Estimating MA parameters through factorization of the autocovariance matrix and an MA-sieve bootstrap

Timothy L. McMurry
University of Virginia

Dimitris N. Politis
University of California, San Diego

October 6, 2017

Abstract

A new method to estimate the moving-average (MA) coefficients of a stationary time series is proposed. The new approach is based on the modified Cholesky factorization of a consistent estimator of the autocovariance matrix. Convergence rates are established, and the new estimates are used in order to implement a MA-type sieve bootstrap. Finite-sample simulations corroborate the good performance of the proposed methodology.

Keywords: ARMA models, Sieve bootstrap, Wold representation.

1 Introduction

Let $X_1, \ldots, X_n$ be a realization of a mean zero, covariance stationary, purely nondeterministic time series $\{X_t,\ t \in \mathbf{Z}\}$. A basic tool in this quite general setup is the Wold representation (Brockwell and Davis, 1991, p. 187), i.e.,

$$X_t = \sum_{k=0}^{\infty} \beta_k \epsilon_{t-k} \tag{1}$$

where $\{\epsilon_t\}$ are the white noise innovations, and $\beta = (\beta_0, \beta_1, \ldots)$ is the sequence of moving-average (MA) coefficients. Under a weak additional condition, $\{X_t\}$ also admits an infinite autoregressive (AR) representation, i.e.,

$$X_t = \sum_{k=1}^{\infty} \phi_k X_{t-k} + \epsilon_t, \tag{2}$$

where $\phi = (\phi_1, \phi_2, \ldots)$ is the sequence of AR coefficients, and $\{\epsilon_t\}$ are the same innovations appearing in (1). Eqs. (1) and (2) describe MA(∞) and AR(∞) models respectively although, strictly speaking, these are just models for the covariance structure of the process. If $\beta_k = 0$ for all $k >$ some q, then $\{X_t\}$ is said to follow an MA(q) model. Similarly, if $\phi_k = 0$ for $k >$ some p, then $\{X_t\}$ is said to follow an AR(p) model.

Faced with real data, practitioners often prefer fitting an AR(p) or AR(∞) model, where in the latter case, the infinite sum in the AR representation (2) is truncated to its first p terms as an approximation. In either case, p is estimated from the data by $\hat{p}$, chosen by a model selection criterion such as AIC. The fitted AR($\hat{p}$) model then reads:

$$X_t = \sum_{k=1}^{\hat{p}} \hat{\phi}_k X_{t-k} + \hat{\epsilon}_t. \tag{3}$$

In (3) the autoregressive coefficient estimates $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$ can be obtained by the Yule-Walker equations, least squares regression, or other closely related approaches; the residuals $\{\hat{\epsilon}_t\}$ are just the error in approximating X_t by $\sum_{k=1}^{\hat{p}} \hat{\phi}_k X_{t-k}$. In what follows, we will assume that $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$ were obtained by the Yule-Walker equations that also ensure causality of the fitted AR model.

The advantages of having an approximation such as (3) are manifold, but we focus on two in particular. First, the best one-step-ahead linear predictor of the unobserved X_{n+1} given the data $X_1, \ldots, X_n$ is immediately approximated by $\sum_{k=1}^{\hat{p}} \hat{\phi}_k X_{n+1-k}$. Second, treating the errors in (3) as if they were i.i.d. (independent, identically distributed) and resampling them, gives rise to a popular residual-based bootstrap usually termed the *AR–sieve bootstrap*; see Kreiss et al. (2011) and the references therein.

A strength of (3) is that by allowing $\hat{p}$ to increase as n increases, it can reasonably model the covariance structure of any stationary time series admitting representation (2), e.g., any ARMA(p, q) process with nonvanishing spectral density. However, if the MA component of the underlying process is significant, $\hat{p}$ might need to be very large in order to provide an accurate approximation. For example, if $\{X_t\}$ happens to follow an MA(q) model, approximation (3) parametrizes the problem with $\hat{p}$ parameters (with $\hat{p}$ diverging) instead of relying on a finite number of q parameters; doing so unnecessarily turns a parametric problem to an infinite-parametric, i.e., nonparametric, one.

The overwhelming reason that practitioners prefer fitting AR as opposed to MA models is that MA models are less straightforward to estimate. The standard tool in the literature is the innovations algorithm (Brockwell and Davis, 1991, p.245) that establishes parametric rates of convergence for the first k MA coefficients for any fixed k. Hence, if $\{X_t\}$ follows an MA(q) model with finite q, then the innovations algorithm can be successfully employed. However, the innovations algorithm does not appear helpful in estimating the whole infinite sequence of MA coefficients under an MA(∞) specification.

