Electronic Journal of Statistics
Vol. 9 (2015) 753–788
ISSN: 1935-7524
DOI: 10.1214/15-EJS1000

High-dimensional autocovariance matrices and optimal linear prediction*

Timothy L. McMurry

Department of Public Health Sciences
University of Virginia
Charlottesville, VA 22908
e-mail: tmcmurry@virginia.edu

and

Dimitris N. Politis

Department of Mathematics
University of California, San Diego
La Jolla, CA 92093
e-mail: dpolitis@ucsd.edu

Abstract: A new methodology for optimal linear prediction of a stationary time series is introduced. Given a sample $X_1, \ldots, X_n$, the optimal linear predictor of X_{n+1} is $\tilde{X}_{n+1} = \phi_1(n)X_n + \phi_2(n)X_{n-1} + \cdots + \phi_n(n)X_1$. In practice, the coefficient vector $\phi(n) \equiv (\phi_1(n), \phi_2(n), \ldots, \phi_n(n))'$ is routinely truncated to its first p components in order to be consistently estimated. By contrast, we employ a consistent estimator of the $n \times n$ autocovariance matrix Γ_n in order to construct a consistent estimator of the optimal, full-length coefficient vector $\phi(n)$. Asymptotic convergence of the proposed predictor to the oracle is established, and finite sample simulations are provided to support the applicability of the new method. As a by-product, new insights are gained on the subject of estimating Γ_n via a positive definite matrix, and four ways to impose positivity are introduced and compared. The closely related problem of spectral density estimation is also addressed.

MSC 2010 subject classifications: Primary 62M20; secondary 62M10.
Keywords and phrases: Autocovariance matrix, time series, prediction, spectral density.

Received July 2014.

1. Introduction

Let $X_1, \ldots, X_n$ be the realization of a covariance stationary time series with mean zero and autocovariance function $\gamma_k = E[X_t X_{t-k}]$. We consider the problem of predicting X_{n+1} based on these observed data. With respect to Mean Squared Error (MSE), the optimal linear predictor is

$$\tilde{X}_{n+1} = \phi_1(n)X_n + \phi_2(n)X_{n-1} + \cdots + \phi_n(n)X_1, \tag{1}$$

*Discussed in 10.1214/15-EJS1010, 10.1214/14-EJS953, 10.1214/15-EJS996 and 10.1214/15-EJS1007; rejoinder at 10.1214/15-EJS1000REJ.

where the coefficients $\phi_i(n)$ are given by

$$\phi(n) \equiv \begin{bmatrix} \phi_1(n) \\ \vdots \\ \phi_n(n) \end{bmatrix} = \Gamma_n^{-1}\gamma(n); \tag{2}$$

(see e.g. p. 167 in Brockwell and Davis, 1991). In equation (2), $\Gamma_n = [\gamma_{|i-j|}]_{i,j=1}^n$ is the autocovariance matrix of $X_1, \ldots, X_n$, and $\gamma(n) = [\gamma_1, \ldots, \gamma_n]'$ is the vector of covariances at lags $1, \ldots, n$. Predictor (1) is an *oracle* because the coefficients $\phi_1(n), \ldots, \phi_n(n)$ are unknown. In practice, the coefficient vector $\phi(n) \equiv (\phi_1(n), \phi_2(n), \ldots, \phi_n(n))'$ is routinely truncated to its first p components in order to be consistently estimated; this procedure is equivalent to fitting an auto-regressive AR(p) process to the data. The resulting predictor is

$$\hat{X}_{n+1}^{AR} = \hat{\phi}_1 X_n + \hat{\phi}_2 X_{n-1} + \cdots + \hat{\phi}_p X_{n-p+1}, \tag{3}$$

where the coefficient vector is typically estimated by the Yule-Walker equations

$$[\hat{\phi}_1, \ldots, \hat{\phi}_p]' = \breve{\Gamma}_p^{-1}\breve{\gamma}(p). \tag{4}$$

In (4), $\breve{\gamma}_k = n^{-1}\sum_{t=1}^{n-|k|} X_t X_{t+|k|}$ is the sample autocovariance at lag k, $\breve{\gamma}(p) = [\breve{\gamma}_1, \ldots, \breve{\gamma}_p]'$, and $\breve{\Gamma}_p = [\breve{\gamma}_{|i-j|}]_{i,j=1}^p$.

Interestingly, $\breve{\Gamma}_p$ is positive definite for any p as long as $\breve{\gamma}_0 > 0$, which is a *sine qua non*. In addition, for any finite p, $\breve{\gamma}(p)$ and $\breve{\Gamma}_p$ are consistent for their respective targets $\gamma(p)$ and Γ_p. Unfortunately, when p is large, problems ensue. For example, when $p = n$, Wu and Pourahmadi (2009) showed that the sample autocovariance matrix $\breve{\Gamma}_n = [\breve{\gamma}_{|i-j|}]_{i,j=1}^n$ is not a consistent estimator of Γ_n in operator norm. Hence, equation (4) cannot be used with $p = n$ to give a consistent estimator of the full coefficient vector $\phi(n)$.

In the present work, we investigate an alternative approach to estimating all n coefficients in the oracle predictor (1); this allows for the complete process history to be used in prediction. The estimated prediction coefficients $\hat{\phi}_1(n), \ldots, \hat{\phi}_n(n)$ are given by the n-dimensional Yule-Walker equations:

$$\hat{\phi}(n) = (\hat{\Gamma}_n^*)^{-1}\hat{\gamma}(n), \tag{5}$$

where $\hat{\Gamma}_n^*$ is a positive definite version of the $n \times n$ banded and tapered estimate of the autocovariance matrix Γ_n introduced in McMurry and Politis (2010), and $\hat{\gamma}(n)$ is the corresponding estimate of the autocovariance vector; see Section 3.2 for details.

It has been widely thought until now that an estimate such as the one in (5) is not feasible. For example, on p. 717 of the recent work by Bickel and Gel (2011) it is stated that *"given n observations, it is impossible to estimate n AR parameters sufficiently well for prediction purposes."* The present work demonstrates that this is not the case. In addition, we discuss an intermediate approach, i.e., an analog of (4) but with p that can be arbitrarily large as long as $p \leq n$.

The remainder of the paper is structured as follows. Section 2 provides the background on the estimators $\hat{\Gamma}_n^*$ and $\hat{\gamma}(n)$ that are required in order to estimate the prediction coefficients. Section 3 contains our main asymptotic results;

in particular, the consistency of $\hat{\phi}(n)$ is shown, and the resulting predictor is shown to be asymptotically equivalent to the oracle predictor (1). Section 4 presents four ways to correct our matrix estimator in order to ensure positive definiteness—and therefore invertibility—in finite samples. Section 5 contains the results of finite-sample simulation studies and a real data experiment. Section 6 summarizes our results. All technical proofs have been placed in Section 7. Our paper concludes with an Appendix that shows how the positive definiteness corrections described in Section 4 can find application in the related problem of spectral density estimation.

2. Estimation set-up

2.1. Estimating the $n \times n$ autocovariance matrix Γ_n

The accuracy of the coefficients estimated by equation (5) rests on the ability to accurately estimate $\Gamma_n = [\gamma_{|i-j|}]_{i,j=1}^n$. However, as mentioned in the introduction, Wu and Pourahmadi (2009) showed that the sample autocovariance matrix $\breve{\Gamma}_n = [\breve{\gamma}_{|i-j|}]_{i,j=1}^n$ is not a consistent estimator of Γ_n in operator norm. In order to achieve consistency, they introduced an l-banded estimate that leaves the $2l+1$ main diagonals of the sample autocovariance matrix intact, and sets the remaining entries to 0. Under conditions on l and short range dependence assumptions on $\{X_t\}_{t\in\mathbb{Z}}$ they established the asymptotic consistency of the banded matrix. Their banded estimator is optimal for MA(q) models as, in that case, it corresponds to the parametric estimator. However, if the autocorrelation does not vanish after a finite lag, the banded estimator can behave erratically (Politis, 2011).

To improve performance when the autocorrelation does not vanish, McMurry and Politis (2010) proposed a banded and tapered matrix estimator in which the $2l+1$ main diagonals of the sample autocovariance matrix are kept intact but the remaining entries are gradually tapered to zero. The gradual taper can substantially improve finite sample performance when the autocorrelation does not vanish, at little cost when it does. The asymptotic convergence rates given in McMurry and Politis (2010) were similar to those in Wu and Pourahmadi (2009).

Remark 1. Recently, Cai, Ren and Zhou (2013) showed that the banded and tapered estimator also enjoys a (slightly) improved rate of convergence as compared to the purely banded estimator; for their proof, they used the trapezoidal taper proposed by Politis and Romano (1995) but it is conjectured that the same holds true for the family of so-called *'flat-top'* tapers as long as they are continuous—see Politis (2001) for more details.

Consequently, we focus on the general matrix estimator proposed by McMurry and Politis (2010) given by

$$\hat{\Gamma}_n = \left[\hat{\gamma}_{|i-j|}\right]_{i,j=1}^n \tag{6}$$

with

$$\hat{\gamma}_s = \kappa(|s|/l)\breve{\gamma}_s \quad \text{for} \quad |s| \le n, \text{ and } \hat{\gamma}(n) = [\hat{\gamma}_1, \ldots, \hat{\gamma}_n]'. \tag{7}$$

In the above, $\kappa(\cdot)$ can be the aforementioned trapezoidal taper, i.e.,

$$\kappa(x) = \begin{cases} 1 & \text{if} \quad |x| \leq 1 \\ 2 - |x| & \text{if} \quad 1 < |x| \leq 2 \\ 0 & \text{if} \quad |x| > 2. \end{cases} \tag{8}$$

More generally, $\kappa(\cdot)$ can be any member of the flat-top family of functions defined in Politis (2001), i.e., $\kappa(\cdot)$ is given as

$$\kappa(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ g(|x|) & \text{if } 1 < |x| \leq c_\kappa \\ 0 & \text{if } |x| > c_\kappa, \end{cases} \tag{9}$$

where the function $g(\cdot)$ satisfies $|g(x)| < 1$, and c_κ is a constant satisfying $c_\kappa \geq 1$.

The matrix estimator (6) has a banding parameter $l \geq 0$. The flat-top tapering leaves the $2l + 1$ main diagonals of the sample autocovariance matrix intact, and gradually down-weights more distant diagonals. In order to cover the possibility of the data at hand being uncorrelated, it is useful to adopt the convention that when $l = 0$, the resulting $\hat{\Gamma}_n$ matrix is given by $\breve{\gamma}_0 I$; this is equivalent to adopting that 0/0=0 in the context of eq. (7).

The trapezoidal taper given in (8) is very convenient, and has been shown to have good performance in different practical settings; we will also employ it in the numerical work in this paper. Nevertheless, our theoretical results apply to a broad class of weight functions including pure banding (no taper) as used in Wu and Pourahmadi (2009), and ultra-smooth tapers such as that suggested in McMurry and Politis (2004). These possibilities are captured by different choices of the function $g(\cdot)$ and the constant c_κ in (9); e.g., letting $c_\kappa = 1$ corresponds to pure banding.

Note that $\hat{\Gamma}_n$ as defined by (6) is asymptotically positive definite, but for finite samples it can have negative eigenvalues. For the remainder of the paper, we assume that it has been corrected to positive definiteness—if needed—as described in Section 4. The positive definite version of matrix $\hat{\Gamma}_n$ will be denoted by $\hat{\Gamma}_n^*$.

2.2. Estimating the length n vector $\phi(n)$

After the above preparatory work, we are able to define the proposed new predictor as

$$\hat{X}_{n+1} = \hat{\phi}_1(n)X_n + \hat{\phi}_2(n)X_{n-1} + \cdots + \hat{\phi}_n(n)X_1, \tag{10}$$

where the coefficients $\hat{\phi}_1(n), \ldots, \hat{\phi}_n(n)$ are given by equation (5) in conjunction with the estimates from equations (6) and (7). We can call predictor (10), the Full-Sample Optimal (FSO) predictor since—as shown in Section 3—it is a consistent proxy for the oracle optimal predictor (1).

By comparison, Bickel and Gel (2011) have recently investigated a predictor for X_{n+1} that uses the upper-left $p_n \times p_n$ submatrix of the banded sample autocovariance matrix $\hat{\Gamma}_n$ with $p_n = o(n)$. Their estimator is designed for an

"on-line" prediction problem that allows for the parameters to be updated after each new observation at relatively low computational cost, and the resulting prediction for X_{n+1} is a linear combination of $X_n, \ldots, X_{n-p_n+1}$. This is still an AR-type predictor as in (3) but they use a higher order p_n than the one obtained by minimizing AIC or a related criterion.

Letting

$$\hat{\gamma}(p_n) = [\hat{\gamma}_1, \ldots, \hat{\gamma}_{p_n}]' \text{ and } \hat{\Gamma}_{p_n} = \left[\hat{\gamma}_{|i-j|}\right]_{i,j=1}^{p_n} \tag{11}$$

we can construct an alternative predictor that is based on a partial sample, i.e., a predictor as in (3) with p_n that can be arbitrarily large as long as $p_n \leq n$. This new predictor is defined as

$$\hat{X}_{n+1}^{p_n} = \hat{\phi}_1^{p_n}(n) X_n + \hat{\phi}_2^{p_n}(n) X_{n-1} + \cdots + \hat{\phi}_{p_n}^{p_n} X_{n-p_n+1} \tag{12}$$

where the length-p_n coefficient vector $\hat{\phi}^{p_n}(n) = [\hat{\phi}_1^{p_n}(n), \ldots, \hat{\phi}_{p_n}^{p_n}(n)]'$ is obtained from

$$\hat{\phi}^{p_n}(n) = (\hat{\Gamma}_{p_n}^*)^{-1} \hat{\gamma}(p_n) \tag{13}$$

where $\hat{\Gamma}_{p_n}^*$ is the matrix that results after $\hat{\Gamma}_{p_n}$ in (11) is corrected to positive definiteness. We can call predictor (12), the Partial-Sample Optimal (PSO) predictor as it will be shown to be a consistent proxy for the oracle optimal Partial-Sample predictor, i.e., the optimal linear predictor of X_{n+1} given the last p_n observations; recall that the oracle predictor is constructed using the (unrealistic) knowledge of the whole autocovariance structure.

2.3. Data-based choice of the banding parameter l

The FSO and PSO predictors of equations (10) and (12) clearly depend on the choice of the banding parameter l. One possible approach to choosing it in a data-dependent way is the following rule, which was introduced for density and spectral density estimation in Politis (2003). McMurry and Politis (2010) further showed this rule produces approximately correct rates for autocovariance matrix estimation and good finite sample performance.

Empirical rule for picking *l* Let $\varrho_k = \gamma_k/\gamma_0$ and $\breve{\varrho}_k = \breve{\gamma}_k/\breve{\gamma}_0$. Let $\hat{l}$ be the smallest positive integer such that $|\breve{\varrho}_{\hat{l}+k}| < c(\log n/n)^{1/2}$ for $k = 1, \ldots, K_n$ where $c > 0$ is a fixed constant, and K_n is a positive, nondecreasing sequence that satisfies $K_n = o(\log n)$.

Remark 2. The empirical rule for picking l remains valid for all $c > 0$ and $1 \leq K_n \leq n$, although different choices of c and K_n can lead to very different finite sample performances. Nonetheless, there are some guidelines for practically useful choices. The factor $(\log n)^{1/2}$ varies slowly, so it has little influence. For example, if log is taken to denote base 10 logarithm, then for sample sizes between 100 and 1000, as is quite typical, $(\log n)^{1/2}$ varies between 1.41 and 1.73. Thus, if c is chosen to be around 2 and K_n about 5, Bonferroni's inequality implies that the bound $\pm c(\log n/n)^{1/2}$ can be used as the critical value of

for an approximate 95% test of the null hypothesis that $\varrho(\hat{l}+1), \ldots, \varrho(\hat{l}+K_n)$ are all simultaneously equal to zero. We have found values in this range work well in practice.

3. Asymptotic results

3.1. Basic assumptions

The convergence of $\hat{\Gamma}_n^*$ to Γ_n, the primary result underpinning our present work, is established in McMurry and Politis (2010) under physical dependence measure conditions (Wu, 2005). In order to define our results, we briefly describe these conditions.

