Available online at www.sciencedirect.com
ScienceDirect
Journal of Multivariate Analysis 99 (2008) 949–967
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

Sample covariance shrinkage for high dimensional dependent data

Alessio Sancetta*

Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DE, UK

Received 3 April 2006
Available online 16 June 2007

Abstract

For high dimensional data sets the sample covariance matrix is usually unbiased but noisy if the sample is not large enough. Shrinking the sample covariance towards a constrained, low dimensional estimator can be used to mitigate the sample variability. By doing so, we introduce bias, but reduce variance. In this paper, we give details on feasible optimal shrinkage allowing for time series dependent observations.

AMS 2000 subject classification: 62H12; 62C12

Keywords: Sample covariance matrix; Shrinkage; Weak dependence

1. Introduction

This paper considers the problem of estimating the variance covariance matrix of high dimensional data sets when the sample size is relatively small and the data exhibit time series dependence. The importance of estimating the covariance matrix in these situations is obvious. The number of applied problems where such an estimate is required is large, e.g. mean–variance portfolio optimization for a large number of assets, generalized method of moments estimation when the number of moment equations is large, etc. However, the estimator based on the sample covariance can be noisy, it can be difficult to find its approximate inverse, hence might perform poorly.

To mitigate this problem, the sample covariance matrix can be shrunk towards a low dimensional constrained covariance matrix. Recently, this approach has been successfully studied by Ledoit

* Fax: +44 1223 335299.
E-mail address: alessio.sancetta@econ.cam.ac.uk.

0047-259X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2007.06.004

and Wolf [6]. These authors assume iid observations, a certain cross-dependence structure for the vector of observations and shrink towards a matrix proportional to the identity. Related references can be found in their work. The idea is to find an optimal convex combination of the sample covariance and the constrained covariance matrix. The parameter defining shrinkage depends on unknown quantities and needs to be estimated consistently. Intuitively, the problem is the usual one of balancing the bias and the variance of the estimator to obtain lower mean square error.

The goal of the present paper is to show that covariance matrix shrinkage can be used in quite general situations, when data are time dependent and are not restricted in their cross-dependence structure. To account for time dependence, the estimator based on iid observations has to be slightly changed. However, an interesting property of the estimator is that accounting for time series dependence is not always crucial. We will make this statement more precise in our simulation study. Extending the theory to more general situations is important when dealing with real data.

The results derived here are weak, as they only hold in probability versus L_2 consistency of Ledoit and Wolf [6]. However, we show that the constrained covariance matrix does not need to be proportional to the identity and can be chosen more generally. In Ledoit and Wolf [5] a constrained covariance matrix based on a one factor model was suggested, assuming that the cross-sectional dimension stays fixed. Our framework covers the case when time and cross-sectional dimensions may grow at the same rate. However, this requires that the constrained covariance matrix is chosen appropriately. In this respect, while the results of this paper cover a lot of cases of interest not covered by Ledoit and Wolf [6], the two papers are still complementary. Details will be given in due course.

The plan of the paper is as follows. Section 2 states the problem and the suggested solution. Section 3 contains a Monte Carlo study of the small sample performance of shrinkage when data series are dependent. Section 4 proves that the procedure is consistent.

We introduce some notation. Given two sequences $a := a_n$ and $b := b_n$, $a \lesssim b$, means that there is a finite absolute constant c such that $a \leqslant cb$; $a \asymp b$ means that $a \lesssim b$ and $b \lesssim a$. We may also use the O and o notation as complement and substitute of the above symbols to describe orders of magnitude, which ever is felt more appropriate. Given two numbers a and b, the symbols $a \vee b$ and $a \wedge b$ mean, respectively, the maximum and minimum between a and b. If A is a countable set, $\#A$ stands for its cardinality. Finally, for a matrix A, A_{ij} stands for the (i, j)th entry.

2. Estimation of the unconditional covariance matrix

Suppose $(Y_t)_{t\in\{1,\dots,T\}}$ are random variables with values in $\mathbb{R}^N$. For simplicity assume the variables are mean zero. The covariance matrix is defined as $\Sigma := T^{-1}\sum_{t=1}^{T}\mathbb{E}Y_tY_t'$ and under second-order stationarity this reduces to $\Sigma := \mathbb{E}Y_tY_t'$. Then, $\hat{\Sigma}_T = \sum_{t=1}^{T} Y_tY_t'/T$ is a sample estimator for Σ. In some cases, we may have that N grows with T. If $N/T \to 0$ the sample covariance matrix $\hat{\Sigma}_T$ is consistent (under an appropriate metric), but the rate of convergence can be arbitrarily slow, moreover, $\hat{\Sigma}_T$ might be singular in finite samples. If $N \asymp T$, $\hat{\Sigma}_T$ is also inconsistent. This paper considers the case where $N/T \to c \in [0, \infty]$, so that we might even have $T = o(N)$.

To be more precise, we define the Frobenious norm in order to measure the distance between matrices.

Definition 1. Suppose A is a square N-dimensional matrix. The Frobenious norm is defined as $\|A\|_2 := \sqrt{\text{Trace}\,(AA')}$.

Remark 1. Note that $\|A\|_2^2 := \sum_{i=1}^N \sum_{j=1}^N A_{ij}^2$. Moreover, when A is symmetric, $\|A\|_2^2 = \sum_{i=1}^N \lambda_{Ai}^2$, where $\lambda_{A1}^2, \ldots, \lambda_{AN}^2$ are the eigenvalues of A. Ledoit and Wolf [6] suggest standardization by N, so that the Frobenious norm of the identity matrix is always one independently of the dimension. This will not be done here.

To mitigate the problem that $\left\|\hat{\Sigma}_T - \Sigma\right\|_2$ is large when N is relatively large, it is suggested that we use a shrunk estimator $\tilde{\Sigma}_T(\alpha) = \alpha F + (1-\alpha)\hat{\Sigma}_T$, where $\alpha \in [0,1]$ and F is a constrained version of Σ. In general, F is chosen to impose stringent restrictions on the unconditional covariance matrix so that $F \neq \Sigma$. Note that F is usually unknown and need to be replaced by an estimator. However, in Theorem 1, we will show that asymptotically, this does not affect the argument if the estimator is low dimensional. On the other hand, $\hat{\Sigma}_T$ is unbiased for Σ, but very noisy, especially in finite sample, where we may even have $N > T$. The shrunk estimator $\tilde{\Sigma}_T$ is preferred to $\hat{\Sigma}_T$ if there exists an $\alpha \in (0,1]$ such that $\mathbb{E}\left\|\tilde{\Sigma}_T(\alpha) - \Sigma\right\|_2^2 < \mathbb{E}\left\|\hat{\Sigma}_T - \Sigma\right\|_2^2$. As done in Ledoit and Wolf [6], we consider the expected squared Frobenious norm and minimize it with respect to α.

Proposition 1. *Suppose $\tilde{\Sigma}_T(\alpha) = \alpha F + (1-\alpha)\hat{\Sigma}_T$. The optimal choice of α under the expected squared Frobenious norm is*

$$\alpha_0 = \frac{\mathbb{E}\left\|\hat{\Sigma}_T - \Sigma\right\|_2^2}{\mathbb{E}\left\|F - \hat{\Sigma}_T\right\|_2^2} \wedge 1 = \arg\min_{\alpha\in[0,1]} \mathbb{E}\left\|\tilde{\Sigma}_T - \Sigma\right\|_2^2, \tag{1}$$

where

$$\mathbb{E}\left\|\hat{\Sigma}_T - \Sigma\right\|_2^2 = \sum_{1\leqslant i,j\leqslant N} Var\left(\hat{\Sigma}_{ijT}\right),$$

and all relevant moments are assumed to exist.

