Electronic Journal of Statistics
Vol. 9 (2015) 811–822
ISSN: 1935-7524
DOI: 10.1214/15-EJS1000REJ

Rejoinder of "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

Received March 2015.

We would like to sincerely thank all discussants for their kind remarks and insightful comments. To start with, we wholeheartedly welcome the proposal of Rob Hyndman for a "better acf" plot based on our vector estimator $\hat{\gamma}^*(n)$ from Section 3.2. As mentioned, the sample autocovariance is not a good estimate for the vector $\gamma(n)$, and this is especially apparent in the wild excursions it takes at higher lags—see the left panel of Figure 1 of Hyndman's piece. Note that these wild (and potentially confusing) excursions are the norm rather than the exception; they are partly explainable by two facts: (a) the identity $\sum_{k=-n}^{n} \breve{\gamma}_k = 0$ implies that $\breve{\gamma}_k$ must misbehave for higher lags to counteract its good behavior for small lags; and (b) the $\breve{\gamma}_k$ are correlated, and therefore their excursions appear smooth (and may be confused for structure). The only saving point of the current `acf` plot in R is that it has a `lag.max` default of $10 \log_{10} n$ so the ugliness occuring at higher lags is masked. Interestingly, showing just the lags up to $10 \log_{10} n$ is tantamount to employing a rectangular lag-window—which is one of the flat-top kernels albeit not the best—with a logarithmic choice for l that is indeed optimal under the exponential decay of γ_k typical of ARMA models.

Rob Hyndman also brings up the question of optimal linear prediction. Here, we would just like to offer a linguistic comment. The statistical term 'optimal estimation' is clear: an optimal estimator is closest (according to some criterion) to its target estimand. However, the term 'optimal prediction' is typically used in a probabilistic context where all parameters are assumed known and only the form of the predictor is in question; in other words, the term 'optimal prediction'

[*] Main article 10.1214/15-EJS1000.

does not entail any claims as to estimation accuracy. We are using the term this way, i.e., we are proposing a new estimator of the theoretically optimal linear predictor of eq. (1). Whether our estimate of the optimal linear predictor is better than the benchmark AR-based estimate is an open question (and is not expected to admit an easy answer). In the next section, we elaborate on the problems in comparing our proposed method to the benchmark AR-based predictor; the main theoretical problem is lack of knowledge of their respective rates of convergence.

Xiaohui Chen proposes another approach to estimating Γ_n^{-1}. His approach has the appealing properties that large eigenvalues of $\hat{\Gamma}_n$ are inverted relatively unperturbed, while those close to 0 are adjusted gracefully. These appealing features are balanced with two less intuitive characteristics. First, the resulting estimates are not positive definite, even though Γ_n is. Second, eigenvalues close to 0 in $\hat{\Gamma}_n$ are converted into eigenvalues close to 0 in $\hat{\Gamma}_n^{-1}$. Consider the almost equivalent problem of estimating the spectral density $f(\omega)$ and its inverse $1/f(\omega)$. If $\hat{f}(\omega)$ drops below 0, it can reasonably be assumed that the $f(\omega)$ is small but positive, which implies that the best available estimate of $1/f(\omega)$ is large and positive, rather than near 0; these distinctions are illustrated in his Fig. 1. Chen also offers an extension to sparse prediction; we welcome this contribution and are eager to better understand the conditions under which the sparse prediction converges to the oracle.

Wei Biao Wu proposes an upper bound of $O_p(r_n)$ on the rate of convergence of our predictors in Theorem 3. He then brings up the question on the performance of the empirical bandwidth rule of Politis (2003) in terms of minimizing this upper bound. Note, however, that this is just an upper bound, not the exact rate of convergence. As mentioned in Remark 3, under stronger conditions r_n can be replaced by r_n' in all theorems, including Wu's upper bound. Choosing which quantity to minimize could create a bit of controversy but fortunately the empirical bandwidth choice $\hat{l}$ is not focused on minimizing any such quantity. Instead, it is inspired by the fact that, under a Moving Average MA(q) model, a flat-top kernel is optimized by letting $l = q$.

Under a high-level assumption on the behavior of the maximal deviation of $\breve{\gamma}_k$ (which incidentally was verified 10 years later by Xiao and Wu (2012)), Politis (2003) was able to show that $\hat{l} \to q$ in probability under an MA(q) model. Hence, both r_n and r_n' are optimized in this case. Of course, an MA(q) assumption seems very restrictive but by the Wold decomposition we can always approximate the autocovariance structure by an MA(q) model provided we choose a large enough q. When γ_k decays exponentially, Politis (2003) showed that $\hat{l}$ increases logarithmically, and thus it again optimizes (up to a log term) both r_n and r_n'. Finally, if γ_k decays polynomially, $\hat{l}$ increases polynomially at a rate that is close—especially in the case of fast decay—to the rate optimizing r_n'. More work is undoubtedly needed here including the quantification of the MSE of the estimators that use the random bandwidth $\hat{l}$.

