Covariance Matrix Estimation in Time Series

Wei Biao Wu and Han Xiao

June 15, 2011

Abstract

Covariances play a fundamental role in the theory of time series and they are critical quantities that are needed in both spectral and time domain analysis. Estimation of covariance matrices is needed in the construction of confidence regions for unknown parameters, hypothesis testing, principal component analysis, prediction, discriminant analysis among others. In this paper we consider both low- and high-dimensional covariance matrix estimation problems and present a review for asymptotic properties of sample covariances and covariance matrix estimates. In particular, we shall provide an asymptotic theory for estimates of high dimensional covariance matrices in time series, and a consistency result for covariance matrix estimates for estimated parameters.

1 Introduction

Covariances and covariance matrices play a fundamental role in the theory and practice of time series. They are critical quantities that are needed in both spectral and time domain analysis. One encounters the issue of covariance matrix estimation in many problems, for example, the construction of confidence regions for unknown parameters, hypothesis testing, principal component analysis, prediction, discriminant analysis among others. It is particularly relevant in time series analysis in which the observations are dependent and the covariance matrix characterizes the second order dependence of the process. If the underlying process is Gaussian, then the covariances completely capture its dependence structure. In this paper we shall provide an asymptotic distributional theory for sample covariances and convergence rates for covariance matrix estimates of time series.

In Section 2 we shall present a review for asymptotic theory for sample covariances of stationary processes. In particular, the limiting behavior of sample covariances at both small and large lags is discussed. The obtained result is useful for constructing consistent covariance matrix estimates for stationary processes. We shall also present a uniform convergence result so that one can construct simultaneous confidence intervals for covariances and perform tests for white noises. In that section we also introduce dependence measures that are necessary for asymptotic theory for sample covariances.

Sections 3 and 4 concern estimation of covariance matrices, the main theme of the paper. There are basically two types of covariance matrix estimation problems: the first one is the estimation of covariance matrices of some estimated finite-dimensional parameters. For example, given a sequence of observations $Y_1, \ldots, Y_n$, let $\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$ be an estimate of the unknown parameter vector $\theta_0 \in \mathbb{R}^d$, $d \in \mathbb{N}$, which is associated with the process (Y_i). For statistical inference of θ_0, one would like to estimate the $d \times d$ covariance matrix $\Sigma_n = \mathrm{cov}(\hat{\theta}_n)$. For example, with an estimate of Σ_n, confidence regions for θ_0 can be constructed and hypotheses regarding θ_0 can be tested. We generically call such problems low-dimensional covariance matrix estimation problem since the dimension d is assumed to be fixed and it does not grow with n.

For the second type, let $(X_1, \ldots, X_p)$ be a p-dimensional random vector with $E(X_i^2) < \infty$, $i = 1, \ldots p$; let $\gamma_{i,j} = \mathrm{cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$, $1 \le i, j \le p$, be its covariance function. The problem is to estimate the $p \times p$ dimensional matrix

$$\Sigma_p = (\gamma_{i,j})_{1 \le i,j \le p}. \tag{1}$$

A distinguished feature of such type of problem is that the dimension p can be very large. Techniques and asymptotic theory for high-dimensional covariance matrix estimates are quite different from the low-dimensional ones. On the other hand, however, we can build the asymptotic theory for both cases based on the same framework of causal processes and the physical dependence measure proposed in Wu (2005).

The problem of low-dimensional covariance matrix estimation is studied in Section 3. In particular, we consider the latter problem in the context of sample means of random vectors and estimates of linear regression parameters. We shall review the classical theory of Heteroskedasticity and Autocorrelation Consistent (HAC) covariance matrix estimates of White (1980), Newey and West (1987), Andrews (1991) and Andrews and Monahan (1992), de Jong and Davidson (2000) among others. In comparison with those traditional result, an interesting feature of our asymptotic theory is that we impose very mild moment conditions. Additionally, we do not need the strong mixing conditions and the cumulant summability conditions which are widely used in the literature (Andrews (1991), Rosenblatt (1985)). For example, for consistency of covariance matrix estimates, we only require the existence of 2 or $(2 + \epsilon)$ moments, where $\epsilon > 0$ can be very small, while in the classical theory one typically needs the existence of fourth moments. The imposed dependence conditions are easily verifiable and they are optimal in certain sense. In the study of the convergence rates of the estimated covariance matrices, since the dimension is finite, all commonly

used norms (for example, the operator norm, the Frobenius norm and the $\mathcal{L}^1$ norm) are equivalent and the convergence rates do not depend on the norm that one chooses.

Section 4 deals with the second type covariance matrix estimation problem in which p can be big. Due to the high dimensionality, the norms mentioned above are no longer equivalent. Additionally, unlike the lower dimensional case, the sample covariance matrix estimate is no longer consistent. Hence suitable regularization procedures are needed so that the consistency can be achieved. In Section 4 we shall use the operator norm: for an $p \times p$ matrix A, let

$$\rho(A) = \sup_{v:|v|=1} |Av| \tag{2}$$

be the operator norm (or spectral radius), where for a vector $v = (v_1, \ldots, v_p)^\top$, its length $|v| = (\sum_{i=1}^p v_i^2)^{1/2}$. Section 4 provides an exact order of the operator norm of the sample auto-covariance matrix, and the convergence rates of regularized covariance matrix estimates. We shall review the regularized covariance matrix estimation theory of Bickel and Levina (2008a, 2008b), the Cholesky decomposition theory in Pourahmadi (1999), Wu and Pourahmadi (2003) among others, and the parametric covariance matrix estimation using generalized linear models. Suppose one has n independent and identically distributed (iid) realizations of $(X_1, \ldots, X_p)$. In many situations p can be much larger than n, which is the so-called "large p small n problem". Bickel and Levina (2008a) showed that the banded covariance matrix estimate is consistent in operator norm if X_i's have a very short tail and the growth speed of the number of replicates n can be such that $\log(p) = o(n)$. In many time series applications, however, there is only one realization available, namely $n = 1$. In Section 4 we shall consider high-dimensional matrix estimation for both one and multiple realizations. In the former case we assume stationarity and use sample auto-covariance matrix. A banded version of the sample auto-covariance matrix can be consistent.

2 Asymptotics of Sample Auto-Covariances

In this section we shall introduce the framework of stationary causal process, its associated dependence measures and an asymptotic theory for sample auto-covariances. If the process (X_i) is stationary, then $\gamma_{i,j}$ can be written as $\gamma_{i-j} = \mathrm{cov}(X_0, X_{i-j})$, and $\Sigma_n = (\gamma_{i-j})_{1 \le i,j \le n}$ is then a Toeplitz matrix. Assuming at the outset that $\mu = EX_i = 0$. To estimate Σ_n, it

is natural to replace γ_k in Σ_n by the sample version

$$\hat{\gamma}_k = \frac{1}{n}\sum_{i=1+|k|}^{n} X_i X_{i-|k|}, \quad 1-n \le k \le n-1. \tag{3}$$

If $\mu = EX_i$ is not known, then we can modify (3) by

$$\tilde{\gamma}_k = \frac{1}{n}\sum_{i=1+|k|}^{n} (X_i - \bar{X}_n)(X_{i-|k|} - \bar{X}_n), \quad \text{where } \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{4}$$

Section 4.4 concerns estimation of Σ_n and asymptotic properties of $\tilde{\gamma}_k$ will be useful for deriving convergence rates of estimates of Σ_n.

There is a huge literature on asymptotic properties of sample auto-covariances. For linear processes this problem has been studied in Priestley (1981), Brockwell and Davis (1991), Hannan (1970), Anderson (1971), Hall and Heyde (1980), Hannan (1976), Hosking (1996), Phillips and Solo (1992), Wu and Min (2005) and Wu, Huang and Zheng (2010). If the lag k is fixed and bounded, then $\hat{\gamma}_k$ is basically the sample average of the stationary process of lagged products $(X_i X_{i-|k|})$ and one can apply the limit theory for strong mixing processes; see Ibragimov and Linnik (1971), Eberlein and Taqqu (1986), Doukhan (1994) and Bradley (2007).

The asymptotic problem for $\hat{\gamma}_k$ with unbounded k is important since, with that, one can assess the dependence structure of the underlying process by examining its autocovariance function (ACF) plot at large lags. For example, if the time series is a moving average process with an unknown order, then as a common way one can estimate the order by checking its ACF plot. However, the latter problem is quite challenging if the lag k can be unbounded. Keenan (1997) derived a central limit theorem under the very restrictive lag condition $k_n \to \infty$ with $k_n = o(\log n)$ for strong mixing processes whose mixing coefficients decay geometrically fast. A larger range of k_n is allowed in Harris, McCabe and Leybourne (2003). However, they assume that the process is linear. Wu (2008) dealt with nonlinear processes and the lag condition can be quite weak.

To study properties of sample covariances and covariance matrix estimates, it is necessary to impose appropriate structural conditions on (X_i). Here we assume that it is of the form

$$X_i = H(\varepsilon_i, \varepsilon_{i-1}, \ldots), \tag{5}$$

where ε_j, $j \in \mathbb{Z}$, are iid and H is a measurable function such that X_i is properly defined. The framework (5) is very general and it includes many widely used linear and

nonlinear processes (Wu, 2005). Wiener (1958) claimed that, for every stationary purely non-deterministic process $(X_j)_{j\in\mathbb{Z}}$, there exists iid uniform$(0,1)$ random variables ε_j, and a measurable function H such that (5) holds. The latter claim, however, is generally not true; see Rosenblatt (2009), Ornstein (1973) and Kalikow (1982). Nonetheless the above construction suggests that the class of processes that (5) represents can be very huge. See Borkar (1993), Tong (1990), Kallianpur (1981), Ornstein (1973) and Rosenblatt (2009) for more historical backgrounds on the above stochastic realization theory. See also Wu (2011) for examples of stationary processes that are of form (5).

