k_0}$	$\hat{\Sigma}_{k_1}$	$\hat{\Sigma}$
10	0.5	0	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.3	0.3	0.3	1.1
10	0.7	5	5.0(1.8)	2.3(1.5)	2.7(2.5)	1.4	1.2	1.1	1.2
10	0.9	9	8.6(0.6)	9.0(0.1)	-0.4(0.6)	1.5	1.5	1.5	1.5
100	0.5	0	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.4	0.4	0.4	10.2
100	0.7	4	4.9(2.2)	4.1(1.6)	0.8(2.9)	5.5	5.5	5.4	10.7
100	0.9	99	82.1(10.9)	85.1(15.5)	-3.1(19.0)	17.6	16.6	16.6	16.6
200	0.5	0	0.0(0.0)	0.0(0.1)	-0.0(0.1)	0.4	0.4	0.4	20.1
200	0.7	3	4.2(2.2)	4.9(2.1)	-0.7(3.4)	7.9	7.7	7.7	20.9
200	0.9	199	164.0(22.7)	139.7(38.9)	24.3(47.4)	37.8	33.3	33.3	33.3

Table 3: FGN: Oracle and estimated k and the corresponding loss values.

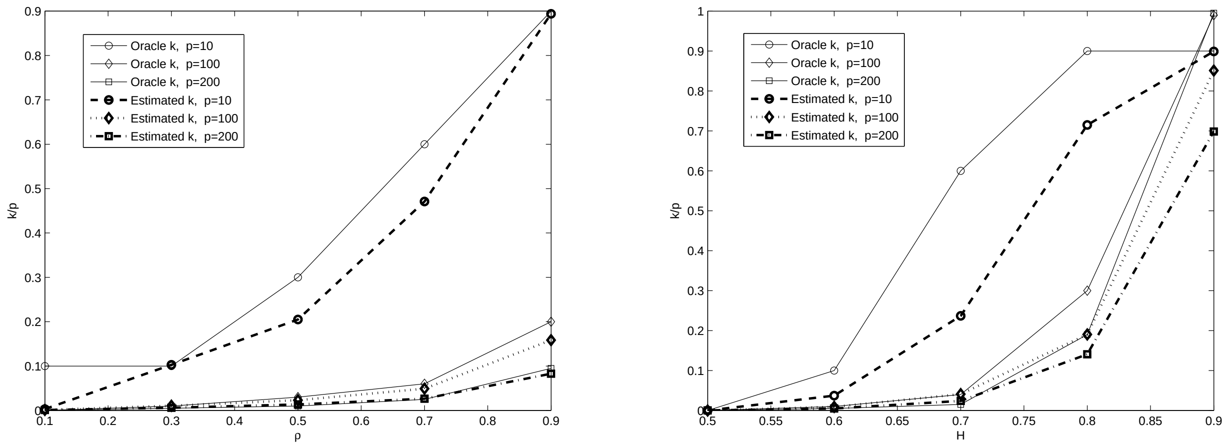
the increment process of fractional Brownian motion, provides a classic example of such a process. The covariance matrix is given by

$$\sigma_{ij} = \frac{1}{2}\left[(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}\right],\ 1 \le i,j \le p,$$

where $H \in [0.5, 1]$ is the Hurst parameter. $H = 0.5$ corresponds to white noise, and the larger H, the more long-range dependence in the process. Values of H up to 0.9 are common in practice, for example, in modeling Internet network traffic. For simulating this process, we used the circulant matrix embedding method (Bardet et al., 2002), which is numerically stable for large p.

Results in Table 3 show that the procedure based on the estimated risk correctly selects a large k ($k \approx p$) when the covariance matrix contains strong long-range dependence ($H = 0.9$). In this case banding cannot help – but it does not hurt, either, since the selection procedure essentially chooses to do no banding. For smaller H, the procedure adapts correctly and selects $k = 1$ for $H = 0.5$ (diagonal estimator for white noise), and a small k for $H = 0.7$.

Another interesting question is how the optimal choice of k depends on dimension p. Figure 3 shows the ratio of optimal k to p, for both oracle k_0 and estimated $\hat{k}$, for AR(1) and FGN (for MA(1), the optimal k is always 1). The plots confirm the intuition that (a) the optimal amount



AR(1) FGN

Figure 3: The ratio of optimal k to dimension p for AR(1) (as a function of ρ) and FGN (as a function of H).

of regularization is model dependent, and the faster the off-diagonal entries decay, the smaller the optimal k; and (b) the same model requires relatively more regularization in higher dimensions.