Recently, Krampe et al. (2016) proposed an alternative approach to estimating MA(∞) models; their approach involves first estimating the spectral density $f(\cdot)$ with a bandwidth chosen by cross-validation, then calculating the Fourier coefficients of $\log \hat{f}$, and finally using these Fourier coefficients to estimate the MA coefficients through a factorization given in Pourahmadi (1983). Krampe et al. (2016) were able to show consistency of their approach for the MA(∞) coefficients although they did not give rates of convergence. Furthermore, they used their fitted MA model (of order say $\hat{q}$) to devise a residual-based bootstrap based on simulating innovations from (1) as if they were i.i.d.; in effect, their procedure is an *MA-sieve* bootstrap since typically $\hat{q}$ would diverge as $n \to \infty$ unless, of course, $\{X_t\}$ follows an MA(q) model with finite q.

McMurry and Politis (2015) showed how to use banded and tapered autocovariance matrix estimates to obtain the aforementioned best one-step-ahead linear predictor of the unobserved X_{n+1} given the data $X_1, \ldots, X_n$ without relying on the AR approximation (3). In the paper at hand we show that by employing banded and tapered autocovariances (McMurry and Politis, 2010), rather than the raw sample autocovariance function used in the innovations algorithm, fitting an MA(q) or MA(∞) model can be re-framed as a factorization of a consistent estimate of the autocovariance matrix; this allow for easier selection of the model complexity parameter, i.e., $\hat{q}$, and provides an approach that is more direct, easier to implement, and has more tractable theoretical properties than the one proposed in Krampe et al. (2016).

It is also possible to construct an MA-sieve bootstrap using the newly estimated MA$(\hat{q})$ coefficients (with $\hat{q}$ diverging); this should be compared to the MA-sieve bootstrap of Krampe et al. (2016) but also to the Linear Process Bootstrap (LPB) of (McMurry and Politis, 2010). The latter is an alternative resampling procedure that is also based on resampling the ϵ_t errors of (1) as if they were i.i.d.; the difference is that the LPB recovers the ϵ_t via a whitening process based on the

banded and tapered estimator of the autocovariance matrix without explicitly producing estimates of AR or MA coefficients.

The remainder of the paper is structured as follows: Section 2 contains the technical background motivating our new estimators. Section 3 contains the precise statement of our results that include convergence rates of the estimated MA(∞) coefficients. The application to MA-sieve bootstrap is given in Section 4. Section 5 contains a small simulation study. Technical proofs have been placed in Section 6.

2 Technical background and construction of the MA estimators

Let $\gamma_k = E[X_t X_{t+k}]$ be the autocovariance function of the mean zero, covariance stationary, purely nondeterministic time series $\{X_t,\ t \in \mathbf{Z}\}$ producing the observed data $X_1, \ldots, X_n$. Further, let $\Gamma_m = [\gamma_{|i-j|}]_{i,j=1}^m$ be the associated $m \times m$ autocovariance matrix; in many cases it is of interest to take $m = n$ but our discussion is more general.

Let

$$\Gamma_m = L_m L_m' \tag{4}$$

where L_m is the lower triangular Cholesky factor. (Pourahmadi, 2001, p.230) additionally defines the modified Cholesky factor, B_m, by the equation $L_m = B_m \Sigma_m^{1/2}$, with $\Sigma_m^{1/2} = \mathrm{diag}(\sigma_0, \ldots, \sigma_{m-1})$. Here, $\sigma_k^2 = \mathrm{var}\left[X_{k+1} - \hat{X}_{k+1}\right]$ are the one-step-ahead prediction error variances associated with the L_2-optimal linear predictor $\hat{X}_{k+1}$ of X_{k+1}; in other words, $\hat{X}_{k+1} = E(X_{k+1}|X_k, \ldots, X_1)$. Equivalently, we may write

$$\begin{aligned} \sigma_k^2 &= |\Gamma_{k+1}|/|\Gamma_k| \\ &= \gamma_0 - \boldsymbol{\gamma}_k' \Gamma_k^{-1} \boldsymbol{\gamma}_k, \end{aligned} \tag{5}$$

where $\boldsymbol{\gamma}_k = (\gamma_1, \ldots, \gamma_k)$ (Pourahmadi, 2001, p.229).

Let $b_{m,k}$ denote the (m,k) entry of the lower triangular matrix B_m; by construction, $b_{m,m} = 1$. Now the role of B_m can be interpreted via the identity

$$X_m = \sum_{k=0}^{m-1} b_{m,m-k}(X_{m-k} - \hat{X}_{m-k}).$$

When m is large, $X_{m-k} - \hat{X}_{m-k}$ well approximates a single innovation, namely ϵ_{m-k}, and therefore $b_{m,m-k}$ estimates β_k, i.e., the corresponding MA coefficient in (1). In particular, when k is not large relative to m this intuition can be made precise, e.g., through bound (2.2) in Brockwell and Davis (1988).

Following this reasoning, if Γ_m can be well estimated, its modified Cholesky factor can be used to estimate the moving average coefficients. If we define the sample autocovariance function $\breve{\gamma}_k = n^{-1} \sum_{t=1}^{n-|k|} X_t X_{t+|k|}$, it would seem natural to estimate Γ_m with the sample autocovariance matrix $\breve{\Gamma}_m = [\breve{\gamma}_{|i-j|}]_{i,j=1}^m$. However when m is not small relative to n, $\breve{\Gamma}_m$ is not a consistent estimate of Γ_m (Wu and Pourahmadi, 2009); this concern is also related to the aforementioned shortcomings in the original innovations algorithm.