The solution shows that we might reduce the error under the Frobenious norm even if $\tilde{\Sigma}_T$ is biased (recall that $F \neq \Sigma$) because we reduce its variance. This is the usual bias–variance trade-off in the mean square error of the estimator. Unfortunately, $\tilde{\Sigma}_T(\alpha_0)$ is based on unknown quantities, but Σ can be replaced by its unbiased estimator $\hat{\Sigma}_T$, and F by an unbiased estimator, say $\hat{F}_T$. Clearly, $\hat{F}_T$ should have low variance in order for the procedure to work well in practice. The choice of shrinkage parameter changes if we replace in (1) the unfeasible estimator $\tilde{\Sigma}_T(\alpha)$ with $\alpha\hat{F}_T + (1-\alpha)\hat{\Sigma}_T$. In particular, from the proof of Proposition 1, deduce that

$$\begin{aligned}\alpha_0' &:= \frac{\mathbb{E}\left\|\hat{\Sigma}_T - \Sigma\right\|_2^2 - \sum_{1\leqslant i,j\leqslant N} Cov\left(\hat{\Sigma}_{ijT}, \hat{F}_{ij}\right)}{\mathbb{E}\left\|\hat{F} - \hat{\Sigma}_T\right\|_2^2} \wedge 1 \\ &= \arg\min_{\alpha\in[0,1]} \mathbb{E}\left\|\alpha\hat{F}_T + (1-\alpha)\hat{\Sigma}_T - \Sigma\right\|_2^2, \end{aligned} \tag{2}$$

under regularity conditions. We shall show that under suitable conditions on $\hat{F}_T$, α_0' and α_0 are asymptotically equivalent. For this reason, we will just consider α_0 as the quantity to estimate. This is reassuring because estimation of $Cov\left(\hat{\Sigma}_{ijT}, \hat{F}_{ij}\right)$ could be a nontrivial exercise. Moreover, note that

$$\mathbb{E}\left\|\hat{\Sigma}_T - \Sigma\right\|_2^2 = \sum_{1 \leqslant i,j \leqslant N} Var\left(\frac{1}{T}\sum_{t=1}^{T} Y_{ti}Y_{tj}\right) \tag{3}$$

so that when the observations are dependent, we should estimate the covariance terms $Cov\left(Y_{ti}Y_{tj}, Y_{si}Y_{sj}\right)$. Define

$$\begin{aligned}
\hat{\Sigma}_{T,ij} &:= \frac{1}{T}\sum_{t=1}^{T}\left(Y_{ti}Y_{tj}\right), \\
\hat{\Gamma}_{T,ij}(s) &:= \frac{1}{T}\sum_{t=1}^{T-s}\left(Y_{t,i}Y_{t,j} - \hat{\Sigma}_{T,ij}\right)\left(Y_{t+s,i}Y_{t+s,j} - \hat{\Sigma}_{T,ij}\right), \\
\hat{\Gamma}^b_{T,ij} &:= \hat{\Gamma}_{T,ij}(0) + 2\sum_{s=1}^{T-1}\kappa(s/b)\,\hat{\Gamma}_{T,ij}(s), \quad b > 0,
\end{aligned} \tag{4}$$

where $\kappa(s)$ is some function decreasing to zero and continuous in the neighborhood of zero and $b > 0$ is a smoothing parameter. With the above notation, under stationarity conditions, an estimator of (3) is given by $T^{-1}\sum_{1 \leqslant i,j \leqslant N}\hat{\Gamma}^b_{T,ij}$. Then we define the sample estimator

$$\hat{\alpha}_T := \frac{T^{-1}\sum_{1 \leqslant i,j \leqslant N}\hat{\Gamma}^b_{T,ij}}{\left\|\hat{\Sigma}_T - \hat{F}_T\right\|_2^2}$$

and will show that it is asymptotically equivalent to use either $\hat{\alpha}_T$ or α_0, where α_0 is as in (1). To make this statement formal, we require some conditions. Comments about the following technical conditions are deferred to the next subsection.

Condition 1. (1) $\mathbb{E}\left|\hat{F}_{T,ij} - F_{ij}\right|^2 = O\left(T^{-1}\right)$ $(\forall i, j)$ where $F = \mathbb{E}\hat{F}_T$;
(2) $\#\left\{1 \leqslant i, j \leqslant N : F_{ij} \neq \hat{F}_{T,ij}\right\} \lesssim N^{\beta}$, $\beta \in [0, 2)$;
(3) $\|F - \Sigma\|_2^2 \asymp N^{\gamma}$, $\gamma > 0$;
(4) $\epsilon = \epsilon_T := N/T$ such that ϵ, β and γ satisfy

$$\left(\epsilon^{1/2}N^{\beta - 1/2} \vee \epsilon^{1/2}N^{(1+\gamma)/2}\right) = o\left(N^{\gamma}\right).$$

Condition 2. Suppose $u, v \in \{1, 2, 3, 4\}$. Consider the u and v tuples $(i_1, \ldots, i_u)$, $(s_1, \ldots, s_u) \in \mathbb{N}^u$ and $(j_1, \ldots, j_v)$, $(t_1, \ldots, t_v) \in \mathbb{N}^v$ such that $s_1 \leqslant \cdots \leqslant s_u < s_u + r \leqslant t_1 \leqslant \cdots \leqslant t_v$ for some $r \in \mathbb{N}$. Then, there exists a sequence $(\theta_r)_{r \in \mathbb{N}}$, where $\theta_r \lesssim r^{-a}$ with $a > 3$ such that

$$\left|Cov\left(Y_{t_1 j_v} \cdots Y_{t_u j_v}, Y_{s_1 i_1}, \ldots, Y_{s_u i_u}\right)\right| \leqslant \theta_r.$$

Condition 3. In (4) above,

(1) $\kappa : \mathbb{R} \to \mathbb{R}$ is a decreasing positive function, continuous from the right with left-hand limits such that $\lim_{s\to 0^+} \kappa(s) = 1$ and $\int_0^\infty [\kappa(s)]^2 \, ds < \infty$.

(2) $b \to \infty$ such that $b = o\left(T^{1/2}\right)$.

Condition 4. $(Y_t)_{t\in\mathbb{N}}$ is an N-dimensional vector of nondegenerate random variables with finite and stationary eighth moment.

Hence, we have consistency of the estimated shrinkage parameter and the feasible shrinkage estimator.

Theorem 1. *Under Conditions* 1–4,

(1)

$$\left[\epsilon N^{1-\gamma}\right]^{-1} \left(\hat{\alpha}_T - \alpha_0\right) = o_p(1),$$

where $\alpha_0 \asymp \epsilon N^{1-\gamma} = o(1)$ *is as in* (1);

(2)

$$\left[\epsilon N^{1-\gamma}\right]^{-1} \left(\alpha_0' - \alpha_0\right) = o(1),$$

where $\alpha_0' \asymp \epsilon N^{1-\gamma} = o(1)$ *is as in* (2);

(3)

$$\left\|\hat{\alpha}_T \hat{F}_T + \left(1 - \hat{\alpha}_T\right) \hat{\Sigma}_T - \Sigma\right\|_2 = \left\|\alpha_0 F + (1 - \alpha_0) \hat{\Sigma}_T - \Sigma\right\|_2 \left\{1 + o_p\left(\epsilon^{1/2} N^{(1-\gamma)/2}\right)\right\}.$$

Below, we provide comments about Theorem 1, and in the subsequent subsection, we remark on the technical conditions of the paper.

2.1. Remarks on Theorem 1

Theorem 1 gives a rate of convergence in probability uniformly in

$$\left\|\alpha_0 F + (1 - \alpha_0) \hat{\Sigma}_T - \Sigma\right\|_2,$$

where $\epsilon^{1/2} N^{(1-\gamma)/2} \to 0$ by Condition 1(4). Note that $\epsilon \to 0$ is not required, but it is allowed. We may also have $\epsilon \to \infty$ as long as Condition 1 is satisfied (remarks about this condition can be found in the next subsection). Note that Theorem 1 is not concerned with consistency of $\hat{\Sigma}_T$, but only assures that with high probability, the shrunk estimator $\hat{\alpha}_T \hat{F}_T + \left(1 - \hat{\alpha}_T\right) \hat{\Sigma}_T$ will perform better than $\hat{\Sigma}_T$ under the Frobenious norm.