6.2 Call center data

Here we apply the banded estimators $\hat{\Sigma}_k$ and $\tilde{\Sigma}_k$ to the call center data used as an example of a large covariance estimation problem by Huang et al. (2006), who also provide a detailed description of the data. Briefly, the data consists of call records from a call center of a major U.S. financial institution. Phone calls were recorded from 7:00am till midnight every day in 2002, and weekends, holidays, and days when equipment was malfunctioning have been eliminated, leaving a total of 239 days. On each day, the 17-hour recording period was divided into 10-minute intervals, and the number of calls in each period, N_{ij}, was recorded for each of the days $i = 1, \dots, 239$ and time periods $j = 1, \dots, 102$. A standard transformation $x_{ij} = (N_{ij} + 1/4)^{1/2}$ was applied to make the data closer to normal.

The goal is to predict arrival counts in the second half of the day from counts in the first half of the day. Let $\boldsymbol{x}_i = (\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)})$, with $\boldsymbol{x}_i^{(1)} = (x_{i1}, \dots, x_{i,51})$, and $\boldsymbol{x}_i^{(2)} = (x_{i,52}, \dots, x_{i,102})$. The mean and the variance of Vx are partitioned accordingly,

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{53}$$

The best linear predictor of $\boldsymbol{x}_i^{(2)}$ from $\boldsymbol{x}_i^{(1)}$ is then given by

$$\hat{\boldsymbol{x}}_i^{(2)} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_i^{(1)} - \mu_1). \tag{54}$$

Different estimators of Σ in (53) can be plugged in to (54). To compare their performance, the data were divided into a training set (January to October, 205 days) and a test set (November and

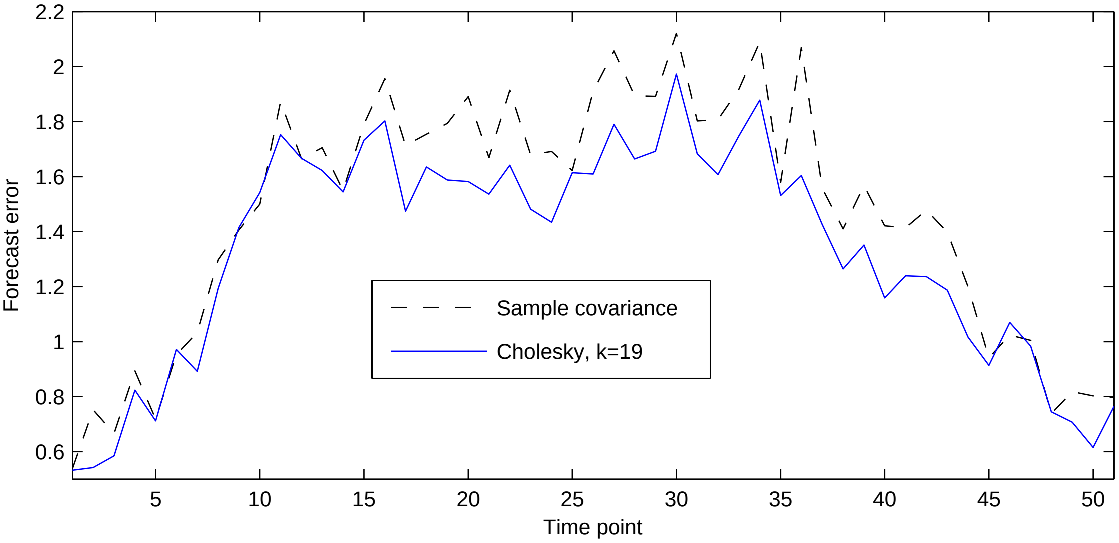


Figure 4: Call center forecast error using the sample covariance $\hat{\Sigma}$ and the best Cholesky-based estimator $\tilde{\Sigma}_k$, $k = 19$.

December, 34 days). For each time interval j, the performance is measured by the average absolute forecast error

$$E_j = \frac{1}{34} \sum_{i=206}^{239} |\hat{x}_{ij} - x_{ij}|.$$

The selection procedure for k described in Section 5 to both $\hat{\Sigma}_k$ and $\tilde{\Sigma}_k$. It turns out that the data exhibits strong long-range dependence, and for $\hat{\Sigma}_k$ the selection procedure picks $k = p = 102$, so banding the covariance matrix is not beneficial here. For $\tilde{\Sigma}_k$, the selected $k = 19$ produces a better prediction for almost every time point than the sample covariance $\hat{\Sigma}$ (see Figure 4).