Yulia Gel and Wilfredo Palma draw distinctions between same realization and independent realization predictions, and further between offline and online problems. While our problem is framed in terms of same realization predictions, we see no reason our methods cannot also be made to work for independent realizations. We can say more about the distinction between the online and offline problems. Optimally implemented, our FSO shrinkage to white noise should run in $O(\max\{n \log^2 n, nl\})$ computational time, although the other corrections to positive definiteness can be slower. The $n \log^2 n$ bound results from solving the $n \times n$ Toeplitz system, while the nl rate reflects calculation of the necessary autocovariances. Naively, in an online problem, either of these rates will eventually overwhelm available computational resources as the data grows. However, the autocovariance calculations can likely be sped up by starting from and updating previous calculations, and the PSO predictor can dramatically reduce the equation solving burden.

Yulia Gel and Wilfredo Palma also offer an intriguing suggestion towards the prediction of long range dependent time series. They propose directly estimating the well-behaved Γ_n^{-1} rather than trying to estimate and invert the poorly behaved Γ_n. From this point, we foresee two further challenges. First, eq. (2) still requires an estimate of $\gamma(n)$, which is closely related to the rows of Γ_n; we therefore expect any estimate of $\gamma(n)$ to have convergence rates similar to the rates at which estimates of Γ_n converge. Second, our Theorem 3, (convergence of the predictor to the oracle) rests on being able to show that the entries of $\phi(n)$ are small for large lags, but we do not necessarily expect this to be the case for long memory processes; the approach outlined by Wei Biao Wu may offer a way forward on this front.

Coming back to the current paper, we have formulated the main part of the rejoinder on the basis of three general themes that were inspired by the discussion pieces.

R.1. AR vs. MA models

In time series analysis, there has long been a tension between Autoregressive (AR) and Moving Average (MA) model fitting. In practice, fitting an AR(p) model typically entails selecting the order p in a data-dependent way, e.g., minimizing AIC or a related criterion; this implies that we are really approximating an AR(∞) model with an AR(p) with $p \to \infty$ as $n \to \infty$, i.e., a *sieve* approximation. Analogously, one can approximate an MA(∞) model with an MA(q) with $q \to \infty$ as $n \to \infty$. However, fitting an MA(q) is practically cumbersome, and this is especially true when q is large. A surrogate for MA model fitting is approximating/estimating the spectral density via kernel smoothing, which is nothing other than tapering the sample autocovariance with a lag-window of compact support; the support, say lc_κ in the notation of our paper, is related to the order q of an underlying fitted MA(q) model, and is allowed to tend to infinity as $n \to \infty$.

To further elaborate on the tension between AR and MA/kernel smoothing consider the following cases:

i. **Spectral estimation.** AR-based spectral estimates have been available for a long time, and have been popular in the engineering literature—

see e.g. Kay (1988); their statistical performance was quantified by Berk (1974) who showed that they have a large-sample variance of $O(p/n)$. Similarly, kernel smoothed spectral estimates were introduced in the late 1940s, and have been studied extensively; see Brillinger (1993) for a historical perspective. The large-sample variance of kernel smoothed spectral estimates is typically $O(l/n)$.

ii. **Bootstrap for time series.** AR(p) and/or AR-sieve bootstrap was one of the earliest methods for time series resampling; see e.g. Kreiss and Paparoditis (2011) for a review. The AR–bootstrap is to be contrasted with (different versions of) the block bootstrap that has at its core an implicit spectral estimator based on kernel smoothing. For example, the original block bootstrap of Künsch (1989) is associated with kernel smoothing using Bartlett's triangular kernel, while the tapered block bootstrap of Paparoditis and Politis (2001) is associated with a kernel given by the self-convolution of the data taper.

iii. **Linear prediction.** As is well-known, AR-based prediction has been the most popular method; it is the subject of the present paper to offer an alternative based on windowing the sample autocovariance sequence which is intimately related to the aforementioned lag-window spectral estimates.

iv. **Autocovariance matrix estimation.** To our knowledge the first consistent estimates of Γ_n based on a time series sample of size n were given by Wu and Pourahmadi (2009); their estimators—as well as the ones that are discussed in the present paper—are based on windowing the sample autocovariance. Up to a log-term, the tapered matrix estimators have a variance that achieves the same rate of convergence as the kernel estimates of the spectral density; see e.g. the quantity r'_n in Remark 3. Interestingly, this is a setting where perhaps the AR method has been overlooked; the next section tries to do it justice.

In all the above settings, the question can be asked: which is better, AR or MA/kernel estimation. There cannot be a sweeping yes/no answer here as it depends on whether the autocovariance and/or spectral density can be better, e.g. more parsimoniously, approximated by an AR(p) or MA(q) model. For example, if the underlying time series has an AR(p) structure (with a finite order p), then one can fit an AR(p) model with finite p and achieve spectral and/or autocovariance estimation with a parametric rate of convergence of $\sqrt{n}$; if one follows the MA/kernel approach in this example, it would be necessary to let $l \to \infty$, and a rate of convergence of $\sqrt{n/l} = o(\sqrt{n})$ ensues. On the other hand, if the underlying time series has an MA(q) structure (with a finite order q), then fitting an AR(p) model necessitates letting $p \to \infty$ that yields a rate of convergence of $\sqrt{n/p} = o(\sqrt{n})$; by contrast, using a flat-top kernel here, the practitioner can use a finite l (that is bigger or equal to q), and therefore achieve spectral and/or autocovariance estimation with a parametric rate of convergence of $\sqrt{n}$.