Following Priestley (1988) and Wu (2005), we can view (X_i) as a physical system with $(\varepsilon_j, \varepsilon_{j-1}, \ldots)$ (resp. X_i) being the input (resp. output) and H being the transform, filter or data-generating mechanism. Let the shift process

$$\mathcal{F}_i = (\varepsilon_i, \varepsilon_{i-1}, \ldots). \tag{6}$$

Let $(\varepsilon_i')_{i\in\mathbb{Z}}$ be an iid copy of $(\varepsilon_i)_{i\in\mathbb{Z}}$. Hence $\varepsilon_i', \varepsilon_j, i, j \in \mathbb{Z}$, are iid. For $l \le j$ define

$$\mathcal{F}_{j,l}^* = (\varepsilon_j, \ldots, \varepsilon_{l+1}, \varepsilon_l', \varepsilon_{l-1}, \ldots).$$

If $l > j$, let $\mathcal{F}_{j,l}^* = \mathcal{F}_j$. Define the projection operator

$$\mathcal{P}_j \cdot = E(\cdot|\mathcal{F}_j) - E(\cdot|\mathcal{F}_{j-1}). \tag{7}$$

For a random variable X, we say $X \in \mathcal{L}^p$ $(p > 0)$ if $\|X\|_p := (E|X|^p)^{1/p} < \infty$. Write the $\mathcal{L}^2$ norm $\|X\| = \|X\|_2$. Let $X_i \in \mathcal{L}^p$, $p > 0$. For $j \ge 0$ define the physical (or functional) dependence measure

$$\delta_p(j) = \|X_j - X_j^*\|_p, \text{ where } X_j^* = H(\mathcal{F}_{j,0}). \tag{8}$$

Note that X_j^* is a coupled version of X_j with ε_0 in the latter being replaced by ε_0'. The dependence measure (8) greatly facilitates asymptotic study of random processes. In many cases it is easy to work with and it is directly related to the underlying data-generating mechanism of the process. For $p > 0$, introduce the p-stability condition

$$\Delta_p := \sum_{i=0}^{\infty} \delta_p(i) < \infty. \tag{9}$$

As explained in Wu (2005), (9) means that the cumulative impact of ε_0 on the process $(X_i)_{i\ge 0}$ is finite, thus suggesting short-range dependence. If the above condition is barely

violated, then the process (X_i) may be long-range dependent and the spectral density no longer exists. For example, let $X_n = \sum_{j=0}^{\infty} a_j \varepsilon_{n-j}$ with $a_j \sim j^{-\beta}$, $1/2 < \beta$, and ε_i are iid, then $\delta_p(k) = |a_k| \|\varepsilon_0 - \varepsilon_0'\|_p$ and (9) is violated if $\beta < 1$. The latter is a well-known long-range dependent process. If K is a Lipschitz continuous function, then for the process $X_n = K(\sum_{j=0}^{\infty} a_j \varepsilon_{n-j})$, its physical dependence measure $\delta_p(k)$ is also of order $O(|a_k|)$. Wu (2011) also provides examples of Volterra processes, nonlinear AR(p) and AR(∞) processes for which $\delta_p(i)$ can be computed and (9) can be verified.

For a matrix A, we denote its transpose by $A^\top$.

Theorem 1. *(Wu, 2008, 2011) Let $k \in \mathbb{N}$ be fixed and $E(X_i) = 0$; let $Y_i = (X_i, X_{i-1}, \ldots, X_{i-k})^\top$ and $\Gamma_k = (\gamma_0, \gamma_1, \ldots, \gamma_k)^\top$. (i) Assume $X_i \in \mathcal{L}^p$, $2 < p \le 4$, and (9) holds with this p. Then for all $0 \le k \le n-1$, we have*

$$\|\hat{\gamma}_k - (1 - k/n)\gamma_k\|_{p/2} \le \frac{4n^{2/p-1}\|X_1\|_p \Delta_p}{p-2}. \tag{10}$$

(ii) Assume $X_i \in \mathcal{L}^4$ and (9) holds with $p = 4$. Then as $n \to \infty$,

$$\sqrt{n}(\hat{\gamma}_0 - \gamma_0, \hat{\gamma}_1 - \gamma_1, \ldots, \hat{\gamma}_k - \gamma_k) \Rightarrow N[0, E(D_0 D_0^\top)] \tag{11}$$

where $D_0 = \sum_{i=0}^{\infty} \mathcal{P}_0(X_i Y_i) \in \mathcal{L}^2$ and $\mathcal{P}_0$ is the projection operator defined by (7). (iii) Let $l_n \to \infty$ and assume (9) with $p = 4$. Then we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [X_i Y_{i-l_n} - E(X_{l_n} Y_0)] \Rightarrow N(0, \Sigma_h), \tag{12}$$

where Σ_h is an $h \times h$ matrix with entries

$$\sigma_{ab} = \sum_{j \in \mathbb{Z}} \gamma_{j+a}\gamma_{j+b} = \sum_{j \in \mathbb{Z}} \gamma_j \gamma_{j+b-a} =: \sigma_{0,a-b}, \quad 1 \le a, b \le h, \tag{13}$$

and if additionally $l_n/n \to 0$, then

$$\sqrt{n}[(\hat{\gamma}_{l_n}, \ldots, \hat{\gamma}_{l_n-h+1})^\top - (\gamma_{l_n}, \ldots, \gamma_{l_n-h+1})^\top] \Rightarrow N(0, \Sigma_h). \tag{14}$$

An attractive feature of Theorem 1 is that it provides an explicit error bound (10), which in many cases is sharp up to a multiplicative constant. This is a significant merit for our framework of causal processes with functional or physical dependence measures. See also other theorems in later sections for a similar flavor. In (11) and (12) we give

explicit forms of the asymptotic covariance matrices, and they can be estimated by using the techniques in Section 3.

Theorem 1 suggests that, at large lags, $\sqrt{n}(\hat{\gamma}_k - E\hat{\gamma}_k)$ behaves asymptotically as $\sum_{j\in\mathbb{Z}} \gamma_j \eta_{k-j}$, where η_j are iid standard normal random variables. Wu (2011) discussed the connection with Bartlett's (1946) approximate expressions of covariances of estimated covariances. This result implies that the sample covariance $\hat{\gamma}_k$ can be a bad estimate for γ_k if γ_k is small, due to the weak signal-to-noise ratio. Specifically, if k_n is such that $\gamma_{k_n} = o(n^{-1/2})$, then the sample covariance $\hat{\gamma}_{k_n}$ has an asymptotic mean squared error (MSE) σ_{00}/n, which is larger than $\gamma_{k_n}^2$. Note that $\gamma_{k_n}^2$ is the MSE of the trivial estimate $\breve{\gamma}_{k_n} = 0$. The MSE of the truncated estimate of the form $\bar{\gamma}_k = \hat{\gamma}_k \mathbf{1}_{|\hat{\gamma}_k| \ge c_n}$, where $c_n = c/\sqrt{n}$ for some constant $c > 0$, can reach the minimum order of magnitude $O[\min(1/n, r_n^2)]$. Similar truncation ideas are used in Lumley and Heagerty (1999) and Bickel and Levina (2008b). The latter paper deals with thresholded covariance matrix estimators; see Section 4.4.

As a popular way to test the existence of correlations of a process, one checks its ACF plot. Testing of correlations involves testing multiple hypotheses $H_0 : \gamma_1 = \gamma_2 = \ldots = 0$. The multiplicity issue should be adjusted if the number of lags is unbounded. To develop a rigorous test, we need to establish a distributional result for $\max_{k\le s_n} |\hat{\gamma}_k - \gamma_k|$, where s_n is the largest lag which can grow to infinity. It turns out that, with the physical dependence measure, we can formulate an asymptotic result for the maximum deviation $\max_{k\le s_n} |\hat{\gamma}_k - E\hat{\gamma}_k|$. Such a result can be used to construct simultaneous confidence intervals for γ_k with multiple lags. Let

$$\Delta_p(m) = \sum_{i=m}^{\infty} \delta_p(i), \;\; \Psi_p(m) = (\sum_{i=m}^{\infty} \delta_p^2(i))^{1/2} \tag{15}$$

and

$$\Phi_p(m) = \sum_{i=0}^{\infty} \min\{\delta_p(i), \, \Psi_p(m)\}. \tag{16}$$

Theorem 2. *(Xiao and Wu, 2011a) Assume that $EX_i = 0$, $X_i \in \mathcal{L}^p$, $p > 4$, $\Delta_p(m) = O(m^{-\alpha})$ and $\Phi_p(m) = O(m^{-\alpha'})$, where $\alpha, \alpha' > 0$. (i) If $\alpha > 1/2$ or $\alpha' p > 2$, then for $c_p = 6(p+4)e^{p/4}\Delta_4 \|X_i\|_4$, we have*

$$\lim_{n\to\infty} P\left(\max_{1\le k<n} |\hat{\gamma}_k - E\hat{\gamma}_k| \le c_p \sqrt{\frac{\log n}{n}} \right) = 1. \tag{17}$$

(ii) If $s_n \to \infty$ satisfies $s_n = O(n^\eta)$ with $0 < \eta < \min(1, \alpha p/2)$ and $\eta \min(2 - 4/p - 2\alpha,\, 1 - 2\alpha') < 1 - 4/p$, then we have the Gumbel convergence: for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\max_{1\le k\le s_n} \sqrt{n}|\hat{\gamma}_k - E\hat{\gamma}_k| \le \sigma_0^{1/2}(a_{2s_n}x + b_{2s_n})\right) = \exp(-\exp(-x)), \tag{18}$$

where $a_n = (2\log n)^{-1/2}$ and $b_n = a_n(4\log n - \log\log n - \log 4\pi)/2$.