This example suggests that a reasonable strategy for choosing between $\hat{\Sigma}_k$ and $\tilde{\Sigma}_k$ in practice is to estimate the optimal k for both and use the one that selects a smaller k. The two estimators are meant to exploit different kinds of sparsity in the data, and a smaller k selected for one of them indicates that that particular kind of sparsity is a better fit to the data.

7 Discussion

I. If $\sigma^{ij} = 0$, $|i - j| > k$ and $\|\Sigma^{-1}\| \leq \varepsilon_0^{-1}$, then $\boldsymbol{X}$ is a kth order auto regressive process and as we might expect, $\tilde{\Sigma}_{k,p}$ is the right estimate. Now suppose $\sigma^{ii} \leq \varepsilon_0^{-1}$ for all i and we only know that $\sigma^{ij} = 0$ for each i and $p - (2k + 1)$ j's. This condition may be interpreted as saying that, for each i there is a set S_i with $|S_i| \leq k$, $i \notin S_0$, such that, X_i is independent of $\{X_t, t \notin S_i, t \neq i\}$ given $\{X_j : j \in S_i\}$. Although banding would not in general give us sparse estimates, the following seem intuitively plausible.

1) Minimize a suitable objective function $\Psi(\hat{P}, \Sigma) \geq 0$ where $\hat{P}$ is the empirical distribution of

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and

$$\Psi(P, \Sigma_p) = 0$$

subject to $\|\Sigma\|_{(1,1)} \leq \gamma_{n,p}$.

2) Let $\gamma_{n,p} \to 0$ "slowly". This approach should yield estimates which consistently estimate sparse covariance structure. Banerjee et al. (2006) and Huang et al. (2006) both use normal or Wishart-based loglikelihoods for Ψ and a Lasso-type penalty in this context. We are currently pursuing this approach more systematically.

II. The connections with graphical models are also apparent. If D is the dependency matrix of Σ^{-1}, with entries 0 and 1 then $\|D\|_{(1,1)}$ is just the maximum degree of the graph vertices. See Meinshausen and Buhlmann (2006) for a related approach in determining covariance structure in this context.

III. A similar interpretation can be attached if we assume Σ is k_0 banded after a permutation of the rows. This is equivalent to assuming that there is a permutation of variables after which X_i is independent of $\{X_j : |j-i| > k\}$ for all i.

A Additional lemmas and proofs

The key to Theorem 2 is the following lemma which substitutes for (26). Consider symmetric matrices M indexed by (a,b), $a, b \in \mathcal{A}$, a finite index set. Suppose for each $a \in \mathcal{A}$ there exist $N_a \leq N$ sets $S_{a,j}$ such that, the $S_{a,j}$ form a partition of $\mathcal{A} - \{a\}$. Define, for any $1 \leq j \leq N$, $M = [m(a,b)]$ as above

$$r(j) = \max\{|m(a,b)| : \; b \in S_{a,j}, \; a \in \mathcal{A}\}$$

and

$$\mu = \max_a |m(a,a)| \; .$$

Lemma 1. *Under assumption A,*

$$\|M\| \leq \mu + \sum_{j=1}^{N} r(j) \; . \tag{A1}$$

Proof of Lemma 1. Apply (18) noting that

$$\sum\{|m(a,b)| : \; b \in \mathcal{A}\} \leq \sum_{j=1}^{N} r(j) + \mu$$

for all $a \in \mathcal{A}$. □

Proof of Corollary 2. An examination of the proof of Theorem 1 will show that the bound of $\|\Sigma_p - B_k(\Sigma_p)\|_{(1,1)}$ was used solely to bound $\|\Sigma_p - B_k(\Sigma_p)\|$. But in the case of Corollary 2 a theorem of Kolmogorov (De Vore and Lorentz (1993), p. 334) has, after the identification (15),

$$\|\Sigma_p - B_k(\Sigma_p)\| \leq \frac{C' \log k}{k^m} \tag{A2}$$

where C' depends on C and m only, for all $\Sigma_p \in \mathcal{L}(\varepsilon_0, m, C)$. The result follows. Note that Corollary 1 would give the same results as the inferior bound $C'k^{-(m-1)}$. □