Our simulations on the prediction problem show a similar phenomenon, i.e., in some models AR-based prediction performs better while in others the kernel/lag-window method performs just as well or better. However, as Rob Hyndman points out, the situation is not as clear cut as compared to the aforementioned estimation settings. There are at least two possible reasons for this:

(a) Rates of convergence of the two predictors have yet to be established and compared. Wei Biao Wu's discussion gives some insights on obtaining an upper bound on the rate of convergence of the lag-window predictor; we do hope that the details can be worked out in order to make these arguments rigorous. Despite the fact that AR-based predictors have been the norm for the last 100 or so years, we are not aware of rigorous results quantifying their rate of convergence.

(b) AR–based prediction is helped by an implicit model selection that is taking place; for example, choosing to leave out all terms X_t having $t < n-p+1$ in the formula for the optimal predictor of X_{n+1} offers big savings in the variance of the predictor at a (hopefully) small cost in its bias. This trade-off could/would be quantified if/when the rate of convergence of the AR-based predictor is made available; if confirmed, it may explain the edge that the AR–predictor seems to have in the real data example where the sample sizes where admittedly quite small.

R.2. AR-based estimation of the matrix Γ_n

The symmetry between AR and MA processes, suggests an alternative estimate of Γ_n via an AR(p) approximation. In other words, fit an AR(p) model to the data (with p chosen by AIC) based on the sample autocovariances $\breve{\gamma}_0, \ldots, \breve{\gamma}_p$; fitting the AR($p$) model via the Yule-Walker equations is convenient as it results in a causal (and therefore stationary) model. Then, estimate the whole autocovariance sequence by the autocovariance implied by the fitted AR model by solving the difference equation as outlined in Brockwell and Davis (1991, Section 3.3); R automates this process through the ARMAacf() function.

Denote by $\hat{\gamma}_k^{AR}$ the lag-k autocovariance of the fitted AR model. As the autocovariance sequence of a stationary time series, the sequence $\hat{\gamma}_k^{AR}$ for $k \in Z$ is positive definite. Hence if we define $\hat{\Gamma}_n^{AR}$ as the $n \times n$ Toeplitz matrix with i, j'th element given by $\hat{\gamma}_{|i-j|}^{AR}$, it then follows that $\hat{\Gamma}_n^{AR}$ will be a positive definite estimator of matrix Γ_n.

We tried this new estimator on the scenarios described in Section 5.5. The AR estimator shows substantial improvements when the time series is indeed an AR process (with a large autoregressive coefficient) but notably worse performance in all other settings; see Table R.1 where $\hat{\Gamma}_n^{AR}$ is compared to the uncorrected flat-top estimator $\hat{\Gamma}_n$, and to the two global shrinkage estimators of eq. (22) and (23). The performance of the latter was recomputed over 1,000 new replications of datasets of size $n = 200$ (as opposed to just copying the relevant entries of Table 6) in order to provide a fair comparison with $\hat{\Gamma}_n^{AR}$.

R.3. A new predictor based on the Model-Free Prediction Principle

A recent development in predictive inference is the Model-Free Prediction Principle of Politis (2013); it is interesting to see if following the Model-Free paradigm

Table R.1
Average operator norm loss of different autocovariance matrix estimators

		$\hat{\Gamma}_n$	WN-Shrink	2o-Shrink	$\hat{\Gamma}_n^{AR}$
AR(1)	$\phi = -0.9$	11.1049	10.0554	10.2395	9.2601
AR(1)	$\phi = -0.5$	0.9581	0.9839	0.9529	0.8532
AR(1)	$\phi = -0.1$	0.2956	0.2955	0.2956	0.3693
AR(1)	$\phi = 0.1$	0.2955	0.2954	0.2955	0.3699
AR(1)	$\phi = 0.5$	0.9242	0.9525	0.9198	0.8070
AR(1)	$\phi = 0.9$	9.4281	9.1549	9.4940	8.6014
MA(1)	$\theta = -0.9$	0.2946	0.2707	0.3240	1.3399
MA(1)	$\theta = -0.5$	0.2606	0.2488	0.2616	0.7212
MA(1)	$\theta = -0.1$	0.2886	0.2886	0.2886	0.3711
MA(1)	$\theta = 0.1$	0.2928	0.2928	0.2928	0.3764
MA(1)	$\theta = 0.5$	0.2570	0.2460	0.2580	0.6918
MA(1)	$\theta = 0.9$	0.2888	0.2695	0.3222	1.2947
ARMA(2,1)		1.3650	1.3570	1.3613	1.4580

can lead to a different predictor. To start with, let us assume the working hypothesis that $\{X_t, t \in \mathbf{Z}\}$ is a linear time series that is causal and invertible, i.e., it satisfies the following two equations:

$$X_t = \sum_{k=0}^{\infty} \psi_k Z_{t-k} \tag{R.1}$$

and

$$X_t = \sum_{k=1}^{\infty} \phi_k X_{t-k} + Z_t \tag{R.2}$$

with respect to innovations $\{Z_t\}$ that are i.i.d. with mean zero and variance σ^2; a typical assumption here is that the sequences ψ_k and ϕ_k are absolutely summable although square-summability of the ψ_k is enough. In such a case, it is easy to see that

$$E(X_{n+1}|X_n, X_{n-1}, \ldots) = \sum_{k=1}^{\infty} \phi_k X_{t-k}$$

where $E(X_{n+1}|X_n, X_{n-1}, \ldots)$ denotes the conditional expectation given the infinite past. In the practical setting, we only observe the finite history $X_n, \ldots, X_1$ but for large n the approximation

$$E(X_{n+1}|X_n, \ldots, X_1) \simeq \sum_{k=1}^{n} \phi_k X_{t-k}$$

holds under regularity conditions. Hence, for causal and invertible linear time series the best (with respect to MSE) predictor is (approximately) linear in its arguments.