Moreover, if F is full rank, then $\alpha F + (1 - \alpha) \hat{\Sigma}_T$ is invertible when $\alpha > 0$ even though $\hat{\Sigma}_T$ is rank deficient. This intuition can be made formal in the special case $F = vI_N$, where I_N is the identity matrix and v a positive constant. Then,

$$\det\left(\alpha v I_N + (1 - \alpha) \hat{\Sigma}_T - \lambda I_N\right) = (1 - \alpha)^N \det\left(\hat{\Sigma}_T - \frac{(\lambda - \alpha v)}{(1 - \alpha)} I_N\right)$$

and $\hat{\Sigma}_T$ has arbitrary eigenvalue $\omega := (\lambda - \alpha v)/(1-\alpha) \geqslant 0$ (because $\hat{\Sigma}_T$ is positive semidefinite), implying $\lambda = (1-\alpha)\omega + \alpha v > 0$ (which is the corresponding eigenvalue of the shrunk estimator). Therefore, the minimum eigenvalue of this shrunk estimator is always larger than the one of the sample covariance matrix. We now turn to specific comments regarding the conditions of the paper.

2.2. Remarks on the technical conditions

2.2.1. Condition 1

Theorem 1(3) holds even if in Condition 1(1) we replace L_2 convergence with $O_p\left(T^{-1/2}\right)$ convergence and we allow for $\beta = 2$ in Condition 1(2). However, part (2) in Theorem 1 requires the present slightly stronger conditions. In practice we might only be interested in knowing that the shrunk estimator performs asymptotically as well as the unfeasible optimal $\alpha F + (1-\alpha)\hat{\Sigma}_T$, in which case milder conditions can be used.

As just mentioned, part (1) in Condition 1 implies that $\hat{F}_{T,ij}$ is L_2 root-n consistent. Part (2) says that $\hat{F}_T$ and F are constrained so that there are at most $O\left(N^{\beta}\right)$ elements to be estimated in F, with the other elements fixed and known. The simplest way to achieve this is by setting at most $O\left(N^{\beta}\right)$ elements to be nonzero in F and estimate them using $\hat{F}_T$. We give some examples.

Example 1. Suppose $F := vI_N$ where $v = \sum_{i=1}^N \Sigma_{ii}/N$. Then, $\hat{F}_T = \hat{v} I_N$ where $\hat{v} = \sum_{i=1}^N \hat{\Sigma}_{iiT}/N$. In this case, $\tilde{\Sigma}_T$ shrinks all the off diagonal elements of $\hat{\Sigma}_T$ towards zero and the diagonal towards the mean of its diagonal elements, in both cases by a factor $(1-\alpha)$. In this case, we need $\left(\epsilon^{1/2}N^{1/2} \vee \epsilon N\right) = o\left(N^{\gamma}\right)$. This is the estimator used in Ledoit and Wolf [6], but with different restrictions on Σ.

Example 2. Suppose the data can be divided in groups and we constraint the correlation between groups to be zero. Controlling for the number of groups and elements in each group would allow us to satisfy Condition 1. Many examples, also based on factor models, can be generated once we restrict between groups correlations to be zero. Details can be left to the interested reader.

Part (3) implies that $F \neq \Sigma$, and γ quantifies how different F and Σ are under the Frobenious norm. Part (4) is the crucial condition of the paper and relates the coefficients β and γ together with the ratio $\epsilon_T := N/T$. A simple example shows that these conditions do not define an empty set.

Example 3. Suppose F and $\hat{F}_T$ are diagonal, then $\beta = 1$. Suppose $\gamma > 1$, which is surely satisfied if, for example, $\|\Sigma\|_2^2 \asymp N^{\gamma}$. Then, Condition 1(4) is satisfied for $\epsilon_T \to c > 0$ and we may even have $\epsilon_T \to \infty$ at, e.g. a logarithmic rate.

It is interesting to note that if $\epsilon_T \to c > 0$, the result of the paper does not cover the case $\beta = 1$ (i.e. F and $\hat{F}_T$ are diagonal) and Σ is diagonal as well. In this case, we do require $\epsilon_T \to 0$. In practice, we would often use shrinkage for a matrix Σ such that Σ and F are different (i.e. $\gamma > 1$) because the number of entries to be estimated in F is relatively small (e.g. $\beta < \frac{3}{2}$). In this case $\epsilon_T \to c > 0$ is allowed.

It is useful to compare with the results in Ledoit and Wolf [6] and in particular with their Assumption 2. We note that a necessary condition for Assumption 2 in Ledoit and Wolf is $\|\Sigma\|_2^2 =$

$O(N)$. We will show this below. Based on another restrictive assumption on the higher order cross-dependence structure, Ledoit and Wolf show that using F proportional to the identity allows for successful shrinkage. Theorem 1 does not cover this case, though it is quite restrictive, as this would imply $\beta = \gamma = 1$ and $\epsilon_T \to c > 0$, as remarked before. In this case, we require $\epsilon_T \to 0$. This is the price one has to pay for lifting the iid condition and restrictive conditions on the higher order cross-dependence structure of the data (Assumption 3 in Ledoit and Wolf). Clearly, the results in Ledoit and Wolf do not allow for, say $\|\Sigma\|_2^2 \asymp N^2$, which is covered by this paper. Hence, the present result and the one in Ledoit and Wolf are somehow complementary. We remark that what makes the approach of Ledoit and Wolf work is that under their conditions, they can show that $(1 - \mathbb{E})\left\|\hat{\Sigma}_T\right\|_2^2 = o_p(N)$ (they actually show it in L_2) while $\left\|\hat{\Sigma}_T - \Sigma\right\|_2^2 = O_p(N)$. This cannot be done under the present, more general conditions, and a different route had to be used.

To see that Assumption 2 in Ledoit and Wolf implies $\|\Sigma\|_2^2 = O(N)$, write $\Sigma = P\Lambda P'$ where Λ is the matrix of eigenvalues and P is the matrix of orthonormal eigenvectors. Define $X_t := P'Y_t$. Assumption 2 in Ledoit and Wolf says that $\sum_{i=1}^N \mathbb{E}X_{ti}^8 = O(N)$ (using our notation). By Jensen inequality, this implies

$$O(N) = \sum_{i=1}^N \mathbb{E}X_{ti}^8 \geqslant \sum_{i=1}^N \left(\mathbb{E}X_{ti}^2\right)^4 = \sum_{i=1}^N \Lambda_{ii}^4 = \text{Trace}\left(\Lambda^4\right). \tag{5}$$

By the properties of L_p norms, $\mathbb{E}|Z| \leqslant \left(\mathbb{E}|Z|^2\right)^{1/2}$ for any random variable Z. Hence, setting $Z = \Lambda_{ii}^2$ and taking expectation with respect to i using the measure with mass $1/N$ at each $i = 1, \ldots, N$, deduce

$$\left(\frac{\text{Trace}\left(\Lambda^2\right)}{N}\right) = \left(N^{-1}\sum_{i=1}^N |\Lambda_{ii}|^2\right) \leqslant \left(N^{-1}\sum_{i=1}^N |\Lambda_{ii}|^4\right)^{1/2} = \left(\frac{\text{Trace}\left(\Lambda^4\right)}{N}\right)^{1/2}. \tag{6}$$

Squaring (6) and multiplying it by N, (5) together with Remark 1 give

$$O(N) = \text{Trace}\left(\Lambda^4\right) \geqslant N^{-1}\left[\text{Trace}\left(\Lambda^2\right)\right]^2 = N^{-1}\left(\|\Sigma\|_2^2\right)^2,$$

so that

$$\|\Sigma\|_2^2 = O(N).$$

Note that Assumption 3 in Ledoit and Wolf also imposes a further restriction on the cross-sectional dependence of the data (not used here), which is satisfied by a restricted class of random variables like Gaussian random variables.