Let $\{\epsilon_i,\, i \in \mathbb{Z}\}$ be a sequence of i.i.d. random variables. Assume that X_i is a causal function of $\{\epsilon_i\}$, i.e.,

$$X_i = f(\ldots, \epsilon_{i-1}, \epsilon_i),$$

where f is a measurable function such that X_i is well defined and $E\left[X_i^2\right] < \infty$. In order to quantify dependence, let ϵ_i' be an independent copy of ϵ_i, $i \in \mathbb{Z}$. Let $\xi_i = (\ldots, \epsilon_{i-1}, \epsilon_i)$, $\xi_i' = (\ldots, \epsilon_{-1}, \epsilon_0', \epsilon_1, \ldots, \epsilon_i)$, and $X_i' = g(\xi_i')$. For $\alpha > 0$, define the physical dependence measure

$$\delta_\alpha(i) := E\left[|X_i - X_i'|^\alpha\right]^{1/\alpha}.$$

Note that the difference between X_i and X_i' is due only to the difference between ϵ_0 and ϵ_0', and therefore $\delta_\alpha(i)$ measures the dependence of X_i on an event i units of time in the past. To measure the cumulative dependence across all time, the quantity

$$\Delta_\alpha := \sum_{i=1}^{\infty} \delta_\alpha(i)$$

is helpful. We will say that $\{X_i\}$ is short-range dependent with moment α if $\Delta_\alpha < \infty$.

These notions of dependence underlie the following assumptions which, in conjunction with further assumptions about the weight function $\kappa(\cdot)$, the bandwidth l, and the underlying process, will be sufficient to establish the consistency of FSO predictor (10).

Assumption 1. $E\left[X_i^4\right]^{1/4} < \infty$ and $\Delta_4 < \infty$.

Assumption 2. The weight function κ is a 'flat-top' taper defined by eq. (9) where the function $g(\cdot)$ and the constant c_κ satisfy $|g(x)| < 1$ for all x, and $c_\kappa \geq 1$.

Assumption 3. The quantity

$$r_n = ln^{-1/2} + \sum_{i=l}^{\infty} |\gamma_i| \tag{14}$$

converges to zero as $n \to \infty$.

All asymptotic results and order notations in the paper will be understood to hold as $n \to \infty$ without explicitly denoting it. In fact, Assumption 3 necessitates that $n \to \infty$; furthermore, the banding parameter l may have to diverge at an appropriate rate to ensure the convergence of (14). However, if it so happens that $\gamma_i = 0$ for all $i >$ some q, e.g., under a moving average MA(q) model, l does not need to diverge; any finite value of l would be acceptable as long as it is at least q.

Assumption 4. The spectral density of $\{X_t\}_{t\in\mathbb{Z}}$, defined as

$$f(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-i\omega k},$$

satisfies $0 < c_1 \leq f(\omega) \leq c_2 < \infty$ for all w, and some positive constants c_1 and c_2.

We now recall one of the main results in McMurry and Politis (2010).

Theorem 1 (McMurry and Politis (2010))**.** *Under Assumptions 1–4,*

$$\left|\left|\hat{\Gamma}_n - \Gamma_n\right|\right|_2 = O_p(r_n) \text{ and } \left|\left|\hat{\Gamma}_n^{-1} - \Gamma_n^{-1}\right|\right|_2 = O_p(r_n),$$

where $||\cdot||_2$ *denotes operator norm and* r_n *is as given in eq.* (14)*.*

Remark 3. Under different regularity conditions, Xiao and Wu (2012) were able to show the sharper result

$$\left|\left|\hat{\Gamma}_n - \Gamma_n\right|\right|_2 = O_p(r'_n)$$

where

$$r'_n = C\sqrt{l \log l/n} + 2 \sum_{i=\lfloor l \rfloor+1}^{\lfloor c_\kappa l \rfloor} \left[1 - \kappa\left(\frac{i}{l}\right)\right] |\gamma_i| + \frac{2}{n} \sum_{i=1}^{\lfloor c_\kappa l \rfloor} i|\gamma_i| + 2 \sum_{i=l+1}^{n-1} |\gamma_i|$$

using our Assumption 2 and some constant $C > 0$. For instance, Xiao and Wu (2012) assume $E|X_t|^{4+\delta} < \infty$ for some $\delta > 0$ whereas we allow for the possibility that $\delta = 0$. Nonetheless, we note that r_n can be replaced by r'_n in all asymptotic results of our paper provided our Assumptions 2 and 4 hold together with the conditions of Theorem 4 in Xiao and Wu (2012).

3.2. Estimating the length n vector $\gamma(n)$

Implicit in the n-dimensional Yule-Walker equations (5) is the need for consistent estimation of the length n vector of auto-covariances $\gamma(n) = [\gamma_1, \ldots, \gamma_n]'$. The vector of sample auto-covariances $\breve{\gamma}(n) = [\breve{\gamma}_1, \ldots, \breve{\gamma}_n]'$ is not a consistent estimator of $\gamma(n)$; in fact, $\breve{\gamma}(n)$ misbehaves. To see why, recall that the periodogram of the centered data vanishes at frequency zero; this implies the identity $\sum_{i=1}^{n} \breve{\gamma}_i = -\breve{\gamma}_0/2$ which, of course, has no reason to hold for the true γ_i.

By contrast, the flat-top weighted estimator $\hat{\gamma}(n) = [\hat{\gamma}_1, \ldots, \hat{\gamma}_n]'$ defined in equation (7) is consistent, as the following Lemma shows. Let $|\vec{v}|_2$ denote the l_2 norm of the vector $\vec{v}$. Then,

Lemma 1. *Under Assumptions 1–4,* $|\hat{\gamma}(n) - \gamma(n)|_2 = O_p(r_n)$.

Notice $\hat{\gamma}(n)$ is closely related to the first row of $\hat{\Gamma}_n$ which is a consistent estimator of Γ_n; the only difference is that while $\hat{\gamma}(n) = [\hat{\gamma}_1, \ldots, \hat{\gamma}_n]'$, the first row of $\hat{\Gamma}_n$ is $[\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_{n-1}]'$. However, the Yule-Walker equations (5) require a positive definite version of $\hat{\Gamma}_n$, denoted $\hat{\Gamma}_n^*$ (see Section 4). By looking at the first row of such a $\hat{\Gamma}_n^*$, we can obtain alternative estimates of $\hat{\gamma}(n)$ that are also consistent as the following Lemma shows.

Lemma 2. *Let $\hat{\Gamma}_n^*$ denote a positive definite version of $\hat{\Gamma}_n$ that satisfies*

$$\left|\left|\hat{\Gamma}_n^* - \Gamma_n\right|\right|_2 = O_p(r_n) \text{ and } \left|\left|(\hat{\Gamma}_n^*)^{-1} - \Gamma_n^{-1}\right|\right|_2 = O_p(r_n).$$

Let $(\hat{\Gamma}_n^*)_{i,j}$ denote the ij'th entry of $\hat{\Gamma}_n^*$, and define

$$\hat{\gamma}^*(n) = [(\hat{\Gamma}_n^*)_{1,2}, \ldots, (\hat{\Gamma}_n^*)_{1,n}, 0]'.$$

Then, under Assumptions 1–4, we have $|\hat{\gamma}^*(n) - \gamma(n)|_2 = O_p(r_n)$.

3.3. Optimal prediction using the full sample

Assumptions 1–4 are sufficient to ensure the vector convergence of the estimated prediction coefficients $\hat{\phi}(n)$ given by (5) to the optimal prediction coefficients $\phi(n)$ given by (2).

Theorem 2. *Under Assumptions 1–4,*

$$|\hat{\phi}(n) - \phi(n)|_2 = O_p(r_n). \tag{15}$$

Corollary 2 of Wu and Pourahmadi (2009) establishes the same rate of convergence for the vector of prediction coefficients resulting from purely banded estimates of Γ_n and $\gamma(n)$. In addition, Corollary 1 of Bickel and Gel (2011) establishes the convergence of a vector of prediction coefficients of length p_n to the optimal vector of the same length; this is similar in spirit to our Theorem 2 but using p_n of smaller order than n.

The vector convergence of estimated prediction coefficients $\hat{\phi}(n)$ to $\phi(n)$ shown in Theorem 2 is important but it is not by itself sufficient to establish the convergence of the resulting predictor to the oracle predictor; this convergence is the subject of our Theorem 3 and is our main theoretical result. To show that the FSO predictor $\hat{X}_{n+1}$ converges to the oracle predictor $\tilde{X}_{n+1}$ we will need two modest additional assumptions.

Assumption 5. There exists a rate $k_n \to \infty$ satisfying

i. $k_n/(ln^{\epsilon}) \to \infty$ for some $\epsilon > 0$.

ii. $r_n k_n^{1/2} \to 0$, where r_n is as given in (14).
iii. $n \sum_{i=k_n+1}^{\infty} \phi_i^2 \to 0$, where ϕ_i are the AR(∞) coefficients of the process $\{X_t\}_{t\in\mathbb{Z}}$.

Assumption 6. $n^{1/2} \sum_{i=n+1}^{\infty} |\phi_i| \to 0$.

Remark 4. The rate k_n described in Assumption 5 is required to exist in order to establish the asymptotic optimality of the FSO predictor (10). However, it is not a tuning parameter, and does not need to be estimated and/or chosen by the practitioner.

Remark 5. It can easily be seen that Assumptions 5 and 6 impose few additional restrictions on the process $\{X_t\}_{t\in\mathbb{Z}}$, as the following discussion shows.

i. Assumption 5i requires that k_n grows slightly faster than l. The optimal l depends on the rate of decay of $|\gamma_i|$ (see Corollary 1 of McMurry and Politis, 2010). If $|\gamma_i| = O(i^{-d})$ for some $d > 1$, then the optimal l is proportional to $n^{1/(2d)}$; if $|\gamma_i|$ decays exponentially, then it is sufficient for l to grow logarithmically.
ii. Assumption 6 is satisfied whenever $|\phi_i| \le C_\phi i^{-k}$ for $i > I_0$, some $k > 3/2$, and some $C_\phi > 0$.
iii. Assumptions 5ii and 5iii require some balancing of convergence rates, but they can be achieved with only modest restrictions on the underlying process. As long as $|\phi_i|$ decays at a rate faster than $i^{-3/2}$ (as required by Assumption 6), the term in Assumption 5iii will converge to 0 provided $k_n > C_{k_n} n^{1/2+\epsilon}$ for some $C_{k_n} > 0$ and some $\epsilon > 0$. As long as $|\gamma_i| < C_\gamma i^{-k}$ for $i > I_0$ for some I_0 and some $k > 2$, this allows for convergence of the prediction when l is the asymptotically optimal bandwidth. If ϕ_i and γ_i decay faster, conditions 5ii and 5iii will be satisfied by a wider range of k_n and continue to allow for the optimal l.

Theorem 3. *Under Assumptions 1–6,*

$$|\hat{X}_{n+1} - \tilde{X}_{n+1}| = o_p(1). \tag{16}$$

In other words, the FSO predictor (10) converges in probability to the theoretically optimal oracle predictor (1).

Remark 6. Our Theorems 2 and 3 are expected to hold true *verbatim* if the estimated autocovariances $\hat{\gamma}_s$ that constitute the entries of matrix (6) are also thresholded as described in Section 2.3 of Paparoditis and Politis (2012). It is less clear how the estimator will perform if the entries of autocovariance matrix are only thresholded without use of the flat-top weight function $\kappa(\cdot)$, i.e., a thresholded version of the sample autocovariance matrix as in Xiao and Wu (2012) and Section 2.2 of Paparoditis and Politis (2012). The reason is that our proof of Theorem 3 relies on the rapid decay of the $\hat{\phi}_i(n)$ as i increases; this is in part ensured because the non-zero diagonals of $\hat{\Gamma}_n$ are constrained to a band of width proportional to l which grows slowly with n.

3.4. Optimal prediction using the partial sample

Theorem 3 demonstrates the asymptotic consistency of the estimated prediction coefficients when the $n \times n$ covariance matrix $\hat{\Gamma}_n^*$ is used. An approach more in line with traditional AR model fitting or the work of Bickel and Gel (2011) would be to fit an AR model of order $p_n < n$, where p_n could potentially grow with n. This would entail one-step ahead prediction that uses only the last p_n observations; the prediction coefficients $\hat{\phi}(p_n)$ are given by (13).

Theorem 2 carries over to this lower order setting without modification, although it should be emphasized that if p_n grows slowly enough, faster convergence rates than the one given below are possible; for example, if p_n is constant, then the actual convergence rate will be $n^{-1/2}$, i.e., $|\hat{\phi}(p_n) - \phi(p_n)|_2 = O_p(n^{-1/2})$.

Corollary 1. *Let $1 \leq p_n \leq n$. Under Assumptions 1–4,*

$$|\hat{\phi}(p_n) - \phi(p_n)|_2 = O_p(r_n).$$

The extension of Theorem 3 requires a strengthening of Assumption 6. Our arguments depend on the closeness of $\phi(p_n)$ to the corresponding AR(∞) coefficients; this closeness improves as p_n increases, necessitating the following assumption.

Assumption 7. Let k_n be as in Assumption 5. Then either
a. $p_n \leq k_n$ for all n, or
b. $p_n > k_n$ and $n^{1/2} \sum_{i=p_n+1}^{\infty} |\phi_i| \to 0$.

Remark 7. In the case where $p_n > k_n$, Assumption 7b is only a modest strengthening of Assumption 5iii.

Corollary 2. *Let $1 \leq p_n \leq n$. Under Assumptions 1–5 and 7,*

$$|\hat{X}_{n+1}^{p_n} - \tilde{X}_{n+1}^{p_n}| = o_p(1),$$

where $\hat{X}_{n+1}^{p_n}$ is the PSO predictor of eq. (12) with coefficients $\hat{\phi}^{p_n}(n)$ obtained from eq. (13), and $\tilde{X}_{n+1}^{p_n}$ is its oracle counterpart of order p_n.

Corollary 1 is quite similar to Corollary 1 in Bickel and Gel (2011); however, their result requires $p_n = o(n)$ whereas ours is valid for all non-negative sequences $p_n \leq n$. In addition, neither Bickel and Gel (2011) nor Wu and Pourahmadi (2009) provide a result similar to Corollary 2.

Remark 8. The FSO predictor (10) and the PSO predictor (12) are based on eq. (5) and (13) respectively that employ the matrix estimator $\hat{\Gamma}_n^*$, and the vector estimator $\hat{\gamma}(n)$. Of course, using the positive definite matrix estimator is necessary because the finite-sample inverse is needed. Note, however, that we could equally have chosen the vector estimator $\hat{\gamma}^*(n)$ of Lemma 2 instead of $\hat{\gamma}(n)$ in the Yule-Walker equations (5) and (13). All our asymptotic results of Section 3 on FSO/PSO predictors remain true *verbatim* with such a choice.

4. Corrections towards positive definiteness

Under Assumptions 1–4, the matrix $\hat{\Gamma}_n$ of eq. (6) will have eigenvalues bounded away from zero with probability tending to one as $n \to \infty$. However, for finite samples, $\hat{\Gamma}_n$ will occasionally have eigenvalues that are negative and/or positive but too small. Since the inverse of $\hat{\Gamma}_n$ is a key element in prediction, the matrix $\hat{\Gamma}_n$ must be corrected to achieve finite-sample positive definiteness and avoid ill-conditioning. In this section, we present four ways to implement such a correction. The method presented in Section 4.1 was originally proposed in McMurry and Politis (2010); we now complete that proposal by observing the need to rescale the matrix after its being corrected. The methods in Sections 4.2, 4.3, and 4.4 are novel.