Recall that the autocovariance matrix of data vector $\underline{X}_n = (X_1, \ldots, X_n)'$ is Γ_n which is a positive definite Toeplitz matrix. Consider a square-root decomposition

$$\Gamma_n = C_n C_n' \tag{R.3}$$

where C_n is positive definite. Now define the new vector $\underline{Z}_n^{(n)} = (Z_1^{(n)}, \ldots, Z_n^{(n)})'$ by

$$\underline{Z}_n^{(n)} = C_n^{-1} \underline{X}_n. \tag{R.4}$$

Eq. (R.4) is a *whitening* filter since the variables $Z_1^{(n)}, \ldots, Z_n^{(n)}$ are mean-zero, variance one, and uncorrelated, i.e., they constitute a *white noise* sample path.

Nevertheless, a stronger result is true if we insist that eq. (R.3) is the Cholesky decomposition of Γ_n, i.e., require that the positive definite matrix C_n is (lower) triangular. In that case, it is not hard to see that the variables $Z_1^{(n)}, \ldots, Z_n^{(n)}$ are approximately i.i.d. as the filter (R.4) gives an approximation to the inversion (R.2). In fact, the whitening filter (R.4) that uses the Cholesky decomposition of Γ_n is equivalent to the well-known innovations algorithm of Brockwell and Davis (1988); see also Rissanen and Barbosa (1969).

To elaborate, letting C_n be the (lower) triangular Cholesky factor of Γ_n implies $Z_j^{(n)} \simeq Z_j/\sigma$ for all $j \geq$ some j_0; the reason we have approximation instead of equality is due to edge effects in initializing the filter. Furthermore, transformation (R.4) is invertible so if we define the transformation $H_n : \underline{X}_n \mapsto \underline{Z}_n^{(n)}$, then H_n satisfies premise (a) of the Model-Free Prediction Principle of Politis (2013). It is easy to see that it also satisfies premise (b) of the Model-Free Prediction Principle in that we can express the (yet unobserved) X_{n+1} as a function of the current data $\underline{X}_n$ and the new (yet unobserved) $Z_{n+1}^{(n+1)}$; the details are given in the sequel—see eq. (R.7).

In order to put the Model-Free Prediction Principle to work, we need to estimate the transformation H_m both for $m = n$ and for $m = n + 1$. Recall that Section 4 developed several estimators of Γ_n that are consistent and positive definite. Let $\hat{\Gamma}_n^*$ denote one of the two global shrinkage estimators, i.e., either estimators (22) or (23). The reason we focus on the two global shrinkage estimators is they yield a matrix $\hat{\Gamma}_n^*$ that is banded and Toeplitz; see Remark 11. In addition to fast computation, the banded Toeplitz property gives us an immediate way of constructing $\hat{\Gamma}_{n+1}^*$ that is needed for transformation H_{n+1} and its inverse.

To elaborate, let us denote by $\hat{\gamma}_{|i-j|}^*$ the i, j'th element of $\hat{\Gamma}_n^*$ for $i, j = 1, \ldots, n$; by construction, the sequence $\hat{\gamma}_s^*$ for $s = 0, 1, \ldots$ is positive definite, and consistent for the true $\hat{\gamma}_s$ for $s = 0, 1, \ldots$. Hence, we define $\hat{\Gamma}_{n+1}^*$ to be the symmetric, banded Toeplitz matrix with ij element given by $\hat{\gamma}_{|i-j|}^*$ for $i, j = 1, \ldots, n + 1$. Recall that $\hat{\Gamma}_n^*$ is banded, so $\hat{\gamma}_{|i-j|}^* = 0$ if $|i - j| > lc_\kappa$. Thus, the two entries of $\hat{\Gamma}_{n+1}^*$ at the upper-right and lower-left, i.e., the i, j'th elements satisfying $|i - j| = n$, are naturally estimated by zeros.

