Statistica Sinica **19** (2009), 1755-1768

BANDING SAMPLE AUTOCOVARIANCE MATRICES OF STATIONARY PROCESSES

Wei Biao Wu and Mohsen Pourahmadi

The University of Chicago and Texas A&M University

Abstract: We consider estimation of covariance matrices of stationary processes. Under a short-range dependence condition for a wide class of nonlinear processes, it is shown that the banded covariance matrix estimates converge in operator norm to the true covariance matrix with explicit rates of convergence. We also establish the consistency of the estimate of the inverse covariance matrix. These results are applied to a prediction problem, and error bounds for the finite predictor coefficients are obtained. A sub-sampling approach is proposed to choose the banding parameter, and simulation results reveal its satisfactory performance for linear and certain nonlinear processes as the procedure is solely based on the second-order characteristics of the underlying process. Selection of the band parameter for nonlinear processes remains an open problem.

Key words and phrases: Covariance matrix, prediction, regularization, short-range dependence, stationary process.

1. Introduction

Nonstationary covariance estimators by banding a sample covariance matrix or its Cholesky factor have been proposed by Wu and Pourahmadi (2003) and Bickel and Levina (2008) in the context of longitudinal and multivariate data, and their consistency was established under some regularity conditions when $m, n \to \infty$ and $m^{-1} \log n \to 0$, where m and n are the number of subjects and variables, respectively. The requirement of m being large ($m \to \infty$) is not always feasible, as for example when one has only one realization ($m = 1$) in the setup of univariate time series and the number of subjects is small. See for instance Li, Wang, Hong, Turner, Lupton and Carroll (2007) and Hall, Marron and Neeman (2005) for examples of high dimension, low sample size data.

The idea of banding a stationary covariance matrix or limiting moving average (MA) and autoregressive (AR) model fitting dates back at least to the 1920's and the works of Slutsky and Yule. Its genesis and later reincarnations, in spectral density estimation via smoothing the periodogram, Burg's spectral density estimator, and prediction based on fitting increasing-order AR and MA models (Berk (1974), Brockwell and Davis (1988), and Ing and Wei (2003)), can be

traced to the implicit and heuristic regularizing assumption that measurements far apart in time are weakly correlated. Given a realization $X_1, \ldots, X_n$ of a mean-zero stationary process $\{X_t\}$, its autocovariance function $\gamma_k = \mathrm{cov}(X_0, X_k)$ can be estimated by

$$\hat{\gamma}_k = \frac{1}{n} \sum_{i=1}^{n-|k|} X_i X_{i+|k|}, \quad k = 0, \pm 1, \ldots, \pm(n-1), \tag{1.1}$$

and it is known that for fixed $k \in \mathbb{Z}$, under the ergodicity condition, $\hat{\gamma}_k \to \gamma_k$ in probability. However, entry-wise convergence does not automatically imply that $\hat{\Sigma}_n = (\hat{\gamma}_{i-j})_{1 \le i,j \le n}$ is a good estimate of $\Sigma_n = (\gamma_{i-j})_{1 \le i,j \le n}$ (Hannan and Deistler (1988, Sec. 5.3)). Indeed, though $\hat{\Sigma}_n$ is positive definite (see Chapter 5 in Pourahmadi (2001)), it is not uniformly close to the population covariance matrix Σ_n, in the sense that the largest eigenvalue or the operator norm of $\hat{\Sigma}_n - \Sigma_n$ does not converge to zero; see Theorem 1 in Section 2. Such uniform convergence is important when studying the rate of convergence of the finite predictor coefficients and performance of various classification methods in time series.

Our, not necessarily positive-definite, covariance matrix estimator is of the form

$$\hat{\Sigma}_{n,l} = (\hat{\gamma}_{i-j} \mathbf{1}_{|i-j| \le l})_{1 \le i,j \le n}, \tag{1.2}$$

where $l \ge 0$ is an integer. It is a truncated version of $\hat{\Sigma}_n$, preserving the diagonal and the $2l$ main sub-diagonals; note that if $l \ge n-1$, then $\hat{\Sigma}_{n,l} = \hat{\Sigma}_n$. Following Bickel and Levina (2008), we call $\hat{\Sigma}_{n,l}$ *the banded covariance matrix estimate* and l its band parameter. The motivation for banding comes from the fact that, for a large lag k, either γ_k is close to zero or $\hat{\gamma}_k$ is an unreliable estimate of γ_k. Thus, prudent use of banding may bring considerable computational economy in the former case, and statistical efficiency in the latter, by keeping small or unreliable $\hat{\gamma}_k$ out of the calculations.

There are important differences between our setup and results and those in Wu and Pourahmadi (2003) and Bickel and Levina (2008). Here, we work with only one ($m = 1$) realization and establish consistency by banding the *sample autocovariance matrix*. This is attractive in time series analysis and application situations where typically only one realization is available. Also we impose very mild moment and dependence conditions on the underlying process.

The rest of the paper is organized as follows. Section 2 introduces a class of nonlinear processes and, using a new concept of short-range dependence (Wu (2005)), convergence properties of the banded covariance matrix and its inverse

are established. Section 3 presents an application of banding to a prediction problem and provides error bounds for the finite predictor coefficients. The selection of the band parameter, discussed in Section 4, is an adaptation of a resampling and risk-minimization approach due to Bickel and Levina (2008). Its performance is assessed via simulations for linear and nonlinear processes. The results are more satisfactory for linear processes, since the procedure relies solely on the second-order characteristics of the underlying processes; its extension to nonlinear processes remains an open problem. The findings confirm our intuition that the optimal band depends on the characteristics of the underlying model and that the faster the autocorrelation functions decay the smaller is the optimal band. Surprisingly, this conclusion is in agreement with that in Bickel and Levina (2008) in spite of the differences in the two setups.