4.1. Eigenvalue thresholding

In the context of the Linear Process Bootstrap, McMurry and Politis (2010) suggested correcting the eigenvalues obtained in the spectral decomposition

$$\hat{\Gamma}_n = T_n D T_n' \tag{17}$$

where T_n is an orthogonal matrix, and D is diagonal with ith entry denoted d_i. Letting $D^\epsilon = \mathrm{diag}(d_1^\epsilon, \ldots, d_n^\epsilon)$ with $d_i^\epsilon = \max\{d_i, \epsilon\hat{\gamma}_0/n^\beta\}$, McMurry and Politis (2010) showed that the adjusted estimate

$$\hat{\Gamma}_n^\epsilon = T_n D^\epsilon T_n' \tag{18}$$

is positive definite but maintains the same asymptotic rate of convergence as $\hat{\Gamma}_n$; in the above, $\epsilon > 0$ and $\beta > 1/2$ are some fixed numbers. For the purposes of Linear Process Bootstrap, it had been found that the simple choices $\epsilon = 1$ and $\beta = 1$ worked well in practice. In the present context, however, we found that $\epsilon = 1$ sometimes produced unstable predictions. A much larger ϵ, of the order of 10 or 20, seems to solve the problem; we used $\epsilon = 20$ and $\beta = 1$ in the simulations.

Note that the average eigenvalue of $\breve{\Gamma}_n$ equals $\breve{\gamma}_0$, which is our best estimator of $\mathrm{var}\,[X_t]$; similarly, the average eigenvalue of $\hat{\Gamma}_n$ equals $\hat{\gamma}_0 = \breve{\gamma}_0$. However, the threshold correction (18) increases the average eigenvalue of the estimated matrix, implicitly suggesting an increased estimate of $\mathrm{var}\,[X_t]$ (see Appendix A for the connection of the eigenvalues of Γ_n to the spectral density, and therefore also to $\mathrm{var}\,[X_t]$). Consequently, it is intuitive to rescale the estimate $\hat{\Gamma}_n^\epsilon$ in order to ensure that its average eigenvalue remains equal to $\hat{\gamma}_0 = \breve{\gamma}_0$.

Another way to motivate rescaling the corrected matrix estimate is to note that the Yule-Walker equations (5) should be *scale invariant*, i.e., invariant upon changes of $\mathrm{var}\,[X_t]$. In fact, they are often defined via a correlation matrix and vector instead of a covariance matrix and vector. To turn $\hat{\gamma}(n)$ into a vector of correlations, we just divide it by $\hat{\gamma}_0$. Dividing $\hat{\Gamma}_n^*$ by $\hat{\gamma}_0$ should then provide a correlation matrix—hence the need for rescaling.

The rescaled estimate is thus given by

$$\hat{\Gamma}_n^* = c\hat{\Gamma}_n^\epsilon \text{ where } c = \hat{\gamma}_0/\bar{d}^\epsilon \tag{19}$$

and $\bar{d}^\epsilon = n^{-1}\sum_{i=1}^n d_i^\epsilon$ is the average eigenvalue of $\hat{\Gamma}_n^\epsilon$.

4.2. Shrinkage of problematic eigenvalues towards positive definiteness

Section 4.1 described a hard-threshold adjustment to the eigenvalues of $\hat{\Gamma}_n$ in order to render it positive definite. An alternative approach is to make the adjustment based on a positive definite estimate of Γ_n; this approach is novel in the literature of estimating large Toeplitz matrices and/or spectral densities—for the latter see Appendix A.

If the flat top weight function (8) is replaced by a weight function with a positive Fourier transform satisfying $\kappa(0) = 1$, such as Parzen's piecewise cubic lag window (Brockwell and Davis, 1991, p. 361), the resulting estimator $\hat{\Gamma}_n^{pd}$ will be positive definite and consistent—albeit with a slower rate of convergence than $\hat{\Gamma}_n$. Since $\hat{\Gamma}_n^{pd}$ and $\hat{\Gamma}_n$ are both Toeplitz, they are asymptotically diagonalized by the same orthogonal matrix (Grenander and Szegő, 1958). Therefore, letting T_n be the orthogonal matrix from equation (17), the matrix defined as

$$\tilde{D} = T_n'\hat{\Gamma}_n^{pd}T_n$$

will be close to diagonal, and its diagonal entries will approximate the eigenvalues of $\hat{\Gamma}^{pd}$. Let $\tilde{d}_1, \ldots, \tilde{d}_n$ be the diagonals of $\tilde{D}$. We then produce adjusted eigenvalues d_i^* of D (as in (17)) by the following shrinkage rule. Let $d_i^+ = \max\{d_i, 0\}$. Then

$$d_i^\star = \begin{cases} d_i & \text{if} \quad d_i \geq \tilde{d}_i \\ (1-\tau_n)d_i^+ + \tau_n\tilde{d}_i & \text{if} \quad d_i < \tilde{d}_i, \end{cases} \tag{20}$$

where $\tau_n = c/n^a$ for constants $c > 0$ and $a > 1/2$. Let $D^\star$ be a diagonal matrix with diagonal elements $d_1^\star, \ldots, d_n^\star$, and define the shrinkage estimator

$$\hat{\Gamma}_n^\star = T_n D^\star T_n'$$

that is positive definite, and maintains the same asymptotic properties as $\hat{\Gamma}_n$ as long as the constant a in (20) is greater than 1/2. However, if a is chosen too large, the shrinkage correction will be ineffective for small samples. Finally, note that a rescaling step as given in eq. (19) must be performed here as well; hence, our final estimator is given by

$$\hat{\Gamma}_n^* = c\hat{\Gamma}_n^\star \text{ where } c = \hat{\gamma}_0/\bar{d}^\star \tag{21}$$

and $\bar{d}^\star = n^{-1}\sum_{i=1}^n d_i^\star$ is the average eigenvalue of $\hat{\Gamma}_n^\star$.

Remark 9. Shrinking the PSO predictor towards a positive definite estimator is not expected to perform well when $p_n << n$; this is because $\hat{\Gamma}_{p_n}^{pd}$ and $\hat{\Gamma}_{p_n}$ are less close to being diagonalizable by the same orthogonal matrix when p_n is not large.

4.3. Shrinkage towards white noise

Section 4.2 proposed shrinking $\hat{\Gamma}_n$ towards the positive definite estimator $\hat{\Gamma}_n^{pd}$. The shrinking was selective: only problematic eigenvalues were corrected as in the threshold method of Section 4.1. We now propose a different correction that is based on shrinking the corresponding spectral density estimate toward that of a white noise with the same variance—in effect adjusting all eigenvalues. This approach is novel in the literature of estimating large Toeplitz matrices and spectral densities, and provides substantial computational benefits. However, the notion of shrinking covariance matrices towards the identity has been previously employed by Ledoit and Wolf (2003, 2004) in a different context, namely as a tool to regularize the sample covariance matrix based on a sample consisting of multiple i.i.d. copies of a random vector.

Recall that, up to a factor of 2π, the eigenvalues of $\hat{\Gamma}_n$ are asymptotically given by the values of the corresponding spectral density estimate evaluated at the Fourier frequencies; see e.g. Gray (2006). Thus, negative eigenvalues correspond to negative values in the estimated spectral density. The estimated spectral density can be made positive—while keeping $\hat{\gamma}_0$ fixed—by shrinking $\hat{\gamma}_i$ (for $i \neq 0$) towards zero by a constant factor $s \in (0, 1]$, chosen to ensure that the minimum of the estimated spectral density is greater or equal to $\epsilon\hat{\gamma}_0/(2\pi n^\beta)$. To elaborate, if the minimum of the estimated spectral density happens to be greater or equal to $\epsilon\hat{\gamma}_0/(2\pi n^\beta)$, then no correction is needed; if not, then s is chosen so that the the minimum of the corrected spectral density is exactly equal to $\epsilon\hat{\gamma}_0/(2\pi n^\beta)$. See Appendix A for more details.

The same adjustment can be applied to the estimated autocovariance matrix, resulting in the shrinkage corrected version of $\hat{\Gamma}_n$ given by

$$\hat{\Gamma}_n^* = s\hat{\Gamma}_n + (1 - s)\hat{\gamma}_0 I_n, \tag{22}$$

where I_n is the identity matrix and $s \in (0, 1]$. If all the eigenvalues d_i are greater or equal to $\epsilon\hat{\gamma}_0/n^\beta$, then we let $s = 1$. Otherwise, we let s be the maximum value that ensures that the minimum eigenvalue of $\hat{\Gamma}_n^*$ is exactly equal to $\epsilon\hat{\gamma}_0/n^\beta$.

Estimator (22) has several appealing properties. First, it keeps the estimated variance of the process fixed to $\hat{\gamma}_0$, i.e., there is no need for rescaling. Second, the shrinkage estimator $\hat{\Gamma}_n^*$ remains banded and Toeplitz, so fast, memory efficient Toeplitz equation solving algorithms (Brent, Gustavson and Yun, 1980), can always be used. Third, the estimate itself does not require numerical diagonalization of $\hat{\Gamma}_n$ since s can be estimated by evaluating the corresponding spectral density estimate.

4.4. Shrinkage towards a 2nd order estimate

Section 4.2 suggested shrinking the smaller eigenvalues of $\hat{\Gamma}_n$ towards a second order target. Section 4.3 introduced the idea of shrinking all the eigenvalues of $\hat{\Gamma}_n$ towards those of a white noise process. An approach that combines the most appealing features of these methods is to shrink the whole of $\hat{\Gamma}_n$ towards a positive definite, 2nd order estimate of Γ_n.

Let $\hat{\Gamma}_n^{pd}$ be as defined in Section 4.2, and define the corrected estimator by

$$\hat{\Gamma}_n^* = s\hat{\Gamma}_n + (1-s)\hat{\Gamma}_n^{pd}. \tag{23}$$

The shrinkage factor $s \in [0,1]$ is chosen to raise the minimum eigenvalue of $\hat{\Gamma}_n$ as close as possible to $\epsilon\hat{\gamma}_0/n^\beta$ while keeping s in the desired range. Our algorithm exploits the connection between Toeplitz matrices and the spectral density and is described in detail in Appendix A.3.

Estimator (23) is particularly appealing because $\hat{\Gamma}_n^*$ remains banded and Toeplitz, and thus can be inverted via a fast algorithm. In addition, $\hat{\Gamma}_n^*$ has no need for rescaling as it has $\breve{\gamma}_0$ on the main diagonal. Finally, using the second order estimator as the target feels less arbitrary than shrinking towards white noise. But the reason that both corrections work well, both asymptotically and in simulations, is that the correction is a small one, i.e., s tends to one in large samples; thus, the target is not meant to be achieved but gives only a general direction for the correction—see the Appendix for more discussion in the spectral density analog.

4.5. Remarks on the four correction methods

Let $\hat{\Gamma}_n^*$ denote the corrected (and rescaled—if needed) matrix according to one of the methods presented in Sections 4.1, 4.2, 4.3, or 4.4. By construction, $\hat{\Gamma}_n^*$ is positive definite but maintains the same fast asymptotic rate of convergence as $\hat{\Gamma}_n$. The proof of the following corollary is similar to the proofs of Corollaries 2 and 3 in McMurry and Politis (2010).

Corollary 3. *Under Assumptions 1–4, the estimator $\hat{\Gamma}_n^*$ satisfies*

$$\left|\left|\hat{\Gamma}_n^* - \Gamma_n\right|\right|_2 = O_p(r_n) \text{ and } \left|\left|(\hat{\Gamma}_n^*)^{-1} - \Gamma_n^{-1}\right|\right|_2 = O_p(r_n)$$

where $||\cdot||_2$ denotes operator norm, and r_n is as given in eq. (14).

In addition, the positive definite estimator $\hat{\Gamma}_n^*$ may find other applications when a consistent estimator of Γ_n^{-1} is needed. For example, in the aforementioned Linear Process Bootstrap of McMurry and Politis (2010) the un-scaled threshold estimator $\hat{\Gamma}_n^\epsilon$ discussed in Section 4.1 was employed. We conjecture that using the rescaled estimator $\hat{\Gamma}_n^*$ of eq. (19) may improve the finite-sample performance of the Linear Process Bootstrap by better capturing/preserving the scale of the problem. In addition, the estimators $\hat{\Gamma}_n^*$ from Sections 4.2, 4.3, and 4.4 are directly applicable to the Linear Process Bootstrap, and may also also improve its performance.

Remark 10. The problem of estimating high-dimensional covariance and/or precision matrices emerges under many settings different from ours; see e.g. Basu and Michailidis (2014), Bickel and Levina (2008a,b), Cai and Liu (2011), Cai and Zhou (2012a,b), and Chen, Xu and Wu (2013). As the method of shrinkage towards the identity has been found useful by Ledoit and Wolf (2003, 2004), it is conjectured that the above four methods of correcting a matrix towards positive definiteness may be found useful in different such contexts.

Remark 11. Among the four correction methods, the two shrinkage estimators, namely estimators (22) and (23), may prove especially useful in the case of very large data sets. The reason is that they both result in a banded Toeplitz matrix that can be calculated easily, stored efficiently, and inverted via fast algorithms. Recall that the system $Tb = z$ with T Toeplitz can be solved in $O(n \log^2 n)$ time using $O(n)$ memory (Brent, Gustavson and Yun, 1980).

5. Simulations and numerical experiments

We tried a variety of simulation experiments. For each simulated data set, we used the first n observations to predict the $n+1$'st observation. Each prediction was made using 16 approaches:

- The FSO predictor with the threshold correction to positive definiteness described in Section 4.1 together with rescaling to keep the average eigenvalue the same. Two versions of $\hat{\gamma}(n)$ were considered: The first was the version given by (7), and the second given by the first row of $\hat{\Gamma}_n^*$, i.e., $([\hat{\Gamma}_n^*]_{1,2}, \ldots, [\hat{\Gamma}_n^*]_{1,n-1}, 0)'$; see Section 3.2 for details. In the simulation tables, these estimates are denoted FSO-Th-Raw and FSO-Th-Shr respectively.
- The FSO predictor with shrinkage to positive definiteness described in Section 4.2. Both raw and shrunken versions of $\hat{\gamma}_n$ were considered. These predictions are denoted respectively FSO-PD-Raw and FSO-PD-Shr.
- The FSO predictor with shrinkage towards white noise, as described in Section 4.3. Both raw and shrunken versions of $\hat{\gamma}_n$ were considered. In the simulation tables, these estimates are denoted respectively FSO-WN-Raw and FSO-WN-Shr.
- The FSO predictor with shrinkage towards a 2nd order estimate described in Section 4.4. Both raw and shrunken versions of $\hat{\gamma}_n$ were considered. In the simulation tables, these estimates are denoted respectively FSO-2o-Raw and FSO-2o-Shr.
- The FSO predictor using a rectangular weight function $\kappa(\cdot)$ along with either the adaptive bandwidth choice (ABC) described in Section 2.3 or the subsampling bandwidth choice (SSBC) described in Wu and Pourahmadi (2009); both bandwidth choices were considered in combination with raw and shrunken versions of $\hat{\gamma}_n$. These estimates are denoted Rect-ABC-Raw, Rect-ABC-Shr, Rect-SSBC-Raw, and Rect-SSBC-Shr. All matrices were corrected to positive definiteness by shrinkage to white noise.
- The PSO predictor with a threshold correction and $p_n = (np_{aic})^{1/2}$ together with the raw estimate of $\hat{\gamma}_n$; this estimator is denoted PSO-Th-Raw. We found the shrunken version of $\hat{\gamma}_n$ preformed erratically in this setting, and the results are omitted. Note that p_{aic} denoted the AR order chosen by minimization of the AIC criterion.
- The PSO predictor with shrinkage towards white noise and $p_n = (np_{aic})^{1/2}$ together with the shrunken estimate of $\hat{\gamma}_n$. This estimator is denoted PSO-WN-Shr.

- An AR(p_{aic}) prediction with p_{aic} chosen by AIC minimization, denoted AR.
- A version of the method described in Bickel and Gel (2011), denoted BG, with $p_n = n^{1/2}$.

Accuracy of all predictions is described by root mean square prediction error (RMSPE) taken across all simulations. The trapezoidal taper of equation (8) was used throughout.