The practical application of the Model-Free Prediction Principle in order to obtain the L_2–optimal predictor of X_{n+1} can be summarized as follows:

i. Let $\hat{C}_n$ be the (lower) triangular Cholesky factor of $\hat{\Gamma}_n^*$, and define

$$\underline{\hat{Z}}_n = \hat{C}_n^{-1} \underline{X}_n \quad \text{and hence} \quad \underline{X}_n = \hat{C}_n \underline{\hat{Z}}_n. \tag{R.5}$$

Table R.2
Root mean square prediction errors for MA(1) processes with $n = 200$, including the Model-Free predictor

	FSO-WN-Raw	FSO-WN-Shr	FSO-2o-Raw	FSO-2o-Shr	MF-WN	AR
$\theta = -0.9$	1.0626	1.0662	1.0635	1.0629	1.0647	1.0614
$\theta = -0.5$	0.9849	0.9839	0.9892	0.9885	0.9840	0.9886
$\theta = -0.1$	0.9869	0.9869	0.9869	0.9869	0.9869	0.9939
$\theta = 0.1$	1.0314	1.0314	1.0314	1.0314	1.0314	1.0348
$\theta = 0.5$	1.0087	1.0070	1.0112	1.0106	1.0070	1.0222
$\theta = 0.9$	1.0481	1.0507	1.0460	1.0484	1.0504	1.0374

Table R.3
Root mean square prediction errors for AR(1) processes with $n = 200$, including the Model-Free predictor

	FSO-WN-Raw	FSO-WN-Shr	FSO-2o-Raw	FSO-2o-Shr	MF-WN	AR
$\phi = -0.9$	1.1481	1.0968	1.0948	1.0633	1.0952	1.0091
$\phi = -0.5$	1.0121	1.0100	1.0239	1.0204	1.0102	0.9978
$\phi = -0.1$	0.9874	0.9874	0.9874	0.9874	0.9875	0.9841
$\phi = 0.1$	0.9975	0.9975	0.9975	0.9975	0.9975	0.9983
$\phi = 0.5$	1.0322	1.0298	1.0489	1.0454	1.0300	1.0093
$\phi = 0.9$	1.0942	1.0866	1.0654	1.0496	1.0849	1.0087

Ignoring the aforementioned edge effects, we have denoted $\underline{\hat{Z}}_n = (\hat{Z}_1, \ldots, \hat{Z}_n)'$ as a simple sequence as opposed to a triangular array.

ii. Let $\underline{X}_{n+1} = (X_1, \ldots, X_n, X_{n+1})'$ that includes the unobserved X_{n+1}, and $\underline{\hat{Z}}_{n+1} = (\hat{Z}_1, \ldots, \hat{Z}_n, \hat{Z}_{n+1})'$. Use the inverse transformation to write

$$\underline{X}_{n+1} = \hat{C}_{n+1}\underline{\hat{Z}}_{n+1} \tag{R.6}$$

where $\hat{C}_{n+1}$ is the (lower) triangular Cholesky factor of $\hat{\Gamma}^*_{n+1}$.

iii. Note that eq. (R.6) implies that

$$X_{n+1} = \underline{\hat{c}}_{n+1}\underline{\hat{Z}}_{n+1} \tag{R.7}$$

where $\underline{\hat{c}}_{n+1} = (\hat{c}_1, \ldots, \hat{c}_n, \hat{c}_{n+1})$ is the last row of $\hat{C}_{n+1}$.

iv. Recall that the prediction is carried out conditionally on $\underline{X}_n$. Due to eq. (R.5), the first n elements of the vector $\underline{\hat{Z}}_{n+1}$ can be treated as fixed (and known) given $\underline{X}_n$. Then, the Model-Free approximation to the L_2–optimal predictor $E(X_{n+1}|X_n, \ldots, X_1)$ is given by

$$\hat{X}_{n+1} = \sum_{i=1}^{n} \hat{c}_i \hat{Z}_i + \hat{c}_{n+1}\overline{\hat{Z}} \tag{R.8}$$

where $\overline{\hat{Z}}$ is an empirical approximation to the expected value of $\hat{Z}_{n+1}$. A natural choice is to let $\overline{\hat{Z}} = n^{-1}\sum_{i=1}^{n} \hat{Z}_i$; this is what we used in our simulations. Alternatively, we can simply estimate $\overline{\hat{Z}}$ by zero using the fact that $\hat{Z}_{n+1} \simeq Z_{n+1}/\sigma$, and $E(Z_{n+1}|X_n, \ldots, X_1) = E(Z_{n+1}) = 0$ by assumption. The two choices for $\overline{\hat{Z}}$ lead to virtually identical results in practice.

Table R.4

Root mean square prediction errors for MA(2) processes with $n = 100$, including the Model-Free predictor