To prove Theorem 3 we will need

Lemma 2. *Under conditions of Theorem 3, uniformly on $\mathcal{U}$,*

$$\max\{\|\tilde{\boldsymbol{a}}_j^{(k)} - \boldsymbol{a}_j^{(k)}\|_\infty : \ 1 \le j \le p\} = O_P\left(n^{-1/2}\log p\right), \tag{A3}$$

$$\max\{|\tilde{d}_{j,k}^2 - d_{j,k}^2| : \ 1 \le j \le p\} = O_P\left(\left(n^{-1/2}\log p\right)^{\frac{\alpha}{\alpha+1}}\right), \tag{A4}$$

and

$$\|A_k\| = \|D_k^{-1}\| = O(1), \tag{A5}$$

where $\tilde{\boldsymbol{a}}_j^{(k)} = \left(\tilde{a}_{j1}^{(k)}, \dots, \tilde{a}_{j,j-1}^{(k)}\right)$ *are the empirical estimates of the vectors* $\boldsymbol{a}_j^{(k)} = (a_{j1}^{(k)}, \dots, a_{j,j-1}^{(k)})$ *defined in* (11), $\tilde{d}_{j,k}^2$, $1 \le j \le p$ *are the empirical estimates of the* $d_{j,k}^2$ *defined in* (12), *and* A_k *and* D_k *are defined in* (13).

To prove Lemma 2 we need an additional

Lemma 3. *Let Z_i be i.i.d. $\mathcal{N}(\mathbf{0}, \Sigma_p)$ and $0 < \lambda_{\min}(\Sigma_p) \le \lambda_{\max}(\Sigma_p) \le \varepsilon_0^{-1} < \infty$. Then, if $\Sigma_p = [\sigma_{ab}]$,*

$$P\big[|\sum_{i=1}^n (Z_{ij}Z_{ik} - \sigma_{jk})| \ge n\nu\big] \le C_1 \exp(-C_2 n\nu) \tag{A6}$$

where C_1, C_2 depend on ε_0 only.

Proof of Lemma 3. By hypothesis, $|\sigma_{ab}| < \varepsilon_0^{-1}$ for all a, b. If $j = k$, we can divide by σ_{jj} inside the sum and the result is well known. In particular, it can be obtained by specializing to the case $p = 1$ in Lemma 4 of Bickel and Levina (2004). If $j = k$, w.l.o.g. we take $\sigma_{jj} \equiv 1$ for all j. Then write

$$Z_{ik} = \sigma_{jk}\, Z_{ij} + (1 - \sigma_{jk})^{1/2} Z'_i,$$

where Z'_i is independent of Z_{ij}. Then

$$P\Big[|\sum_{i=1}^n (Z_{ij}Z_{ik} - \sigma_{jk})| \ge n\nu\Big] \le P\Big[|\sum_{i=1}^n (Z_{ij}^2 - 1)| \ge |\sigma_{jk}|^{-1}\frac{n\nu}{2}\Big] + P\Big[|\sum_{i=1}^n Z_{ij}Z'_i| \ge \frac{n\nu}{2}\Big] \tag{A7}$$

and the result (A6) follows from the case $p = 1$ in Lemma 4 of Bickel and Levina (2004). □

Proof of Lemma 2. Note first that,

$$\|\mathrm{Var}\boldsymbol{X} - \widehat{\mathrm{Var}}\boldsymbol{X}\|_\infty = O_P\big(n^{-1/2}\log p\big)\ , \tag{A8}$$

by Lemma 3. Hence,

$$\max_j \big\|\widehat{\mathrm{Var}}^{-1}(\boldsymbol{Z}_j^{(k)}) - \mathrm{Var}^{-1}(\boldsymbol{Z}_j^{(k)})\big\|_\infty = O_P(n^{-1/2}\log p). \tag{A9}$$

To see this, note that the entries of $\widehat{\mathrm{Var}}\boldsymbol{X} - \mathrm{Var}\boldsymbol{X}$ can be bounded by $n^{-1}|\sum_{i=1}^n X_{ia}X_{ib} - \sigma_{ab}| + n^{-2}|\sum_{i=1}^n X_{ia}|\,|\sum_{i=1}^n X_{ib}|$, where w.l.o.g. we assume $E\boldsymbol{X} = \mathbf{0}$. Lemma 3 ensures that

$$P\big[\max_{a,b}|n^{-1}\sum_{i=1}^n (X_{ia}X_{ib} - \sigma_{ab})| \ge \nu\big] \le C_1 p^2 \exp(-C_2 n\nu).$$

The second term is similarly bounded.