2. Main Results

We introduce some structural assumptions on the process $\{X_t\}$ and work within the framework of nonlinear stationary processes that includes the standard linear processes (Brockwell and Davis (1991), and Ing and Wei (2003)). Hannan and Deistler (1988) have considered certain linear ARMA processes and obtained the uniform bound $\|\hat{\Sigma}_{n,\ell} - \Sigma_n\|_\infty = O\left(\sqrt{\log\log n}/\sqrt{n}\right)$, $\ell \le (\log n)^\alpha$, $\alpha < \infty$; see Theorem 5.3.2 therein. Here we obtain comparable results for nonlinear processes and allow a wider range of ℓ, see Theorem 2 below.

Let ε_i, $i \in \mathbb{Z}$, be independent and identically distributed (iid) random variables. Assume that $\{X_i\}$ is a causal process of the form

$$X_i = g(\ldots, \varepsilon_{i-1}, \varepsilon_i), \tag{2.1}$$

where g is a measurable function such that X_i is well-defined and $\mathrm{E}\,(X_i^2) < \infty$. Many stationary processes fall within the framework of (2.1); see Tong (1990) and Wu (2005). To introduce the dependence structure, let $(\varepsilon_i')_{i\in\mathbb{Z}}$ be an independent copy of $(\varepsilon_i)_{i\in\mathbb{Z}}$ and $\xi_i = (\ldots, \varepsilon_{i-1}, \varepsilon_i)$. Following Wu (2005), for $i \ge 0$ let $\xi_i' = (\ldots, \varepsilon_{-1}, \varepsilon_0', \varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_i)$ and $X_i' = g(\xi_i')$. For $\alpha > 0$ define the physical dependence measure

$$\delta_\alpha(i) = \|X_i - X_i'\|_\alpha. \tag{2.2}$$

Here, for a random variable Z, we write $Z \in \mathcal{L}^\alpha$, if $\|Z\|_\alpha := [\mathrm{E}\,(|Z|^\alpha)]^{1/\alpha} < \infty$, and write $\|\cdot\| = \|\cdot\|_2$. Observe that $X_i' = g(\xi_i')$ is a coupled version of $X_i = g(\xi_i)$ with ε_0 in the latter replaced by an iid copy ε_0'. The quantity $\delta_p(i)$ measures the dependence of X_i on ε_0. We say that $\{X_i\}$ is short-range dependent with moment α if

$$\Delta_\alpha := \sum_{i=0}^{\infty} \delta_\alpha(i) < \infty. \tag{2.3}$$

Namely, the cumulative impact of ε_0 on future values of the process or $\{X_i\}_{i\geq 0}$ is finite, thus suggesting a short-range dependence. In many applications it is easy to work with $\delta_\alpha(i)$ which is directly related to the data generating mechanism of the underlying process, as indicated in the next two examples.

Example 1. Let $X_j = g(\sum_{i=0}^{\infty} a_i\varepsilon_{j-i})$, where a_i are real coefficients with $\sum_{i=0}^{\infty}|a_i| < \infty$, ε_i are iid with $\varepsilon_i \in \mathcal{L}^\alpha$, $\alpha \geq 1$, and g is a Lipschitz continuous function. Then $\sum_{i=0}^{\infty} a_i\varepsilon_{j-i}$ is a well-defined random variable and $\delta_\alpha(i) = O(|a_i|)$. Hence we have (2.3).

Example 2. Let ε_i be iid random variables and set $X_i = g(X_{i-1}, \varepsilon_i)$, where g is a bivariate function. Many nonlinear time series models follow this framework. Let $L_\varepsilon = \sup_{x\neq x'} |g(x,\varepsilon) - g(x',\varepsilon)|/|x - x'|$. Assuming that $\mathrm{E}\,(L_{\varepsilon_0}^\alpha) < 1$, $\alpha > 0$, and for some x_0, $g(x_0, \varepsilon_0) \in \mathcal{L}^\alpha$. Then $\delta_\alpha(i) = O(r^i)$, $0 < r < 1$, so (2.3) follows, see Theorem 5.1 in Shao and Wu (2007). The latter paper also gives more examples.

In the sequel, for an $n \times n$ matrix A with real entries the operator norm $\rho(A)$ is defined by $\rho(A) = \max_{x\in\mathbb{R}^n:|x|=1} |Ax|$ where, for an n-dimensional real vector $x = (x_1, \ldots, x_n)'$, $|x| = (\sum_{i=1}^n x_i^2)^{1/2}$. Hence $\rho^2(A)$ is the largest eigenvalue of $A'A$, where $'$ denotes the matrix transpose. Theorem 1 below shows that $\hat{\Sigma}_n$ is not a consistent estimate of Σ_n in the sense that the operator norm or the largest eigenvalue of $\hat{\Sigma}_n - \Sigma_n$ does not converge to zero. On the positive side, we are able to show the convergence to zero and obtain an explicit upper bound for $\rho(\hat{\Sigma}_{n,l} - \Sigma_n)$ for the banded estimate $\hat{\Sigma}_{n,l}$ in our Theorem 2. We define the projection operator $\mathcal{P}_k$ as $\mathcal{P}_k\cdot = \mathrm{E}\,(\cdot|\xi_k) - \mathrm{E}\,(\cdot|\xi_{k-1})$, $k \in \mathbb{Z}$.