The predictor described in Bickel and Gel (2011) has coefficients given by

$$\hat{\phi}^{BG}(n) = \left(\hat{\Gamma}^k_{p_n}\right)^{-1} \breve{\gamma}(p_n),$$

where $\hat{\Gamma}^k_{p_n}$ is the k-banded version of the $p_n \times p_n$ autocovariance matrix with $p_n = o(n)$, and $\breve{\gamma}(p_n)$ is the vector of autocovariances at lags $1, \ldots, p_n$. In our simulations, we found that $\hat{\Gamma}^k_{p_n}$ was frequently not positive definite. Bickel and Gel (2011) recommend either considering a reduced set of banding parameters k or tapering $\breve{\Gamma}_{p_n}$ with a positive definite taper (Xiao and Wu, 2012). We found the first approach to produce unstable predictions, so we focused on the second. In particular, we tapered the entries of $\breve{\Gamma}_{p_n}$ by the Parzen piecewise cubic lag window (Brockwell and Davis, 1991, p. 361), and chose the width of the lag window by cross-validation over the values from 1 to $3p_n$ plus ∞ (no tapering).

For the implementation of the FSO predictor (10), the employed threshold correction used constants $\epsilon = 20$ and $\beta = 1$. When shrinking towards a positive definite estimator, we used constants $c = 6$ and $a = 0.55$. For shrinkage towards white noise, we scaled the off-diagonals of $\hat{\Gamma}_n$ until the smallest eigenvalue was at least $\max\{10\hat{\gamma}_0/n, 0.5 \times \lambda_{\min}(\hat{\Gamma}^{pd}_n)\}$. Shrinkage towards a 2nd order estimate used threshold of $10\hat{\gamma}_0/n$. All second order estimates used the Parzen piecewise cubic lag window and bandwidth chosen by the plug-in approach proposed by Politis (2003).

5.1. AR(1) prediction

For the first experiment, we simulated AR(1) time series of length 201 and used the first 200 data points to predict the 201'st. Each simulation was repeated 1000 times, and the root mean square prediction errors are shown in Table 1. This simulation should favor the AR predictor (3) since it directly fits an AR model.

The banding of $\hat{\Gamma}_n$ implies that the FSO predictor (10) is based on a model whose autocovariances vanish for lags bigger than $2l$, in effect an MA model of order $2l$. Hence, if a dataset can be well approximated by a low-order AR model, it is expected that the AR predictor (3) will have an advantage over a method that is trying to approximate the low-order AR by a high-order MA.

Table 1 shows that FSO predictor (10) is competitive (and even better) than the AR(p) model for small values of the AR coefficient but becomes less competitive as the AR coefficient becomes larger; this is not surprising since accurate approximation of such models by a moving average would require a very high order MA model.

TABLE 1
Root mean square prediction errors (standard error in parentheses) for AR(1) processes

	$\phi = -0.9$	$\phi = -0.5$	$\phi = -0.1$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
FSO-Th-Raw	1.126 (0.027)	1.037 (0.022)	0.992 (0.021)	1.045 (0.023)	1.052 (0.024)	1.117 (0.028)
FSO-Th-Shr	1.091 (0.024)	1.035 (0.022)	0.992 (0.021)	1.045 (0.023)	1.050 (0.024)	1.081 (0.026)
FSO-PD-Raw	1.129 (0.024)	1.027 (0.022)	0.992 (0.021)	1.045 (0.023)	1.037 (0.024)	1.101 (0.026)
FSO-PD-Shr	1.105 (0.024)	1.021 (0.021)	0.991 (0.021)	1.045 (0.023)	1.032 (0.023)	1.078 (0.025)
FSO-WN-Raw	1.112 (0.027)	1.024 (0.021)	0.992 (0.021)	1.045 (0.023)	1.029 (0.023)	1.106 (0.028)
FSO-WN-Shr	1.114 (0.024)	1.023 (0.022)	0.992 (0.021)	1.045 (0.023)	1.028 (0.023)	1.090 (0.025)
FSO-2o-Raw	1.090 (0.024)	1.031 (0.022)	0.992 (0.021)	1.045 (0.023)	1.051 (0.024)	1.051 (0.024)
FSO-2o-Shr	1.066 (0.023)	1.029 (0.022)	0.992 (0.021)	1.045 (0.023)	1.049 (0.024)	1.037 (0.024)
Rect-ABC-Raw	1.649 (0.060)	1.041 (0.022)	0.992 (0.021)	1.045 (0.023)	1.042 (0.023)	1.455 (0.040)
Rect-ABC-Shr	1.360 (0.031)	1.038 (0.022)	0.992 (0.021)	1.045 (0.023)	1.039 (0.023)	1.316 (0.031)
Rect-SSBC-Raw	1.906 (0.070)	1.035 (0.022)	0.992 (0.021)	1.046 (0.023)	1.037 (0.023)	2.060 (0.062)
Rect-SSBC-Shr	1.396 (0.031)	1.036 (0.022)	0.992 (0.021)	1.045 (0.023)	1.033 (0.023)	1.466 (0.035)
PSO-Th-Raw	1.046 (0.023)	1.036 (0.022)	0.992 (0.021)	1.045 (0.023)	1.052 (0.024)	1.049 (0.025)
PSO-Sh-Shr	1.079 (0.024)	1.023 (0.022)	0.992 (0.021)	1.045 (0.023)	1.027 (0.023)	1.071 (0.025)
AR	1.023 (0.023)	1.009 (0.021)	0.991 (0.021)	1.046 (0.023)	1.008 (0.022)	1.004 (0.023)
BG	1.109 (0.030)	1.043 (0.022)	1.024 (0.022)	1.074 (0.024)	1.051 (0.022)	1.108 (0.028)

TABLE 2
Root mean square prediction errors (standard error in parentheses) for MA(1) processes

	$\theta = -0.9$	$\theta = -0.5$	$\theta = -0.1$	$\theta = 0.1$	$\theta = 0.5$	$\theta = 0.9$
FSO-Th-Raw	1.079 (0.026)	1.057 (0.024)	0.991 (0.023)	0.980 (0.022)	1.004 (0.023)	1.052 (0.024)
FSO-Th-Shr	1.075 (0.026)	1.056 (0.024)	0.991 (0.023)	0.980 (0.022)	1.004 (0.023)	1.049 (0.024)
FSO-PD-Raw	1.084 (0.026)	1.054 (0.024)	0.991 (0.023)	0.980 (0.022)	1.000 (0.023)	1.068 (0.025)
FSO-PD-Shr	1.076 (0.026)	1.052 (0.024)	0.991 (0.023)	0.980 (0.022)	0.998 (0.022)	1.056 (0.024)
FSO-WN-Raw	1.059 (0.026)	1.053 (0.024)	0.991 (0.023)	0.980 (0.022)	0.998 (0.023)	1.038 (0.023)
FSO-WN-Shr	1.059 (0.026)	1.053 (0.024)	0.991 (0.023)	0.980 (0.022)	0.997 (0.023)	1.042 (0.023)
FSO-2o-Raw	1.060 (0.026)	1.055 (0.024)	0.991 (0.023)	0.980 (0.022)	1.003 (0.023)	1.040 (0.023)
FSO-2o-Shr	1.060 (0.026)	1.055 (0.024)	0.991 (0.023)	0.980 (0.022)	1.002 (0.023)	1.041 (0.023)
Rect-ABC-Raw	1.061 (0.026)	1.054 (0.024)	0.991 (0.023)	0.980 (0.022)	0.998 (0.023)	1.042 (0.023)
Rect-ABC-Shr	1.060 (0.026)	1.053 (0.024)	0.991 (0.023)	0.980 (0.022)	0.996 (0.023)	1.047 (0.023)
Rect-SSBC-Raw	1.064 (0.026)	1.055 (0.024)	0.991 (0.023)	0.983 (0.022)	0.996 (0.023)	1.040 (0.023)
Rect-SSBC-Shr	1.067 (0.027)	1.055 (0.024)	0.990 (0.023)	0.983 (0.022)	0.996 (0.023)	1.044 (0.023)
PSO-Th-Raw	1.079 (0.026)	1.057 (0.024)	0.991 (0.023)	0.980 (0.022)	1.004 (0.023)	1.052 (0.024)
PSO-Sh-Shr	1.059 (0.026)	1.053 (0.024)	0.991 (0.023)	0.980 (0.022)	0.997 (0.023)	1.042 (0.023)
AR	1.070 (0.026)	1.059 (0.024)	0.993 (0.023)	0.987 (0.022)	1.000 (0.022)	1.047 (0.023)
BG	1.071 (0.025)	1.081 (0.024)	1.007 (0.023)	1.012 (0.023)	1.029 (0.023)	1.054 (0.023)

Standard errors for the RMSPE estimates are shown in parentheses in Table 1. The standard errors for the differences in RMSPE between our methods and the AR predictions tend to decrease with the magnitude of the AR parameter. When the AR parameter is -0.1 or 0.1, the standard errors for these differences tend to be around 0.005. When the AR parameter is -0.5 or 0.5, the standard errors for these differences tend to be around 0.009. When the AR parameter is -0.9 or 0.9, the standard errors for these differences tend to be around 0.014.

5.2. MA(1) prediction

For the second experiment, we simulated time series of length 201 and used the first 200 data points to predict the 201'st. Each simulation was repeated 1000 times, and the root mean square prediction errors are shown in Table 2. This simulation should favor the FSO predictor (10) since it estimates the correlation structure of an MA model directly.

Note that the FSO predictor (10) is competitive with the AR(p) model for all values of the MA coefficient, and frequently shows slightly better performance.

Standard errors for the RMSPEs are shown in parentheses in Table 2. Standard errors for the differences in RMSPEs between our methods and the AR(p) predictions were approximately 0.004 when the MA parameters were -0.1 or 0.1, 0.006 when the MA parameters were -0.5 or 0.5, and 0.010 when the MA parameters were -0.9 or 0.9.

Remark 12. The AR(1) simulations suggest that the shrunken estimate of γ_n tends to outperform the raw estimate when the AR parameter is large. The improvement seems to come at little to no cost, so the shrunken estimate seems advisable in practice; further evidence to support this point is provided in Section 5.6.

5.3. MA(2) prediction

In the next simulation, we considered a wide range of MA(2) processes, with coefficients θ_1 and θ_2 ranging from -1 to 1 in steps of 1/3. Several of these combinations of parameters, for example $\theta_1 = -1$ and $\theta_2 = 0$, have MA polynomials with roots on the unit circle causing the spectral density to have a corresponding zero. These simulations are expected to cause trouble for all approaches to prediction since Γ_n is not invertible for large n, but the troubles have been somewhat masked by the correction to positive definiteness.

For the simulation, 1000 data sets of sizes 101 and 501 were generated for each combination of θ_1 and θ_2, and the final observation was predicted using all preceding observations; results are given in Tables 3 and 4. Note that for the larger sample size, one of our estimators was the best performing in more than half of the MA(2) cases under consideration. For the smaller sample size, our estimators were consistently competitive except for the case $\theta_2 = 1$.

Standard errors for the RMSPEs in the MA(2) case were again consistently close to 0.024. The five-number summary for the simulations of size 100 was: min $= 0.021$, $Q_1 = 0.024$, $Q_2 = 0.026$, $Q_3 = 0.028$, max $= 0.051$. Standard errors for the simulations of size 500 were almost identical.

Remark 13. The shrinkage corrected FSO and PSO predictors produced very similar results across most of the simulations; this is not surprising since in general $\hat{\phi}_i(n)$ decays rapidly as i increases. As long as $p_n \gg l$, the coefficients of the FSO and PSO predictors that are significantly different from zero agree almost exactly.

5.4. Real data experiment

For our final prediction experiment, we tried our methods on real data using time series from the M3 competition database (Hyndman, Akram and Bergmeir, 2013). In order to avoid the need for seasonal adjustment, we first selected only those time series measured yearly. We then further restricted to those time series which were not found to be nonstationary, i.e., for which the test of Kwiatkowski et al. (1992) could not reject the null hypothesis of absence of a unit root at the $\alpha = 0.05$ level. The end result was 105 time series of lengths between 20 and 47.

TABLE 3. *Root mean square prediction errors for MA(2) process with* $n = 100$

	$\theta_2 = -1$	$\theta_2 = -2/3$	$\theta_2 = -1/3$	$\theta_2 = 0$	$\theta_2 = 1/3$	$\theta_2 = 2/3$	$\theta_2 = 1$
FSO-Th-Raw $\theta_1 = -1$	1.703	1.615	1.390	1.166	$\boxed{1.045}$	1.185	1.275
FSO-Th-Shr $\theta_1 = -1$	1.699	1.613	1.389	1.164	1.054	1.144	1.224
FSO-PD-Raw $\theta_1 = -1$	1.699	1.613	1.391	1.163	1.271	1.637	1.573
FSO-PD-Shr $\theta_1 = -1$	1.695	1.609	1.386	1.151	1.133	1.384	1.382
FSO-WN-Raw $\theta_1 = -1$	1.695	1.614	1.388	1.148	1.062	1.184	1.234
FSO-WN-Shr $\theta_1 = -1$	1.690	1.612	1.387	1.153	1.119	1.198	1.234
FSO-2o-Raw $\theta_1 = -1$	1.696	1.616	1.389	$\boxed{1.145}$	1.056	1.206	1.280
FSO-2o-Shr $\theta_1 = -1$	1.691	1.613	1.385	1.151	1.067	1.125	1.193
Rect-ABC-Raw $\theta_1 = -1$	1.692	1.617	1.389	1.153	1.079	1.194	1.244
Rect-ABC-Shr $\theta_1 = -1$	$\boxed{1.684}$	1.613	1.387	1.157	1.134	1.209	1.242
Rect-SSBC-Raw $\theta_1 = -1$	1.708	1.563	1.364	1.187	1.139	1.159	1.248
Rect-SSBC-Shr $\theta_1 = -1$	1.692	$\boxed{1.549}$	1.358	1.188	1.146	1.174	1.250
PSO-Th-Raw $\theta_1 = -1$	1.702	1.614	1.389	1.165	1.046	1.184	1.272
PSO-Sh-Shr $\theta_1 = -1$	1.689	1.611	1.387	1.153	1.119	1.197	1.233
AR $\theta_1 = -1$	1.698	1.558	$\boxed{1.347}$	1.153	1.051	$\boxed{1.089}$	$\boxed{1.135}$
BG $\theta_1 = -1$	1.714	1.569	1.360	1.145	1.088	1.091	1.159
FSO-Th-Raw $\theta_1 = -2/3$	1.402	1.289	1.195	1.051	1.123	1.174	1.202
FSO-Th-Shr $\theta_1 = -2/3$	1.398	1.288	1.197	1.051	1.085	1.140	1.200
FSO-PD-Raw $\theta_1 = -2/3$	1.407	1.290	1.193	1.057	1.188	1.216	1.193
FSO-PD-Shr $\theta_1 = -2/3$	1.404	1.288	1.189	1.054	1.127	1.166	1.176
FSO-WN-Raw $\theta_1 = -2/3$	1.400	1.288	1.194	1.046	1.081	1.135	1.183
FSO-WN-Shr $\theta_1 = -2/3$	$\boxed{1.397}$	1.289	1.196	1.049	1.075	1.122	1.186
FSO-2o-Raw $\theta_1 = -2/3$	1.404	1.287	1.193	$\boxed{1.044}$	1.116	1.167	1.193
FSO-2o-Shr $\theta_1 = -2/3$	1.401	1.289	1.194	1.045	1.071	1.122	1.187
Rect-ABC-Raw $\theta_1 = -2/3$	1.405	1.287	1.196	1.045	1.091	1.137	1.192
Rect-ABC-Shr $\theta_1 = -2/3$	1.404	1.289	1.198	1.047	1.083	1.125	1.194
Rect-SSBC-Raw $\theta_1 = -2/3$	1.427	1.286	1.159	1.083	1.066	1.112	1.209
Rect-SSBC-Shr $\theta_1 = -2/3$	1.414	$\boxed{1.276}$	1.157	1.082	1.061	1.103	1.208
PSO-Th-Raw $\theta_1 = -2/3$	1.404	1.287	1.195	1.051	1.121	1.173	1.202
PSO-Sh-Shr $\theta_1 = -2/3$	1.397	1.288	1.196	1.048	1.075	1.122	1.186
AR $\theta_1 = -2/3$	1.414	1.290	$\boxed{1.136}$	1.062	$\boxed{1.035}$	$\boxed{1.079}$	$\boxed{1.122}$
BG $\theta_1 = -2/3$	1.420	1.297	1.140	1.077	1.068	1.101	1.129
FSO-Th-Raw $\theta_1 = -1/3$	1.242	1.109	1.129	1.009	1.044	1.068	1.190
FSO-Th-Shr $\theta_1 = -1/3$	1.243	1.106	1.128	1.008	1.042	1.063	1.196
FSO-PD-Raw $\theta_1 = -1/3$	1.242	1.112	1.135	1.009	1.034	1.072	1.183
FSO-PD-Shr $\theta_1 = -1/3$	1.239	1.107	1.132	1.006	1.026	1.065	1.186
FSO-WN-Raw $\theta_1 = -1/3$	1.236	1.099	1.127	1.008	1.036	1.057	1.188
FSO-WN-Shr $\theta_1 = -1/3$	1.244	1.103	1.126	1.008	1.033	1.053	1.193
FSO-2o-Raw $\theta_1 = -1/3$	$\boxed{1.234}$	1.100	1.128	1.009	1.043	1.057	$\boxed{1.182}$