2.2.2. Condition 2

Condition 2 can be verified by deriving the weak dependence coefficients of Doukhan and Louhichi [4]. Condition 2 is satisfied by a wide range of time series models. It is weaker and much easier to derive than strong mixing, and Doukhan and Louhichi [4] give important examples of processes satisfying conditions of this type. Ledoit and Wolf [6] assume independence.

2.2.3. Condition 3

Condition 3 is standard for the estimation of the spectral density at frequency zero. Note that there are other alternatives for the estimation of the variance of the sample mean of dependent observations: block bootstrap, sieve bootstrap, subsampling, etc. (see [3,7] for reviews). Clearly any of these other approaches could be used as an estimator of (3) in place of (4).

2.2.4. Condition 4

From the proofs it is evident that stationarity is mainly used to simplify the notation in the definition of $\hat{\Gamma}_{T,ij}(s)$. Under suitable conditions, we could allow $(Y_t)_{t\in\mathbb{N}}$ to be heterogeneous and interpret Σ to be the arithmetic average of $\left(\mathbb{E}Y_tY_t'\right)_{t\in\{1,\dots,T\}}$ and similarly for other quantities that will be defined in the next section. Details can be left to the interested reader.

3. Simulation study

Ledoit and Wolf [6] carry out a simulation study to verify the small sample properties of their estimator. Theorem 1 says that we need to account for time series dependence. However, it is interesting to see what is the effect of dependence in practice. In the simulation examples we carry out below we can see that there is no substantial gain unless there is some moderate time series dependence across all the (i, j) terms. Here is an explanation for this. To keep it simple, suppose that in Condition 1 $\gamma = 1$ and $\epsilon = 1$. Then, by Lemma 2

$$\alpha_0 = \frac{T^{-1}\sum_{1\leqslant i,j\leqslant N} Var\left(T^{-1/2}\sum_{t=1}^{T} Y_{ti}Y_{tj}\right)}{\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2} \asymp (NT)^{-1}\sum_{1\leqslant i,j\leqslant N} Var\left(T^{-1/2}\sum_{t=1}^{T} Y_{ti}Y_{tj}\right),$$

which is the average over the variances of the (i, j) sample covariances. Hence, if

$$\sum_{t=2}^{T-1} Cov\left(Y_{1i}Y_{1j}, Y_{t,i}Y_{t,j}\right) \simeq 0$$

for many i and j's, then, averaging over (i, j) would considerably decrease the impact of dependence on the estimator. Hence, optimal shrinkage can be thought to be somehow robust to departures from independence, especially in the positive dependent case. In fact, (2) implies that if $\hat{\Sigma}_T$ and $\hat{F}_T$ are positively correlated, α_T is upward biased for (2) in finite samples, and not accounting for positive dependence might counterbalance this bias.

The Monte Carlo study is carried out as follows. Simulate several sequences of vectors and compute their covariance using the covariance shrinkage proposed here. In particular we choose the constrained estimator to be as in Example 1. This way, results can be compared with the shrunk estimator used for iid observations and proposed by Ledoit and Wolf [6]. We want to verify if in practice we should worry too much about weak dependence. For all the simulated data, we compute $\mathbb{E}\left\|\hat{\Sigma}_T^*\left(\hat{\alpha}_T\right)-\Sigma\right\|_2^2$ where $\hat{\Sigma}_T^*(\alpha) := \alpha\hat{F}_T + (1-\alpha)\hat{\Sigma}_T$, and as usual Σ is the true covariance matrix. We also compute the percentage relative improvement in average loss (PRIAL), i.e.

$$PRIAL\left(\Sigma_T^*\left(\hat{\alpha}_T\right)\right) = 100\frac{\mathbb{E}\left\|\hat{\Sigma}_T^*\left(\hat{\alpha}_T\right)-\hat{\Sigma}_T\right\|_2^2 - \mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}.$$

The expectations are computed (approximated) using 1000 Monte Carlo replications, and standard errors are also computed.

The smoothing function in Condition 3 is chosen to be $\kappa(s) = (1 - |s|) I_{\{|s| \leqslant 1\}}$, which is the Bartlett kernel. For $b = 1$ it corresponds to the case when no time series dependence is accounted for, as the covariance terms all drop. We consider the cases $b = 1, 5$, so that in the second case, fourth-order autocovariance terms are retained in the estimation of $\hat{\Gamma}^b_{T,ij}$. When $b = 1$, we just recover the exact estimator considered in Ledoit and Wolf [6]. For comparison reasons, we also compute $\hat{\Sigma}^*_T(\alpha)$ for $\alpha = 0, 1$, i.e. $\hat{\Sigma}_T$ and $\hat{F}_T$.

Details on the simulated data are as follows. The sample is $T = 40$ from an $N = 20$ dimensional vector autoregressive process (VAR) of order one. The matrix of autoregressive coefficients is diagonal with diagonal entries in [0, .8], and [.5, .8] in a second simulation example. These coefficients were obtained by simulating an N-dimensional vector of [0, .8] and [.5, .8] uniform random variables. The innovations are iid Gaussian vectors with diagonal covariance matrix, whose coefficients were simulated from a lognormal with mean one and scaling parameter $\sigma = .25, .5, 1, 2$ (σ is the standard deviation of the logs of the observations). Different values of σ allows us to assess changes in performance as the diagonal entries becomes less concentrated around their mean equal to one. As σ increases, $\hat{F}_T$ becomes noisier and more biased for Σ so that shrinkage is less justified. Results show that when the scaling parameter σ is equal to 1, the PRIAL is quite small becoming negative when $\sigma = 2$. We also report the same results when $\sigma = 1$ but $N = 40, 80, 160$ to see if there is a relative improvement, which indeed happens to be substantial, despite the increased variability and bias in $\hat{F}_T$. Simulations carried out by the author, but not reported here, show that a similar improvement is obtained when $\sigma = 2$, leading to positive PRIAL as soon as $N \geqslant 40$ (i.e. $N/T \geqslant 1$). A larger N implies that $\hat{\Sigma}_T$ is noisier, and we can argue more strongly for shrinkage despite the bias in $\hat{F}_T$. Finally, in a third simulation, we briefly consider the case of nondiagonal true covariance matrix. To this end we use the same VAR model with autoregressive coefficients in [0, .8] and [.5, .8], but we set the covariance matrix of the innovations to be one along the diagonal and .25 off the diagonal. In this case, we only consider $N = 20, 40$. The results seem to be representative of the behavior of the covariance shrinkage estimator in the presence of exponentially decaying time series dependence. Note that for a VAR(1), Condition 2 is satisfied for any $a > 0$. All the results are reported in Table 1, Panels A, B, C and D. The first two columns in Table 1 refer to the shrunk estimator with estimated α, the third and fourth column refer to the estimator with fixed $\alpha = 0, 1$, i.e. $\hat{\Sigma}_T$ and $\hat{F}_T$, while the last two columns give values of $\hat{\alpha}_T$.

Remark 2. For the experimental results in Panel A and B we have $\beta = \gamma = 1$, using the notation in Condition 1. By Theorem 1, the estimator might not be consistent, in this case, unless $\epsilon \to 0$. For the experiments in Panel C and D, we have $\beta = 1$ and $\gamma = 2$ and the estimator is consistent also for $\epsilon \to c > 0$.