3 Low-dimensional Covariance Matrix Estimation

The problem of low-dimensional covariance matrix estimation often arises when one wants to estimate unknown parameters that are associated with a time series. Let θ_0 be an unknown parameter associated with the process (Y_i). Given observations $Y_1, \ldots, Y_n$, we estimate θ_0 by $\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$. For example, if (Y_i) is a d-dimensional process with a common unknown mean vector $\mu_0 = EY_i$, then we can estimate it by the sample mean vector

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{19}$$

Under appropriate conditions on the process (Y_i), we expect that the central limit theorem for $\hat{\theta}_n$ holds:

$$\Sigma_n^{-1/2}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, \mathrm{Id}_d), \tag{20}$$

where Id_d is the d-dimensional identity matrix. With (20), one can construct confidence regions for θ_0. In particular, let $\hat{\Sigma}_n$ be an estimate of Σ_n. Then the $(1-\alpha)$th, $0 < \alpha < 1$, confidence ellipse for θ_0 is

$$\{\nu \in \mathbb{R}^d :\, (\hat{\theta}_n - \nu)^\top \hat{\Sigma}_n^{-1}(\hat{\theta}_n - \nu) = |\hat{\Sigma}_n^{-1/2}(\hat{\theta}_n - \nu)|^2 \le \chi^2_{d,1-\alpha}\}, \tag{21}$$

where $\chi^2_{d,1-\alpha}$ is the $(1-\alpha)$th quantile of a χ^2 distribution with degree of freedom d. The key question in the above construction now becomes the estimation of Σ_n. The latter question is closely related to the long-run variance estimation problem.

In the derivation of the central limit theorem (20), one typically needs to establish an asymptotic expansion of the type

$$\hat{\theta}_n - \theta_0 = \sum_{i=1}^{n} X_i + R_n, \tag{22}$$

where R_n is negligible in the sense that $\Sigma_n^{-1/2} R_n = o_{\mathbb{P}}(1)$ and (X_i) is a random process associated with (Y_i) satisfying the central limit theorem

$$\Sigma_n^{-1/2} \sum_{i=1}^{n} X_i \Rightarrow N(0, \mathrm{Id}_d).$$

Sometimes the expansion (22) is called the Bahadur representation (Bahadur, 1966). For iid random variables $Y_1, \ldots, Y_n$, Bahadur obtained an asymptotic linearizing approximation for its αth ($0 < \alpha < 1$) sample quantile. Such an approximation greatly facilitates an asymptotic study. Note that the sample quantile depends on Y_i in a complicated nonlinear manner. The asymptotic expansion (22) can be obtained from the maximum likelihood, quasi maximum likelihood, or general method of moments estimation procedures. The random variables X_i in (22) are called scores or estimating functions. As another example, assume that (Y_i) is a stationary Markov process with transition density $p_{\theta_0}(Y_i|Y_{i-1})$, where θ_0 is an unknown parameter. Then given the observations $Y_0, \ldots, Y_n$, the conditional maximum likelihood estimate $\hat{\theta}_n$ maximizes

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(Y_i|Y_{i-1}). \tag{23}$$

As is common in the likelihood estimation theory, let $\dot{\ell}_n(\theta) = \partial \ell_n(\theta)/\partial\theta$ and $\ddot{\ell}_n(\theta) = \partial^2 \ell_n(\theta)/\partial\theta\partial\theta^\top$ be a $d \times d$ matrix. By the ergodic theorem, $\ddot{\ell}_n(\theta_0)/n \to E\ddot{\ell}_1(\theta_0)$ almost surely. Since $\dot{\ell}_n(\hat{\theta}_n) = 0$, under suitable conditions on the process (Y_i), we can perform the Taylor expansion $\dot{\ell}_n(\hat{\theta}_n) \approx \dot{\ell}_n(\theta_0) + \ddot{\ell}_n(\theta_0)(\hat{\theta}_n - \theta_0)$. Hence the representation (22) holds with

$$X_i = n^{-1}(E\ddot{\ell}_1(\theta_0))^{-1} \frac{\partial}{\partial\theta} \log p_\theta(Y_i|Y_{i-1})|_{\theta=\theta_0}. \tag{24}$$

A general theory for establishing (22) is presented in Amemiya (1985) and Heyde (1997) and various special cases are considered in Hall and Hyde (1980), Hall and Yao (2003), Wu (2007), He and Shao (1996), Klimko and Nelson (1978) and Tong (1990), among others.

For the sample mean estimate (19), it is also of form (22) by writing $\hat{\mu}_n - \mu_0 = n^{-1}\sum_{i=1}^{n}(Y_i - \mu_0)$ and $X_i = (Y_i - \mu_0)/n$. Therefore, to estimate the covariance matrix of an estimated parameter, in view of (22), we typically need to estimate the covariance matrix Σ_n of the sum $S_n = \sum_{i=1}^{n} X_i$. Clearly,

$$\Sigma_n = \sum_{1 \le i,j \le n} \mathrm{cov}(X_i, X_j), \tag{25}$$

where $\mathrm{cov}(X_i, X_j) = E(X_i X_j^\top) - E(X_i)E(X_j^\top)$. Sections 3.1, 3.2 and 3.3 concern convergence rates of estimates of Σ_n based on observations $(X_i)_{i=1}^n$ which can be independent, uncorrelated, non-stationary and weakly dependent. In the estimation of the covariance matrix for $S_n = \sum_{i=1}^n X_i$ for $\hat{\theta}_n$ based on the representation (22), the estimating functions X_i may depend on the unknown parameter θ_0, hence $X_i = X_i(\theta_0)$ may not be observed. For example, for the sample mean estimate (19), one has $X_i = (Y_i - \mu_0)/n$ while for the conditional MLE, X_i in (24) also depends on the unknown parameter θ_0. Heagerty and Lumley (2000) considered estimation of covariance matrices for estimated parameters for strong mixing processes; see also Newey and West (1987) and Andrews (1991). In Corollary 1 of Section 3.2 and Section 3.4 we shall present asymptotic results for covariance matrix estimates with estimated parameters.

3.1 HC Covariance Matrix Estimators

For independent but not necessarily identically distributed random vectors X_i, $1 \le i \le n$, White (1980) proposed a heteroskedasticity-consistent (HC) covariance matrix estimator for $\Sigma_n = \mathrm{var}(S_n)$, $S_n = \sum_{i=1}^n X_i$. Other contributions can be found in Eicker (1963) and MacKinnon and White (1985). If $\mu_0 = EX_i$ is known, we can estimate Σ_n by

$$\hat{\Sigma}_n^\circ = \sum_{i=1}^n (X_i - \mu_0)(X_i - \mu_0)^\top. \tag{26}$$

If μ_0 is unknown, we shall replaced it by $\hat{\mu}_n = \sum_{i=1}^n X_i/n$, and form the estimate

$$\begin{aligned} \hat{\Sigma}_n &= \frac{n}{n-1} \sum_{i=1}^n (X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^\top \\ &= \frac{n}{n-1} \sum_{i=1}^n (X_i X_i^\top - \hat{\mu}_n \hat{\mu}_n^\top). \end{aligned} \tag{27}$$

Both $\hat{\Sigma}_n^\circ$ and $\hat{\Sigma}_n$ are unbiased for Σ_n. To this end, assume without loss of generality $\mu = 0$, then by independence, $n^2 E(\hat{\mu}_n \hat{\mu}_n^\top) = \sum_{i=1}^n E(X_i X_i^\top)$, hence

$$\begin{aligned} E\hat{\Sigma}_n &= \frac{n}{n-1} \left[\sum_{i=1}^n E(X_i X_i^\top) - E(n \hat{\mu}_n \hat{\mu}_n^\top) \right] \\ &= \sum_{i=1}^n E(X_i X_i^\top) = \Sigma_n. \end{aligned} \tag{28}$$

Theorem 3 below provides a convergence rate of $\hat{\Sigma}_n^\circ$. We omit its proof since it is an easy consequence of the Rothenthal inequality.

Theorem 3. *Assume that X_i are independent $\mathbb{R}^d$ random vectors with $EX_i = 0, X_i \in \mathcal{L}^p$, $2 < p \le 4$. Then there exists a constant C, only depending on p and d, such that*

$$\|\hat{\Sigma}_n^\circ - \Sigma_n\|_{p/2}^{p/2} \le C \sum_{i=1}^n \|X_i\|_p^p. \tag{29}$$

As an immediate consequence of Theorem 3, if $\Sigma := \mathrm{cov}(X_i)$ does not depend on i and Σ is positive definite (namely $\Sigma > 0$) and $\sup_i \|X_i\|_p < \infty$, then

$$\|\hat{\Sigma}_n^\circ \Sigma_n^{-1} - \mathrm{Id}_d\|_{p/2} = O(n^{2/p-1})$$

and the confidence ellipse in (21) has an asymptotically correct coverage probability. Simple calculation shows that the above relation also holds if $\hat{\Sigma}_n^\circ$ is replaced by $\hat{\Sigma}_n$.