Table 3. *(continued)*

	$\theta_2 = -1$	$\theta_2 = -2/3$	$\theta_2 = -1/3$	$\theta_2 = 0$	$\theta_2 = 1/3$	$\theta_2 = 2/3$	$\theta_2 = 1$
FSO-2o-Shr $\theta_1 = -1/3$	1.237	$\boxed{1.097}$	1.126	1.008	1.041	1.051	1.192
Rect-ABC-Raw $\theta_1 = -1/3$	1.242	1.103	1.130	1.010	1.038	1.057	1.196
Rect-ABC-Shr $\theta_1 = -1/3$	1.251	1.107	1.128	1.008	1.034	$\boxed{1.050}$	1.199
Rect-SSBC-Raw $\theta_1 = -1/3$	1.276	1.120	1.098	1.004	1.029	1.069	1.229
Rect-SSBC-Shr $\theta_1 = -1/3$	1.278	1.118	$\boxed{1.094}$	0.999	1.020	1.061	1.229
PSO-Th-Raw $\theta_1 = -1/3$	1.242	1.109	1.129	1.009	1.044	1.068	1.190
PSO-Sh-Shr $\theta_1 = -1/3$	1.244	1.103	1.126	1.008	1.032	1.054	1.194
AR $\theta_1 = -1/3$	1.269	1.121	1.095	$\boxed{0.998}$	$\boxed{1.003}$	1.052	1.188
BG $\theta_1 = -1/3$	1.258	1.111	1.108	1.017	1.043	1.060	1.190
FSO-Th-Raw $\theta_1 = 0$	1.181	1.025	1.031	0.996	1.021	1.059	1.140
FSO-Th-Shr $\theta_1 = 0$	1.187	1.023	1.030	0.996	1.021	1.059	1.147
FSO-PD-Raw $\theta_1 = 0$	1.167	1.035	1.031	0.996	1.019	1.058	1.135
FSO-PD-Shr $\theta_1 = 0$	$\boxed{1.163}$	1.032	1.029	0.996	$\boxed{1.018}$	1.060	1.136
FSO-WN-Raw $\theta_1 = 0$	1.175	$\boxed{1.022}$	1.029	0.996	1.020	1.050	1.131
FSO-WN-Shr $\theta_1 = 0$	1.186	1.027	1.028	$\boxed{0.995}$	1.020	1.052	1.139
FSO-2o-Raw $\theta_1 = 0$	1.170	1.024	1.031	0.996	1.021	$\boxed{1.049}$	$\boxed{1.127}$
FSO-2o-Shr $\theta_1 = 0$	1.177	1.024	1.030	0.996	1.021	1.050	1.132
Rect-ABC-Raw $\theta_1 = 0$	1.190	1.026	1.032	0.996	1.023	1.050	1.143
Rect-ABC-Shr $\theta_1 = 0$	1.200	1.032	1.029	0.996	1.021	1.050	1.147
Rect-SSBC-Raw $\theta_1 = 0$	1.226	1.045	1.043	1.015	1.032	1.085	1.186
Rect-SSBC-Shr $\theta_1 = 0$	1.228	1.047	1.035	1.013	1.026	1.077	1.189
PSO-Th-Raw $\theta_1 = 0$	1.181	1.023	1.030	0.996	1.021	1.059	1.141
PSO-Sh-Shr $\theta_1 = 0$	1.186	1.027	1.028	0.995	1.020	1.052	1.140
AR $\theta_1 = 0$	1.189	1.057	$\boxed{1.015}$	1.009	1.024	1.080	1.146
BG $\theta_1 = 0$	1.179	1.050	1.040	1.038	1.052	1.089	1.138
FSO-Th-Raw $\theta_1 = 1/3$	1.190	1.114	1.091	1.037	1.087	1.068	1.172
FSO-Th-Shr $\theta_1 = 1/3$	1.191	1.114	1.091	1.036	1.085	1.065	1.177
FSO-PD-Raw $\theta_1 = 1/3$	1.187	1.109	1.092	1.036	1.077	1.073	1.167
FSO-PD-Shr $\theta_1 = 1/3$	1.185	1.106	1.089	1.036	1.068	1.068	1.162
FSO-WN-Raw $\theta_1 = 1/3$	1.185	1.106	1.089	1.037	1.078	1.062	1.164
FSO-WN-Shr $\theta_1 = 1/3$	1.190	1.107	1.090	1.036	1.074	1.061	1.173
FSO-2o-Raw $\theta_1 = 1/3$	$\boxed{1.183}$	1.106	1.088	1.038	1.087	1.062	1.163
FSO-2o-Shr $\theta_1 = 1/3$	1.188	1.109	1.088	1.037	1.085	$\boxed{1.058}$	1.171
Rect-ABC-Raw $\theta_1 = 1/3$	1.194	1.112	1.090	1.040	1.084	1.071	1.173
Rect-ABC-Shr $\theta_1 = 1/3$	1.199	1.113	1.091	1.040	1.080	1.069	1.179
Rect-SSBC-Raw $\theta_1 = 1/3$	1.221	1.117	1.070	1.033	1.050	1.103	1.197
Rect-SSBC-Shr $\theta_1 = 1/3$	1.219	1.116	1.062	$\boxed{1.028}$	1.045	1.091	1.205
PSO-Th-Raw $\theta_1 = 1/3$	1.188	1.114	1.091	1.037	1.087	1.067	1.172

TABLE 3. *(continued)*

	$\theta_2 = -1$	$\theta_2 = -2/3$	$\theta_2 = -1/3$	$\theta_2 = 0$	$\theta_2 = 1/3$	$\theta_2 = 2/3$	$\theta_2 = 1$
PSO-Sh-Shr $\theta_1 = 1/3$	1.190	1.107	1.090	1.036	1.074	1.061	1.173
AR $\theta_1 = 1/3$	1.223	1.113	1.061	1.029	1.032	1.069	1.164
BG $\theta_1 = 1/3$	1.210	1.104	1.068	1.052	1.056	1.085	1.156
FSO-Th-Raw $\theta_1 = 2/3$	1.505	1.313	1.209	1.011	1.122	1.150	1.260
FSO-Th-Shr $\theta_1 = 2/3$	1.502	1.313	1.210	1.009	1.093	1.130	1.255
FSO-PD-Raw $\theta_1 = 2/3$	1.506	1.314	1.207	1.012	1.176	1.171	1.249
FSO-PD-Shr $\theta_1 = 2/3$	1.503	1.312	1.201	1.006	1.122	1.139	1.231
FSO-WN-Raw $\theta_1 = 2/3$	1.499	1.307	1.205	1.002	1.092	1.126	1.243
FSO-WN-Shr $\theta_1 = 2/3$	1.497	1.305	1.207	1.001	1.093	1.122	1.246
FSO-2o-Raw $\theta_1 = 2/3$	1.501	1.307	1.204	1.002	1.120	1.155	1.250
FSO-2o-Shr $\theta_1 = 2/3$	1.497	1.306	1.204	0.999	1.088	1.121	1.238
Rect-ABC-Raw $\theta_1 = 2/3$	1.499	1.310	1.206	1.001	1.097	1.130	1.253
Rect-ABC-Shr $\theta_1 = 2/3$	1.496	1.307	1.208	1.000	1.095	1.122	1.253
Rect-SSBC-Raw $\theta_1 = 2/3$	1.524	1.302	1.156	1.027	1.107	1.100	1.250
Rect-SSBC-Shr $\theta_1 = 2/3$	1.518	1.293	1.154	1.026	1.104	1.091	1.251
PSO-Th-Raw $\theta_1 = 2/3$	1.503	1.312	1.208	1.012	1.120	1.150	1.260
PSO-Sh-Shr $\theta_1 = 2/3$	1.497	1.305	1.207	1.001	1.093	1.122	1.246
AR $\theta_1 = 2/3$	1.503	1.302	1.141	1.016	1.071	1.078	1.171
BG $\theta_1 = 2/3$	1.536	1.306	1.137	1.030	1.098	1.098	1.175
FSO-Th-Raw $\theta_1 = 1$	1.639	1.502	1.391	1.147	1.057	1.146	1.379
FSO-Th-Shr $\theta_1 = 1$	1.635	1.500	1.391	1.156	1.053	1.108	1.327
FSO-PD-Raw $\theta_1 = 1$	1.640	1.505	1.397	1.124	1.267	1.619	1.594
FSO-PD-Shr $\theta_1 = 1$	1.635	1.503	1.393	1.124	1.132	1.346	1.445
FSO-WN-Raw $\theta_1 = 1$	1.634	1.502	1.390	1.139	1.046	1.146	1.350
FSO-WN-Shr $\theta_1 = 1$	1.630	1.502	1.390	1.147	1.083	1.166	1.354
FSO-2o-Raw $\theta_1 = 1$	1.639	1.501	1.390	1.132	1.046	1.165	1.391
FSO-2o-Shr $\theta_1 = 1$	1.635	1.500	1.390	1.143	1.049	1.091	1.298
Rect-ABC-Raw $\theta_1 = 1$	1.637	1.503	1.399	1.143	1.063	1.157	1.359
Rect-ABC-Shr $\theta_1 = 1$	1.631	1.502	1.397	1.150	1.097	1.175	1.366
Rect-SSBC-Raw $\theta_1 = 1$	1.647	1.513	1.361	1.195	1.099	1.127	1.290
Rect-SSBC-Shr $\theta_1 = 1$	1.630	1.500	1.355	1.192	1.107	1.141	1.302
PSO-Th-Raw $\theta_1 = 1$	1.638	1.501	1.391	1.146	1.058	1.145	1.377
PSO-Sh-Shr $\theta_1 = 1$	1.629	1.501	1.390	1.147	1.084	1.166	1.354
AR $\theta_1 = 1$	1.629	1.499	1.347	1.135	1.049	1.056	1.204
BG $\theta_1 = 1$	1.666	1.516	1.345	1.138	1.073	1.071	1.195

TABLE 4. *Root mean square prediction errors for MA(2) process with $n = 500$*

	$\theta_2 = -1$	$\theta_2 = -2/3$	$\theta_2 = -1/3$	$\theta_2 = 0$	$\theta_2 = 1/3$	$\theta_2 = 2/3$	$\theta_2 = 1$
FSO-Th-Raw $\theta_1 = -1$	1.631	1.443	1.240	1.106	1.346	1.045	1.097
FSO-Th-Shr $\theta_1 = -1$	1.631	1.443	1.239	1.101	1.216	1.041	1.096
FSO-PD-Raw $\theta_1 = -1$	1.631	1.440	1.238	1.128	2.108	1.066	1.104
FSO-PD-Shr $\theta_1 = -1$	1.631	1.439	1.236	1.115	1.756	1.058	1.100
FSO-WN-Raw $\theta_1 = -1$	1.631	1.439	1.230	1.081	1.020	1.030	1.095
FSO-WN-Shr $\theta_1 = -1$	1.631	1.439	1.231	1.085	1.032	1.028	1.099
FSO-2o-Raw $\theta_1 = -1$	1.631	1.442	1.234	1.075	1.066	1.036	1.084
FSO-2o-Shr $\theta_1 = -1$	1.631	1.442	1.234	1.079	1.018	1.034	1.086
Rect-ABC-Raw $\theta_1 = -1$	1.630	1.439	1.230	1.086	1.037	1.050	1.108
Rect-ABC-Shr $\theta_1 = -1$	1.631	1.438	1.230	1.090	1.048	1.050	1.116
Rect-SSBC-Raw $\theta_1 = -1$	1.633	1.440	1.230	1.079	1.037	1.050	1.104
Rect-SSBC-Shr $\theta_1 = -1$	1.633	1.440	1.230	1.085	1.051	1.050	1.111
PSO-Th-Raw $\theta_1 = -1$	1.631	1.443	1.240	1.106	1.345	1.045	1.097
PSO-Sh-Shr $\theta_1 = -1$	1.631	1.439	1.231	1.085	1.032	1.028	1.099
AR $\theta_1 = -1$	1.651	1.457	1.233	1.071	0.999	1.021	1.079
BG $\theta_1 = -1$	1.670	1.472	1.234	1.075	1.023	1.024	1.077
FSO-Th-Raw $\theta_1 = -2/3$	1.415	1.266	1.056	1.023	1.029	0.998	1.083
FSO-Th-Shr $\theta_1 = -2/3$	1.415	1.265	1.056	1.021	1.017	0.996	1.081
FSO-PD-Raw $\theta_1 = -2/3$	1.410	1.264	1.062	1.024	0.996	1.001	1.100
FSO-PD-Shr $\theta_1 = -2/3$	1.409	1.262	1.061	1.021	0.987	0.997	1.094
FSO-WN-Raw $\theta_1 = -2/3$	1.406	1.260	1.069	1.013	0.962	0.981	1.085
FSO-WN-Shr $\theta_1 = -2/3$	1.403	1.261	1.073	1.013	0.963	0.981	1.088
FSO-2o-Raw $\theta_1 = -2/3$	1.415	1.258	1.055	1.017	0.987	0.985	1.069
FSO-2o-Shr $\theta_1 = -2/3$	1.414	1.259	1.058	1.015	0.978	0.984	1.069
Rect-ABC-Raw $\theta_1 = -2/3$	1.411	1.262	1.073	1.014	0.963	0.986	1.089
Rect-ABC-Shr $\theta_1 = -2/3$	1.408	1.263	1.077	1.013	0.965	0.983	1.092
Rect-SSBC-Raw $\theta_1 = -2/3$	1.410	1.262	1.068	1.015	0.966	0.986	1.087
Rect-SSBC-Shr $\theta_1 = -2/3$	1.407	1.263	1.072	1.014	0.965	0.984	1.091
PSO-Th-Raw $\theta_1 = -2/3$	1.415	1.266	1.056	1.023	1.029	0.998	1.084
PSO-Sh-Shr $\theta_1 = -2/3$	1.403	1.261	1.073	1.013	0.963	0.981	1.088
AR $\theta_1 = -2/3$	1.407	1.270	1.042	1.021	0.967	0.993	1.063
BG $\theta_1 = -2/3$	1.419	1.288	1.046	1.036	0.981	0.997	1.064
FSO-Th-Raw $\theta_1 = -1/3$	1.235	1.073	0.985	0.989	0.971	1.037	1.117
FSO-Th-Shr $\theta_1 = -1/3$	1.230	1.071	0.985	0.989	0.971	1.036	1.115
FSO-PD-Raw $\theta_1 = -1/3$	1.228	1.080	0.983	0.989	0.971	1.030	1.117
FSO-PD-Shr $\theta_1 = -1/3$	1.223	1.074	0.983	0.989	0.970	1.027	1.113
FSO-WN-Raw $\theta_1 = -1/3$	1.201	1.060	0.978	0.989	0.971	1.018	1.121
FSO-WN-Shr $\theta_1 = -1/3$	1.203	1.062	0.977	0.989	0.971	1.016	1.125