As we mentioned above, when time dependence is moderate across all the series (Panel B and D), accounting for time dependence can be advantageous. However, the difference between the estimators based on $b = 5$ and 1 decreases as shrinkage becomes less desirable. The results also suggest that for both estimators, the PRIAL increases in N/T. Simulations carried out by the author, but not reported here confirm this finding in a variety of situations, like the one contemplated in Ledoit and Wolf [6, Fig. 6]. If there is moderate time dependence, accounting for it in the estimation could be advantageous even when T is small and N large (e.g. $N/T = 80/10$) despite noise in estimation of $\hat{\Gamma}^b_{T,ij}$.

Table 1

		Shrunk estimator				Estimated value of alpha	
		Estimated alpha		Fixed alpha			
		$b = 5$	$b = 1$	Alpha $= 0$	Alpha $= 1$	$b = 5$	$b = 1$
Panel A. Diagonal covariance, VAR coefficients in [0,0.8]							
$N = 20$							
Sigma $= .25$	MEAN	9.9	11.6	35.3	9.2	0.72	0.61
	SE	0.12	0.16	0.33	0.02	0.005	0.004
	PRIAL (%)	71.9	67.3	0.0	74.1		
Sigma $= .5$	MEAN	18.2	19.4	39.2	22.0	0.60	0.51
	SE	0.19	0.23	0.42	0.02	0.005	0.004
	PRIAL (%)	53.5	50.4	0.0	43.8		
Sigma $= 1$	MEAN	43.9	43.5	50.1	94.2	0.36	0.29
	SE	0.56	0.59	0.80	0.04	0.004	0.003
	PRIAL (%)	12.5	13.2	0.0	−87.9		
Sigma $= 2$	MEAN	71.0	68.7	63.1	301.2	0.17	0.13
	SE	1.54	1.57	1.70	0.08	0.003	0.002
	PRIAL (%)	−12.5	−8.9	0.0	−377.3		
$N = 40$							
Sigma $= 1$	MEAN	76.7	79.4	128.4	119.6	0.49	0.41
	SE	0.53	0.63	1.20	0.04	0.003	0.003
	PRIAL (%)	40.2	38.2	0.0	6.9		
$N = 80$							
Sigma $= 1$	MEAN	117.0	123.0	284.6	154.1	0.60	0.53
	STD	0.47	0.63	1.74	0.03	0.003	0.003
	PRIAL (%)	58.9	56.8	0.0	45.8		
$N = 160$							
Sigma $= 1$	MEAN	574.4	665.7	1796.1	634.2	0.63	0.52
	STD	3.04	4.84	11.07	0.08	0.003	0.002
	PRIAL (%)	68.0	62.9	0.0	64.7		
Panel B. Diagonal covariance, VAR coefficients in [0.5,0.8]							
$N = 20$							
Sigma $= .25$	MEAN	19.6	31.2	82.1	11.6	0.66	0.46
	SE	0.36	0.51	0.79	0.05	0.004	0.003
	PRIAL (%)	76.1	62.0	0.0	85.9		
Sigma $= .5$	MEAN	34.2	44.3	90.2	31.4	0.59	0.41
	SE	0.47	0.65	0.99	0.06	0.004	0.003
	PRIAL (%)	62.1	50.9	0.0	65.3		
Sigma $= 1$	MEAN	82.3	85.5	109.2	136.2	0.40	0.28
	SE	1.00	1.18	1.65	0.09	0.004	0.003
	PRIAL (%)	24.7	21.8	0.0	−24.6		
Sigma $= 2$	MEAN	131.7	125.7	120.1	424.7	0.21	0.14
	SE	2.53	2.61	2.92	0.15	0.003	0.002
	PRIAL (%)	−9.6	−4.7	0.0	−253.7		

Table 1 (*continued*)

$N = 40$							
Sigma = 1	MEAN	143.8	169.9	293.2	174.8	0.50	0.35
	SE	1.12	1.64	2.69	0.09	0.003	0.002
	PRIAL (%)	50.9	42.1	0.0	40.4		
$N = 80$							
Sigma = 1	MEAN	240.4	318.9	713.6	234.8	0.57	0.41
	STD	1.59	2.58	4.58	0.08	0.003	0.002
	PRIAL (%)	66.3	55.3	0.0	67.1		
$N = 160$							
Sigma = 1	MEAN	1082.7	1647.1	3991.0	832.5	0.59	0.41
	STD	8.51	13.81	23.06	0.18	0.002	0.002
	PRIAL (%)	72.9	58.7	0.0	79.1		
Panel C. Nondiagonal covariance, VAR coefficients in [0,0.8]							
Sigma = 1							
$N = 20$							
	MEAN	24.3	24.5	33.4	44.9	0.43	0.37
	SE	0.23	0.24	0.39	0.02	0.004	0.003
	PRIAL (%)	27.1	26.6	0.0	−34.6		
$N = 40$							
	MEAN	92.7	93.8	131.4	177.8	0.42	0.36
	SE	0.72	0.73	1.20	0.04	0.003	0.003
	PRIAL (%)	29.4	28.6	0.0	−35.3		
Panel D. Nondiagonal covariance, VAR coefficients in [0.5,0.8]							
Sigma = 1							
$N = 20$							
	MEAN	52.6	55.8	77.8	82.6	0.44	0.31
	SE	0.49	0.59	0.93	0.05	0.004	0.003
	PRIAL (%)	32.3	28.2	0.0	−6.2		
$N = 40$							
	MEAN	206.9	220.9	309.8	335.8	0.42	0.30
	SE	1.64	1.99	3.12	0.10	0.004	0.003
	PRIAL (%)	33.2	28.7	0.0	−8.4		

4. Asymptotics for covariance shrinkage estimators

Proof of Proposition 1. Differentiating with respect to α,

$$\begin{aligned}
&\frac{d\mathbb{E}\left\|\alpha F + (1-\alpha)\hat{\Sigma}_T - \Sigma\right\|_2^2}{d\alpha} \\
&= 2\sum_{1\leqslant i,j\leqslant N} \mathbb{E}\left(\alpha F_{ij} + (1-\alpha)\hat{\Sigma}_{T,ij} - \Sigma_{ij}\right)\left(F_{ij} - \hat{\Sigma}_{T,ij}\right) \\
&= 2\sum_{1\leqslant i,j\leqslant N} \left[\alpha\mathbb{E}\left(F_{ij} - \hat{\Sigma}_{T,ij}\right)^2 + Cov\left(F_{ij}, \hat{\Sigma}_{T,ij}\right) - Var\left(\hat{\Sigma}_{T,ij}\right)\right],
\end{aligned}$$

which, imposing the constraint, implies the result because $Cov\left(F_{ij}, \hat{\Sigma}_{T,ij}\right) = 0$, as F is nonstochastic. □

We introduce some notation.

Notation 1. $\Gamma_{ij}(s) := Cov\left(Y_{t,i}Y_{t,j}, Y_{t+s,i}Y_{t+s,j}\right)$, $\Gamma_{T,ij} := \Gamma_{ij}(0) + 2\sum_{s=1}^{T-1}(1 - s/T)\,\Gamma_{ij}(s)$. Moreover, $\|\ldots\|_{p,\mathbb{P}}$ is the L_p norm ($p = 1, 2$).

The following lemmata are used to prove Theorem 1.