In order to address this problem, we propose instead to estimate Γ_m by $\hat{\Gamma}_m^*$, a positive definite version of the banded and tapered autocovariance matrix estimator of McMurry and Politis (2010). Let $\hat{\gamma}_k = \kappa(|k|/l)\breve{\gamma}_k$, where $l \geq 0$ is a banding parameter and $\kappa(\cdot)$ is any member of the flat-top

family of functions defined in Politis (2001), i.e., $\kappa(\cdot)$ is given as

$$\kappa(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ g(|x|) & \text{if } 1 < |x| \leq c_\kappa \\ 0 & \text{if } |x| > c_\kappa, \end{cases} \tag{6}$$

where $g(\cdot)$ is some function satisfying $|g(x)| < 1$, and c_κ is a constant satisfying $c_\kappa \geq 1$. In the terminology of the pioneering work of Parzen (1957, 1958, 1961), the flat-top function $\kappa(x)$ has infinite order as all its derivatives vanish at the origin. Nevertheless, the exact constancy of $\kappa(x)$ in a neighborhood of the origin gives an additional important benefit, namely an intuitive way to choose the associated bandwidth (Politis, 2003). A simple flat-top example is the trapezoidal taper:

$$\kappa(x) = \begin{cases} 1 & \text{if } \ |x| \leq 1 \\ 2 - |x| & \text{if } \ 1 < |x| \leq 2 \\ 0 & \text{if } \ |x| > 2. \end{cases} \tag{7}$$

The banded and tapered autocovariance matrix estimate is then given by

$$\hat{\Gamma}_m = [\hat{\gamma}_{|i-j|}]_{i,j=1}^m. \tag{8}$$

In (8), $\kappa(\cdot)$ and l work together to form $\hat{\Gamma}_m$ from $\breve{\Gamma}_m$ by leaving the $2l+1$ main diagonals of $\breve{\Gamma}_m$ intact, while gradually tapering more distant diagonals towards zero. Unfortunately, while Γ_m and $\breve{\Gamma}_m$ are positive definite, $\hat{\Gamma}_m$ may not be. Since positive definiteness is required for the Cholesky decomposition (4), $\hat{\Gamma}_m$ needs to be appropriately adjusted. There are many possible approaches (McMurry and Politis, 2015), each of which have the same asymptotic performance as $\hat{\Gamma}_m$ since, in effect, all these adjustments are asymptotically vanishing.

In what follows, we will insist on a correction method that yields an adjusted estimator $\hat{\Gamma}_m^*$ that is banded and Toeplitz; this allows us to define $\hat{\Gamma}_m^*$ for general values of m (even larger than n) and show its consistency for Γ_m; see McMurry and Politis (2015). Such an adjustment is given by the shrinkage to white noise estimator which is defined as:

$$\hat{\Gamma}_m^* = s\hat{\Gamma}_m + (1-s)\hat{\gamma}_0 I_m, \tag{9}$$

where $s \in (0, 1]$ is chosen to raise the smallest eigenvalue of $\hat{\Gamma}_m^*$ to the asymptotically vanishing positive threshold $\epsilon\hat{\gamma}_0/n^\beta$, where $\epsilon > 0$ and $\beta > 1/2$ are user defined tuning parameters.

Based on the above reasoning, we define the Cholesky decomposition $\hat{\Gamma}_m^* = \hat{L}_m\hat{L}_m'$, and we let $\hat{L}_m = \hat{B}_m\hat{\Sigma}_m^{1/2}$ be the modified Cholesky decomposition, where $\hat{\Sigma}_m^{1/2} = \text{diag}(\hat{\sigma}_0, \ldots, \hat{\sigma}_{m-1})$,

$$\hat{\sigma}_k^2 = \hat{\gamma}_0 - \hat{\boldsymbol{\gamma}}_k'(\hat{\Gamma}_k^*)^{-1}\hat{\boldsymbol{\gamma}}_k, \tag{10}$$

and where

$$\hat{\boldsymbol{\gamma}}_k = (\hat{\gamma}_1, \ldots, \hat{\gamma}_k).$$

Remark 1. Equation (10) is theoretically useful but computationally slow. A faster alternative is to compute $\hat{L}_m$, and then find $\hat{\Sigma}_m$ via $\hat{\Sigma}_m = \text{diag}(1/\hat{L}_{11}^2, 1/\hat{L}_{22}^2, \ldots, 1/\hat{L}_{mm}^2)$, where the $\hat{L}_{ii}$ are the diagonal entries of $\hat{L}_m$.