If X_i are uncorrelated, using the computation in (28), it is easily seen that the estimates $\hat{\Sigma}_n^\circ$ in (26) and $\hat{\Sigma}_n$ in (27) are still unbiased. However, one no longer has (29) if X_i are only uncorrelated instead of being independent. To establish an upper bound, as in Wu (2011), we assume that (X_i) has the form

$$X_i = H_i(\varepsilon_i, \varepsilon_{i-1}, \ldots), \tag{30}$$

where ε_i are iid random variables and H_i is a measurable function such that X_i is a proper random variable. If the function H_i does not depend on i, then (30) reduces to (5). In general (30) defines a non-stationary process. According to the stochastic representation theory, any finite dimensional random vector can be expressed in distribution as functions of iid uniform random variables; see Wu (2011) for a review. As in (8), define the physical dependence measure

$$\delta_p(k) = \sup_i \|X_i - X_{i,k}\|_p, \ k \ge 0, \tag{31}$$

where $X_{i,k}$ is a couple process of X_i with ε_{i-k} in the latter being replaced by ε'_{i-k}. For stationary processes of form (5), (8) and (31) are identical.

Theorem 4. *Assume that X_i are uncorrelated with form (30) and $EX_i = 0, X_i \in \mathcal{L}^p$, $2 < p \le 4$. Let $\kappa_p = \sup_i \|X_i\|_p$. Then there exists a constant $C = C_{p,d}$ such that*

$$\|\hat{\Sigma}_n^\circ - \Sigma_n\|_{p/2} \le C n^{2/p} \kappa_p \sum_{k=0}^\infty \delta_p(k). \tag{32}$$

Proof. Let $\alpha = p/2$. Since $X_iX_i^\top - EX_iX_i^\top = \sum_{k=0}^\infty \mathcal{P}_{i-k}(X_iX_i^\top)$ and $\mathcal{P}_{i-k}(X_iX_i^\top)$, $i = 1, \ldots, n$ are martingale differences, by the Burkholder and Minkowski inequalities, we have

$$\begin{aligned} \|\hat{\Sigma}_n^\circ - \Sigma_n\|_\alpha &\le \sum_{k=0}^\infty \|\sum_{i=1}^n \mathcal{P}_{i-k}(X_iX_i^\top)\|_\alpha \\ &\le C\sum_{k=0}^\infty \left[\sum_{i=1}^n \|\mathcal{P}_{i-k}(X_iX_i^\top)\|_\alpha^\alpha\right]^{1/\alpha}. \end{aligned}$$

Observe that $E[(X_{k,0}X_{k,0}^\top)|\mathcal{F}_0] = E[(X_kX_k^\top)|\mathcal{F}_{-1}]$. By Scharwz inequality, $\|\mathcal{P}_0(X_kX_k^\top)\|_\alpha \le \|X_{k,0}X_{k,0}^\top - X_kX_k^\top\|_\alpha \le 2\kappa_p\delta_p(k)$. Hence we have (32). ♢

3.2 Long-run Covariance Matrix Estimation for Stationary Vectors

If X_i are correlated, then the estimate (27) is no longer consistent for Σ_n and auto-covariances need to be taken into consideration. Recall $S_n = \sum_{i=1}^n X_i$. Assume $EX_i = 0$. Using the idea of lag window spectral density estimate, we estimate the covariance matrix $\Sigma_n = \mathrm{var}(S_n)$ by

$$\tilde{\Sigma}_n = \sum_{1\le i,j\le n} K(\frac{i-j}{B_n})X_iX_j^\top, \tag{33}$$

where K is a window function satisfying $K(0) = 1$, $K(u) = 0$ if $|u| > 1$, K is even and differentiable on the interval $[-1, 1]$, and B_n is the lag sequence satisfying $B_n \to \infty$ and $B_n/n \to 0$. The former condition is for including unknown order of dependence, while the latter is for the purpose of consistency.

If (X_i) is a scalar process, then (33) is the lag-window estimate for the long-run variance $\sigma_\infty^2 = \sum_{k\in\mathbb{Z}} \gamma_i$, where $\gamma_i = \mathrm{cov}(X_0, X_i)$. Note that $\sigma_\infty^2/(2\pi)$ is the value of the spectral density of (X_i) at zero frequency. There is a huge literature on spectral density estimation; see the classical textbooks Anderson (1971), Brillinger (1975), Brockwell and Davis (1991), Grenander and Rosenblatt (1957), Priestley (1981), Rosenblatt (1985) and the third volume Handbook of Statistics "*Time Series in the Frequency Domain*" edited by Brillinger and Krishnaiah (1983). Rosenblatt (1985) showed the asymptotic normality for lag-window spectral density estimates for strong mixing processes under a summability condition of eighth-order joint cumulants.

Liu and Wu (2010) present an asymptotic theory for lag-window spectral density estimates under minimal moment and natural dependence conditions. Their results can be easily extended to the vector-valued processes. Assume $EX_i = 0$. then $\Sigma_n = \mathrm{var}(S_n)$ satisfies

$$\frac{1}{n}\Sigma_n = \sum_{k=1-n}^{n-1} (1 - |k|/n) E(X_0 X_k^\top) \to \sum_{k=-\infty}^{\infty} E(X_0 X_k^\top) =: \Sigma^\dagger. \tag{34}$$

Let vec be the vector operator. We have the following consistency and central limit theorem for $\mathrm{vec}(\tilde{\Sigma}_n)$. Its proof can be similarly carried out by using the argument in Liu and Wu (2010). Details are omitted.

Theorem 5. *Assume that the d-dimensional stationary process (X_i) is of form (5), and $B_n \to \infty$ and $B_n = o(n)$. (i) If the short-range dependence condition (9) holds with $p \geq 2$, then $\|\tilde{\Sigma}_n/n - \Sigma_n/n\|_{p/2} = o(1)$ and, by (34), $\|\tilde{\Sigma}_n/n - \Sigma^\dagger\|_{p/2} = o(1)$. (ii) If (9) holds with $p = 4$, then there exists a matrix Γ with $\rho(\Gamma) < \infty$ such that*

$$(nB_n)^{-1/2}[\mathrm{vec}(\tilde{\Sigma}_n) - E\mathrm{vec}(\tilde{\Sigma}_n)] \Rightarrow N(0, \Gamma), \tag{35}$$

and the bias

$$n^{-1} \left\| E\mathrm{vec}(\tilde{\Sigma}_n) - \Sigma_n \right\| \leq \sum_{k=-B_n}^{B_n} |1 - K(k/B_n)| \, \gamma_2(k) + 2 \sum_{k=B_n+1}^{n} \gamma_2(k), \tag{36}$$

where $\gamma_2(k) = \left\| E(X_0 X_{i+k}^\top) \right\| \leq \sum_{i=0}^{\infty} \delta_2(i)\delta_2(i+k)$.

An interesting feature of Theorem 5(i) is that, under the minimal moment condition $X_i \in \mathcal{L}^2$ and the very mild weak dependence condition $\Delta_2 < \infty$, the estimate $\tilde{\Sigma}_n/n$ is consistent for Σ_n/n. This property substantially extends the range of applicability of lag-window covariance matrix estimates. For consistency Andrews (1991) requires finite fourth moment and a fourth-order joint cumulant summability condition, while for computing the asymptotic mean square error, he needs finite eighth moment and an eighth-order cumulant summability condition. For nonlinear processes it might be difficult to verify those cumulant summability conditions. Our framework of physical dependence measure seems quite convenient and useful for long-run covariance matrix estimation and it is no longer needed to work with joint cumulants.

In many situations X_i depends on unknown parameters and thus is not directly observable. For example, X_i in (24) depends on the unknown parameter θ_n. Then it is natural

to modify the $\tilde{\Sigma}_n$ in (33) by the following estimate

$$\tilde{\Sigma}_n(\hat{\theta}_n) = \sum_{1\le i,j\le n} K(\frac{i-j}{B_n}) X_i(\hat{\theta}_n) X_j(\hat{\theta}_n)^\top, \tag{37}$$

where $\hat{\theta}_n$ is an estimate of θ_0, so that $X_i(\hat{\theta}_n)$ are estimates of $X_i(\theta_0) = X_i$. Note that $\tilde{\Sigma}_n(\theta_0) = \tilde{\Sigma}_n$. As in Newey and West (1987) and Andrews (1991), appropriate continuity conditions on the random function $X_i(\cdot)$ can imply the consistency of the estimate $\tilde{\Sigma}_n(\hat{\theta}_n)$. The following Corollary 1 is a straightforward consequence of Theorem 5.