Lemma 1. *Under Conditions* 2–4,

$$\mathbb{E}\left\|\hat{\Sigma}_T - \Sigma\right\|_2^2 \asymp \epsilon N$$

and

$$\mathbb{E}\left|\left(\sum_{1\leqslant i,j\leqslant N}\hat{\Gamma}^b_{T,ij}/T\right) - \mathbb{E}\left\|\hat{\Sigma}_T - \Sigma\right\|_2^2\right| = o(\epsilon N).$$

Proof. By Condition 2,

$$\begin{aligned}\mathbb{E}\left\|\hat{\Sigma}_T - \Sigma\right\|_2^2 &= \sum_{1\leqslant i,j\leqslant N} Var\left(\frac{1}{T}\sum_{t=1}^{T} Y_{t,i}Y_{t,j}\right)\\ &\leqslant 2\frac{N^2}{T^2}\max_{1\leqslant i,j\leqslant N}\sum_{1\leqslant t_1\leqslant t_2\leqslant T} Cov\left(Y_{t_1,i}Y_{t_1,j}, Y_{t_2,i}Y_{t_2,j}\right)\\ &\lesssim \frac{N^2}{T} = \epsilon N. \end{aligned} \tag{7}$$

By Condition 4, $Y_{t,i}Y_{t,j}$ is nondegenerate ($\forall i, j$) hence we must also have

$$\sum_{1\leqslant i,j\leqslant N} Var\left(\frac{1}{T}\sum_{t=1}^{T} Y_{t,i}Y_{t,j}\right) \gtrsim \frac{N^2}{T}\min_{1\leqslant i,j\leqslant N} Var\left(Y_{t,i}Y_{t,j}\right) \asymp \epsilon N,$$

implying the first part of Lemma 1. For arbitrary, but fixed i, j, define $S_t := (1 - \mathbb{E})\,Y_{t,i}Y_{t,j}$. Condition 2 implies (e.g. [4]),

$$\sum_{1\leqslant r_1\leqslant r_2\leqslant r_3\leqslant\infty}\mathbb{E}S_tS_{t+r_1}S_{t+r_2}S_{t+r_3} \leqslant \sum_{r=1}^{\infty}(r+1)^2\theta_r < \infty, \tag{8}$$

which implies that the fourth mixed cumulant of $\left(S_t, S_{t+r_1}, S_{t+r_2}, S_{t+r_3}\right)$ is summable in (r_1, r_2, r_3). Noting

$$Var\left(\frac{1}{T}\sum_{t=1}^{T} Y_{t,i}Y_{t,j}\right) = \frac{\Gamma_{T,ij}}{T}$$

by Condition 3 and (8), we deduce,

$$\max_{1\leqslant i,j\leqslant N}\left\|\hat{\Gamma}^b_{T,ij} - \Gamma_{T,ij}\right\|_{1,\mathbb{P}} = o(1)$$

using Theorem 1 in Andrews [2] and the results in Anderson [1, Chapter 8]. □

We give the rate of growth of $\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2$.

Lemma 2. *Under Conditions* 1, 2 *and* 4

$$\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2=\|F-\Sigma\|_2^2+\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2\asymp N^\gamma.$$

Proof. Adding and subtracting Σ_{ij}^2,

$$\begin{aligned}\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2&=\sum_{1\leqslant i,j\leqslant N}\left[\left(F_{ij}^2-2F_{ij}\mathbb{E}\hat{\Sigma}_{T,ij}+\Sigma_{ij}^2\right)+\left(\mathbb{E}\hat{\Sigma}_{T,ij}^2-\Sigma_{ij}^2\right)\right]\\&=\|F-\Sigma\|_2^2+\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2\\&\asymp N^\gamma\end{aligned}$$

because, by Lemma 1,

$$\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2=O\left(\epsilon N\right)=o\left(N^\gamma\right),$$

because Condition 1(4) gives $\epsilon^{1/2}N^{(1+\gamma)/2}=o\left(N^\gamma\right)$ which implies $\epsilon N=o\left(N^\gamma\right)$ and because $\|F-\Sigma\|_2^2\asymp N^\gamma$ by Condition 1(3). □

We show convergence of $\left\|\hat{F}_T-\hat{\Sigma}_T\right\|_2^2$.

Lemma 3. *Under Conditions* 1, 2, *and* 4,

$$\left\|\hat{F}_T-\hat{\Sigma}_T\right\|_2^2-\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2=o_p\left(N^\gamma\right).$$

Proof. By direct calculation,

$$\begin{aligned}&\left\|\hat{F}_T-\hat{\Sigma}_T\right\|_2^2\\&\quad=\left\|\left(\hat{F}_T-F\right)+(F-\Sigma)+\left(\Sigma-\hat{\Sigma}_T\right)\right\|_2^2\\&\qquad\text{[adding and subtracting } F \text{ and } \Sigma]\\&\quad=\sum_{1\leqslant i,j\leqslant N}\left[\left(\hat{F}_{T,ij}-F_{ij}\right)^2+\left(F_{ij}-\Sigma_{ij}\right)^2+\left(\Sigma_{ij}-\hat{\Sigma}_{T,ij}\right)^2\right.\\&\qquad+2\left(\hat{F}_{T,ij}-F_{ij}\right)\left(F_{ij}-\Sigma_{ij}\right)+2\left(\hat{F}_{T,ij}-F_{ij}\right)\left(\Sigma_{ij}-\hat{\Sigma}_{T,ij}\right)\\&\qquad\left.+2\left(F_{ij}-\Sigma_{ij}\right)\left(\Sigma_{ij}-\hat{\Sigma}_{T,ij}\right)\right]\\&\qquad\text{[expanding the square]}\\&\quad=\|F-\Sigma\|_2^2+o_p\left(N^\gamma\right),\end{aligned}$$

using the following, which are easily derived using Condition 1,

$$\sum_{1\leqslant i,j\leqslant N}\left(\hat{F}_{T,ij}-F_{ij}\right)^2=O_p\left(N^\beta T^{-1}\right)=O_p\left(\epsilon N^{\beta-1}\right)=o_p\left(N^\gamma\right),\tag{9}$$

because there are N^β nonzero elements in the sum and $\hat{F}_{T,ij}$ is root-n consistent:

$$\sum_{1\leqslant i,j\leqslant N}\left(\Sigma_{ij}-\hat{\Sigma}_{T,ij}\right)^2 = O_p\left(\epsilon N\right) = o_p\left(N^\gamma\right),$$

by Lemma 1,

$$\sum_{1\leqslant i,j\leqslant N}\left(\hat{F}_{T,ij}-F_{ij}\right)\left(F_{ij}-\Sigma_{ij}\right) = O_p\left(N^\beta T^{-1/2}\right) = O_p\left(\epsilon^{1/2}N^{\beta-1/2}\right) = o_p\left(N^\gamma\right), \tag{10}$$

by similar reasoning as for (9),

$$\begin{aligned}\sum_{1\leqslant i,j\leqslant N}\left(\hat{F}_{T,ij}-F_{ij}\right)\left(\Sigma_{ij}-\hat{\Sigma}_{T,ij}\right) &\leqslant \left\|\hat{F}_T-F\right\|_2\left\|\hat{\Sigma}_T-\Sigma\right\|_2\\ &= O_p\left(\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2\right) = o_p\left(N^\gamma\right),\end{aligned}$$

by Lemma 1,

$$\begin{aligned}\sum_{1\leqslant i,j\leqslant N}\left(F_{ij}-\Sigma_{ij}\right)\left(\Sigma_{ij}-\hat{\Sigma}_{T,ij}\right) &\leqslant \|F-\Sigma\|_2\left\|\hat{\Sigma}_T-\Sigma\right\|_2\\ &= O_p\left(\epsilon^{1/2}N^{(1+\gamma)/2}\right) = o_p\left(N^\gamma\right),\end{aligned} \tag{11}$$

by Lemma 1 and Condition 1(3). By Lemmata 2 and 1,

$$\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2 = \|F-\Sigma\|_2^2 + O\left(\epsilon N\right).$$

Hence,

$$\left\|\hat{F}_T-\hat{\Sigma}_T\right\|_2^2 - \mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2 = o_p\left(N^\gamma\right). \qquad \square$$

The following two lemmata will be used to show adaptiveness with respect to $\hat{F}$.