We now propose using the bottom row of $\hat{B}_m$, i.e., $\{\hat{b}_{m,k},\, k = m, m-1, \ldots, 1\}$ as our estimates of the first m MA coefficients in the Wold representation (1). In particular, letting $\hat{b}_{m,k}$ denote the (m,k) entry of $\hat{B}_m$, we propose estimating β_k by $\hat{b}_{m,m-k}$ for $k = 0, 1, \ldots, m-1$. Note that if the shrinkage to white noise estimator is used, then $\hat{B}_m$ retains a banded structure, making $\hat{b}_{m,m-\lfloor c_\kappa l \rfloor - 1} = \hat{b}_{m,m-\lfloor c_\kappa l \rfloor - 2} = \ldots = \hat{b}_{m,1} = 0$. It is therefore natural to extend the definition of our estimators to cover the infinite-dimensional parameter setting of (1) by estimating β_k to be zero whenever $k \geq m$.

Remark 2. The proposed algorithm is closely related to the innovations algorithm. The innovations algorithm is a modified Cholesky decomposition of the the un-tapered $q \times q$ autocovariance matrix (Brockwell and Davis, 1991, p.254–255), with q small compared to the sample size n; this allows estimation of an MA(q) process with finite order. By working with a much larger autocovariance matrix and tapering off-diagonals towards zero, our approach allows us to consistently estimate the coefficients of an MA(∞) process.

3 Consistency and rates of convergence of the MA estimators

The convergence of $\hat{\Gamma}_m^*$ to Γ_m is the primary result underpinning the current work, and is described in detail in McMurry and Politis (2010). While McMurry and Politis (2010) only presents the case $m = n$, a similar proof holds for all values of m, and is outlined in Lemma 1, although the resulting rate is conservative when $m << n$. This convergence is established under physical dependence measure conditions (Wu, 2005). In order to define our results, we briefly describe these conditions.

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

$$X_i = G(\ldots, \zeta_{i-1}, \zeta_i),$$

where $G(\cdot)$ 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 ζ_i' be an independent copy of ζ_i, $i \in \mathbb{Z}$. Let $\xi_i = (\ldots, \zeta_{i-1}, \zeta_i)$, $\xi_i' = (\ldots, \zeta_{-1}, \zeta_0', \zeta_1, \ldots, \zeta_i)$, 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 ζ_0 and ζ_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, define the quantity

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

is helpful.

The technical assumptions we need are as follows.

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, i.e., it satisfies equation (6).

Assumption 3. The quantity

$$r_n = l n^{-1/2} + \sum_{i=l}^{\infty} |\gamma_i| \tag{11}$$

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

All results in this paper are asymptotic and 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 need to diverge at an appropriate rate to ensure the convergence of (11). However, if $\gamma_k = 0$ for all $k >$ some q, e.g., under a moving average MA(q) model, l does not need to diverge; see Remark 1 (b) in what follows.

Assumption 4. Assume that $\sum_{k=-\infty}^{\infty} |\gamma_k| < \infty$. Also assume that for some positive constants c_1 and c_2, we have $0 < c_1 \leq f(\omega) \leq c_2 < \infty$ for all w where $f(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \gamma_k e^{-i\omega k}$ is the spectral density of $\{X_t\}$.

Note that the condition $\sum_{k=-\infty}^{\infty} |\gamma_k| < \infty$ in Assumption 4 ensures that $f(\cdot)$ is continuous, and therefore satisfies the upper bound $f(\omega) \leq c_2$; the essential requirement of Assumption 4 is, therefore, the strict positivity of the spectral density.

For a sequence $a = (a_1, a_2, \ldots)$ denote by $|a|_1 = \sum_{k=1}^{\infty} |a_k|$ and $|a|_2 = \left(\sum_{k=1}^{\infty} a_k^2\right)^{1/2}$ its ℓ_1 and ℓ_2 norm respectively; similar notation will be used when a is a finite-dimensional vector. Assumption 4 is sufficient to ensure that both the AR representation (2) and the Wold representation (1) hold true with absolutely summable coefficients, i.e., $|\phi|_1 < \infty$ and $|\beta|_1 < \infty$; for a proof, see Ch. 2 of Kreiss and Paparoditis (2017).

Recall that our estimated MA coefficients are given by the last row of $\hat{B}_m$, i.e., $e_m' \hat{B}_m$, where $e_m = (0, \ldots, 0, 1)'$ is the m'th standard basis column vector. Our first result establishes the convergence of this vector to the optimal finite predictor, given by the bottom row of B_m.

Theorem 1. *Under Assumptions 1–4, the estimated finite (of order m) predictor MA coefficients converge as a vector to the true coefficients at the rate*

$$|e_m'(\hat{B}_m - B_m)|_2 = O_p(r_n \log m).$$

Theorem 1 establishes that—as long as m is not exponentially larger than n—we can consistently estimate the finite (of order m) predictor coefficients based on the innovations method. Showing that the bottom row of $\hat{B}_m$ also provides reasonable estimates for the infinite sequence of coefficients in (1) requires additional assumptions and notation. In order to focus attention on the estimated infinite sequence in the infinite predictor problem, we propose to estimate the coefficient β_k in the Wold representation (1) by the estimates

$$\hat{\beta}_{k,m} = \begin{cases} \hat{b}_{m,m-k} & \text{if } k < m \\ 0 & \text{if } k \geq m \end{cases} \tag{12}$$

where $\hat{b}_{m,k}$ is the (m,k) entry of $\hat{B}_m$.