Corollary 1. *Assume that $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimate of θ_0, namely $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_{\mathbb{P}}(1)$. Further assume that there exists a constant $\delta_0 > 0$ such that the local maximal function $X_i^* = \sup\{|\partial X_i(\theta)/\partial\theta| : |\theta - \theta_0| \le \delta_0\} \in \mathcal{L}^2$. Assume $B_n \to \infty$, $B_n = o(\sqrt{n})$ and (9) holds with $p = 2$. Then $\tilde{\Sigma}_n(\hat{\theta}_n)/n - \Sigma_n/n \to 0$ in probability.*

3.3 HAC Covariance Matrix Estimators

Recall (31) for the definition of physical dependence measures for non-stationary processes. If (X_i) is non-stationary and correlated, then under a similar short-range dependence condition as (9), we can also obtain a convergence rate for the estimate $\tilde{\Sigma}_n$ defined in (33). Results of similar type were given in Newey and West (1987) and Andrews (1991). Andrews and Monahan (1992) improved the estimate by using a pre-whitening procedure.

Theorem 6. *Let $B_n \to \infty$ and $B_n/n \to 0$. Assume that the non-stationary process (X_i) is of form (30), $X_i \in \mathcal{L}^p$, $p > 2$, and the short-range dependence condition (9). (i) If $2 < p < 4$, then $\|\tilde{\Sigma}_n/n - \Sigma_n/n\|_{p/2} = o(1)$. (ii) If $p \ge 4$, then there exists a constant C depending on p and d only such that*

$$\|\tilde{\Sigma}_n - E\tilde{\Sigma}_n\|_{p/2} \le C\Delta_p^2 B_n, \tag{38}$$

and the bias

$$n^{-1}\left\|E\tilde{\Sigma}_n - \Sigma_n\right\| \le \sum_{k=-B_n}^{B_n} |1 - K(k/B_n)|\,\gamma_2(k) + 2\sum_{k=B_n+1}^{n} \gamma_2(k), \tag{39}$$

where $\gamma_2(k) = \sup_i \left\|E(X_i X_{i+k}^\top)\right\| \le \sum_{i=0}^{\infty} \delta_2(i)\delta_2(i+k)$.

Remark 1. We emphasize that in this section, since the dimension of the covariance matrix Σ is fixed, all matrix norms are essentially equivalent and the relations (29), (32), (36) and (38) also hold if we use other types of matrix norms, such as the Frobenius norm and the maximum entry norm. This feature is no longer present for high dimensional matrix estimation where the dimension can be unbounded; see Section 4. $\diamondsuit$

As Theorem 5, the proof of Theorem 6 can be similarly carried out by using the argument in Liu and Wu (2010). A crucial step in applying Theorem 6 is how to choose the smoothing parameter. See Zeileis (2004) for an excellent account for the latter problem.

3.4 Covariance Matrix Estimation for Linear Models

Consider the linear model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i, \quad 1 \le i \le n, \tag{40}$$

where $\boldsymbol{\beta}$ is an $s \times 1$ unknown regression coefficient vector, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{is})'$ are $s \times 1$ known (non-stochastic) design vectors. Let $\hat{\boldsymbol{\beta}}$ be the least square estimate of $\boldsymbol{\beta}$. Here we consider the estimation of $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ under the assumption that (e_i) is a non-stationary process of form (30). As a special case, if there is only one covariate and $x_i = 1$ for each $1 \le i \le n$, then $\hat{\boldsymbol{\beta}} = S_n/n$ with $S_n = \sum_{i=1}^n e_i$, so the estimation of the covariance matrix of S_n in Section 3.3 is a special case here. Assume that for large n, $T_n := \mathbf{X}_n^\top \mathbf{X}_n$ is positive definite. It is more convenient to consider the rescaled model

$$y_i = \boldsymbol{z}_i^\top \boldsymbol{\theta} + e_i \quad \text{with} \quad \boldsymbol{z}_i = \boldsymbol{z}_{i,n} = T_n^{-1/2} \boldsymbol{x}_i \text{ and } \boldsymbol{\theta} = \boldsymbol{\theta}_n = T_n^{1/2} \boldsymbol{\beta}, \tag{41}$$

under which the least square estimate $\hat{\boldsymbol{\theta}} = \sum_{i=1}^n \boldsymbol{z}_i e_i$. If e_i were known, we can estimate $\Sigma_n := \mathrm{cov}(\hat{\boldsymbol{\theta}})$ by

$$V_n = \sum_{1 \le i,j \le n} K\left(\frac{i-j}{B_n}\right) \boldsymbol{z}_i e_i \boldsymbol{z}_j^\top e_j, \tag{42}$$

which is in the similar fashion as (33). Since e_i are unknown, we should replace e_i in V_n by the estimated residuals $\hat{e}_i$ and employ the following estimate

$$\hat{V}_n = \sum_{1 \le i,j \le n} \boldsymbol{z}_i \hat{e}_i \boldsymbol{z}_j^\top \hat{e}_j c_{ij}, \tag{43}$$

where $c_{ij} = K((i-j)/B_n)$. We have the following convergence rate of the estimate $\hat{V}_n$, which can be derived using similar arguments as those in Liu and Wu (2010).

Theorem 7. *Assume that the non-stationary process (e_i) is of form (30), $e_i \in \mathcal{L}^p$ with $p \geq 4$, and $\Delta_p < \infty$. Let $c_k := K(k/B_n)$. Then there exists a constant C depending only on p and s such that*

$$\left\| \hat{V}_n - EV_n \right\|_{p/2} \leq C\Delta_p^2 \left(\sum_{1 \leq i,j \leq n} c_{i-j}^2 |\boldsymbol{z}_i|^2 |\boldsymbol{z}_j|^2 \right)^{1/2}, \tag{44}$$

and the bias

$$\|EV_n - \Sigma_n\| \leq s \sum_{k=1-n}^{n-1} |1 - c_k| \gamma_2(k), \tag{45}$$

where $\gamma_2(k) = \sup_{i \in \mathbb{Z}} |E(e_i e_{i+k})| \leq \sum_{i=0}^{\infty} \delta_2(i)\delta_2(i + |k|)$.

In Example 1 of Section 4.4, we shall obtain a best linear unbiased estimate for β by estimating the high-dimensional covariance matrix of $(e_1, \ldots, e_n)$. It illustrates the different natures of two types of covariance matrix estimation.

4 High Dimensional Covariance Matrix Estimation

In this section we shall consider estimation of high-dimensional covariance matrices in time series in which the dimensions can grow to infinity. This setting is quite different from the one in (25) where the dimension is fixed and does not grow. During the last decade the problem of high-dimensional covariance matrix estimation has attracted considerable attention. See Pourahmadi (2011) for an excellent review. The problem is quite challenging since, for estimating Σ_p given in (1), one has to estimate $p(p+1)/2$ unknown parameters. Additionally, those parameters must follow the highly nontrivial positive-definiteness constraint. In the multivariate setting in which one has multiple iid p-variates random variables, the problem has been extensively studied; see Meinshausen and Bühlman (2006), Yuan and Lin (2007), Rothman et al (2009), Bickel and Levina (2008a, 2008b), Cai et al (2010), Lam and Fan (2009) and Ledoit and Wolf (2004) among others. As commented in Bickel and Gel (2011), the same problem in longitudinal and time series setting has been much less investigated. In comparison with the matrix estimation problem in the context of multivariate statistics, there are several distinguished features when we consider time series:

(i) The order information among observations is highly relevant;

(ii) Variables that are far apart are weakly dependent;

(iii) The number of replicates is very small, and in many cases there is only one realization available.

In multivariate statistics, in many cases one can permute the variables without sacrificing the interpretability and the permutation-invariance property of a covariance matrix estimate can be quite appealing. For covariance matrix estimation in time series, however, the permutation-invariance property is not a must.

Section 4.1 reviews the Cholesky decomposition of covariance matrices. As argued in Pourahmadi (1999), the Cholesky decomposition based covariance matrix estimate is inherently positive definite and its entries have a nice interpretation of being autoregressive coefficients. In Section 4.2 we briefly review parametric covariance matrix estimation where the target covariance matrix is of certain parametric forms. Thus it suffices to estimate the governing parameters. Sections 4.3 and 4.4 concern the nonparametric covariance matrix estimation problem for two different settings: in the first setting we assume that there are multiple iid realizations of the underlying process, while in the second one only one realization is available. For the latter we assume that the underlying process is stationary.

4.1 Cholesky Decomposition

Assume that $X_1, \ldots, X_p$ is a mean zero Gaussian process with covariance matrix Σ_p given in (1). As in Pourahmadi (1999), we perform successive auto-regression of X_t on its predecessors $X_1, \ldots, X_{t-1}$ in the following manner:

$$X_t = \sum_{j=1}^{t-1} \phi_{tj} X_j + \eta_t =: \hat{X}_t + \eta_t, \ t = 1, \ldots, p, \tag{46}$$

where ϕ_{tj} are the auto-regressive coefficients such that $\hat{X}_t$ is the projection of X_t onto the linear space spanned by $X_1, \ldots, X_{t-1}$. Then $\eta_1 \equiv X_1$ and $\eta_t = X_t - \hat{X}_t$, $t = 2, \ldots, n$, are independent. Let $\sigma_t^2 = \mathrm{var}(\eta_t)$ be the innovation variance, $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ and

$$L = \begin{pmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \\ -\phi_{11} & -\phi_{p2} & \ldots & -\phi_{p,p-1} & 1 \end{pmatrix} \tag{47}$$

be a lower triangle matrix. Then Σ_p has the representation

$$L\Sigma_p L^\top = D^2, \tag{48}$$

which implies the useful fact that the inverse, or the precision matrix,

$$\Sigma_p^{-1} = LD^{-2}L^\top. \tag{49}$$

An important feature of the representation (48) is that the coefficients in L are unconstrained and, if an estimate of Σ_p is computed based on estimated L and D, then it is guaranteed to be nonnegative-definite. The Cholesky method is particularly suited for covariance and precision matrix estimation in time series, and the entries in L can be interpreted as auto-regressive coefficients.