TABLE 4. *(continued)*

	$\theta_2 = -1$	$\theta_2 = -2/3$	$\theta_2 = -1/3$	$\theta_2 = 0$	$\theta_2 = 1/3$	$\theta_2 = 2/3$	$\theta_2 = 1$
FSO-2o-Raw $\theta_1 = -1/3$	1.203	1.053	0.983	0.989	0.971	1.030	1.104
FSO-2o-Shr $\theta_1 = -1/3$	1.201	1.055	0.983	0.989	0.971	1.028	1.106
Rect-ABC-Raw $\theta_1 = -1/3$	1.206	1.066	0.979	0.989	0.976	1.020	1.121
Rect-ABC-Shr $\theta_1 = -1/3$	1.210	1.069	0.978	0.989	0.976	1.017	1.126
Rect-SSBC-Raw $\theta_1 = -1/3$	1.209	1.063	0.978	0.990	0.975	1.018	1.123
Rect-SSBC-Shr $\theta_1 = -1/3$	1.213	1.067	0.977	0.990	0.975	1.015	1.127
PSO-Th-Raw $\theta_1 = -1/3$	1.233	1.072	0.985	0.989	0.971	1.037	1.117
PSO-Sh-Shr $\theta_1 = -1/3$	1.203	1.062	0.977	0.989	0.971	1.016	1.125
AR $\theta_1 = -1/3$	1.199	1.063	0.982	0.990	0.972	1.036	1.104
BG $\theta_1 = -1/3$	1.201	1.056	0.996	1.015	0.991	1.051	1.098
FSO-Th-Raw $\theta_1 = 0$	1.124	0.981	0.981	1.011	0.988	1.072	1.155
FSO-Th-Shr $\theta_1 = 0$	1.121	0.979	0.981	1.011	0.988	1.071	1.151
FSO-PD-Raw $\theta_1 = 0$	1.118	0.975	0.981	1.011	0.988	1.067	1.146
FSO-PD-Shr $\theta_1 = 0$	1.112	0.973	0.981	1.011	0.988	1.066	1.138
FSO-WN-Raw $\theta_1 = 0$	1.111	0.965	0.981	1.011	0.988	1.055	1.123
FSO-WN-Shr $\theta_1 = 0$	1.120	0.966	0.981	1.011	0.988	1.055	1.123
FSO-2o-Raw $\theta_1 = 0$	1.094	0.972	0.981	1.011	0.988	1.065	1.112
FSO-2o-Shr $\theta_1 = 0$	1.097	0.971	0.981	1.011	0.988	1.063	1.110
Rect-ABC-Raw $\theta_1 = 0$	1.117	0.966	0.982	1.011	0.989	1.055	1.126
Rect-ABC-Shr $\theta_1 = 0$	1.129	0.968	0.982	1.011	0.989	1.056	1.127
Rect-SSBC-Raw $\theta_1 = 0$	1.114	0.965	0.982	1.011	0.990	1.056	1.126
Rect-SSBC-Shr $\theta_1 = 0$	1.126	0.968	0.982	1.011	0.990	1.056	1.127
PSO-Th-Raw $\theta_1 = 0$	1.124	0.980	0.981	1.011	0.988	1.072	1.155
PSO-Sh-Shr $\theta_1 = 0$	1.120	0.966	0.981	1.011	0.988	1.055	1.123
AR $\theta_1 = 0$	1.068	0.976	0.986	1.012	0.990	1.070	1.112
BG $\theta_1 = 0$	1.070	0.982	1.008	1.032	1.010	1.079	1.113
FSO-Th-Raw $\theta_1 = 1/3$	1.234	1.058	1.003	0.988	1.051	1.018	1.099
FSO-Th-Shr $\theta_1 = 1/3$	1.229	1.057	1.002	0.988	1.051	1.017	1.095
FSO-PD-Raw $\theta_1 = 1/3$	1.237	1.055	1.001	0.988	1.051	1.011	1.099
FSO-PD-Shr $\theta_1 = 1/3$	1.230	1.052	1.001	0.988	1.050	1.008	1.092
FSO-WN-Raw $\theta_1 = 1/3$	1.205	1.050	0.996	0.988	1.051	0.998	1.086
FSO-WN-Shr $\theta_1 = 1/3$	1.206	1.052	0.996	0.988	1.051	0.998	1.089
FSO-2o-Raw $\theta_1 = 1/3$	1.210	1.044	1.000	0.988	1.051	1.008	1.071
FSO-2o-Shr $\theta_1 = 1/3$	1.207	1.044	1.000	0.988	1.051	1.007	1.073
Rect-ABC-Raw $\theta_1 = 1/3$	1.214	1.056	0.995	0.988	1.056	0.995	1.091
Rect-ABC-Shr $\theta_1 = 1/3$	1.214	1.060	0.994	0.988	1.055	0.995	1.093
Rect-SSBC-Raw $\theta_1 = 1/3$	1.213	1.054	0.994	0.988	1.056	0.995	1.090

TABLE 4. *(continued)*

	$\theta_2 = -1$	$\theta_2 = -2/3$	$\theta_2 = -1/3$	$\theta_2 = 0$	$\theta_2 = 1/3$	$\theta_2 = 2/3$	$\theta_2 = 1$
Rect-SSBC-Shr $\theta_1 = 1/3$	1.214	1.058	0.994	0.988	1.056	0.995	1.093
PSO-Th-Raw $\theta_1 = 1/3$	1.234	1.058	1.002	0.988	1.051	1.019	1.099
PSO-Sh-Shr $\theta_1 = 1/3$	1.206	1.052	0.996	0.988	1.051	0.998	1.089
AR $\theta_1 = 1/3$	1.217	1.057	1.007	0.991	1.052	1.002	1.065
BG $\theta_1 = 1/3$	1.233	1.052	1.020	1.005	1.071	1.007	1.063
FSO-Th-Raw $\theta_1 = 2/3$	1.438	1.297	1.043	1.016	1.068	1.002	1.111
FSO-Th-Shr $\theta_1 = 2/3$	1.438	1.296	1.042	1.014	1.057	1.001	1.109
FSO-PD-Raw $\theta_1 = 2/3$	1.431	1.296	1.057	1.011	1.031	1.002	1.113
FSO-PD-Shr $\theta_1 = 2/3$	1.431	1.294	1.054	1.008	1.022	1.001	1.109
FSO-WN-Raw $\theta_1 = 2/3$	1.429	1.283	1.054	1.002	0.995	0.993	1.094
FSO-WN-Shr $\theta_1 = 2/3$	1.428	1.283	1.058	1.001	0.993	0.993	1.095
FSO-2o-Raw $\theta_1 = 2/3$	1.436	1.288	1.042	1.008	1.033	0.999	1.081
FSO-2o-Shr $\theta_1 = 2/3$	1.436	1.286	1.045	1.007	1.023	0.998	1.083
Rect-ABC-Raw $\theta_1 = 2/3$	1.425	1.286	1.059	1.003	0.999	1.001	1.112
Rect-ABC-Shr $\theta_1 = 2/3$	1.423	1.284	1.064	1.002	0.996	1.002	1.111
Rect-SSBC-Raw $\theta_1 = 2/3$	1.429	1.285	1.059	1.002	1.002	1.004	1.102
Rect-SSBC-Shr $\theta_1 = 2/3$	1.427	1.284	1.063	1.001	1.000	1.003	1.102
PSO-Th-Raw $\theta_1 = 2/3$	1.439	1.297	1.043	1.015	1.068	1.002	1.111
PSO-Sh-Shr $\theta_1 = 2/3$	1.428	1.283	1.058	1.001	0.993	0.993	1.095
AR $\theta_1 = 2/3$	1.427	1.296	1.035	1.011	0.992	0.998	1.090
BG $\theta_1 = 2/3$	1.437	1.302	1.035	1.024	1.000	1.011	1.092
FSO-Th-Raw $\theta_1 = 1$	1.659	1.450	1.295	1.093	1.301	0.995	1.089
FSO-Th-Shr $\theta_1 = 1$	1.659	1.450	1.295	1.088	1.193	0.988	1.087
FSO-PD-Raw $\theta_1 = 1$	1.659	1.449	1.296	1.107	1.948	1.034	1.116
FSO-PD-Shr $\theta_1 = 1$	1.658	1.450	1.294	1.097	1.636	1.013	1.107
FSO-WN-Raw $\theta_1 = 1$	1.659	1.450	1.287	1.069	1.030	0.984	1.084
FSO-WN-Shr $\theta_1 = 1$	1.659	1.450	1.287	1.073	1.038	0.984	1.087
FSO-2o-Raw $\theta_1 = 1$	1.659	1.450	1.292	1.062	1.049	0.987	1.069
FSO-2o-Shr $\theta_1 = 1$	1.659	1.450	1.292	1.064	1.022	0.984	1.070
Rect-ABC-Raw $\theta_1 = 1$	1.660	1.451	1.284	1.070	1.035	1.002	1.104
Rect-ABC-Shr $\theta_1 = 1$	1.659	1.451	1.284	1.073	1.044	0.999	1.104
Rect-SSBC-Raw $\theta_1 = 1$	1.661	1.451	1.280	1.069	1.044	1.006	1.095
Rect-SSBC-Shr $\theta_1 = 1$	1.660	1.451	1.280	1.073	1.053	1.003	1.096
PSO-Th-Raw $\theta_1 = 1$	1.659	1.450	1.295	1.094	1.300	0.995	1.090
PSO-Sh-Shr $\theta_1 = 1$	1.659	1.450	1.287	1.073	1.038	0.984	1.087
AR $\theta_1 = 1$	1.667	1.461	1.282	1.058	1.023	0.993	1.052
BG $\theta_1 = 1$	1.678	1.470	1.288	1.055	1.046	1.003	1.045

TABLE 5
Root mean square prediction errors for M3 competition data and reversed M3 competition data

	Forward	Reversed
FSO-Th-Raw	0.8693	0.8421
FSO-Th-Shr	0.9746	1.0021
FSO-PD-Raw	0.8828	0.8485
FSO-PD-Shr	0.8930	0.8835
FSO-WN-Raw	0.8821	0.8509
FSO-WN-Shr	0.9831	1.0237
FSO-2o-Raw	0.8894	0.8640
FSO-2o-Shr	0.8916	0.8877
Rect-ABC-Raw	0.8785	0.8561
Rect-ABC-Shr	0.9941	1.0413
Rect-SSBC-Raw	0.8572	0.7971
Rect-SSBC-Shr	0.9405	0.9661
PSO-Th-Raw	0.8804	0.8490
PSO-Sh-Shr	0.9864	1.0246
AR	0.8356	0.7852
BG	0.8682	0.7916

Each of these time series was then rescaled to have variance 1 so that prediction errors would have approximately the same scale. The experiment consisted of predicting the second to last and the last value in each series using all preceding observations, i.e., having a testing set of size 2 for each series; this resulted in 210 total predictions. Root mean square prediction errors (obtained as an average of the 210 predictions) corresponding to the different methods are shown in the first column of Table 5.

To corroborate and add weight to these findings, we then reversed time and used the later times in each series to predict the first and second observations. Time reversal produces stationary time series with the same covariance structures, allowing us to perform an additional 210 predictions. Results are shown in the second column of Table 5.

In both the original and time reversed real data experiments, the benchmark AR prediction outperformed all other methods. Our procedure is analogous to nonparametric spectral estimation, so in some sense it is unsurprising that a nonparametric procedure is uncompetitive with smaller sample sizes. More surprisingly, subsampling bandwidth choice outperformed adaptive bandwidth choice, and the estimators using shrunken estimates of $\gamma(n)$ underperformed those using raw estimates.

In both the original and time reversed real data experiments, the benchmark AR prediction outperformed all other methods. Our procedure is analogous to nonparametric spectral estimation, so in some sense it is not surprising that it is not competitive with sample sizes between 20 to 50. More surprisingly, subsampling bandwidth choice outperformed adaptive bandwidth choice, and the predictors using shrunken estimates of $\gamma(n)$ underperformed compared to those using raw estimates. Both these phenomena seem to contradict the findings of the simulations; we suspect the sample sizes were so small that they did not allow the asymptotics to kick in.

TABLE 6
Average operator norm loss for autocovariance matrix estimates using various corrections to positive definiteness

		$\hat{\Gamma}_n$	Thresh	Thresh+Scl	PD-Shrink	WN-Shrink	2o-Shrink	Rect-ABC-WN	Rect-SSBC-WN
AR(1)	$\phi = -0.9$	10.9824	10.9822	10.7257	10.8039	9.8317	10.0831	9.4513	10.4375
AR(1)	$\phi = -0.5$	0.9154	0.9153	0.9172	0.9010	0.9429	0.9098	0.9483	0.9108
AR(1)	$\phi = -0.1$	0.2992	0.2992	0.2992	0.2982	0.2992	0.2992	0.2977	0.2857
AR(1)	$\phi = 0.1$	0.2930	0.2930	0.2930	0.2923	0.2930	0.2930	0.2924	0.2863
AR(1)	$\phi = 0.5$	0.9454	0.9451	0.9465	0.9352	0.9713	0.9417	0.9948	0.9133
AR(1)	$\phi = 0.9$	9.7730	9.7708	9.7022	9.7519	9.3799	9.8239	10.6984	13.2768
MA(1)	$\theta = -0.9$	0.3185	0.3248	0.3143	0.3206	0.2909	0.3501	0.2804	0.3686
MA(1)	$\theta = -0.5$	0.2488	0.2486	0.2474	0.2517	0.2377	0.2500	0.2284	0.3103
MA(1)	$\theta = -0.1$	0.2883	0.2883	0.2883	0.2877	0.2883	0.2883	0.2877	0.2835
MA(1)	$\theta = 0.1$	0.2816	0.2816	0.2816	0.2809	0.2816	0.2816	0.2812	0.2833
MA(1)	$\theta = 0.5$	0.2567	0.2565	0.2552	0.2593	0.2452	0.2581	0.2368	0.3085
MA(1)	$\theta = 0.9$	0.2999	0.3059	0.2976	0.3042	0.2794	0.3327	0.2784	0.3776
ARMA(2,1)		1.3995	1.3988	1.3977	1.3845	1.3882	1.3926	1.3953	1.3855

5.5. Relative performance of different matrix estimators

Rescaling of the threshold corrected matrix given in eq. (19) is a new proposal in the literature. Similarly, the shrinkage corrected matrices described in Sections 4.2, 4.3, and 4.4 are also novel. For this reason, we also carried out a small simulation study demonstrating their ability to improve estimates of Γ_n. Data sets of size $n = 200$ were used throughout using some simple AR(1) and MA(1) models, along with the ARMA(2,1) model $X_t - 0.7X_{t-1} + 0.5X_{t-2} = \epsilon_t - 0.3\epsilon_{t-1}$. Each data set was used to estimate the autocovariance matrix, and the estimate was then compared to the true autocovariance matrix in operator norm. 1,000 replications were performed for each model.

Average differences in operator norm are shown in Table 6. The estimators that have been corrected to positive definiteness and then scaled to keep the average eigenvalue unchanged show a consistent advantage over the initial estimate $\hat{\Gamma}_n$ and the unadjusted threshold corrected matrix. Shrinkage to white noise and to a second order estimate both show strong performance, with the former particularly strong for the MA processes, and the latter stronger for the AR processes.

5.6. Relative performance of autocovariance vector estimators

Section 3.2 introduced two estimates of the autocovariance vector $\gamma(n)$. The first, $\hat{\gamma}(n)$ is the raw or unadjusted banded and tapered estimate. The second, $\hat{\gamma}^*(n)$ is the shrunken version taken from the first row of the estimated autocovariance matrix after correction to positive definiteness. These two estimators have similar theoretical performance, but it is unclear which is preferable in application.

In order to compare performance, we conducted a small simulation study using a selection of AR(1) and MA(1) models, along with the ARMA(2,1) model $X_t - 0.7X_{t-1} + 0.5X_{t-2} = \epsilon_t - 0.3\epsilon_{t-1}$, each with a sample size of $n = 200$. l_2 norm errors are shown in Table 7. All the shrinkage type estimators, except possibly the selective shrinkage towards a positive definite estimate, seem to consistently improve on the raw estimate $\hat{\gamma}(n)$; shrinkage towards white noise is a particularly strong performer here.