We are now ready to state our main result, keeping in mind that the matrix $\hat{B}_m$ is banded by construction and $\hat{\beta}_{k,m} = 0$ for $k \geq c_\kappa l$.

Theorem 2. *Under Assumptions 1–4, and picking m such that $m \geq \lfloor c_\kappa l \rfloor$, the estimated MA coefficients converge in ℓ_2 norm at the rate*

$$\left(\sum_{k=0}^{\infty} [\hat{\beta}_{k,m} - \beta_k]^2\right)^{1/2} = O_p(r_n \log m) + O\left(l^{1/2} \sum_{j=m-\lfloor c_\kappa l \rfloor + 1}^{\infty} |\phi_j|\right) + \left(\sum_{k=\lfloor c_\kappa l \rfloor + 1}^{\infty} |\beta_k|^2\right)^{1/2}, \tag{13}$$

where ϕ_k are the AR coefficients described in (2).

Remark 3. Theorem 2 helps illustrate the tradeoffs associated with different values of m and l. In the first term of (13) there is a small $\log m$ penalty for making $\hat{\Gamma}_m$ larger; in addition, m enters in the middle term leading to the requirement $m \geq \lfloor c_\kappa l \rfloor$. It is easier to elaborate by giving some concrete examples:

(a) *AR(p) model with p finite.* In this case, $\phi_j = 0$ when $j > p$, the β_k decay exponentially fast, and the middle term at the right-hand-side of (13) vanishes whenever $m - \lfloor c_\kappa l \rfloor + 1 > p$. Hence, taking $l \sim C \log n$ for some $C > 0$, and $m = \lfloor c_\kappa l \rfloor + p$ yields a parametric rate of convergence in (13) up to a $\log \log n$ term; if we were to take $m \sim n$, we would still attain the same parametric rate of convergence in (13) up to a logarithmic term.

(b) *MA(q) model with q finite.* In this case, $\beta_k = 0$ if $k > q$, the ϕ_j decay exponentially fast, and the last term of (13) vanishes whenever $\lfloor c_\kappa l \rfloor \geq q$. Letting $l = q/c_\kappa$, i.e., a constant, we again have a parametric rate of convergence in (13) up to a $\log \log n$ or logarithmic term according to the choices $m \sim \log n$ or $m \sim n$ respectively.

(c) *ARMA(p, q) model with p, q finite.* In this case, both the ϕ_j and the β_k decay exponentially fast. Taking $l \sim C \log n$ for some $C > 0$, we again have a parametric rate of convergence in (13) up to a $\log \log n$ or logarithmic term according to the choices $m = \lfloor c_\kappa l \rfloor \sim c_\kappa C \log n$ or $m \sim n$ respectively.

(d) *Polynomial decay.* If the autocovariance $\gamma(k)$ only decays polynomially fast, then the same is true for the ϕ_j and β_k coefficients; see e.g. Ch. 2 of Kreiss and Paparoditis (2017). To fix ideas, if the AR coefficients satisfy $|\phi_k| \leq Ck^{-d}$ for some $d > 1$, then the middle term at the right-hand-side of (13) is of order $l(m-l)^{-d+1}$. The optimal choice of l increases as a power of n, and the same is true for m. Since the penalty regarding using a large value of m is logarithmic in m, one might as well take $m \sim n$ here.

Hence, a conservative choice for m is $m = n$. Choice of l is more difficult as it is in essence a bandwidth choice problem. Luckily, the adaptive bandwidth rule of Politis (2003), that is elaborated upon in Politis (2011), is tailor-made for the flat-top kernels, and manages to automatically adapt to the underlying (unknown) strength of dependence. In particular, the choice of l via the adaptive bandwidth rule of Politis (2003) increases as $\log n$ in the above cases (a) and (c), it increases as a power of n in case (d), and converges in probability to q/c_κ in case (b). In the latter case, the adaptive bandwidth rule can be used as a model identification rule in fitting an MA (q) model via our novel estimation procedure.

The ℓ_1 convergence of the $\hat{\beta}_{k,m}$ is also of interest; this is also ensured under the conditions of Theorem 2 and the ℓ_1/ℓ_2 norm inequality.