Theorem 1. *Assume that the process $\{X_t\}$ in (2.1) satisfies*

$$\sum_{i=0}^{\infty} \|\mathcal{P}_0 X_i\| < \infty. \tag{2.4}$$

If $\|\sum_{i=0}^{\infty} \mathcal{P}_0 X_i\| > 0$, then, $\rho(\hat{\Sigma}_n - \Sigma_n) \not\to 0$ in probability.

Proof. By (2.4), $\sigma := \|\sum_{i=0}^{\infty} \mathcal{P}_0 X_i\| < \infty$. Let $c = (1, \ldots, 1)'/\sqrt{n}$. Then $|c| = 1$ and

$$|c'(\hat{\Sigma}_n - \Sigma_n)c| \leq \rho(\hat{\Sigma}_n - \Sigma_n). \tag{2.5}$$

Under (2.4), since $\mathcal{P}_k$, $k \in \mathbb{Z}$, are orthogonal projections, we have $X_k = \sum_{j\in\mathbb{Z}} \mathcal{P}_j X_k$ and

$$\gamma(k) = \mathrm{E}\left[\sum_{i\in\mathbb{Z}} \mathcal{P}_i X_0 \sum_{j\in\mathbb{Z}} \mathcal{P}_j X_k\right] = \sum_{i\in\mathbb{Z}} \mathrm{E}\left[(\mathcal{P}_i X_0)(\mathcal{P}_i X_k)\right].$$

Let $\Delta = \sum_{i=0}^{\infty} \|\mathcal{P}_0 X_i\|$. By Schwarz's inequality and stationarity,

$$\sum_{k\in\mathbb{Z}} |\gamma(k)| \le \sum_{k\in\mathbb{Z}} \sum_{i\in\mathbb{Z}} \|\mathcal{P}_i X_0\| \|\mathcal{P}_i X_k\| = \Delta^2 < \infty.$$

We also have $\sigma^2 = \sum_{k\in\mathbb{Z}} \gamma(k) \le \Delta^2$. Hence, as $n \to \infty$,

$$c'\Sigma_n c = \gamma(0) + 2\sum_{i=1}^{n} \Big(1 - \frac{i}{n}\Big)\gamma(i) \to \sigma^2. \tag{2.6}$$

On the other hand, let $S_i = X_1 + \ldots + X_i$, then

$$\begin{aligned} c'\hat{\Sigma}_n c &= \sum_{i=1-n}^{n-1} \hat{\gamma}(i) - \frac{2}{n}\sum_{j=1}^{n-1} j\hat{\gamma}(j) \\ &= \frac{S_n^2}{n} - \frac{2}{n^2}\sum_{j=1}^{n-1} X_{j+1}\sum_{i=1}^{j}(j+1-i)X_i \\ &= \frac{S_n^2}{n} - \frac{2}{n^2}\sum_{j=1}^{n-1}(S_{j+1} - S_j)\sum_{i=1}^{j} S_i \\ &= \frac{S_n^2}{n} + \frac{2}{n^2}\sum_{j=1}^{n} S_j^2 - \frac{2}{n^2} S_n \sum_{j=1}^{n} S_j. \end{aligned}$$

Under (2.4), by Hannan (1979) (see also Wu (2005)), we have the invariance principle

$$\left\{ \frac{S_{\lfloor nu \rfloor}}{\sqrt{n}},\ 0 \le u \le 1 \right\} \Rightarrow \sigma\{I\!B(u),\ 0 \le u \le 1\}, \tag{2.7}$$

where $I\!B$ is the standard Brownian motion. With elementary manipulations, we have by (2.7) and the Continuous Mapping Theorem that

$$\frac{c'\hat{\Sigma}_n c}{\sigma^2} \Rightarrow I\!B^2(1) + 2\int_0^1 I\!B^2(t)dt - 2I\!B(1)\int_0^1 I\!B(t)dt.$$

Hence, by (2.5) and (2.6),

$$\frac{c'(\hat{\Sigma}_n - \Sigma_n)c}{\sigma^2} \Rightarrow \int_0^1 \Big\{ I\!B^2(t) + [I\!B(1) - I\!B(t)]^2 \Big\} dt - 1 =: Z_0. \tag{2.8}$$

So the theorem follows and, asymptotically, $\rho(\hat{\Sigma}_n - \Sigma_n)/\sigma^2$ has a lower bound which is distributed as Z_0.

It is very difficult to find the asymptotic distribution of the largest eigenvalue $\rho(\hat{\Sigma}_n - \Sigma_n)$. Recently, Bryc, Dembo and Jiang (2006) studied spectral measures of Toeplitz matrices with sub-diagonals being independent. In our case the matrix $\hat{\Sigma}_n - \Sigma_n$ is Toeplitz where the sub-diagonals are dependent, so their results are not directly applicable. For other contributions to inconsistency of the largest eigenvalues of sample covariance matrices see Johnstone (2001) and El Karoui (2007), and references therein.

The following result is needed in the proof of Theorem 2.