Also,

$$\|\Sigma^{-1}\| = \|(\mathrm{Var}X)^{-1}\| \le \varepsilon_0^{-1} \ .$$

Claim (A3) and the first part of (A5) follow from (5), (A8), and (A9). Since

$$\tilde{d}_{jk}^2 = \widehat{\mathrm{Var}}X_j - \widehat{\mathrm{Var}}\big(\sum_{t=j-k}^{j-1} \tilde{a}_{jt}^{(k)} X_t\big),$$

$$d_{jk}^2 = \mathrm{Var}X_j - \mathrm{Var}\big(\sum_{t=j-k}^{j-1} a_{jt}^{(k)} X_t\big)$$

and the covariance operator is linear,

$$\begin{aligned} \big|\tilde{d}_{jk}^2 - d_{jk}^2\big| \ &\le \ |\mathrm{Var}(X_j) - \widehat{\mathrm{Var}}X_j| + \big|\widehat{\mathrm{Var}} \sum_{t=j-k}^{j-1} \big(\tilde{a}_{jt}^{(k)} - a_{jt}^{(k)}\big) X_t\big| \\ &+ \ \big|\widehat{\mathrm{Var}} \sum_{t=j-k}^{j-1} a_{jt}^{(k)} X_t - \mathrm{Var} \sum_{t=j-k}^{j-1} a_{jt}^{(k)} X_t\big| \ . \end{aligned} \tag{A10}$$

The sum $\sum_{t=j-k}^{j-1}$ is understood to be $\sum_{t=\max(1,j-k)}^{j-1}$. The maximum over j of the first term is $O_P(\frac{\log p}{n^{1/2}})$ by Lemma 3. The second can be written as

$$\begin{aligned} &\big|\sum\big\{\big(\tilde{a}_{js}^{(k)} - a_{js}^{(k)}\big)\big(\tilde{a}_{jt}^{(k)} - a_{jt}^{(k)}\big)\widehat{\mathrm{cov}}(X_s, X_t) : \ j-k \le s, \ t \le j-1\big\}\big| \\ &\le \Big(\sum_{t=j-k}^{j-1} \big|\tilde{a}_{jt}^{(k)} - a_{jt}^{(k)}\big|\widehat{\mathrm{Var}}^{\frac{1}{2}}(X_t)\Big)^2 \\ &\le k^2 \max_t (\tilde{a}_{jt}^{(k)} - a_{jt}^{(k)})^2 \max_t \widehat{\mathrm{Var}}(X_t) \\ &= O_P(k^2 n^{-1}(\log p)^2) = O_P\left(\left(n^{-1/2}\log p\right)^{\frac{2\alpha}{\alpha+1}}\right) = O_P\left(\left(n^{-1/2}\log p\right)^{\frac{\alpha}{\alpha+1}}\right) \end{aligned} \tag{A11}$$

by (A3) and $\|\Sigma_p\| \le \varepsilon_0^{-1}$. The last equality is the only place where we use the assumption $n^{-1/2}\log p = o_P(1)$. The third term in (A10) is bounded similarly. Thus (A4) follows. Further, for $1 \le j \le p$,

$$\begin{aligned} d_{jk}^2 &= \mathrm{Var}\Big(X_j - \sum\big\{a_{jt}^{(k)} X_t : \ \max(1, j-k) \le t \le j-1\big\}\Big) \\ &\ge \varepsilon_0 (1 + \sum (a_{jt}^{(k)})^2) \ge \varepsilon_0 \end{aligned} \tag{A12}$$

and the lemma follows. □

Proof of Theorem 3. We parallel the proof of Theorem 1. We need only check that

$$\big\|\tilde{\Sigma}_{k,p}^{-1} - \Sigma_{k,p}^{-1}\big\|_\infty = O_P(n^{-1/2}\log p) \tag{A13}$$

and

$$\|\Sigma_{k,p}^{-1} - B_k(\Sigma_p^{-1})\| = O(k^{-\alpha}) \ . \tag{A14}$$

We first prove (A13). By definition,

$$\tilde{\Sigma}_{k,p}^{-1} - \Sigma_{k,p}^{-1} = (I - \tilde{A}_k)\tilde{D}_k^{-1}(I - \tilde{A}_k)^T - (I - A_k)D_k^{-1}(I - A_k)^T \tag{A15}$$