Corollary 1. *Under Assumptions 1–4, and picking m such that $m \geq \lfloor c_\kappa l \rfloor$, we have*

$$\sum_{k=0}^{\infty} |\hat{\beta}_{k,m} - \beta_k| = O_p(l^{1/2} r_n \log m) + O\left(l \sum_{k=m-\lfloor c_\kappa l \rfloor + 1}^{\infty} |\phi_k| \right) + \sum_{k=\lfloor c_\kappa l \rfloor + 1}^{\infty} |\beta_k|. \tag{14}$$

Remark 4. Under slightly stronger regularity conditions, Xiao and Wu (2012) were able to show a sharper matrix convergence result:

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

where $||\cdot||_2$ denotes the standard matrix 2-norm, and where

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

In all results of our present paper, r_n can be replaced with r'_n provided our Assumptions 2 and 4 hold together with the conditions of Theorem 4 of Xiao and Wu (2012).

4 MA-sieve bootstrap

As already mentioned, the MA coefficients estimated via Theorem 2 can be used in carrying out an MA-sieve bootstrap procedure. The procedure is analogous to the one proposed by Krampe et al. (2016); the only difference is the way the MA coefficients are being estimated.

In this section, we assume the availability of data $Y_1, \ldots, Y_n$ from a covariance stationary, purely nondeterministic time series $\{Y_t,\ t \in \mathbf{Z}\}$ that does not necessarily have mean zero; denote $\mu = EY_t$. We can then work with the centered time series $X_t = Y_t - \mu$. In practice, the centering will take place using a data-based estimate of μ; this has an asymptotically negligible effect to our methodology as long as the estimate of μ is $\sqrt{n}$–convergent, e.g. the sample mean under our Assumption 4.

MA-SIEVE BOOTSTRAP ALGORITHM.

1. Based on data $Y_1, \ldots, Y_n$, define $X_t = Y_t - \bar{Y}$ for $t = 1, \ldots, n$ where $\bar{Y} = n^{-1} \sum_{t=1}^{n} Y_t$.
2. Use the new data $X_1, \ldots, X_n$ to estimate $\hat{B}_n$ and the corresponding MA coefficients from eq. (12) with some choice of m (which could be equal to n).
3. Simulate a bootstrap innovations sequence $\{\epsilon_t^*\}_{t=-\lfloor c_\kappa l \rfloor + 1}^{n}$ in an i.i.d. manner from a distribution F_m^* that has mean zero and standard deviation $\hat{\sigma}_{m-1}$.
4. Generate a bootstrap time series $\{X_t^*\}_{t=1}^{n}$ by $X_t^* = \epsilon_t^* + \sum_{k=1}^{\lfloor c_\kappa l \rfloor} b_{m,m-k} \epsilon_{t-k}^*$ for $t = 1, \ldots, n$.
5. Create a bootstrap time series $\{Y_t^*\}_{t=1}^{n}$ by letting $Y_t^* = X_t^* + \bar{Y}$ for $t = 1, \ldots, n$. Denote $\bar{Y}^* = n^{-1} \sum_{t=1}^{n} Y_t^*$.

In what follows, we show the asymptotic validity of the above MA-sieve bootstrap for the sample mean. Nevertheless, we expect that the MA-sieve bootstrap will be valid for more general statistics whose asymptotic distribution only depends on the mean and covariance structure of the data; this is in analogy to the AR–sieve bootstrap results of Kreiss et al. (2011). For example, we fully expect that the MA-sieve bootstrap will be valid in order to approximate the distribution of a kernel-smoothed estimator of the spectral density; see Jentsch and Politis (2015) for the LPB bootstrap analog. For another example, recall that if the data series $\{Y_t\}$ is *linear*, the large-sample distribution of the sample autocorrelations only depends on the covariance structure; see e.g. Bartlett's formula (Brockwell and Davis, 1991, p. 221). Hence, our MA-sieve bootstrap should work in this case as well as shown for the corresponding procedure by Krampe et al. (2016).

Theorem 3. *Assume Assumptions 1–4, and rates of increase in l and m ensuring that the left-hand-side of* (14) *converges to zero in probability. Then,*

$$\sup_x |P^*\{\sqrt{n}(\bar{Y}^* - \bar{Y}) \leq x\} - P\{\sqrt{n}(\bar{Y} - \mu) \leq x\}| \xrightarrow{P} 0 \tag{15}$$

and

$$Var^*(\sqrt{n}\bar{Y}^*) - Var(\sqrt{n}\bar{Y}) \xrightarrow{P} 0 \tag{16}$$

where $\xrightarrow{P}$ denotes convergence in probability, and P^* and Var^* denote the probability measure and variance operator respectively under the bootstrap mechanism which is conditional on the data $Y_1, \ldots, Y_n$.

Remark 5. The distribution F_m^* in step 3 of the Algorithm can well be the empirical distribution of the mean-centered residuals from the fitted MA model, i.e., Wold decomposition, of order m. These residuals can be obtained as the errors in one-step-ahead linear prediction where the predictor is obtained using the flat-top estimator Γ_m^*; (see McMurry and Politis, 2015). Alternatively, an arbitrary distribution with mean zero and standard deviation $\hat{\sigma}_{m-1}$ could be employed. For example, one may use a distribution that additionally achieves a desired kurtosis level as discussed in Krampe et al. (2016) in the hope to cover more general statistics, e.g., the sample autocovariances, under a linear process setting.