Lemma 1. *Assume that* $\{X_i\}$ *in* (2.1) *satisfies* (2.3) *with* $2 < \alpha \le 4$. *Then, for any* $j \in \mathbb{Z}$,

$$\begin{aligned} &\|\textstyle\sum_{i=1}^n X_i X_{i+j} - n\gamma_j\|_{\alpha/2} \le 2B_{\alpha/2} n^{1/q} \|X_1\|_\alpha \Delta_\alpha, \\ &\text{where } \ B_q = \begin{cases} \frac{18q^{3/2}}{(q-1)^{1/2}}, & \text{if } q \ne 2; \\ 1, & \text{if } q = 2. \end{cases} \end{aligned} \tag{2.9}$$

Proof. Let $q = \alpha/2$. Without loss of generality assume $j \ge 0$, and write $T_n = \sum_{i=1}^n X_i X_{i+j} - n\gamma_j$. By Theorem 1 in Wu (2007),

$$\|T_n\|_q \le B_q n^{1/q} \sum_{i=-j}^{\infty} \|\mathcal{P}_0 X_i X_{i+j}\|_q. \tag{2.10}$$

Recall that $X_i' = g(\xi_i')$ and, for $i < 0$, we have $X_i' = X_i$ and $\mathrm{E}\,(X_i X_{i+j}|\xi_{-1}) = \mathrm{E}\,(X_i' X_{i+j}'|\xi_{-1}) = \mathrm{E}\,(X_i' X_{i+j}'|\xi_0)$. By Jensen's and Schwarz's inequalities,

$$\begin{aligned} \|\mathcal{P}_0 X_i X_{i+j}\|_q &= \|\mathrm{E}\,(X_i X_{i+j} - X_i' X_{i+j}'|\xi_0)\|_q \\ &\le \|X_i X_{i+j} - X_i' X_{i+j}'\|_q \\ &\le \|X_i(X_{i+j} - X_{i+j}')\|_q + \|(X_i - X_i')X_{i+j}'\|_q \\ &\le \|X_i\|_\alpha \|X_{i+j} - X_{i+j}'\|_\alpha + \|X_i - X_i'\|_\alpha \|X_{i+j}'\|_\alpha \end{aligned}$$

which, by (2.10), implies (2.9) since $X_i - X_i' = 0$ if $i < 0$.

Theorem 2. *Let* $2 < \alpha \le 4$ *and* $q = \alpha/2$. *Assume* (2.3) *and* $0 \le l < n - 1$. *Then*

$$\|\rho(\hat{\Sigma}_{n,l} - \Sigma_n)\|_q \le c_\alpha (l+1) n^{1/q-1} \|X_1\|_\alpha \Delta_\alpha + \frac{2}{n} \sum_{j=1}^{l} j|\gamma_j| + 2 \sum_{j=l+1}^{n} |\gamma_j|, \tag{2.11}$$

where $c_\alpha > 0$ *is a constant depending only on* α.

Proof. By Lemma 1, for $i \geq 0$, $\|\hat{\gamma}_i - \mathrm{E}\,\hat{\gamma}_i\|_q \leq c_q(n-i)^{1/q}\|X_1\|_\alpha \Delta_\alpha/n$. Note that $\hat{\Sigma}_{n,l} - \Sigma_n$ is a symmetric Toeplitz matrix. From Golub and Van Loan (1989), we have

$$\begin{aligned}\rho(\hat{\Sigma}_{n,l} - \Sigma_n) &\leq \max_{1\leq j\leq n} \sum_{i=1}^{n} |\hat{\gamma}_{i-j}\mathbf{1}_{|i-j|\leq l} - \gamma_{i-j}| \\ &\leq \sum_{i=1-n}^{n-1} |\hat{\gamma}_i \mathbf{1}_{|i|\leq l} - \gamma_i| \leq 2\sum_{i=0}^{l} |\hat{\gamma}_i - \gamma_i| + 2\sum_{i=1+l}^{n} |\gamma_i|.\end{aligned}$$

So the theorem follows since the bias $|\mathrm{E}\,\hat{\gamma}_i - \gamma_i| \leq i|\gamma_i|/n$.

If $\gamma_m = O(\theta^m)$ for some $\theta \in (0,1)$, by letting $l = \lfloor c \log n \rfloor$ for sufficiently large $c > 0$, elementary calculations show that the bound in (2.11) is $O(n^{1/q-1}\log n)$. If $\gamma_m = O(m^{-\beta})$ with $\beta > 1$, we similarly have the bound $O(n^{(1/q-1)(1/\beta-1)})$ by letting $l = \lfloor n^{(1/q-1)/\beta} \rfloor$.

Next, we turn to the convergence properties of the inverse $\hat{\Sigma}_{n,l}^{-1}$.

Corollary 1. *For $2 < \alpha \leq 4$, assume (2.3), $l \to \infty$, and $l = o(n^{1-2/\alpha})$. Further assume that $f(\theta) = (2\pi)^{-1}\sum_{k\in\mathbb{Z}} \gamma_k e^{ik\theta}$, the spectral density of $\{X_t\}$, satisfies $0 < c_1 \leq f(\theta) \leq c_2 < \infty$ for some positive constants c_1 and c_2.*

(i) *$\hat{\Sigma}_{n,l}$ is positive definite with probability approaching to 1 and*

$$\rho(\hat{\Sigma}_{n,l}^{-1} - \Sigma_n^{-1}) = O_p(r_n), \text{ where } r_n = l n^{2/\alpha - 1} + \sum_{j=l}^{\infty} |\gamma_j|. \tag{2.12}$$

(ii) *Let $\mathcal{K}(\lambda) = \{l : 1 \leq l \leq n, \hat{\Sigma}_{n,1} - \lambda \mathrm{I}_n \geq 0\}$, where $0 < \lambda < 2\pi c_1$ and I_n is the $n \times n$ identity matrix. For $q = \alpha/2$,*

$$\|\rho(\hat{\Sigma}_{n,l}^{-1} - \Sigma_n^{-1})\mathbf{1}_{l\in\mathcal{K}(\lambda)}\|_q = O(r_n). \tag{2.13}$$