TABLE 7
Average l_2 norm differences between estimates and true values of vector $\gamma(n)$

		Raw	Thresh	PD-Shrink	WN-Shrink	2o-Shrink	Rect-ABC	Rect-ABC-WN	Rect-SSBC	Rect-SSBC-WN
AR(1)	$\phi = -0.9$	5.6285	5.4900	5.5622	5.1675	5.2324	5.9317	5.2304	5.9948	5.8240
AR(1)	$\phi = -0.5$	0.3088	0.3086	0.2973	0.3127	0.3056	0.3300	0.3234	0.3021	0.2950
AR(1)	$\phi = -0.1$	0.1059	0.1059	0.1054	0.1058	0.1059	0.1057	0.1057	0.0984	0.0983
AR(1)	$\phi = 0.1$	0.1063	0.1063	0.1057	0.1063	0.1063	0.1062	0.1062	0.0984	0.0983
AR(1)	$\phi = 0.5$	0.3084	0.3082	0.2983	0.3121	0.3059	0.3317	0.3268	0.2981	0.2918
AR(1)	$\phi = 0.9$	5.4163	5.3440	5.3906	5.1878	5.3164	5.7428	5.7425	6.3400	6.9718
MA(1)	$\theta = -0.9$	0.1513	0.1443	0.1508	0.1277	0.1581	0.1499	0.1243	0.1975	0.1789
MA(1)	$\theta = -0.5$	0.0939	0.0934	0.0946	0.0864	0.0934	0.0931	0.0849	0.1248	0.1190
MA(1)	$\theta = -0.1$	0.1038	0.1038	0.1034	0.1038	0.1038	0.1037	0.1036	0.0962	0.0961
MA(1)	$\theta = 0.1$	0.1034	0.1034	0.1031	0.1034	0.1034	0.1033	0.1033	0.0966	0.0965
MA(1)	$\theta = 0.5$	0.0973	0.0968	0.0985	0.0904	0.0969	0.0967	0.0894	0.1264	0.1213
MA(1)	$\theta = 0.9$	0.1525	0.1476	0.1558	0.1353	0.1588	0.1509	0.1323	0.1879	0.1715
ARMA(2,1)		0.3467	0.3463	0.3367	0.3425	0.3445	0.3754	0.3534	0.3616	0.3415

6. Conclusions

The thrust of this paper was to demonstrate the viability and asymptotic consistency of the FSO linear predictor (10) that uses the complete process history. A key element here is an accurate estimate of the full $n \times n$ autocovariance matrix given a sample of size n. As a by-product, we also show the consistency of the PSO linear predictor (12) which is an AR(p) predictor based on the last p data values for any $p \leq n$; this is a substantial strengthening of previous results which had required $p = o(n)$. In simulations, it is shown that the FSO and PSO predictors are competitive as compared to the state-of-the-art linear predictor which amounts to fitting an AR(p) model with p chosen by AIC minimization.

As part of our investigations, we have introduced several refinements to the current state of the art in estimating large autocovariance matrices under the restriction that they are finite-sample positive definite and not ill-conditioned. In particular, when using the eigenvalue threshold correction, we noted the necessity of rescaling the matrix so that the mean eigenvalue remains unchanged. In addition, we introduced three new corrections to positive definiteness, namely shrinking towards positive definiteness, shrinking towards the (rescaled) identity/white noise, and shrinking towards a 2nd order estimate. All three corrections are shown to work well with the shrinkage towards white noise appearing to have a small finite sample performance advantage over shrinking towards a 2nd order estimate.

In particular, note that the estimators resulting from shrinkage towards either white noise or a second order estimate both result in a banded Toeplitz matrix. As such, they can be calculated easily, stored efficiently, and inverted via fast algorithms; this property is especially important in the case of very large sample sizes.

Finally, in Appendix A we use these insights into large covariance matrix estimation to refine flat-top kernel spectral density estimates in order to ensure their positivity.

7. Technical proofs

Let $||A||_2 \equiv \max\{|A\vec{x}|_2 : \vec{x} \in \mathbb{R}^n \text{ with } |\vec{x}|_2 = 1\}$ denote the matrix 2-norm of an $n \times n$ matrix A.

Proof of Lemma 1.

$$
\begin{aligned}
|\hat{\gamma}(n) - \gamma(n)|_2 &= \left\{ \sum_{i=1}^{n} [\kappa(i/l)\breve{\gamma}_i - \gamma_i]^2 \right\}^{1/2} \\
&\leq \sum_{i=1}^{n} |\kappa(i/l)\breve{\gamma}_i - \gamma_i| \\
&\leq \sum_{i=1}^{l} |\breve{\gamma}_i - \gamma_i| + \sum_{i=l+1}^{\lfloor c_\kappa l \rfloor} |\kappa(i/l)\breve{\gamma}_i - \gamma_i| + \sum_{i=\lfloor c_\kappa l \rfloor + 1}^{n} |\gamma_n|
\end{aligned}
$$

Bounds for the above three terms are obtained in the proof of Theorem 1 in McMurry and Politis (2010); using those bounds, we have

$$
E\left[|\hat{\gamma}(n) - \gamma(n)|_2^2\right]^{1/2} \leq d_2(\lfloor c_\kappa l \rfloor + 1)n^{-1/2} + \frac{1}{n}\sum_{i=1}^{\lfloor c_\kappa l \rfloor} i|\gamma_i| + \sum_{i=l+1}^{n} |\gamma_i|,
$$

where d_2 is a constant depending on $E\left[X_i^4\right]$ and Δ_4 but not l or n; this establishes the convergence of $\hat{\gamma}(n)$ to $\gamma(n)$ with the same rates as $\hat{\Gamma}_n$ converges to Γ_n, described in Corollary 1 of McMurry and Politis (2010). □

Proof of Theorem 2.

$$
\begin{aligned}
\hat{\phi}(n) - \phi(n) &= (\hat{\Gamma}_n^*)^{-1}\hat{\gamma}(n) - \Gamma_n^{-1}\gamma(n) \\
&= (\hat{\Gamma}^*)_n^{-1}\hat{\gamma}(n) - \hat{\Gamma}_n^{-1}\gamma(n) + (\hat{\Gamma}_n^*)^{-1}\gamma(n) - \Gamma_n^{-1}\gamma(n) \\
&= (\hat{\Gamma}_n^*)^{-1}[\hat{\gamma}(n) - \gamma(n)] + [(\hat{\Gamma}_n^*)^{-1} - \Gamma_n^{-1}]\gamma(n)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
|\hat{\phi}(n) - \phi(n)|_2 &\leq \left|\left|(\hat{\Gamma}_n^*)^{-1}\right|\right|_2 |\hat{\gamma}(n) - \gamma(n)|_2 + \left|\left|(\hat{\Gamma}_n^*)^{-1} - \Gamma_n^{-1}\right|\right|_2 |\gamma(n)|_2 \\
&= A_1 + A_2.
\end{aligned}
$$

We investigate term A_1 first. With probability tending to one, $||(\hat{\Gamma}_n^*)^{-1}||_2$ is bounded.

$$
\begin{aligned}
|\hat{\gamma}(n) - \gamma(n)|_2 &= \left\{ \sum_{i=1}^{n} [\breve{\gamma}_i\kappa(i/l) - \gamma_i]^2 \right\}^{1/2} \\
&\leq \sum_{i=1}^{n} |\breve{\gamma}_i\kappa(i/l) - \gamma_i| \\
&= O_p(r_n)
\end{aligned}
$$

The final equality is established in the proof of Theorem 1 in McMurry and Politis (2010); see also Section 3.2.

We now turn our attention to A_2. By Corollary 3 in McMurry and Politis (2010),

$$\left|\left|(\hat{\Gamma}_n^*)^{-1} - \Gamma_n^{-1}\right|\right|_2 = O_p(r_n).$$

Since $\sum_{i=1}^{\infty} |\gamma(i)| < \infty$, the result follows. □

Proof of Theorem 3. We compare the FSO predictor $\hat{X}_{n+1}$ to the oracle optimal prediction $\tilde{X}_{n+1}$ based on the following decomposition:

$$\begin{aligned}\hat{X}_{n+1} - \tilde{X}_{n+1} &= \sum_{i=1}^{k_n}[\hat{\phi}_i(n) - \phi_i(n)]X_{n-i+1} + \sum_{i=k_n+1}^{n} \hat{\phi}_i(n)X_{n-i+1} \\ &\quad - \sum_{i=k_n+1}^{n} \phi_i(n)X_{n-i+1} \\ &= A + B + C. \end{aligned} \tag{24}$$

The basic idea of the proof is that we can let $k_n \to \infty$ slowly enough that the first term goes to 0 by the Cauchy-Schwarz inequality. Since the coefficients $\hat{\phi}_i(n)$ and $\phi_i(n)$ decay quickly as i increases, by allowing k_n to grow fast enough the second two terms can also be shown to converge to 0, again by Cauchy-Schwarz.

We begin with term A. By the Cauchy-Schwarz inequality

$$\begin{aligned}\left|\sum_{i=1}^{k_n}[\hat{\phi}_i(n) - \phi_i(n)]X_{n-i+1}\right| &\le |\hat{\phi}(n) - \phi(n)|_2 \left[\sum_{i=1}^{k_n} X_{n-i+1}^2\right]^{1/2} \\ &= O_p(r_n k_n^{1/2})\end{aligned}$$

By Assumption 5ii this term tends to 0.

Term B will be handled by Proposition 2.2 of Demko, Moss and Smith (1984) which shows that as long as $\hat{\Gamma}_n^*$ is a banded matrix, which it will be with probability tending to 1, (for small samples $\hat{\Gamma}_n^*$ may be corrected to positive definiteness, and depending on the technique used, no-longer banded)

$$|(\hat{\Gamma}_n^*)_{ij}^{-1}| \le C_2 \lambda^{|i-j|/l}, \tag{25}$$

where C_2 and $\lambda < 1$ depend only on the largest and smallest eigenvalues of $\hat{\Gamma}_n^*$. Since with probability tending to 1, these are bounded away from 0 and from above, for large enough n, C_2 and λ can be chosen independent of n with (25) holding with probability tending to 1.

Since by Assumption 5i, k_n grows faster than l, there is no loss in considering only $i > 2l$. In this case

$$|\hat{\phi}_i(n)| = \left|\sum_{j=1}^{c_\kappa l}(\hat{\Gamma}_n^*)_{ij}^{-1}\hat{\gamma}_j\right| \le C_3 \sum_{j=1}^{c_\kappa l} \lambda^{(i-j)/l} \le C_4 l \lambda^{i/l}, \tag{26}$$

where the bound above holds with probability tending to 1.

By (26),

$$\left\{\sum_{i=k_n}^{n} |\hat{\phi}_i(n)|^2\right\}^{1/2} \leq C_5 l^{3/2} \lambda^{(k_n-1)/(2l)}. \tag{27}$$

By (27) and the Cauchy-Schwarz inequality, term B converges to 0 by Assumption 5i.

By the Cauchy-Schwarz inequality, Term C can be bounded by

$$\begin{aligned} |C| = \left|\sum_{i=k_n+1}^{n} \phi_i(n) X_{n-i+1}\right| &= \left|\sum_{i=k_n+1}^{n} [\phi_i(n) - \phi_i + \phi_i] X_{n-i+1}\right| \\ &\leq \left[\left(\sum_{i=k_n+1}^{n} |\phi_i(n) - \phi_i|\right) + \left(\sum_{i=k_n+1}^{n} \phi_i^2\right)^{1/2}\right] \\ &\quad \times O_p(n^{1/2}) \\ &\leq \left[C_7 \sum_{i=n+1}^{\infty} |\phi_i| + \left(\sum_{i=k_n+1}^{\infty} \phi_i^2\right)^{1/2}\right] O_p(n^{1/2}) \end{aligned} \tag{28}$$

where ϕ_i denotes the corresponding AR(∞) coefficient, and inequality (28) follows from the variant of Baxter's inequality (Baxter, 1962, 1963) given in Lemma 2.2 of Kreiss, Paparoditis and Politis (2011) and holds for all $n > N_0$ for some positive N_0. The first term in (28) converges by Assumption 6. The second term in (28) converges by Assumption 5iii. □

Proof of Corollary 1. For any sequence $p_n < n$, $\hat{\Gamma}^*_{p_n}$ converges to Γ_{p_n} as fast or faster than the convergence of the larger $n \times n$ matrices; this is because the absolute row sum norm of the difference of the smaller matrices is bounded from above by the maximum absolute row sum norm of the difference of the larger matrices. Similarly, the convergence of $\hat{\gamma}(p_n)$ to $\gamma(p_n)$ is not made worse; see Section 3.2. Finally, the eigenvalues of $\hat{\Gamma}^*_{p_n}$ and Γ_{p_n} have the same positive upper and lower bounds as their larger counterparts; see Lemma 4.1 in Gray (2006). Therefore, the proof of Theorem 2 carries over directly. □

Proof of Corollary 2. In the case that $p_n \leq k_n$, terms B and C in (24) are 0, and the Cauchy-Schwarz inequality can be used directly on term A, giving the desired result. If $p_n > k_n$, term B in decomposition (24) is again handled by Proposition 2.2 of Demko, Moss and Smith (1984) with the only change being that the sum (27) stops at p_n. The challenge comes with term C, where Baxter's inequality is now used to compare $\phi_i(p_n)$ and ϕ_i. Since $p_n < n$, this approximation becomes worse, and the first half of term C becomes

$$n^{1/2} C_7' \sum_{i=p_n+1}^{\infty} |\phi_i|,$$

which converges to 0 by Assumption 7. □

Acknowledgements

The authors would like to thank Yulia Gel, Marco Meyer, and Stathis Paparoditis for helpful conversations and guidance. Special thanks are due to Rob Hyndman for suggesting the M3 competition data as a test case, and for discovering an error in our original implementation. The constructive comments of two reviewers should also be acknowledged as they led to several improvements. Research of the second author was partially supported by NSF grants DMS 13-08319 and DMS 12-23137.

Appendix A: Positivity corrections in spectral density estimation

Let

$$\hat{f}_{io}(\omega) = \frac{1}{2\pi} \sum_{s=-n}^{n} \kappa(s/l)\breve{\gamma}_j e^{-i\omega s} \equiv \frac{1}{2\pi} \sum_{s=-n}^{n} \hat{\gamma}_j e^{-i\omega s} \tag{29}$$

denote the infinite-order estimate of $f(\omega)$ using flat-top weight function $\kappa(\cdot)$. As is well-known, $\hat{f}_{io}(\omega)$ achieves the fastest rate of convergence possible in a given smoothness class; see Politis (2011) and the references therein. However, although $f(\omega) \geq 0$ for all ω, the same is not guaranteed to be true for $\hat{f}_{io}(\omega)$. The usual correction is to clip the negative values, i.e., define the corrected estimator

$$\hat{f}^{+}_{io}(\omega) = \max\{0, \hat{f}_{io}(\omega)\}$$

that is nonnegative while maintaining the same fast rate of convergence of $\hat{f}_{io}(\omega)$.

Nevertheless, in situations where an estimate of the inverse of $f(\omega)$ is needed, a more dramatic correction must take place. For example, recall that the large-sample variance of the sample mean $n^{-1}\sum_{t=1}^{n} X_t$ is given by $2\pi f(0)/n$ under standard conditions. Hence, to create a t-statistic for testing and/or confidence intervals, the practitioner must be able to divide by an estimate of $f(0)$.

In this Appendix we discuss analogs of the three matrix corrections given in Section 4 as they apply to the problem of spectral density estimation. The analogy is made possible due to the aforementioned fact that the eigenvalues of Γ_n are asymptotically given by the values of the spectral density function evaluated on the Fourier frequencies; see e.g. Gray (2006).

A.1. Selective shrinkage to positive definiteness

As in Section 4.2, we can employ a second order kernel estimator to provide a target lower bound for the estimated spectral density. Recall that a positive definite spectral estimator is by necessity based on a second order kernel, and is therefore asymptotically inefficient. Let

$$\hat{f}_{2o}(\omega) = \frac{1}{2\pi} \sum_{s=-n}^{n} \kappa^{2o}(s/l)\breve{\gamma}_j e^{-i\omega s}$$

denote a second-order, positive definite spectral density estimate such as the one that results when the weight function $\kappa^{2o}(\cdot)$ is chosen to be Parzen's piecewise cubic lag window.