Another popular method is the eigen-decomposition $\Sigma_p = Q\Lambda Q^\top$, where Q is an orthonormal matrix, namely $QQ^\top = \mathrm{Id}_p$ and Λ is a diagonal matrix which consists of eigenvalues of Σ_p. The eigen-decomposition is related to the principal component analysis. It is generally not easy to work with the orthonormality constraint. See Pourahmadi (2011) for more discussion.

4.2 Parametric Covariance Matrix Estimation

In the parametric covariance matrix estimation problem, one assumes that Σ_n has a known form $\Sigma_n(\theta)$ indexed by a finite dimensional parameter. To estimate Σ_n, it would then suffice if we can find a good estimate of θ. Anderson (1970) assumed that Σ_n is a linear combination of some known matrices. Burg, Luenberger and Wenger (1982) applied the maximum likelihood estimation method; see also Quang (1984), Dembo (1986), Fuhrmann and Miller (1988), Jansson and Ottersten (2000) and Dietrich (2008). Chiu et al (1996) used a log-linear covariance matrix parametrization.

Based on the Cholesky decomposition (48), Pourahmadi (1999) considered parametric modelling for the auto-regressive coefficients ϕ_{ij} and the innovation variance σ_i^2, thus substantially reducing the number of parameters. See also Pan and MacKenzie (2003) and Zimmerman and Núñez-Antón (2010).

4.3 Covariance Matrix Estimation with Multiple iid Realizations

Assume that $(X_{l,1}, X_{l,2}, \ldots, X_{l,p})$, $l = 1, \ldots, m$, are iid random vectors identically distributed as $(X_1, \ldots, X_p)$. If the means $\mu_j = EX_{l,j}$, $j = 1, \ldots, p$ are known, then the

covariance $\gamma_{i,j} = \mathrm{cov}(X_{l,i}, X_{l,j})$, $1 \le i, j \le p$, can be estimated by

$$\hat{\gamma}_{i,j} = \frac{1}{m} \sum_{l=1}^{m} (X_{l,i} - \mu_i)(X_{l,j} - \mu_j) \tag{50}$$

and the sample covariance matrix estimate is

$$\hat{\Sigma}_p = (\hat{\gamma}_{i,j})_{1 \le i,j \le p}. \tag{51}$$

If μ_j is unknown, one can naturally estimate it by the sample mean $\bar{\mu}_j = m^{-1} \sum_{l=1}^{m} X_{l,j}$ and $\hat{\gamma}_{i,j}$ and $\hat{\Sigma}_p$ in (50) and (51) can then be modified accordingly.

According to the modern random matrix theory, under the assumption that all entries $X_{l,i}, 1 \le l \le m, 1 \le i \le p$, are independent, $\hat{\Sigma}_p$ is a bad estimate of Σ_p in the sense that it is inconsistent in operator norm. Such inconsistency results for sample covariance matrices in multivariate analysis have been discussed in Stein (1975), Bai and Silverstein (2010), El Karoui (2007), Paul (2007), Johnstone (2001), Geman (1980), Wachter (1978), Anderson, Guionnet and Zeitouni (2010) among others. Note that if $m < p$, $\hat{\Sigma}_p$ is a singular matrix. It is known that, under appropriate moment conditions of $X_{l,i}$, if $p/m \to c$, then the empirical distribution of eigenvalues of $\hat{\Sigma}_p$ follows the Marcenko-Pastur law which has the support $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$ and a point mass at zero if $c > 1$; and the largest eigenvalue, after proper normalization, follows the Tracy-Widom law. All those results suggest the inconsistency of sample covariance matrices.

For an improved and consistent estimator, various regularization methods have been proposed. Assuming that the correlations are weak if the lag $i - j$ is large, Bickel and Levina (2008a) proposed the banded covariance matrix estimate

$$\hat{\Sigma}_{p,B} = (\hat{\gamma}_{i,j} \mathbf{1}_{|i-j| \le B})_{1 \le i,j \le p}, \tag{52}$$

where $B = B_p$ is the band parameter, and more generally the tapered estimate

$$\hat{\Sigma}_{p,B} = (\hat{\gamma}_{i,j} K(|i-j|/B))_{1 \le i,j \le p}, \tag{53}$$

where K is a symmetric window function with support on $[-1, 1]$, $K(0) = 1$ and K is continuous on $(-1, 1)$. Here we assume that $B_p \to \infty$ and $B_p/p \to 0$. The former condition ensures that $\hat{\Sigma}_{p,B}$ can include dependencies at unknown orders while the latter aims to circumvent the weak signal-to-noise ratio issue that $\hat{\gamma}_{i,j}$ is a bad estimate of $\gamma_{i,j}$ if $|i-j|$ is big. In particular, Bickel and Levina (2008a) considered the class

$$\mathcal{U}(\epsilon_0, \alpha, C) = \left\{ \Sigma : \max_j \sum_{i:\, |i-j|>k} |\gamma_{i,j}| \le C k^{-\alpha},\ \rho(\Sigma) \le \epsilon_0^{-1}, \rho(\Sigma^{-1}) \le \epsilon_0 \right\}. \tag{54}$$

This condition quantifies issue (ii) mentioned in the beginning of this section. They proved that, (i) if $\max_j E\exp(uX_{l,i}^2) < \infty$ for some $u > 0$ and $k_n \asymp (m^{-1}\log p)^{-1/(2\alpha+2)}$, then

$$\rho(\hat{\Sigma}_{p,k_p} - \Sigma_p) = \mathcal{O}_P[(m^{-1}\log p)^{\alpha/(2\alpha+2)}]; \tag{55}$$

(ii) if $\max_j E|X_{l,i}|^\beta < \infty$ and $k_n \asymp (m^{-1/2}p^{2/\beta})^{c(\alpha)}$, where $c(\alpha) = (1+\alpha+2/\beta)^{-1}$, then

$$\rho(\hat{\Sigma}_{p,k_p} - \Sigma_p) = \mathcal{O}_P[(m^{-1/2}p^{2/\beta})^{\alpha c(\alpha)}]. \tag{56}$$

In the tapered estimate (53), if we choose K such that the matrix $W_p = (K(|i-j|/l))_{1\le i,j\le p}$ is positive definite, then $\tilde{\Sigma}_{p,l}$ is the Hadamard (or Schur) product of $\hat{\Sigma}_n$ and W_p, and by the Schur Product Theorem in matrix theory (Horn and Johnson, 1990), it is also non-negative definite since $\hat{\Sigma}_n$ is nonnegative definite. For example, W_n is positive definite for the triangular window $K(u) = \max(0,\, 1-|u|)$, or the Parzen window $K(u) = 1 - 6u^2 + 6|u|^3$ if $|u| < 1/2$ and $K(u) = \max[0,\, 2(1-|u|)^3]$ if $|u| \ge 1/2$.

Based on the Cholesky decomposition (48), Wu and Pourahmadi (2003) proposed a nonparametric estimator for the precision matrix Σ_p^{-1} for locally stationary processes (Dahlhaus, 1997) which are time-varying AR processes

$$X_t = \sum_{j=1}^{k} f_j(t/p)X_{t-j} + \sigma(t/p)\eta_t^0. \tag{57}$$

Here η_t^0 are iid random variables with mean 0 and variance 1, $f_j(\cdot)$ and $\sigma(\cdot)$ are continuous functions. Hence $\phi_{t,t-j} = f_j(t/p)$ if $1 \le j \le k$, and $\phi_{t,t-j} = 0$ if $j > k$. Wu and Pourahmadi (2003) applied a two-step method for estimating $f_j(\cdot)$ and $\sigma(\cdot)$: the first step is that, based on the data $(X_{l,1}, X_{l,2}, \ldots, X_{l,p})$, $l = 1, \ldots, m$, we perform a successive linear regression and obtain the least squares estimate $\hat{\phi}_{t,t-j}$ and the prediction variance $\hat{\sigma}^2(t/p)$; in the second step we do a local linear regression on the raw estimates $\hat{\phi}_{t,t-j}$ and obtain smoothed estimates $\hat{f}_j(\cdot)$. Then we piece those estimates together and obtain an estimate for the precision matrix Σ_p^{-1} by (49). The lag k can be chosen by AIC, BIC or other information criteria. Huang et al (2006) applied a penalized likelihood estimator which is related to LASSO and ridge regression.

4.4 Covariance Matrix Estimation with One Realization

If there is only one realization available, then it is necessary to impose appropriate structural assumptions on the underlying process and otherwise it would not be possible to

estimate its covariance matrix. Here we shall assume that the process is stationary, hence Σ_n is Toeplitz and $\gamma_{i,j} = \gamma_{i-j}$ can be estimated by the sample auto-covariance (3) or (4), depending on whether the mean μ is known or not.

Covariance matrix estimation of stationary processes has been widely studied in the engineering literature. Lifanov and Likharev (1983) performed maximum likelihood estimation with applications in radio engineering. Christensen (2007) applied an EM-algorithm for estimating band-Toeplitz covariance matrices. Other contributions for estimating Toeplitz covariance matrices can be founded in Jansson and Ottersten (2000), Burg, Luenberger and Wenger (1982). See also Chapter 3 in the excellent monograph of Dietrich (2008). However, in most of those papers it is assumed that multiple iid realizations are available.