Following Remark 1, some additional discussion regarding choice of the parameters m and l is in order. Choosing $m = n$ is a conservative choice with little downside. Choosing l is more difficult but, as already mentioned, the adaptive bandwidth rule of Politis (2003) proves useful here. However, note that with respect to Theorem 3, l has to additionally satisfy a condition of the type $l^{1/2} r_n \log m \to 0$ (with r_n potentially replaceable by r'_n as discussed in Remark 2). This additional restriction on how large l can be should have little effect in practice as long as γ_k decays to zero fast enough.

The finite-sample performance of the MA-sieve algorithm is investigated in simulations presented in Section 5.

5 Simulations

We undertook a small simulation study to compare our MA estimators to the ones presented in Krampe et al. (2016). For the simulation, our estimator was implemented using the Trapezoid taper (7), matrix dimension $m = n$, the shrinkage to white noise correction to positive definiteness using $\beta = 1$ and $\epsilon = 20$ (see McMurry and Politis, 2015), and the Politis (2003) adaptive bandwidth choice rule. The Krampe et al. (2016) estimator was implemented as the authors describe, using the approximate Parzen spectral window (Brockwell and Davis, 1991, p.361), with smoothing parameter chosen by cross-validation as in Beltrão and Bloomfield (1987).

We compared performance across four models

Model 1 $X_t = 0.9X_{t-1} + \epsilon_t$,

Model 2 $X_t = 1.34X_{t-1} - 1.88X_{t-2} + 1.32X_{t-3} - 0.8X_{t-4} + \epsilon_t + 0.71\epsilon_{t-1} + 0.25\epsilon_{t-2}\epsilon_n$,

Model 3 $X_t = 0.2X_{t-1} - 0.5X_{t-2} + \epsilon_t + \sum_{j=1}^{20}(-0.95)^j \epsilon_{t-j}$,

Model 4 $X_t = \epsilon_t + 0.9\epsilon_{t-1}$,

with ϵ_t i.i.d. $N(0, 1)$.

Models 1–3 are used in Krampe et al. (2016), and provide challenging scenarios for both estimators. Model 1 requires many MA coefficients to reasonably capture the structure generated by the single large AR coefficient. Model 2 has a very large spike in the spectral density, which again can only be captured with a very large MA order. Model 3 has a very slowly decaying MA component ($0.95^{20} = 0.36$) which necessitates large values of the smoothing parameter l. In order to ensure reasonable estimates, l was restricted to be at most $n/10$, the upper limit currently implemented in the iosmooth R package (McMurry and Politis, 2017).

Table 1 contains the average squared ℓ^2 norm differences between the true and estimated MA coefficients over 1,000 simulated data sets from each model, compared across the proposed covariance matrix factorization, Krampe et al.'s estimator, and the maximum likelihood estimator with order chosen by AIC. Performance across the 3 estimators varied by model. The MLE performed

the best in model 4 at both sample sizes, and model 2 at the smaller sample size, but performed poorly for models 1 and 3. The covariance matrix estimator performed the best for models 1 and 3 at the smaller sample size and model 1 at the larger, but had trouble with model 2. Krampe et al.'s estimator performed the best for models 2 and 3 at the larger sample size, but had particular trouble with model 4.

n	Model	Cov. Matrix	Krampe et al.	MLE
128	1	0.913	0.9772	1.567
	2	42.19	39.46	38.74
	3	0.733	0.7733	1.723
	4	0.05988	0.1115	0.01352
512	1	0.3162	0.3288	0.7743
	2	68.65	52.71	64.41
	3	0.301	0.2097	0.9218
	4	0.02321	0.03991	0.002875

Table 1: Comparison of sum of square errors in MA coefficient estimates.

Krampe et al. (2016) also recommend pre-whitening the data using an AR(p) model with p selected by AIC. If the data are pre-whitened, then the optimal MA coefficients depend on p, and are therefore no-longer unique. In order to compare the effects of pre-whitening, instead of focusing on the MA coefficients, we compare the mean integrated square errors of the corresponding spectral density estimates.

We also investigated an additional pre-whitening strategy. Pre-whitening with AR order selected by AIC should tend to select too many AR parameters when the MA component of the model is substantial, leaving little work for the autocovariance matrix when it could be most valuable. In the new strategy, we considered pre-whitening with all AR orders up to the order selected by AIC, and we then selected the model with the smallest overall number of parameters, as estimated by $p + (3/2)l$, where the 3/2 is due to the fact that the trapezoidal taper effectively increases the number of parameters beyond the l autocovariances which are not downweighted.

Results of the second simulation are shown in Table 2. Columns labeled 'CM' indicated the proposed method based on covariance matrix factorization. 'New PW' indicates the proposed new pre-whitening method as discussed above, while 'PW' indicates pre-whitening with AR component order selected by AIC.