	$\theta_2 = -1$	$\theta_2 = -2/3$	$\theta_2 = -1/3$	$\theta_2 = 0$	$\theta_2 = 1/3$	$\theta_2 = 2/3$	$\theta_2 = 1$
FSO-WN-Raw $\theta_1 = -1$	1.693	1.522	1.372	1.144	1.051	1.129	1.279
FSO-WN-Shr $\theta_1 = -1$	$\boxed{1.691}$	1.520	1.372	1.147	1.095	1.150	1.275
FSO-2o-Raw $\theta_1 = -1$	1.697	1.524	1.370	1.143	1.049	1.144	1.325
FSO-2o-Shr $\theta_1 = -1$	1.694	1.522	1.369	1.145	1.044	1.091	1.235
MF-WN $\theta_1 = -1$	1.691	1.521	1.371	1.145	1.089	1.148	1.274
AR $\theta_1 = -1$	1.708	$\boxed{1.506}$	$\boxed{1.340}$	$\boxed{1.139}$	$\boxed{1.032}$	$\boxed{1.075}$	$\boxed{1.154}$
FSO-WN-Raw $\theta_1 = -2/3$	1.465	1.302	1.157	$\boxed{1.034}$	1.090	1.071	1.211
FSO-WN-Shr $\theta_1 = -2/3$	$\boxed{1.459}$	1.301	1.158	1.036	1.086	1.057	1.213
FSO-2o-Raw $\theta_1 = -2/3$	1.466	1.303	1.157	1.035	1.119	1.103	1.218
FSO-2o-Shr $\theta_1 = -2/3$	1.461	$\boxed{1.299}$	1.156	1.034	1.081	1.055	1.208
MF-WN $\theta_1 = -2/3$	1.459	1.300	1.157	1.034	1.086	1.058	1.213
AR $\theta_1 = -2/3$	1.475	1.313	$\boxed{1.093}$	1.051	$\boxed{1.053}$	$\boxed{1.004}$	$\boxed{1.155}$
FSO-WN-Raw $\theta_1 = -1/3$	$\boxed{1.198}$	1.061	1.065	1.040	1.043	$\boxed{1.033}$	1.166
FSO-WN-Shr $\theta_1 = -1/3$	1.200	1.065	1.064	1.039	1.041	1.034	1.176
FSO-2o-Raw $\theta_1 = -1/3$	1.200	$\boxed{1.058}$	1.065	1.040	1.052	1.033	$\boxed{1.164}$
FSO-2o-Shr $\theta_1 = -1/3$	1.202	1.060	1.065	1.040	1.050	1.035	1.175
MF-WN $\theta_1 = -1/3$	1.198	1.062	1.063	1.039	1.041	1.034	1.175
AR $\theta_1 = -1/3$	1.228	1.076	$\boxed{1.045}$	$\boxed{1.016}$	$\boxed{1.015}$	1.038	1.173
FSO-WN-Raw $\theta_1 = 0$	1.072	$\boxed{1.073}$	1.044	$\boxed{1.020}$	1.033	1.025	1.159
FSO-WN-Shr $\theta_1 = 0$	1.084	1.078	$\boxed{1.042}$	1.020	1.032	1.025	1.167
FSO-2o-Raw $\theta_1 = 0$	$\boxed{1.066}$	1.076	1.047	1.020	1.034	1.023	$\boxed{1.156}$
FSO-2o-Shr $\theta_1 = 0$	1.072	1.079	1.045	1.020	1.033	$\boxed{1.022}$	1.161
MF-WN $\theta_1 = 0$	1.081	1.076	1.042	1.020	1.032	1.025	1.166
AR $\theta_1 = 0$	1.104	1.101	1.042	1.020	$\boxed{1.016}$	1.061	1.163
FSO-WN-Raw $\theta_1 = 1/3$	1.266	1.137	1.076	1.041	1.086	1.054	1.122
FSO-WN-Shr $\theta_1 = 1/3$	1.274	1.136	1.075	1.040	1.085	$\boxed{1.050}$	1.131
FSO-2o-Raw $\theta_1 = 1/3$	$\boxed{1.265}$	$\boxed{1.135}$	1.075	1.042	1.091	1.056	$\boxed{1.121}$
FSO-2o-Shr $\theta_1 = 1/3$	1.276	1.137	1.072	1.041	1.089	1.052	1.128
MF-WN $\theta_1 = 1/3$	1.272	1.136	1.075	1.040	1.085	1.051	1.131
AR $\theta_1 = 1/3$	1.294	1.141	$\boxed{1.066}$	$\boxed{1.031}$	$\boxed{1.047}$	1.067	1.130
FSO-WN-Raw $\theta_1 = 2/3$	1.397	1.245	1.230	1.054	1.052	1.166	1.260
FSO-WN-Shr $\theta_1 = 2/3$	$\boxed{1.391}$	1.247	1.230	1.054	1.048	1.157	1.254
FSO-2o-Raw $\theta_1 = 2/3$	1.401	$\boxed{1.244}$	1.230	$\boxed{1.052}$	1.082	1.198	1.278
FSO-2o-Shr $\theta_1 = 2/3$	1.395	1.246	1.227	1.053	1.046	1.160	1.261
MF-WN $\theta_1 = 2/3$	1.392	1.247	1.230	1.054	1.048	1.157	1.254
AR $\theta_1 = 2/3$	1.421	1.245	$\boxed{1.147}$	1.085	$\boxed{1.041}$	$\boxed{1.119}$	$\boxed{1.177}$
FSO-WN-Raw $\theta_1 = 1$	1.723	1.457	1.391	1.168	1.026	1.127	1.270
FSO-WN-Shr $\theta_1 = 1$	1.718	1.455	1.389	1.180	1.069	1.146	1.268
FSO-2o-Raw $\theta_1 = 1$	1.727	1.457	1.391	1.165	$\boxed{1.012}$	1.161	1.305
FSO-2o-Shr $\theta_1 = 1$	1.724	1.454	1.388	1.172	1.023	1.092	1.225
MF-WN $\theta_1 = 1$	1.718	1.455	1.389	1.179	1.068	1.145	1.267
AR $\theta_1 = 1$	$\boxed{1.716}$	$\boxed{1.440}$	$\boxed{1.352}$	$\boxed{1.160}$	1.024	$\boxed{1.070}$	$\boxed{1.135}$

Finally, recall our working hypothesis that $\{X_t, t \in \mathbf{Z}\}$ is a linear time series that is causal and invertible. Under this hypothesis, the conditional expectation $E(X_{n+1}|X_n, \ldots, X_1)$ is (approximately) linear in $\underline{X}_n$, and the same is true for its Model-Free estimate (R.8). However, if the working hypothesis of linearity is *not* true, then predictor (R.8) gives a novel approximation to the best *linear* predictor of X_{n+1} on the basis of $\underline{X}_n$, i.e., the orthogonal projection of X_{n+1} onto the linear span of $(X_n, \ldots, X_1)$.