For a stationary process (X_i), Wu and Pourahmadi (2009) proved that the sample auto-covariance matrix $\hat{\Sigma}_p$ is not a consistent estimate of Σ_p. A refined result is obtained in Xiao and Wu (2011b) and they derived the exact order of $\rho(\hat{\Sigma}_p - \Sigma_p)$.

Theorem 8. *(Xiao and Wu, 2011b). Assume that* $X_i \in \mathcal{L}^\beta$, $\beta > 2$, $EX_i = 0$, $\Delta_\beta(m) = o(1/\log m)$ *and* $\min_\theta f(\theta) > 0$. *Then*

$$\lim_{n\to\infty} P\left[\frac{\pi \min_\theta f^2(\theta)}{12\Delta_2^2}\log p \le \rho(\hat{\Sigma}_p) \le 10\Delta_2^2 \log p\right] = 1. \tag{58}$$

To obtain a consistent estimate of Σ_p, following the idea of lag window spectral density estimation and tapering, we define the tapered covariance matrix estimate

$$\hat{\Sigma}_{p,B} = [K((i-j)/B)\hat{\gamma}_{i-j}]_{1\le i,j\le p} = \hat{\Sigma}_p \star W_p, \tag{59}$$

where $B = B_p$ is the bandwidth satisfying $B_p \to \infty$ and $B_p/p \to 0$, and $K(\cdot)$ is a symmetric kernel function with

$$K(0) = 1, \quad |K(x)| \le 1, \quad \text{and} \quad K(x) = 0 \text{ for } |x| > 1. \tag{60}$$

Estimate (59) has the same form as Bickel and Levina's (52) with the sample covariance matrix replaced by the sample auto-covariance matrix. The form (59) is also considered in McMurry and Politis (2010). Toeplitz (1911) studied the infinite dimensional matrix $\Sigma_\infty = (a_{i-j})_{i,j\in\mathbb{Z}}$ and proved that its eigenvalues coincide with the image set $\{g(\theta) : \theta \in [0, 2\pi)\}$, where

$$g(\theta) = \sum_{j\in\mathbb{Z}} a_j e^{\sqrt{-1}j\theta}. \tag{61}$$

Note that $2\pi g(\theta)$ is the Fourier transform of (a_j). For a finite $p\times p$ matrix $\Sigma_p = (a_{i-j})_{1\le i,j\le p}$, its eigenvalues are approximately equally distributed as $\{g(\theta_j),\, j = 0,\ldots,p-1\}$, where $\theta_j = 2\pi j/p$ are the Fourier frequencies. See the excellent monograph by Grenander and Szegö (1958) for a detailed account. Hence the eigenvalues of the matrix estimate $\hat{\Sigma}_{p,B}$ in (59) are expected to be close to the image set of the lag window estimate

$$\hat{f}_{p,B}(\theta) = \frac{1}{2\pi}\sum_{k=-B}^{B} K(k/B)\hat{\gamma}_k \cos(k\theta). \tag{62}$$

Using an asymptotic theory for lag window spectral density estimates, Xiao and Wu (2011b) derived a convergence rate for $\rho(\hat{\Sigma}_{p,B} - \Sigma_p)$. Recall (15) and (16) for $\Delta_p(m)$ and $\Phi_p(m)$.

Theorem 9. *(Xiao and Wu, 2011b) Assume $X_i \in \mathcal{L}^\beta$, $\beta > 4$, $EX_i = 0$, and $\Delta_p(m) = O(m^{-\alpha})$. Assume $B \to \infty$, and $B = O(p^\gamma)$, where $0 < \gamma < \min(1, \alpha\beta/2)$ and $(1-2\alpha)\gamma < 1 - 4/\beta$. Let $c_\beta = (\beta+4)e^{\beta/4}$. Then*

$$\lim_{n\to\infty} P\left[\rho(\hat{\Sigma}_{p,B} - E\hat{\Sigma}_{p,B}) \le 12c_\beta\Delta_4^2\sqrt{\frac{B\log B}{p}}\right] = 1. \tag{63}$$

In particular, if $K(x) = \mathbf{1}_{\{|x|\le 1\}}$ is the rectangular kernel, and $B \asymp (p/\log p)^{1/(2\alpha+1)}$, then

$$\rho(\hat{\Sigma}_{p,B} - \Sigma_p) = O_P\left[\left(\frac{\log p}{p}\right)^{\frac{\alpha}{2\alpha+1}}\right]. \tag{64}$$

The uniform convergence result in Theorem 2 motivates the following thresholded estimate:

$$\hat{\Sigma}^{\ddagger}_{p,T} = (\hat{\gamma}_{i-j}\mathbf{1}_{|\hat{\gamma}_{i-j}|\ge T})_{1\le i,j\le p}. \tag{65}$$

It is a shrinkage estimator. Note that $\hat{\Sigma}^{\ddagger}_{p,T}$ may not be positive. Bickel and Levina (2008b) considered the above estimate under the assumption that one has multiple iid realizations.

Theorem 10. *(Xiao and Wu, 2011b) Assume $X_i \in \mathcal{L}^\beta$, $\beta > 4$, $EX_i = 0$, $\Delta_p(m) = O(m^{-\alpha})$ and $\Phi_p(m) = O(m^{-\alpha'})$, $\alpha \ge \alpha' > 0$. Let $T = 6c_\beta\|X_0\|_4\Delta_2\sqrt{p^{-1}\log p}$. If $\alpha > 1/2$ or $\alpha'\beta > 2$, then*

$$\rho(\hat{\Sigma}^{\ddagger}_{p,T} - \Sigma_p) = O_P\left[\left(\frac{\log p}{p}\right)^{\frac{\alpha}{2\alpha+2}}\right]. \tag{66}$$

Example 1. Here we shall show how to obtain a BLUE (best linear unbiased estimate) for linear models with dependent errors. Consider the linear regression model (40)

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i, \quad 1 \le i \le p, \tag{67}$$

where now we assume that (e_i) is stationary. If the covariance matrix Σ_p of $(e_1, \ldots, e_p)$ is known, then the BLUE for $\boldsymbol{\beta}$ is of the form

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \Sigma_p^{-1} \mathbf{X})^{-1} \Sigma_p^{-1/2} \boldsymbol{y}, \tag{68}$$

where $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ and $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)^\top$. If Σ_p is unknown, we estimate $\boldsymbol{\beta}$ by a two-step method. Using the ordinary least squares approach, we obtain a preliminary estimate $\bar{\boldsymbol{\beta}}$, and compute the estimated residuals $\hat{e}_i = y_i - \boldsymbol{x}_i^T \bar{\boldsymbol{\beta}}$. Based on the latter, using the tapered estimate $\tilde{\Sigma}_p$ of form (59) for Σ_p, a refined estimate of $\tilde{\boldsymbol{\beta}}$ can be obtained via (68) by using the weighted least squares with the weight matrix $\tilde{\Sigma}_p$. Due to the consistency of $\tilde{\Sigma}_p$, the resulting estimate for $\boldsymbol{\beta}$ is asymptotically BLUE. $\diamondsuit$

ACKNOWLEDGEMENTS

This work was supported in part from DMS-0906073 and DMS-1106970. We thank a reviewer for his/her comments that lead to an improved version.

REFERENCES

Amemiya, T. (1985) *Advanced Econometrics.* Cambridge, Harvard University Press.

Anderson, G. W. and O. Zeitouni (2008) A CLT regularized sample covariance matrices. *Ann. Statistics* **36** 2553–2576.

Anderson, Greg W. and Guionnet, Alice and Zeitouni, Ofer (2010) *An introduction to random matrices*, Cambridge Studies in Advanced Mathematics, **118**, Cambridge University Press, Cambridge.

Anderson, T. W. (1968) Statistical Inference for Covariance Matrices with Linear Structure. In: *Proc. of the Second International Symposium on Multivariate Analysis,* **2,** 55–66

Anderson, T. W. (1970) Estimation of Covariance Matrices which are Linear Combinations or whose Inverses are Linear Combinations of Given Matrices. In: *Essays in in Probability and Statistics,* pp. 1-24. The University of North Carolina Press

Anderson, T. W. (1971). *The Statistical Analysis of Time Series.* Wiley, New York.

Andrews, D. W. K., 1991. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica,* **59**, 817–858.

Andrews, D. W. K. and Monahan, J. C., 1992. An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator, *Econometrica,* **60,** 953–966.

Bai, Z. D. (1999) Methodologies in spectral analysis of large-dimensional random matrices, a review. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author. *Statist. Sinica* **9** 611–677.

Bai, Zhidong; Silverstein, Jack W. (2010) *Spectral analysis of large dimensional random matrices.* Second edition. Springer, New York.

Bahadur, R. R. (1966). A Note on Quantiles in Large Samples. *Annals of Mathematical Statistics* **37**, 577–580.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Suppl. J. Roy. Statist. Soc.* **8**:27–41.

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

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

Bickel P. and Gel, Y. (2011) Banded regularization of covariance matrices in application to parameter estimation and forecasting of time series. *To appear in the Journal of the Royal Statistical Society, Series B.*

Borkar, V. S. (1993). White-noise representations in stochastic realization theory. *SIAM J. Control Optim.* **31**:1093–1102.

Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions.* Kendrick Press, Utah.

Brillinger, D. R. and Krishnaiah, P. R. (eds.) (1983) *Time series in the frequency domain*, Handbook of Statistics, 3, North-Holland Publishing Co., Amsterdam.