Proof. (i) By Theorem 2, $\rho(\hat{\Sigma}_{n,l} - \Sigma_n) = O_p(r_n)$, which converges to 0 in probability. Note that all eigenvalues of Σ_n lie in the interval $[2\pi c_1, 2\pi c_2]$ (cf Section 5.2 in Grenander and Szegö (1958)) which is both bounded from above and below. Hence the probability that $\hat{\Sigma}_{n,l}$ is positive definite tends to 1. Let $A_n = \Sigma_n^{-1/2}$ and $\Gamma_n = A_n \hat{\Sigma}_{n,l} A_n$. Then $\rho(\Gamma_n - \mathrm{I}_n) = O_p(r_n)$ and, since $r_n \to 0$, we have $\rho(\Gamma_n^{-1} - \mathrm{I}_n) = O_p(r_n)$. So (2.12) follows. To prove (ii), we note that $\mathcal{K}(\lambda)$ is asymptotically nonempty by Theorem 2. Let $d_1 \geq \ldots \geq d_n$ be eigenvalues of Γ_n, then using Theorem 2, we obtain $\mathrm{E}\,[\max_{i\leq n} |d_i - 1|^q] = O(r_n^q)$. If $l \in \mathcal{K}(\lambda)$, then $d_n \geq t_0 > 0$ for some constant t_0. So $|d_i^{-1} - 1| \leq |d_i - 1|/t_0$ and (ii) follows after elementary manipulations.

3. Banding and the Finite Predictor Coefficients

It is known that computing the linear least squares predictor of a future value of a stationary process $\{X_t\}$, based on the knowledge of its infinite past and the covariance matrix $\Sigma = (\gamma_{i-j})$, amounts to replacing Σ by $\Sigma_n = (\gamma_{i-j})_{1\le i,j\le n}$ and then studying the effect of such truncation (Pourahmadi (2001, pp.69-71)). In this section, we propose an even simpler approximation where instead of Σ_n, its banded version is used. The impact of this simplification on reducing the computational complexity of, say, the Durbin-Levinson algorithm due to sparseness is evident. Here, we study the impact of banding on the accuracy of the finite predictor coefficients.

Let $\mathbf{a}_n = (a_{1n}, \ldots, a_{nn})'$ be the coefficients of the finite predictor of length n and $\boldsymbol{\gamma}_n = (\gamma_1, \ldots, \gamma_n)'$. Then, $\mathbf{a}_n$ satisfies the Yule-Walker equations

$$\Sigma_n \mathbf{a}_n = \boldsymbol{\gamma}_n \quad \text{or} \quad \mathbf{a}_n = \Sigma_n^{-1}\boldsymbol{\gamma}_n. \tag{3.1}$$

Since the γ_k's are unknown, one resorts to solving the sample version of the Yule-Walker equations (3.1), i.e., $\hat{\mathbf{a}}_n = \hat{\Sigma}_n^{-1}\hat{\boldsymbol{\gamma}}_n$ where $\hat{\boldsymbol{\gamma}}_n = (\hat{\gamma}_1, \ldots, \hat{\gamma}_n)'$. It turns out that $\hat{\mathbf{a}}_n$ may perform poorly both statistically and numerically due to poor quality of $\hat{\gamma}_k$ for k large. One remedy is to regularize or use a banded covariance matrix estimate

$$\hat{\mathbf{a}}_{n,l} = \hat{\Sigma}_{n,l}^{-1}\tilde{\boldsymbol{\gamma}}_n, \text{ where } \tilde{\boldsymbol{\gamma}}_n = (\tilde{\gamma}_1, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_n)' \text{ and } \tilde{\gamma}_i = \hat{\gamma}_i \mathbf{1}_{|i|\le l}. \tag{3.2}$$

Here $\hat{\mathbf{a}}_{n,l}$ can be viewed as a shrinkage or regularized estimate of $\mathbf{a}_n$ replacing $(\hat{\gamma}_i)_{|i|>l}$ in $\hat{\mathbf{a}}_n$ by 0. Surprisingly, this simple regularization could improve its performance without sacrificing numerical accuracy of the predictor coefficients, as we indicate next.

Corollary 2. *Let $2 < \alpha \le 4$. Assume that the conditions of Corollary 1 are satisfied. Then $|\hat{\mathbf{a}}_{n,l} - \mathbf{a}_n| = O_p(r_n)$.*

Proof. Let $q = \alpha/2$. By Lemma 1, $\max_{i\le n} \|\tilde{\gamma}_i - \mathrm{E}\,\tilde{\gamma}_i\|_q = O(n^{1/q-1})$ and

$$|\tilde{\boldsymbol{\gamma}}_n - \mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n|^2 = \sum_{i=1}^{l}(\hat{\gamma}_i - \mathrm{E}\,\hat{\gamma}_i)^2 \le \Big(\sum_{i=1}^{l}|\hat{\gamma}_i - \mathrm{E}\,\hat{\gamma}_i|\Big)^2 = O_p\Big[(ln^{1/q-1})^2\Big]. \tag{3.3}$$