Lemma 4. *Under Conditions* 1, 2, *and* 4,

$$\mathbb{E}\left\|\hat{F}-\hat{\Sigma}_T\right\|_2^2 \asymp N^\gamma$$

and

$$\mathbb{E}\left\|\hat{F}-\hat{\Sigma}_T\right\|_2^2 - \mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2 = o\left(N^\gamma\right).$$

Proof. We have the following chain of equalities

$$\begin{aligned}&\mathbb{E}\left\|\hat{F}_T-\hat{\Sigma}_T\right\|_2^2\\ &\quad= \sum_{1\leqslant i,j\leqslant N}\mathbb{E}\left[\left(\hat{F}_{T,ij}-F_{ij}\right)^2 + 2\left(\hat{F}_{T,ij}-F_{ij}\right)\left(F_{ij}-\hat{\Sigma}_{T,ij}\right) + \left(F_{ij}-\hat{\Sigma}_{T,ij}\right)^2\right]\\ &\qquad\text{[adding and subtracting } F \text{ and expanding the square]}\\ &\quad= \mathbb{E}\left\|\hat{F}_T-F\right\|_2^2 + \mathbb{E}\left\|F-\hat{\Sigma}\right\|_2^2 + 2\sum_{1\leqslant i,j\leqslant N}\mathbb{E}\left(\hat{F}_{T,ij}-F_{ij}\right)\left(F_{ij}-\hat{\Sigma}_{T,ij}\right).\end{aligned} \tag{12}$$

By Lemma 2, $\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2 \asymp N^\gamma$, and mutatis mutandis from (9) using Condition 1(1), $\mathbb{E}\left\|\hat{F}_T-F\right\|_2^2 = o\left(N^\gamma\right)$. By Holder inequality

$$\sum_{1\leqslant i,j\leqslant N}\mathbb{E}\left(\hat{F}_{T,ij}-F_{ij}\right)\left(F_{ij}-\hat{\Sigma}_{T,ij}\right) \leqslant \sum_{1\leqslant i,j\leqslant N}\left[\mathbb{E}\left(\hat{F}_{T,ij}-F_{ij}\right)^2\mathbb{E}\left(F_{ij}-\hat{\Sigma}_{T,ij}\right)^2\right]^{1/2} = O\left(N^\beta T^{-1/2}\right)$$

because by Condition 1(1), $\mathbb{E}\left(\hat{F}_{T,ij}-F_{ij}\right)^2 = O\left(T^{-1}\right)$, by Condition 4, $\mathbb{E}\left(F_{ij}-\hat{\Sigma}_{T,ij}\right)^2 = O(1)$, and by Condition 1(2) there are at most $O\left(N^\beta\right)$ nonzero elements in the double sum. As in (10), $O\left(N^\beta T^{-1/2}\right) = o\left(N^\gamma\right)$, implying that $\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2 \asymp N^\gamma$ is the dominating term. Substituting these orders of magnitude in (12), we have the result. □

This is the final lemma before the proof of Theorem 1.

Lemma 5. *Under Conditions* 1, 2, *and* 4

$$\sum_{1\leqslant i,j\leqslant N} Cov\left(\hat{\Sigma}_{T,ij},\hat{F}_{ij}\right) = o(\epsilon N).$$

Proof. By Holder inequality,

$$\sum_{1\leqslant i,j\leqslant N} Cov\left(\hat{\Sigma}_{T,ij},\hat{F}_{ij}\right) \leqslant \sum_{1\leqslant i,j\leqslant N}\left[\mathbb{E}\left(\hat{\Sigma}_{T,ij}-\Sigma_{ij}\right)^2\mathbb{E}\left(\hat{F}_{ij}-F_{ij}\right)^2\right]^{1/2} = O\left(N^\beta/T\right)$$

because, by Condition 1(2), there are at most $O\left(N^\beta\right)$ nonzero elements in the double sum, by Condition 1(1) $\mathbb{E}\left(\hat{F}_{ij}-F_{ij}\right)^2 = O\left(T^{-1}\right)$, and by Condition 2, $\mathbb{E}\left(\hat{\Sigma}_{T,ij}-\Sigma_{ij}\right)^2 = O\left(T^{-1}\right)$ as shown in (7). Since $\beta<2$, $N^\beta/T = o(\epsilon N)$. □

We can now prove Theorem 1.

Proof of Theorem 1(1). By the triangle inequality,

$$\left|\frac{\sum_{1\leqslant i,j\leqslant N}\hat{\Gamma}^b_{T,ij}/T}{\left\|\hat{F}_T-\hat{\Sigma}_T\right\|_2^2}-\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}{\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2}\right| \leqslant \left|\frac{\sum_{1\leqslant i,j\leqslant N}\hat{\Gamma}^b_{T,ij}/T}{\left\|\hat{\Sigma}_T-\hat{F}_T\right\|_2^2}-\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}{\left\|\hat{\Sigma}_T-\hat{F}_T\right\|_2^2}\right| + \left|\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}{\left\|\hat{\Sigma}_T-\hat{F}_T\right\|_2^2}-\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}{\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2}\right| = \mathrm{I}+\mathrm{II}.$$

Control over I: By Lemma 3, the Continuous Mapping Theorem and Lemma 2,

$$\left(\left\|\hat{F}_T-\hat{\Sigma}_T\right\|_2^2\Big/ N^\gamma\right)^{-1}\xrightarrow{p}\left(\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2\Big/ N^\gamma\right)^{-1}\asymp N^{-\gamma}N^\gamma=O(1). \tag{13}$$

By (13) and Lemma 1,

$$\begin{aligned}\mathrm{I}&=\frac{1}{\left\|\hat{F}_T-\hat{\Sigma}_T\right\|_2^2\Big/ N^\gamma}\left|\sum_{1\leqslant i,j\leqslant N}\hat{\Gamma}^b_{T,ij}/T-\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2\right|\Big/ N^\gamma\\&=O_p(1)\left|\sum_{1\leqslant i,j\leqslant N}\hat{\Gamma}^b_{T,ij}/T-\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2\right|\Big/ N^\gamma\\&=o_p\left(\epsilon N^{1-\gamma}\right).\end{aligned}$$

Control over II: By direct calculation, Lemma 1 and (13),

$$\begin{aligned}\mathrm{II}&=\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}{\left\|\hat{\Sigma}_T-\hat{F}_T\right\|_2^2\,\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2}\left|\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2-\left\|\hat{\Sigma}_T-\hat{F}_T\right\|_2^2\right|\\&=\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2\Big/ N^\gamma}{\left\|\hat{\Sigma}_T-\hat{F}_T\right\|_2^2\Big/ N^\gamma\,\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2\Big/ N^\gamma}\left|\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2-\left\|\hat{\Sigma}_T-\hat{F}_T\right\|_2^2\right|\Big/ N^\gamma\\&=o_p\left(\epsilon N^{1-\gamma}\right).\end{aligned}$$

Hence, $\mathrm{I}+\mathrm{II}=o_p\left(\epsilon N^{1-\gamma}\right)$, which gives $\left[\epsilon N^{1-\gamma}\right]^{-1}\left(\hat{\alpha}_T-\alpha_0\right)=o_p(1)$. To see that $\alpha_0\asymp\epsilon N^{1-\gamma}$, we just use Lemmata 1 and 2. Then, Condition 1(4) shows that $\epsilon N^{1-\gamma}=o(1)$. □

Proof of Theorem 1(2). By the triangle inequality

$$\begin{aligned}&\left|\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2-\sum_{1\leqslant i,j\leqslant N}Cov\left(\hat{\Sigma}_{T,ij},\hat{F}_{ij}\right)}{\mathbb{E}\left\|\hat{F}-\hat{\Sigma}_T\right\|_2^2}-\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}{\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2}\right|\\&\quad\leqslant\left|\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}{\mathbb{E}\left\|\hat{F}-\hat{\Sigma}_T\right\|_2^2}-\frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2}{\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2}\right|+\left|\frac{\sum_{1\leqslant i,j\leqslant N}Cov\left(\hat{\Sigma}_{T,ij},\hat{F}_{ij}\right)}{\mathbb{E}\left\|\hat{F}-\hat{\Sigma}_T\right\|_2^2}\right|\\&\quad=\mathrm{I}+\mathrm{II}.\end{aligned}$$

Control over I:

$$\begin{aligned}\mathrm{I} &\leqslant \frac{\mathbb{E}\left\|\hat{\Sigma}_T-\Sigma\right\|_2^2 \Big/ N^{\gamma}}{\mathbb{E}\left\|\hat{F}-\hat{\Sigma}_T\right\|_2^2 \Big/ N^{\gamma}\,\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2 \Big/ N^{\gamma}}\left|\mathbb{E}\left\|\hat{F}-\hat{\Sigma}_T\right\|_2^2-\mathbb{E}\left\|\hat{\Sigma}_T-F\right\|_2^2\right| \Big/ N^{\gamma} \\ &= o\left(\epsilon N^{1-\gamma}\right),\end{aligned}$$

by Lemmata 1, 2 and 4.