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

Burg, J.P., Luenberger, D.G. and Wenger, D.L. (1982) Estimation of structured covariance matrices. *Proceedings of the IEEE* **70** 963–974

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

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

Christensen, Lars P. B. (2007) An EM-algorithm for Band-Toeplitz Covariance Matrix Estimation, In *IEEE International Conference on Acoustics, Speech and Signal Processing* III: 1021–1024

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **36**, 1–37.

de Jong, R. M. and Davidson, J., 2000. Consistency of Kernel Estimators of Heteroscedastic and Autocorrelated Covariance Matrices, *Econometrica,* **68**, 407–423.

Dembo, A. (1986) The Relation Between Maximum Likelihood Estimation of Structured Covariance Matrices and Periodograms. *IEEE Trans. Acoust., Speech, Signal Processing* **34**(6), 1661–1662.

Deng, Xinwei and Ming Yuan (2009). Large Gaussian Covariance Matrix Estimation With Markov Structures *Journal of Computational and Graphical Statistics.* **18** 640–657.

Dietrich, Frank A. (2008) *Robust Signal Processing for Wireless Communications.* Springer, Berlin.

Doukhan, P. (1994). *Mixing: Properties and Examples.* Springer, New York.

Eberlein, E. and Taqqu, M. (ed.) (1986). *Dependence in Probability and Statistics: A Survey of Recent Results.* Birkhauser, Boston.

Eicker F (1963). Asymptotic Normality and Consistency of the Least Squares Estimator for Families of Linear Regressions. *Annals of Mathematical Statistics,* **34,** 447–456.

El Karoui, Noureddine (2007) Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Annals of Probability* **35** 663–714.

El Karoui, Noureddine (2008) Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197.

Fuhrmann, D.R., Miller, M.I. (1988) On the Existence of Positive-Definite Maximum- Likelihood Estimates of Structured Covariance Matrices. *IEEE Trans. Inform. Theory* **34**(4), 722–729

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

Grenander, U. and Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series.* New York: Wiley.

Grenander, Ulf and Szegö, Gabor (1958) *Toeplitz forms and their applications*, University of California Press, Berkeley, CA.

Hall, P and C.C. Heyde (1980). *Martingale Limit Theorem and its Application.* Academic Press, New York.

Hall, P. and Yao, Q. W. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* **71**, 285–317.

Hannan, E. J. (1970). *Multiple Time Series.* Wiley, New York.

Hannan, E. J. (1976). The asymptotic distribution of serial covariances. *Ann. Statist.* **4**: 396–399.

Hansen, B. E., 1992. Consistent Covariance Matrix Estimation for Dependent Heterogeneous Processes, *Econometrica,* **60**, 967–972.

Harris, D., McCabe, B. and Leybourne, S. (2003). Some limit theory for autocovariances whose order depends on sample size. *Econometric Theory* **19** 829–864.

He, X. and Shao, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630.

Heagerty, Patrick J.; Lumley, Thomas (2000) Window subsampling of estimating functions with application to regression models. *J. Amer. Statist. Assoc.* **95** 197–211.

Heyde, C. C. (1997) *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*, Springer, New York.

Horn, Roger A. and Johnson, Charles R. (1990) *Matrix analysis.* Corrected reprint of the 1985 original. Cambridge University Press, Cambridge, UK.

Hosking, J. R. M. (1996). Asymptotic distributions of the samplemean, autocovariances, and autocorrelations of long-memory timeseries. *J. Econom.* **73**:261–284.

Huang, Jianhua Z.; Liu, Naiping; Pourahmadi, Mohsen; Liu, Linxu (2006) Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98.

Ibragimov, I. A. and Linnik, Yu. V. (1971). *Independent and stationary sequences of random variables.* Groningen, Wolters-Noordhoff.

Jansson, M. and B. Ottersten (2000) Structured covariance matrix estimation: a parametric approach, *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, **5** 3172–3175.

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

Kalikow, S. A. (1982). T , T^{-1} transformation is not loosely Bernoulli. *Ann. Math.* **115**:393–409.

Kallianpur, G. (1981). Some ramifications of Wieners ideas on nonlinear prediction. In: *Norbert Wiener, Collected Works with Commentaries.* MIT Press, Mass., 402–424.

Keenan, D. M. (1997). A central limit theorem for m(n) autocovariances. *J. Time Ser. Anal.* **18** 61–78.

Klimko, L. A. and Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes. *Annals of Statistics* **6**, 629–642.

Lam, Clifford and Jianqing Fan (2009) Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37**, 4254–4278

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis,* **88**, 365–411.

Lifanov, E. I. and Likharev, V. A. (1983) Estimation of the covariance matrix of stationary noise. *Radiotekhnika*, **5** 53–55.

Liu, W. and Wei Biao Wu (2010) Asymptotics of Spectral Density Estimates, *Econometric Theory,* **26**: 1218–1245

Lumley, T. and Heagerty, P. (1999). Empirical Adaptive Variance Estimators for Correlated Data Regression. *Journal of the Royal Statistical Society B,* **61**, 459–477.

MacKinnon JG, White H (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics*, **29**, 305–325.

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

McMurry, Timothy L. and Dimitris N. Politis. (2010) Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *J. Time Series Anal.,* **31** 471–482

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

Newey, W. K. and West, K. D., 1987. A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica,* **55**, 703–708.

Newey, W. K. and West, K. D., 1994. Automatic Lag Selection in Covariance Matrix Estimation, *Review of Economic Studies*, **61**, 631–653.

Ornstein, D. S. (1973). An example of a Kolmogorov automorphism that is not a Bernoulli shift. *Advances in Math* **10**:49–62.

Pan, J. and MacKenzie, G. (2003). On modelling mean-covariance structure in longitudinal studies. *Biometrika*, **90**, 239-244.

Paul, D. (2007). Asymptotics of the leading sample eigenvalues for a spiked covariance model. *Statist. Sinica*. **17** 1617–1642

Phillips, P. C. B. and Solo, V. (1992). Asymptotics for linear processes. *Ann. Statist.* **20** 971–1001.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**(3):677–690.

Pourahmadi, Mohsen (2001) *Foundations of time series analysis and prediction theory.* Wiley, New York.

Pourahmadi, Mohsen (2011) Modeling Covariance Matrices: The GLM and Regularization Perspectives. *Statistical Science, To Appear.*

Priestley, M. B. (1981). *Spectral Analysis and Time Series 1.* Academic Press. MR0628735

Priestley, M. B. (1988). *Nonlinear and Nonstationary Time Series Analysis.* Academic Press.

Quang, A.N. (1984) On the Uniqueness of the Maximum-Likeliwood Estimate of Structured Covariance Matrices. *IEEE Trans. Acoust., Speech, Signal Processing* **32**(6), 1249–1251.

Rosenblatt, M. (1985). *Stationary Sequences and Random Fields.* Birkhäuser, Boston.

Rosenblatt, M. (2009) A comment on a conjecture of N. Wiener. *Statist. Probab. Letters,* **79**, 347–348

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc. (Theory and Methods),* **104**, 177–186.

Rothman, A.J., Levina, E. and Zhu, J. (2010). A new approach to Cholesky-based estimation of high-dimensional covariance matrices. *Biometrika*, **97**, 539–550.

Stein, C. (1975). Estimation of a covariance matrix. In Reitz lecture, Atlanta, Georgia, 1975. 39th annual meeting IMS.

Toeplitz, O. (1911). Zur theorie der quadratischen und bilinearan Formen von unendlichvielen, Veranderlichen. *Math. Ann.* **70**:351–376

Tong, H. (1990) *Non-linear Time Series: A Dynamic System Approach,* Oxford University Press, Oxford.

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

White, Halbert (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838.

Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series.* Wiley, New York.

Wiener, N. (1958) *Nonlinear Problems in Random Theory.* MIT Press, MA.

Wu, W. B. (2005) Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences USA.* **102** (40), 14150–14154.

Wu, W. B. (2007): M-estimation of linear models with dependent errors. *Ann. Stat.* **35**, 495-521.

Wu, W. B. (2009) An asymptotic theory for sample covariances of Bernoulli shifts. *Stochastic Process. Appl.* **119** 453–467.

Wu, W. B. (2011) Asymptotic theory for stationary processes. *Statistics and Its Interface*, To appear.

Wu, W. B., Huang, Yinxiao and Zheng, Wei (2010). Covariances estimation for long-memory processes. *Adv. in Appl. Probab.* **42**(1): 137–157

Wu, W. B. and Min, Wanli (2005). On linear processes with dependent innovations. *Stochastic Process. Appl.* **115**(6): 939–959

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

Wu, W. B.; Pourahmadi, Mohsen (2009) Banding sample autocovariance matrices of stationary processes. *Statist. Sinica* **19** 1755–1768

Xiao, Han; Wu, Wei Biao (2011a). Asymptotic inference of autocovariances of stationary processes. *preprint*, available at `http://arxiv.org/abs/1105.3423`

Xiao, Han; Wu, Wei Biao (2011b). Covariance matrix estimation for stationary time series. *preprint*, available at `http://arxiv.org/abs/1105.4563`

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19-35.

Zeileis, Achim. (2004) Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, **11**(10):117, URL `http://www.jstatsoft.org/v11/i10/`.

Zimmerman, D.L. and Núñez-Antón, V. (2010). *Antedependence Models for Longitudinal Data*, CRC Press, New York.