Then $|\tilde{\boldsymbol{\gamma}}_n - \mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n| = O_p(ln^{1/q-1})$. From Corollary 1(i), the largest eigenvalue of $\hat{\Sigma}_{n,l}^{-1}$ is bounded by $\rho(\Sigma_n^{-1}) + O_p(r_n)$. Observe that $|\mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n|^2 \le \sum_{i=1}^{\infty}\gamma_i^2 < \infty$. Hence we have

$$\begin{aligned}
|\hat{\Sigma}_{n,l}^{-1}\tilde{\boldsymbol{\gamma}}_n - \Sigma_n^{-1}\mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n| &\le |\hat{\Sigma}_{n,l}^{-1}(\tilde{\boldsymbol{\gamma}}_n - \mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n)| + |(\hat{\Sigma}_{n,l}^{-1} - \Sigma_n^{-1})\mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n| \\
&\le \rho(\hat{\Sigma}_{n,l}^{-1})|\tilde{\boldsymbol{\gamma}}_n - \mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n| + \rho(\hat{\Sigma}_{n,l}^{-1} - \Sigma_n^{-1})|\mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n| \\
&= \Big[\rho(\Sigma_n^{-1}) + O_p(r_n)\Big]O_p(r_n) + O_p(r_n) = O_p(r_n).
\end{aligned} \tag{3.4}$$

On the other hand, since $\mathrm{E}\,\hat{\gamma}_i = \gamma_i(n-|i|)/n$,

$$|\boldsymbol{\gamma}_n - \mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n|^2 = \sum_{i=1}^{l}(\mathrm{E}\,\hat{\gamma}_i - \gamma_i)^2 + \sum_{i=l+1}^{n}\gamma_i^2$$
$$\leq n^{-2}\sum_{i=1}^{l} i^2\gamma_i^2 + r_n^2 = O(n^{-2}l^2) + r_n^2 = O(r_n^2).$$

Since the eigenvalues of Σ_n are bounded from above and below, we have $|\Sigma_n^{-1}(\boldsymbol{\gamma}_n - \mathrm{E}\,\tilde{\boldsymbol{\gamma}}_n)| = O(r_n)$ which, together with (3.4) implies that $|\hat{\mathbf{a}}_{n,l} - \mathbf{a}_n| = O_p(r_n)$.

Analogues of Corollary 2 can be proved for the coefficients of the multi-step ahead predictors and interpolators of stationary processes, see Pourahmadi (2001, p.232).

4. Band Selection and A Simulation Study

The selection of the band ℓ is intuitively related to order selection for fitting MA models to the data (Brockwell and Davis (1988)). It is also related to estimating the spectral density function at zero, or the quantity $\sigma^2 = \sum_{k=-\infty}^{+\infty}\gamma_k$ by $\hat{\sigma}_n^2 = \sum_{k=-(n-1)}^{n-1}\hat{\gamma}_k$, which is known to be inconsistent, while the banded/truncated estimate $\hat{\sigma}_{n,l}^2 = \sum_{k=-l}^{l}\hat{\gamma}_k$ with $l = l_n \to \infty$ and $l/n \to 0$ can be consistent (Politis, Romano and Wolf (1999, Chap. 9)).

4.1. Band selection

Our Theorem 2 suggests that l should satisfy

$$l \to \infty, \quad l n^{1/q-1} \to 0, \quad \text{or} \quad l n^{1/q-1} \asymp \sum_{j=l+1}^{\infty}|\gamma_j|. \tag{4.1}$$

As a data-driven choice of l, one could propose the following naive algorithm.

1. Choose l such that $\sum_{k=-l}^{l}\hat{\gamma}(k)$ is a "good" estimate of σ^2
2. Check whether $\hat{\Sigma}_{n,l}$ is positive definite. If so, let $l^* = l$.
3. Otherwise, let $l^* = l-1$ and go to Step 2.

With the chosen band l^*, $\hat{\Sigma}_{n,l^*}$ is positive definite. The finer details for implementing this method is worked out in this section using the idea of resampling and risk-minimization as in Bickel and Levina (2008, Sec. 5). While they show that "nonoverlapped" splitting of the data works well for band selection in the multivariate data framework, our preliminary numerical experiments showed this scheme to be unsatisfactory for time series data. Instead, the technique of

subsampling (Politis, Romano and Wolf (1999)) that amounts to "overlapped" splitting of the data proved to be more suitable.

For linear processes, a natural way to select the band parameter l in (1.2) is to minimize the risk

$$R(l) = \mathrm{E}\,\|\hat{\Sigma}_{n,l} - \Sigma_n\|_{(1,1)}, \tag{4.2}$$

where for two $n \times n$ matrices A and B, $\|A - B\|_{(1,1)} = \max_i \sum_{j=1}^n |a_{ij} - b_{ij}|$ is the norm used in Bickel and Levina (2008). Here the "oracle" l is given by $l_0 = \arg\min_l R(l)$. The following subsampling scheme will be used to estimate the risk in (4.2), and hence l_0. An asymptotic justification for it can be found by focusing on the estimation of the vector of parameters $\theta = (\gamma_0, \ldots, \gamma_K)'$, $K \geq 1$, for a stationary process, and using Theorem 3.3.1 and the results related to Example 3.3.4 in Politis, Romano and Wolf (1999, pp.83-85)).