where $\tilde{A}_k$, $\tilde{D}_k$ are the empirical versions of A_k and D_k. Apply the standard inequality

$$\begin{aligned}\|A^{(1)}A^{(2)}A^{(3)} - B^{(1)}B^{(2)}B^{(3)}\| &\le \sum_{j=1}^{3} \|A^{(j)} - B^{(j)}\|\Pi_{k\neq j}\|B^{(k)}\| \\ + \sum_{j=1}^{3} \|B^{(j)}\|\, \|\Pi_{k\neq j}\|A^{(k)} - B^{(k)}\| &+ \Pi_{j=1}^{3}\|A^{(j)} - B^{(j)}\| \ .\end{aligned} \tag{A16}$$

Take $A^{(1)} = [A^{(3)}]^T = I - \tilde{A}_k$, $B^{(1)} = [B^{(3)}]^T = I - A_k$, $A^{(2)} = \tilde{D}_k^{-1}$, $B^{(2)} = D_k^{-1}$ in (A16) and (A13) follows from Lemma 2. For (A14), we need only note that for any matrix M,

$$\|MM^T - B_k(M)B_k(M^T)\| \le 2\|M\|\ \|B_k(M) - M\| + \|B_k(M) - M\|^2$$

and (A14) and the theorem follows from our definition of U^{-1}. □

Lemma 4. *Suppose $\Sigma = [\rho(j-i)]$ is a Toeplitz covariance matrix; $\rho(k) = \rho(-k)$ for all k, $\Sigma \in \mathcal{L}(\varepsilon_0, m, C)$. Then, if f is the spectral density of Σ,*

(i) $\Sigma^{-1} = [\tilde{\rho}(j-i)]$, $\tilde{\rho}(k) = \tilde{\rho}(-k)$.

(ii) Σ^{-1} *has spectral density* $\frac{1}{f}$.

(iii) $\Sigma^{-1} \in \mathcal{L}\big(\varepsilon_0, m, C'(m, \varepsilon_0, C)\big)$.

Proof of Lemma 4. That $\big\|\big(\frac{1}{f}\big)^{(m)}\big\|_\infty \le C'(m, \varepsilon_0, C)$ and $\varepsilon_0 \le \big\|\frac{1}{f}\big\|_\infty \le \varepsilon_0^{-1}$ is immediate. The claims (i) and (ii) follow from the identity, $\frac{1}{f} = \sum_{k=-\infty}^{\infty} \tilde{\rho}(k)e^{2\pi iku}$ in the L_2 sense and

$$1 = \sum_{k=-\infty}^{\infty} \delta_{0k}e^{2\pi iku} = f(u)\frac{1}{f}(u) \ .$$

Proof of Corollary 2. Note that $\Sigma \in \mathcal{L}(\varepsilon_0, m, C_0)$ implies that

$$f_\Sigma^{-\frac{1}{2}}(u) = a_0 + \sum_{j=1}^{\infty} a_k \cos(2\pi ju) \tag{A17}$$

is itself m times differentiable and

$$\big\|\big(f_\Sigma^{-\frac{1}{2}}\big)^{(m)}\big\|_\infty \le C'(C_0, \varepsilon_0) \ . \tag{A18}$$

But then,

$$\begin{aligned}f_{\Sigma^{-1}}(u) &= b_0 + \sum_{j=1}^{\infty} b_j \cos(2\pi ju) \\ &= \big(a_0 + \sum_{j=1}^{\infty} a_j \cos 2\pi ju\big)^2\end{aligned} \tag{A19}$$

where $b_i = \sum_{j=0}^{i} a_j a_{i-j}$. All formal operations are justified since $\sum_{j=0}^{\infty} |a_j| < \infty$ follows from Zygmund (1959), p.138. But (A19) can be reinterpreted in view of Lemma 4 as,

$$\Sigma^{-1} = AA^T$$

where $A = [a_{i-j} 1(i \geq j)]$ and a_j are real and given by (A19). Then, if $A_k \equiv B_k(A)$, $B_k(A)B_k^T(A)$ has spectral density,

$$f_{\Sigma_{k,p}^{-1}}(u) = \big(\sum_{j=0}^{k} a_j \cos 2\pi j u\big)^2 \ . \tag{A20}$$