Table R.5

Root mean square prediction errors for MA(2) processes with $n = 500$, including the Model-Free predictor

	$\theta_2 = -1$	$\theta_2 = -2/3$	$\theta_2 = -1/3$	$\theta_2 = 0$	$\theta_2 = 1/3$	$\theta_2 = 2/3$	$\theta_2 = 1$
FSO-WN-Raw $\theta_1 = -1$	$\boxed{1.687}$	$\boxed{1.432}$	1.283	1.102	1.054	1.042	1.106
FSO-WN-Shr $\theta_1 = -1$	1.687	1.432	1.283	1.105	1.064	1.044	1.108
FSO-2o-Raw $\theta_1 = -1$	1.687	1.437	1.288	1.093	1.091	1.040	1.095
FSO-2o-Shr $\theta_1 = -1$	1.687	1.436	1.287	1.095	1.055	$\boxed{1.039}$	1.095
MF-WN $\theta_1 = -1$	1.687	1.432	1.283	1.105	1.064	1.044	1.108
AR $\theta_1 = -1$	1.694	1.444	$\boxed{1.274}$	$\boxed{1.069}$	$\boxed{1.052}$	1.041	$\boxed{1.089}$
FSO-WN-Raw $\theta_1 = -2/3$	1.395	$\boxed{1.236}$	1.058	0.996	1.009	1.000	1.049
FSO-WN-Shr $\theta_1 = -2/3$	$\boxed{1.394}$	1.236	1.059	$\boxed{0.995}$	1.009	0.999	1.052
FSO-2o-Raw $\theta_1 = -2/3$	1.397	1.246	$\boxed{1.048}$	1.006	1.056	1.007	1.033
FSO-2o-Shr $\theta_1 = -2/3$	1.397	1.244	1.050	1.004	1.044	1.006	1.033
MF-WN $\theta_1 = -2/3$	1.394	1.236	1.059	0.995	1.009	0.999	1.051
AR $\theta_1 = -2/3$	1.404	1.243	1.058	1.002	$\boxed{0.999}$	$\boxed{0.997}$	$\boxed{1.024}$
FSO-WN-Raw $\theta_1 = -1/3$	1.153	1.058	0.979	$\boxed{0.988}$	$\boxed{1.012}$	$\boxed{0.984}$	1.048
FSO-WN-Shr $\theta_1 = -1/3$	$\boxed{1.151}$	1.061	$\boxed{0.978}$	0.988	1.012	0.984	1.049
FSO-2o-Raw $\theta_1 = -1/3$	1.167	$\boxed{1.050}$	0.987	0.988	1.012	0.989	1.041
FSO-2o-Shr $\theta_1 = -1/3$	1.163	1.051	0.987	0.988	1.012	0.988	$\boxed{1.039}$
MF-WN $\theta_1 = -1/3$	1.151	1.060	0.978	0.988	1.012	0.984	1.049
AR $\theta_1 = -1/3$	1.166	1.050	0.984	0.996	1.016	0.993	1.045
FSO-WN-Raw $\theta_1 = 0$	1.103	1.002	$\boxed{0.976}$	0.992	0.988	$\boxed{0.992}$	1.123
FSO-WN-Shr $\theta_1 = 0$	1.107	$\boxed{1.000}$	0.976	0.992	0.988	0.993	1.127
FSO-2o-Raw $\theta_1 = 0$	1.090	1.011	0.976	0.992	0.988	0.995	1.107
FSO-2o-Shr $\theta_1 = 0$	1.091	1.010	0.976	0.992	0.988	0.994	1.108
MF-WN $\theta_1 = 0$	1.106	1.001	0.976	0.992	0.988	0.993	1.127
AR $\theta_1 = 0$	$\boxed{1.074}$	1.001	0.980	$\boxed{0.991}$	$\boxed{0.985}$	1.003	$\boxed{1.101}$
FSO-WN-Raw $\theta_1 = 1/3$	1.212	1.042	$\boxed{1.011}$	$\boxed{1.004}$	$\boxed{0.967}$	$\boxed{1.008}$	1.101
FSO-WN-Shr $\theta_1 = 1/3$	1.214	1.045	1.011	1.004	0.967	1.008	1.102
FSO-2o-Raw $\theta_1 = 1/3$	1.215	$\boxed{1.034}$	1.014	1.004	0.967	1.017	1.089
FSO-2o-Shr $\theta_1 = 1/3$	1.214	1.038	1.014	1.004	0.967	1.016	$\boxed{1.088}$
MF-WN $\theta_1 = 1/3$	1.213	1.045	1.011	1.004	0.967	1.008	1.102
AR $\theta_1 = 1/3$	$\boxed{1.201}$	1.055	1.023	1.009	0.974	1.023	1.090
FSO-WN-Raw $\theta_1 = 2/3$	$\boxed{1.360}$	$\boxed{1.227}$	1.068	$\boxed{1.008}$	0.992	1.017	1.120
FSO-WN-Shr $\theta_1 = 2/3$	1.360	1.227	1.072	1.008	0.990	$\boxed{1.016}$	1.121
FSO-2o-Raw $\theta_1 = 2/3$	1.362	1.234	$\boxed{1.058}$	1.014	1.030	1.023	$\boxed{1.107}$
FSO-2o-Shr $\theta_1 = 2/3$	1.362	1.234	1.060	1.013	1.018	1.023	1.108
MF-WN $\theta_1 = 2/3$	1.361	1.227	1.072	1.008	0.990	1.016	1.121
AR $\theta_1 = 2/3$	1.371	1.246	1.073	1.014	$\boxed{0.984}$	1.022	1.116
FSO-WN-Raw $\theta_1 = 1$	$\boxed{1.654}$	1.386	1.310	1.076	1.026	0.997	1.123
FSO-WN-Shr $\theta_1 = 1$	1.654	$\boxed{1.385}$	1.310	1.078	1.031	0.996	1.127
FSO-2o-Raw $\theta_1 = 1$	1.654	1.386	1.314	1.071	1.077	1.006	1.114
FSO-2o-Shr $\theta_1 = 1$	1.654	1.386	1.314	1.070	1.028	0.999	1.115
MF-WN $\theta_1 = 1$	1.654	1.385	1.310	1.078	1.031	0.996	1.127
AR $\theta_1 = 1$	1.675	1.390	$\boxed{1.287}$	$\boxed{1.063}$	$\boxed{1.012}$	$\boxed{0.995}$	$\boxed{1.092}$