Given the stationary, centered time series data $X_1, X_2, \ldots, X_n$ of length n, the $\hat{\gamma}_k$ in (1.1) is usually computed for $k = 0, 1, \ldots, K$; the choice of K is important in practice, since $\hat{\gamma}_k$ is not a good estimate of γ_k for k large. A useful guide that is part of the folklore of time series analysis is to use $K \leq n/4$ (Box and Jenkins (1976, p.33)), but the default value in the SAS software is $K = 24$, and in R it is $K = 10\log(10n)$. In our calculations we fix it at $K = 30$, and when using subsampling to estimate (4.2), the unknown Σ_n is replaced by the $K \times K$ sample autocovariance matrix $\hat{\Sigma}_K$ as the "target" and the whole data $X_1, \ldots, X_n$ is used to estimate its entries. The $\hat{\Sigma}_{n,l}$ is replaced by the $K \times K$ banded matrix $\hat{\Sigma}_{b,l,\nu}$ whose entries are computed using the ν^{th} block (subseries) of length b, i.e. $\{X_\nu, \ldots, X_{\nu+b-1}\}$, $\nu = 1, \ldots, n - b + 1$. Finally, (4.2) is estimated by

$$\hat{R}(l) = \frac{1}{n-b+1} \sum_{\nu=1}^{n-b+1} \|\hat{\Sigma}_{b,l,\nu} - \hat{\Sigma}_K\|_{(1,1)}, \tag{4.3}$$

and $\hat{l}$ is selected to minimize $\hat{R}(\cdot)$. Note that whereas l_0 is the best choice in terms of the risk (4.2), $\hat{l}$ tries to adapt to the time series data at hand via (4.3). The optimal choice of the block size b plays a crucial role in selecting the band l. As a general guide, Politis, Romano and Wolf (1999, Chaps 3, 9) show that for consistency in estimation of a parameter, the block size b must grow to infinity while $b/n \to 0$ with a rate like $n^{1/3}$. Note that this requirement is similar to (4.1), corresponding to the choice of $q = 3/2$ or $\alpha = 3$ in Theorem 2. For the computations here, we take $b > K$, and it is fixed at $b = 40$.

Only in a simulation setup where Σ_n is known, it is possible and useful to compare $\hat{l}$ from above to the best band choice for the time series data, i.e.,

$$l_1 = \arg\min_l \|\hat{\Sigma}_{K,l} - \Sigma_K\|_{(1,1)}, \tag{4.4}$$

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

		Mean (SD)			Losses				
n	l_0	l_1	$\hat{l}$	$l_1 - \hat{l}$	$\hat{\Sigma}_{\hat{l}}$	$\hat{\Sigma}_{l_0}$	$\hat{\Sigma}_{l_1}$	$\hat{\Sigma}_K$	$\hat{\Sigma}_n$
250	1	1(0)	1.2(0.8)	-0.2(0.8)	2.7	2.7	0.2	2.4	15.4
500	1	1(0)	1(0)	0(0)	2.2	2.2	0.2	1.8	22.3
750	1	1(0)	1(0)	0(0)	1.9	1.9	0.1	1.4	27.5

where Σ_K is the first $K \times K$ principal minor of Σ_n and $\hat{\Sigma}_{K,l}$ is the l-banded version of $\hat{\Sigma}_K$ in (4.3). Also, the losses of the $K \times K$ and $n \times n$ sample autocovariance matrices, i.e., $\|\hat{\Sigma}_K - \Sigma_K\|_{(1,1)}$ and $\|\hat{\Sigma}_n - \Sigma_n\|_{(1,1)}$, do serve as useful guides on the merits of these estimators and the relevance of (4.2)−(4.4) for band selection.

4.2. Simulations

We have investigated the performance of the above band estimators through a simulation study for several simple linear and nonlinear stationary time series of lengths $n = 250, 500$ and 750; the number of replications used to estimate the risk in (4.2) via (4.3) was $N = 100$. Our simulations below show that l_1 and $\hat{l}$ generally agreed very well with the oracle l_0 whenever it existed for the underlying model. Surprisingly, our findings for linear models are similar to those in Bickel and Levina (2008, Sec. 6.1), in spite of the differences in the setups. However, the situation for nonlinear models is different and requires more research, and possibly a different predictor-based procedure for band selection.

Example 3. The $MA(1)$ Model: The autocovariance matrix Σ_n of the $MA(1)$ model $X_t = \varepsilon_t + \theta\varepsilon_{t-1}$ is banded, and the oracle $l_0 = 1$ for all n. We take $\theta = 0.5$ and, throughout this section, $\{\varepsilon_t\}$ is an iid $N(0, 1)$ sequence. In Table 1, we present the oracle values l_0, l_1, and the estimated $\hat{l}$, and their losses, along with the losses of the sample autocovariance matrices $\hat{\Sigma}_K$ and $\hat{\Sigma}_n$. In this case, the estimation procedure based on (4.3) picks the right banding parameter $l = 1$ more often for larger n, and performs nearly as well as the oracle. The $K \times K$ sample autocovariance matrix does better than $\hat{\Sigma}_{\hat{l}}$; this can be explained partially by our choice of K which is far smaller than n. However, this gain disappears for the $n \times n$ sample autocovariance matrix $\hat{\Sigma}_n$.

Example 4. The $AR(1)$ Model: The autocovariance matrix Σ_n of the $AR(1)$ model $X_t = \phi X_{t-1} + \varepsilon_t$ is not sparse as that in Example 3, but since $\gamma_{i-j} = \phi^{|i-j|}$, its entries decay exponentially as $|i - j|$ gets large. The results (not shown) corresponding to $\phi = 0.1$, 0.5, and 0.9, show that the oracle l_0 and other band estimators were smaller for the smaller values of ϕ, as expected. We note that l_1 generally overestimated l_0, but $\hat{l}$ stayed much closer to the oracle, and the

The 'absolute' AR(1): Estimated band $\hat{l}$ and its loss.

n	ϕ	$\hat{l}(SD)$	$\hat{\Sigma}_{\hat{l}}$
250	0.1	0.2 (0.6)	1.7
250	0.5	0.8 (0.7)	2.2
250	0.9	4.9 (2.6)	14.3
500	0.1	0.0 (0.0)	1.3
500	0.5	0.8 (0.4)	1.8
500	0.9	4.8 (1.4)	13.4
750	0.1	0.0 (0.0)	1.1
750	0.5	0.8 (0.4)	1.5
750	0.9	4.8 (1.2)	14.6

variability of $\hat{l}$ and l_1 increased when the true autocovariance matrix was far from being banded.