Moreover, from (A19) and (A20)

$$\|f_{\Sigma_{k,p}^{-1}} - f_{\Sigma_p^{-1}}\|_\infty \leq \|\sum_{j=k+1}^{\infty} a_j \cos 2\pi j u\|_\infty \left(\|f_{\Sigma_p}^{-\frac{1}{2}}\|_\infty + \|\sum_{j=k+1}^{\infty} a_j \cos 2\pi j u\|_\infty \right) .$$

By (A18)

$$|a_j| \leq C' j^{-m}$$

hence finally,

$$\|\Sigma_{k,p}^{-1} - \Sigma_p^{-1}\| = \|f_{\Sigma_{k,p}^{-1}} - f_{\Sigma^{-1}}\|_\infty \leq C k^{-(m-1)} \ . \tag{A21}$$

Corollary 2 now follows from (A21) and (A3) and (A4) by minimizing

$$C_1 \frac{k^3 \log pk}{n^{1/2}} + C_2 k^{-(m-1)} \ .$$

□

Acknowledgments

E. Levina's research is partially supported by grants from the NSF (DMS-0505424) and the NSA (MSPF-04Y-120). We thank Adam Rothman (University of Michigan) for help with simulations, and Jianhua Huang (Texas A&M) for sharing the call center data.

References

Anderson, T. W. (1958). *An Introduction to Multivairate Statistical Analysis.* Wiley, New York.

Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294.

Banerjee, O., d'Aspremont, A., and El Ghaoui, L. (2006). Sparse covariance selection via robust maximum likelihood estimation. In *Proceedings of ICML.*

Bardet, J.-M., Lang, G., Oppenheim, G., Philippe, A., and Taqqu, M. (2002). Generators of long-range dependent processes. In Doukhan, P., Oppenheim, G., and Taqqu, M., editors, *Theory and Applications of Long-Range Dependence*, pages 579–623. Birkhauser, Boston.

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., Ritov, Y., and Zakai, A. (2006). Some theory for generalized boosting algorithms. *J. Machine Learning Research*, 7:705–732.

Böttcher, A. (1996). Infinite matrices and projection methods. In Lancaster, P., editor, *Lectures on Operator Theory and Its Applications*, volume 3 of *Fields Institute Monographs*, pages 1–72. Amer. Math. Soc., Providence, RI.

De Vore, R. and Lorentz, G. (1993). *Constructive Approximation.* Springer-Verlag, Berlin, New York.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Pickard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc., Ser. B*, 57:301–369.

Friedman, J. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, 84:165–175.

Furrer, R. and Bengtsson, T. (2006). Estimation of high-dimensional prior and posteriori covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis.* Under revision.

Geman, S. (1980). A limit theorem for the norm of random matrices. *Ann. Prob.*, 8:252–261.

Golub, G. H. and Van Loan, C. F. (1989). *Matrix Computations.* The John Hopkins University Press, Baltimore, Maryland, 2nd edition.

Grenander, U. and Szegö, G. (1984). *Toeplitz Forms and Their Applications.* Chelsea Publishing Company, New York.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables.* Wolters-Noordholf, Groningen.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327.

Johnstone, I. M. and Lu, A. Y. (2006). Sparse principal components analysis. *J. Amer. Statist. Assoc.* To appear.

Kato, T. (1949). On the convergence of the perturbation method, I. *Progr. Theor. Phys.*, 4:514–523.

Kato, T. (1966). *Perturbation Theory for Linear Operators.* Springer, Berlin.

Ledoit, O. and Wolf, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.

Marĉenko, V. A. and Pastur, L. A. (1967). Distributions of eigenvalues of some sets of random matrices. *Math. USSR-Sb*, 1:507–536.

Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462.

Paul, D. (2004). Asymptotics of the leading sample eigenvalues for a spiked covariance model. Technical report, Department of Statistics, Stanford University.

Riesz, F. and Sz-Nagy, B. (1955). *Functional Analysis*. Ungar, New York.

Sz.-Nagy, B. (1946). Perturbations des transformations autoadjointes dans l'espace de Hilbert. *Comment. Math. Helv.*, 19:347–366.

Tracy, C. A. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.*, 177(3):727–754.

Wachter, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Annals of Probability*, 6:1–18.

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

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

Zygmund, A. (1959). *Trigonometric series*. Cambridge Univ. Press.