We explored the performance of the model free predictor in the scenarios considered in the main manuscript against several of the predictors investigated therein. MA(1) simulations are given in Table R.2, AR(1) simulations in Table R.3, MA(2) with $n = 100$ are in Table R.4, MA(2) with $n = 500$ are in Table R.5, and the real data example is re-analyzed in Table R.6. For simplicity, we considered the Model-Free approach where $\hat{\Gamma}_n$ was corrected to positive

TABLE R.6

Root mean square prediction errors for M3 competition data and reversed M3 competition data, including the Model-Free predictor

	Forward	Reversed
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
MF-WN	0.9809	1.0189
AR	0.8356	0.7852

definiteness using shrinkage to white noise, i.e., the matrix $\hat{\Gamma}_n^*$ used corresponded to estimator (22); hence, the notation MF-WN for the Model-Free method. Interestingly, the MF-WN method generates predictions that are of very similar quality to the FSO-WN-Shr approach; see Tables 2–6. The Model-Free approach using matrix $\hat{\Gamma}_n^*$ obtained from estimator (23), i.e., shrinkage towards a 2nd order estimator, would generate predictions that are of similar quality to the FSO-2o-Shr approach.

Thus, it looks like the Model-Free approach in essense gives a different way to compute the FSO predictor based on $\hat{\Gamma}_n^*$ in connection with the shrunk autocovariance estimator $\hat{\gamma}^*(n)$ whatever the choice of $\hat{\Gamma}_n^*$ might be. This is corroborated by the fact that, as mentioned before, the construction of predictor (R.8) was motivated by the Model-Free Prediction Principle but it is similar in spirit to the innovations algorithm of Brockwell and Davis (1988). The latter, however, assumes knowledge of Γ_n; the crucial difference is that the Model-Free predictor uses the consistent, positive definite estimator $\hat{\Gamma}_n^*$ in place of the unknown Γ_n.

In closing, we would like to reiterate our thanks to all discussants, and in particular to the Editor, George Michailidis, for making all this possible.

References

BERK, K. N. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics* 489–502. MR0421010

BRILLINGER, D. R. (1993). The digital rainbow: Some history and applications of numerical spectrum analysis. *Canadian Journal of Statistics* **21** 1–19. MR1221853

BROCKWELL, P. J. and DAVIS, R. A. (1988). Simple consistent estimation of the coefficients of a linear filter. *Stochastic Processes and Their Applications* **28** 47–59. MR0936372

BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series: Theory and Methods*, 2nd ed., Springer, New York. MR1093459

KAY, S. M. (1988). *Modern Spectral Estimation: Theory and Application*. Prentice Hall, Englewood Cliffs, NJ.

KREISS, J. P. and PAPARODITIS, E. (2011). Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society* **40** 357–378. MR2906623

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* 1217–1241. MR1015147

Paparoditis, E. and Politis, D. N. (2001). Tapered block bootstrap. *Biometrika* **88** 1105–1119. MR1872222

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

Politis, D. N. (2013). Model-free model-fitting and predictive distributions (with discussion). *TEST* **22** 183–250. MR3062250

Rissanen, J. and Barbosa, L. (1969). Properties of infinite covariance matrices and stability of optimum predictors. *Information Sciences* **1** 221–236. MR0243711

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