For the stationary and invertible ARMA (1, 1) model $X_t = \phi X_{t-1}+\varepsilon_t+\theta\varepsilon_{t-1}$, our simulation results (not shown here) corresponding to $\phi = 0.9$ and $\theta = 0.5$ confirm that the band selectors for this more general model mimicked more closely those of the AR(1) model with $\phi = 0.9$ than the MA(1) model with $\theta = 0.5$ discussed in Example 3.

Example 5. Nonlinear MA(1) and AR(1) Models. As a way to assess the suitability and performance of our band selection procedure for nonlinear time series models, we repeated the earlier simulation design for the following nonlinear MA(1) models: $X_t = \varepsilon_t + \theta|\varepsilon_{t-1}|$, $X_t = \varepsilon_t + \theta\varepsilon_{t-1}|\varepsilon_{t-1}|$ with $\theta = 0.5$, and the 'absolute' AR(1) model (Tong (1990, p.140)) $X_t = \phi|X_{t-1}| + \varepsilon_t$ for $\phi = 0.1, 0.5, 0.9$. Note that these are mildly nonlinear versions of the models in Examples 3 and 4, where the first two processes are 1-dependent but the successive values of the first process are uncorrelated when ε_t has a symmetric distribution with finite second moment. In fact, in this case $\gamma_k = cov(X_{t+k}, X_t) = (1+\theta^2)\delta_{k,0}$, so that Σ_n is a diagonal matrix. However, Σ_n for the second process is tridiagonal, just like the covariance matrix of a linear MA(1) model, with diagonal entries equal to $1 + 3\theta^2$ and the first subdiagonal entries equal to $4\theta/\sqrt{2\pi}$. Our simulation results for the first nonlinear MA(1) model (not reported here) were similar to those in Table 1, and the selected band based on (4.3) was always $\hat{l} = 0$ (which is smaller than the intuitively appealing band $l_0 = 1$ for this case). The simulation results for $X_t = \varepsilon_t + 0.5\varepsilon_{t-1}|\varepsilon_{t-1}|$ were similar to those for the linear MA(1) model in Table 1, except that the losses were larger for the former model.

For the 'absolute' AR(1) process, its autocovariance function does not have a simple closed form so that (4.2) and (4.4) cannot be minimized. However, the selected bands $\hat{l}$, their standard deviations and losses, based on (4.3) for various

values of ϕ and n presented in Table 2 were smaller than their counterparts for the linear AR(1) models. This might be due to the 'smoothness' of the time series plot of the former.

Acknowledgements

We are grateful to the three referees and an associate editor for their very valuable comments which greatly improved the paper. The first author's research was supported by the NSF grant DMS-0448704, and the second author's research was supported by the NSF grant DMS-0505696.

References

Berk, K. (1974). Consistent autoregressive spectral estimates. *Ann. Statist.* **2**, 489-502.

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

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, Revised Edition. Holden-Day, San Francisco.

Brockwell, P. J. and Davis, R. A. (1988). Simple consistent estimation of the coefficients of a linear filter. *Stoch. Processes and Their Applications* **22**, 47-59.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, 2nd Ed. Springer-Verlag, New York.

Bryc, W., Dembo, A. and Jiang, T. (2006). Spectral measure of large random Hankel, Markov and Toeplitz matrices. *Ann. Probab.* **34**, 1-8

El Karoui, N. (2007). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Ann. Probab.* **35** 663-714.

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

Grenander, U. and Szegö, G. (1958). *Toeplitz Forms and Their Applications.* Cambridge University Press, London.

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. Roy. Statist. Soc. Ser. B* **67**, 427-444.

Hannan, E. J. (1979). The central limit theorem for time series regression. *Stochastic Process. Appl.* **9**, 281-289.

Hannan, E. J. and Deistler, M. (1988). *The Statistical Theory of Linear Systems.* Wiley, New York.

Ing, C. K. and Wei, C. Z. (2003). On same-realization prediction in an infinite-order autoregressive process. *J. Multivariate Statist.* **85**, 130-155.

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

Li, Y., Wang, N., Hong, M., Turner, N. D., Lupton, J. R. and Carroll, R. J. (2007). Nonparametric estimation of correlation functions in longitudinal and spatial data, with applications to colon carcinogenesis experiments. *Ann. Statist.* **35**, 1608-1643.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling.* Springer Series in Statistics, New York.

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

Shao, X. and Wu, W. B. (2007). Asymptotic spectral theory for nonlinear time series. *Ann. Statist.* **35**, 1773-1801.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford Scientific Publications.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences USA* **102**, 14150-14154.

Wu, W. B. (2007). Strong invariance principles for dependent random variables. *Ann. Probab.* **35**, 2294-2320.

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

Department of Statistics, 5734 S. University Avenue, Chicago, IL 60637, U.S.A.

E-mail: wbwu@galton.uchicago.edu

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, U.S.A.

E-mail: pourahm@stat.tamu.edu

(Received December 2007; accepted August 2008)