Control over II. By Lemmata 4 and 5,

$$\mathrm{II}=\frac{\sum_{1\leqslant i,j\leqslant N} Cov\left(\hat{\Sigma}_{T,ij},\hat{F}_{ij}\right)\Big/ N^{\gamma}}{\mathbb{E}\left\|\hat{F}-\hat{\Sigma}_T\right\|_2^2 \Big/ N^{\gamma}}=o\left(\epsilon N^{1-\gamma}\right). \qquad \square$$

Proof of Theorem 1(3). We have the following chain of inequalities,

$$\begin{aligned}&\left\|\hat{\alpha}_T\hat{F}_T+\left(1-\hat{\alpha}_T\right)\hat{\Sigma}_T-\Sigma\right\|_2 \\ &\quad=\left\|\hat{\alpha}_T\left(\hat{F}_T-F\right)+\hat{\alpha}_T F+\left(1-\hat{\alpha}_T\right)\hat{\Sigma}_T-\Sigma\right\|_2 \\ &\qquad\text{[adding and subtracting } \hat{\alpha}_T F\text{]} \\ &\quad\leqslant\left\|\hat{\alpha}_T F+\left(1-\hat{\alpha}_T\right)\hat{\Sigma}_T-\Sigma\right\|_2+\hat{\alpha}_T\left\|\hat{F}_T-F\right\|_2 \\ &\qquad\text{[by Minkowski inequality]} \\ &\quad=\left\|\left(\hat{\alpha}_T-\alpha_0\right)\left(F-\hat{\Sigma}_T\right)+\alpha_0 F+\left(1-\alpha_0\right)\hat{\Sigma}_T-\Sigma\right\|_2+\hat{\alpha}_T\left\|\hat{F}_T-F\right\|_2 \\ &\qquad\left[\text{adding and subtracting } \alpha_0\left(F-\hat{\Sigma}_T\right)\right] \\ &\quad\leqslant\left\|\alpha_0 F+\left(1-\alpha_0\right)\hat{\Sigma}_T-\Sigma\right\|_2+\left(\hat{\alpha}_T-\alpha_0\right)\left\|F-\hat{\Sigma}_T\right\|_2+\hat{\alpha}_T\left\|\hat{F}_T-F\right\|_2 \\ &\qquad\text{[by Minkowski inequality]} \\ &\quad=\mathrm{I}+\mathrm{II}+\mathrm{III}.\end{aligned}$$

We shall bound the three terms above. First, note that by Theorem 1(1),

$$\left(\hat{\alpha}_T-\alpha_0\right)=o_p\left(\epsilon N^{1-\gamma}\right), \tag{14}$$

and

$$\hat{\alpha}_T=O_p\left(\alpha_0\right)\asymp\epsilon N^{1-\gamma}=o\left(1\right). \tag{15}$$

Control over II: By Lemmata 3 and 2,

$$\left\|F-\hat{\Sigma}_T\right\|_2^2=\mathbb{E}\left\|F-\hat{\Sigma}_T\right\|_2^2+o_p\left(N^{\gamma}\right)=O_p\left(N^{\gamma}\right),$$

hence using (14),

$$\mathrm{II}=o_p\left(\epsilon N^{1-\gamma/2}\right).$$

Control over III: Using (9) and (15)

$$\mathrm{III} = o_p\left(\epsilon N^{1-\gamma/2}\right).$$

Control over I: For the bound to be uniform, we only need to show that the following holds in probability:

$$\left\|\alpha_0 F + (1-\alpha_0)\hat{\Sigma}_T - \Sigma\right\|_2 \gtrsim \epsilon N^{1-\gamma/2}.$$

Using $\overset{p}{\gtrsim}$ to mean that $\gtrsim$ holds in probability and similarly for $\overset{p}{\asymp}$,

$$\begin{aligned}
&\left\|\alpha_0 F + (1-\alpha_0)\hat{\Sigma}_T - \Sigma\right\|_2^2 \\
&\quad = \left\|\alpha_0 (F-\Sigma) + (1-\alpha_0)\left(\hat{\Sigma}_T - \Sigma\right)\right\|_2^2 \\
&\qquad \text{[adding and subtracting } \alpha_0\Sigma\text{]} \\
&\quad = \alpha_0^2 \|F-\Sigma\|_2^2 + (1-\alpha_0)^2 \left\|\hat{\Sigma}_T - \Sigma\right\|_2^2 + \sum_{1\leqslant i,j\leqslant N} 2\alpha_0\left(F_{ij}-\Sigma_{ij}\right)\left(\hat{\Sigma}_{T,ij}-\Sigma_{ij}\right) \\
&\qquad \text{[expanding the square]} \\
&\quad \overset{p}{\gtrsim} \epsilon N, \qquad (16)
\end{aligned}$$

because

$$\begin{aligned}
&(1-\alpha_0)^2 \left\|\hat{\Sigma}_T - \Sigma\right\|_2^2 \overset{p}{\asymp} \epsilon N \\
&\qquad \text{[by (15) and Lemma1],} \\
&\alpha_0^2 \|F-\Sigma\|_2^2 = O\left(\epsilon^2 N^{2-\gamma}\right) = o(\epsilon N) \\
&\qquad \text{[by (15) and Condition 1(3)],} \\
&\sum_{1\leqslant i,j\leqslant N} \alpha_0\left(F_{ij}-\Sigma_{ij}\right)\left(\hat{\Sigma}_{T,ij}-\Sigma_{ij}\right) = O_p\left(\epsilon^{3/2} N^{3/2-\gamma/2}\right) = o(\epsilon N),
\end{aligned}$$

by (15), (11) and Condition 1(4). By Condition 1(4), $\epsilon N^{1-\gamma/2} = o\left([\epsilon N]^{1/2}\right)$, so that

$$\mathrm{I} \overset{p}{\gtrsim} [\epsilon N]^{1/2} \gtrsim \epsilon N^{1-\gamma/2}.$$

To write II and III in terms of I times an $o(1)$ quantity we solve II+III$= x\,[\epsilon N]^{1/2} = o_p\left(\epsilon N^{1-\gamma/2}\right)$ for x to find $x = o_p\left(\epsilon^{1/2} N^{(1-\gamma)/2}\right)$, which implies the result. □

Acknowledgments

I thank an associate editor and a referee for useful comments that improved the content and presentation of the paper.

References

[1] T.W. Anderson, The Statistical Analysis of Time Series, Wiley, New York, 1971.
[2] D. Andrews, Heteroskedasticity and autocorrelation consistent covariance matrix estimation, Econometrica 59 (1991) 817–858.

[3] P. Bühlmann, Bootstraps for time series, Statist. Sci. 17 (2002) 52–72.
[4] P. Doukhan, S. Louhichi, A new weak dependence condition and applications to moment inequalities, Stochastic Process. Appl. 84 (1999) 313–342.
[5] O. Ledoit, M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, J. Empirical Finance 10 (2003) 603–621.
[6] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2004) 365–411.
[7] D.N. Politis, The impact of bootstrap methods on time series analysis, Statist. Sci. 18 (2003) 219–230.