Then we can define a corrected flat-top spectral density estimator as

$$\hat{f}^{\star}_{io}(\omega) = \begin{cases} \hat{f}_{io}(\omega) & \text{if} \quad \hat{f}_{io}(\omega) \geq \hat{f}_{2o}(\omega) \\ (1-\tau_n)\hat{f}^{+}_{io}(\omega) + \tau_n \hat{f}_{2o}(\omega) & \text{if} \quad \hat{f}_{io}(\omega) < \hat{f}_{2o}(\omega) \end{cases}$$

where $\tau_n = c/n^a$ for constants $c > 0$ and $a > 1/2$. Since $a > 1/2$, the correction by factor τ_n is asymptotically negligible so that $\hat{f}^{\star}_{io}(\omega)$ enjoys the same fast rate of convergence as $\hat{f}_{io}(\omega)$.

Using the formula for the Fourier coefficients and noting that $\kappa(0) = \kappa^{2o}(0) = 1$, it follows that

$$\breve{\gamma}_0 = \hat{\gamma}_0 = \int_{-\pi}^{\pi} \hat{f}_{io}(\omega) d\omega = \int_{-\pi}^{\pi} \hat{f}_{2o}(\omega) d\omega,$$

i.e., the area under any choice of spectral density estimate equals the sample autocovariance at lag zero which is our best estimate of $\text{var}\,[X_t]$. Note, however, that the shrinkage estimator $f^{\star}_{io}(\omega)$ has an area that is larger than $\hat{\gamma}_0$, therefore implying a bigger estimate for $\text{var}\,[X_t]$. This is not intuitive, and hence $\hat{f}^{\star}(\omega)$ should be appropriately rescaled. Our final, rescaled shrinkage estimator is given by

$$\hat{f}^{*}_{io}(\omega) = c\hat{f}^{\star}_{io}(\omega) \text{ where } c = \hat{\gamma}_0 / \int_{-\pi}^{\pi} \hat{f}^{\star}_{io}(\omega) d\omega. \tag{30}$$

A.2. Shrinkage toward white noise

As described in Section 4.3, we may shrink $\hat{\gamma}_i$ (for $i \neq 0$) towards zero by a factor $s \in (0, 1]$ chosen to ensure that the minimum of the estimated spectral density is greater or equal to $\epsilon\hat{\gamma}_0/(2\pi n^{\beta})$. The resulting estimator

$$\hat{f}^{*}_{io}(\omega) \equiv (1-s)\frac{\hat{\gamma}_0}{2\pi} + s\hat{f}_{io}(\omega)$$

is positive definite while maintaining the same fast asymptotic rate of convergence as $\hat{f}_{io}(\omega)$. Note that by construction, $\hat{f}^{*}_{io}(\omega) \geq \epsilon\hat{\gamma}_0/(2\pi n^{\beta})$ for all ω. By analogy to Section 4.3, the estimator $\hat{f}^{*}_{io}(\omega)$ has no need for rescaling as it maintains the same area under the curve as $\hat{f}_{io}(\omega)$, and therefore is associated with an estimate of $\text{var}\,[X_t]$ given by $\hat{\gamma}_0 = \breve{\gamma}_0$.

A.3. Shrinkage towards a 2nd order estimate

A spectral density estimator can also be corrected to non-negativity by shrinking it towards a positive definite, 2nd order estimate as described for matrices in Section 4.4. The resulting estimator is

$$\hat{f}^{*}_{io}(\omega) \equiv s\hat{f}_{io}(\omega) + (1-s)\hat{f}_{2o}(\omega).$$

Note that the above amounts to shrinking $\hat{f}_{io}(\omega)$ towards $\hat{f}_{2o}(\omega)$ for all $\omega \in [-\pi, \pi]$; thus, it should be contrasted with the method of Section A.1 where it was proposed to shrink $\hat{f}^{+}_{io}(\omega)$ towards $\hat{f}_{2o}(\omega)$ only for ω such that $\hat{f}^{+}_{io}(\omega) < \hat{f}_{2o}(\omega)$.

The shrinkage factor $s \in [0,1]$ is chosen to be the minimum of $s(\omega)$, where $s(\omega)$ is a "pointwise" shrinkage factor, calculated as follows. If $f_{io}(\omega) \geq f_{2o}(\omega)$ or if $f_{io}(\omega) \geq \epsilon\hat{\gamma}_0/(2\pi n^\beta)$, then $s(\omega) = 1$; in words $f_{io}(\omega)$ is bigger than either the threshold or the second order estimator, so it is left untouched. Otherwise, we calculate the shrinkage factor needed to raise $f_{io}(\omega)$ to the minimum of the threshold and the second order estimator, making $s(\omega)$ the maximum of $\left[\epsilon\hat{\gamma}_0/(2\pi n^\beta) - f_{2o}(\omega)\right]/[f_{io}(\omega) - f_{2o}(\omega)]$ and 0.

The reason that both shrinking towards white noise and shrinking towards a 2nd order estimate work well—asymptotically and in finite samples—is explained in the following remark.

Remark 14. Spectral estimators such as $\hat{f}_{io}$ and $\hat{f}_{2o}$ can be alternatively expressed as weighted local averages of the periodogram (see Brockwell and Davis, 1991). Since the periodogram is (approximately) unbiased, the bias in spectral estimation is due to the local averaging that, in effect, "trims the hills, and fills the valleys". The fact that $\hat{f}_{io}(\omega)$ is less biased than $\hat{f}_{2o}(\omega)$ implies that $\hat{f}_{io}(\omega)$ can follow "the hills and the valleys" better than $\hat{f}_{2o}(\omega)$. In that sense, shrinking $\hat{f}_{io}(\omega)$ towards the spectral density of a white noise is tantamount to shrinking $\hat{f}_{io}(\omega)$ towards $\hat{f}_{2o}(\omega)$ for all $\omega \in [-\pi, \pi]$; the goal of shrinkage towards either target is a flatter version of $\hat{f}_{io}(\omega)$. Of course these targets are not meant to be achieved—just to give a general direction for the correction.

A.4. Thresholding correction

Politis (2011) proposed a threshold correction for the spectral density that is analogous to the eigenvalue thresholding of Section 4.1. To elaborate, the threshold corrected spectral density estimate is $\hat{f}^{\epsilon}_{io}(\omega) = \max\{\hat{f}_{io}(\omega), \epsilon\hat{\gamma}_0/(2\pi n^\beta)\}$ for some $\epsilon > 0$ and $\beta > 1/2$. Note, however, that this threshold estimator could also benefit from rescaling due to the arguments leading to eq. (30). We may thus propose a new *rescaled* threshold corrected flat-top spectral density estimator given by

$$\hat{f}^{*}_{io}(\omega) = c_\epsilon \hat{f}^{\epsilon}_{io}(\omega) \text{ where } c_\epsilon = \hat{\gamma}_0 / \int_{-\pi}^{\pi} \hat{f}^{\epsilon}_{io}(\omega) d\omega. \tag{31}$$

A.5. Numerical illustrations

Although asymptotically negligible, the corrections discussed in Sections A.1–A.4 can dramatically improve finite sample performance. Figure 1 provides an illustration using a dataset simulated from the ARMA(2,1) model $X_t - 0.7X_{t-1} + 0.5X_{t-2} = \epsilon_t - 0.3\epsilon_{t-1}$ with $n = 100$. Notably, this was not a dataset selected at random; it was chosen among many realizations of datasets from this ARMA model because for this particular dataset $\hat{f}_{io}$ behaves poorly at $\omega = 0$ necessitating substantial correction.

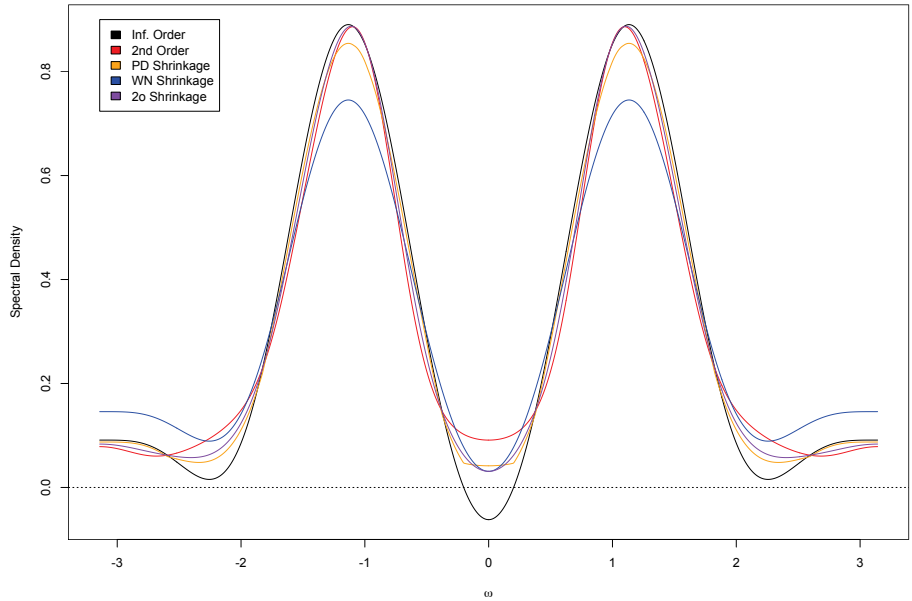

FIG 1. *The 2nd order, infinite order, and shrinkage-corrected spectral density estimates.*

TABLE 8
Mean integrated square errors for spectral density estimates

		$\hat{f}_{io}$	Thresh	PD-Shrink	WN-Shrink	2o-Shrink	Rect-ABC	Rect-ABC-WN	Rect-SSBC	Rect-SSBC-WN
AR(1)	$\phi = -0.9$	15.3518	12.8751	14.2757	11.8498	11.9982	15.6966	9.2717	14.2054	10.4643
AR(1)	$\phi = -0.5$	0.0426	0.0419	0.0396	0.0434	0.0418	0.0449	0.0398	0.0423	0.0404
AR(1)	$\phi = -0.1$	0.0055	0.0055	0.0055	0.0055	0.0055	0.0055	0.0055	0.0059	0.0055
AR(1)	$\phi = 0.1$	0.0055	0.0055	0.0054	0.0055	0.0055	0.0055	0.0055	0.0057	0.0054
AR(1)	$\phi = 0.5$	0.0418	0.0412	0.0392	0.0428	0.0410	0.0441	0.0404	0.0403	0.0406
AR(1)	$\phi = 0.9$	11.3374	10.6823	11.0332	10.4805	10.7259	12.2597	11.1381	13.7981	13.2870
MA(1)	$\theta = -0.9$	0.0232	0.0225	0.0220	0.0200	0.0220	0.0223	0.0189	0.0298	0.0206
MA(1)	$\theta = -0.5$	0.0081	0.0075	0.0078	0.0073	0.0080	0.0080	0.0072	0.0116	0.0075
MA(1)	$\theta = -0.1$	0.0052	0.0052	0.0051	0.0052	0.0052	0.0052	0.0052	0.0052	0.0052
MA(1)	$\theta = 0.1$	0.0052	0.0052	0.0052	0.0052	0.0052	0.0052	0.0052	0.0053	0.0052
MA(1)	$\theta = 0.5$	0.0084	0.0080	0.0082	0.0079	0.0084	0.0083	0.0077	0.0120	0.0080
MA(1)	$\theta = 0.9$	0.0217	0.0214	0.0208	0.0192	0.0208	0.0210	0.0183	0.0310	0.0208
ARMA(2,1)		0.0493	0.0481	0.0464	0.0483	0.0489	0.0540	0.0425	0.0503	0.0441

In addition, we tried a formal simulation experiment to compare the various corrections to positive definiteness in spectral density estimation using difference AR(1) and MA(1) models, as well as the aforementioned ARMA(2,1) model of Figure 1. For each simulated dataset of size $n = 200$, we estimated the spectral density using the uncorrected infinite order estimate and the methods described in Sections A.1–A.4. The thresholds for correction were the same as those used in the corresponding autocovariance matrix simulations.

Mean integrated square errors are shown in Table 8. All of the new correction methods show substantial improvement over $\hat{f}_{io}$. As in the matrix estimation set-up of Section 5.5, shrinkage towards white noise and towards a 2nd order estimator appear particularly powerful. Shrinkage towards white noise seems to perform better for MA processes, while shrinkage towards the 2nd order estimator appears to have a small edge for AR processes.

References

Basu, S. and Michailidis, G. (2014). Regularized estimation in sparse high-dimensional time series models. Preprint.

Baxter, G. (1962). An asymptotic result for the finite predictor. *Mathematica Scandinavica* **10** 137–144. MR0149584

Baxter, G. (1963). A norm inequality for a "finite-section" Wiener-Hopf equation. *Illinois Journal of Mathematics* **7** 97–103. MR0145285

Bickel, P. J. and Gel, Y. R. (2011). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. *Journal of the Royal Statistical Society: Series B* **73** 711–728. MR2867455

Bickel, P. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics* **36** 2577–2604. MR2485008

Bickel, P. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36** 199–227. MR2387969

Brent, R. P., Gustavson, F. G. and Yun, D. Y. (1980). Fast solution of Toeplitz systems of equations and computation of Padé approximants. *Journal of Algorithms* **1** 259–295. MR0604867

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer, New York. MR1093459

Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106** 672–684. MR2847949

Cai, T. T., Ren, Z. and Zhou, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probability Theory and Related Fields* **156** 101–143. MR3055254

Cai, T. T. and Zhou, H. H. (2012a). Minimax estimation of large covariance matrices under l1 norm. *Statistica Sinica* **22** 1319–1378. MR3027084

Cai, T. T. and Zhou, H. H. (2012b). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics* **40** 2389–2420. MR3097607

Chen, X., Xu, M. and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics* **41** 2994–3021. MR3161455

Demko, S., Moss, W. F. and Smith, P. W. (1984). Decay rates for inverses of band matrices. *Mathematics of Computation* **43** 491–499. MR0758197

Gray, R. M. (2006). *Toeplitz and Circulant Matrices: A Review*. Now publishers Inc., Boston, MA.

Grenander, U. and Szegő, G. (1958). *Toeplitz Forms and Their Applications*. University of California Press, Berkeley. MR0094840

Hyndman, R., Akram, M. and Bergmeir, C. (2013). Mcomp: Data from the M-competitions. R package version 2.05.

Kreiss, J. P., Paparoditis, E. and Politis, D. N. (2011). On the range of validity of the autoregressive sieve bootstrap. *The Annals of Statistics* **39** 2103–2130. MR2893863

Kwiatkowski, D., Phillips, P., Schmidt, P. and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* **54** 159–178.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **10** 603–621.

Ledoit, O. and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management* **30** 110–119.

McMurry, T. L. and Politis, D. N. (2004). Nonparametric regression with infinite order flat-top kernels. *Journal of Nonparametric Statistics* **16** 549–562. MR2073041

McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis* **31** 471–482. Corrigendum, *J. Time Ser. Anal.* **33** 2012. MR2732601

Paparoditis, E. and Politis, D. N. (2012). Nonlinear spectral density estimation: Thresholding the correlogram. *Journal of Time Series Analysis* **33** 386–397. MR2915091

Politis, D. N. (2001). On nonparametric function estimation with infinite-order flat-top kernels. In *Probability and Statistical Models with Applications* (C. A. Charalambides, M. V. Koutras and N. Balakrishnan, eds.) 469–483. Chapman & Hall/CRC, Boca Raton.

Politis, D. N. (2003). Adaptive bandwidth choice. *Journal of Nonparametric Statistics* **15** 517–533. MR2017485

Politis, D. N. (2011). Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. *Econometric Theory* **27** 703–744. MR2822363

Politis, D. N. and Romano, J. P. (1995). Bias-corrected nonparametric spectral estimation. *Journal of Time Series Analysis* **16** 67–103. MR1323618

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences* **102** 14150–14154. MR2172215

Wu, W. B. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica* **19** 1755–1768. MR2589209

Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *The Annals of Statistics* **40** 466–493. MR3014314