In all but one of the scenarios, one of the proposed estimators was the best performer, but the optimal estimator depended on the scenario. The benefits of pre-whitening are also inconsistent. In models 1 and 2, with the strongest AR components, pre-whitening appears beneficial. In model 3, which has a strong MA component, pre-whitening was associated with a noticeable decrease in performance. In model 4, which is a pure MA process, the new pre-whitening proposal performed competitively, while pre-whitening using a model selected by AIC was counterproductive. We had hoped the new pre-whitening approach would provide the best of both worlds, but it did not offer consistent improvements, indicating that this idea needs additional refinement.

Finally we examined the performance of the MA bootstrap using the proposed MA coefficient estimator with both a pure MA approach and the basic pre-whitening approach with an AR parameter chosen by AIC. As in Krampe et al. (2016), we use the bootstrap to estimate confidence intervals for the sample mean and for $\rho(2)$, the autocorrelation at lag 2. Results are presented as boxplots showing the bootstrap estimates of the 10'th and 90'th percentiles of $\sqrt{n}(\bar{X} - \mu)$ and $\sqrt{n}(\hat{\rho}(2) - \rho(2))$, and are shown in Figures 1 and 2. The dashed horizontal lines indicate the true

n	Model	CM – New PW	CM	CM PW	Krampe	Krampe PW
128	1	11.99	13.81	13.02	18.55	16.15
	2	8.362×10^4	9.851×10^4	7.98×10^4	9.763×10^4	8.280×10^4
	3	11.26	5.526	9.822	6.872	12.74
	4	0.04208	0.03202	0.1390	0.1086	0.1599
512	1	3.857	6.001	4.297	6.339	4.827
	2	6.057×10^4	9.279×10^4	5.691×10^4	8.698×10^4	5.815×10^4
	3	4.821	2.815	3.919	2.472	3.996
	4	0.006281	0.007138	0.05429	0.03869	0.05653

Table 2: Comparison of mean integrated square errors in spectral density estimates.

10'th and 90'th percentiles, estimated by simulation.

Figures 1 and 2 can be directly compared to Figures 1 and 2 in Krampe et al. (2016). The pure MA bootstrap is closely related to the LPB, which they assessed for models 1–3; the two approaches produced very similar results. The pre-whitened bootstrap uses the same pre-whitening strategy they used for their MA bootstrap, and the results are again similar.

In simulations for the mean (top rows of Figures 1 and 2), both the pure MA bootstrap and the pre-whitened version exhibit the same under-coverage exhibited in Krampe et al.'s simulations, although pre-whitening significantly corrects the worst flaws they noted as associated with the LPB. Surprisingly, with model 1, the two approaches perform similarly, although the pre-whitened bootstrap has coverage slightly closer to nominal. Models 2 and 3 benefit much more substantially from pre-whitening, with the pure MA bootstrap unable to consistently capture autocorrelations to a high enough lag, while keeping $\hat{\Gamma}_n$ invertible. Unsurprisingly, with model 4, the pure MA bootstrap shows the best performance. Interestingly, the pre-whitened bootstrap for the autocorrelation function (bottom rows of Figures 1 and 2) outperformed the pure MA bootstrap across all models and both sample sizes, and it performed well for models 1, 3, and 4. Model 2 caused significant trouble for both approaches.

6 Technical Proofs

Proof of Theorem 1. Since the corrections of $\hat{\Gamma}_m$ to positive definiteness are asymptotically vanishing (McMurry and Politis, 2010, Corollaries 2 and 3), we work with the raw, uncorrected estimate $\hat{\Gamma}_m$ assuming a large enough sample size. Note that

$$\begin{aligned} e_m'(\hat{B}_m - B_m) &= e_m'(\hat{L}_m\hat{\Sigma}_m^{-1/2} - L_m\Sigma_m^{-1/2}) \\ &= e_m'\left[(\hat{L}_m - L_m)(\hat{\Sigma}_m^{-1/2} - \Sigma_m^{-1/2}) + (\hat{L}_m - L_m)\Sigma_m^{-1/2} + L_m(\hat{\Sigma}_m^{-1/2} - \Sigma_m^{-1/2})\right] \\ &= e_m'(R_1 + R_2 + R_3) \end{aligned}$$

By Lemmas 2 and 3 $||R_1||_2 = O_p(r_n^2 \log m)$, where $||\cdot||_2$ is the standard matrix 2-norm. Since the diagonal entries values of Σ_m are bounded away from 0, $||R_2||_2 = O_p(r_n \log m)$ by Lemma (2). To bound $||R_3||_2$, we note $||L_m||_2^2 \leq ||\Gamma_m||_2 \leq 2\pi c_2$. Therefore, by Lemma 3, $||R_3||_2 = O_p(r_n \log m)$. Since $|e_m'|_2 = 1$, this completes the proof.

□

Lemma 1. *Under Assumptions 1–4,*

$$\left|\left|\hat{\Gamma}_m - \Gamma_m\right|\right|_2 = O_p(r_n), \tag{17}$$

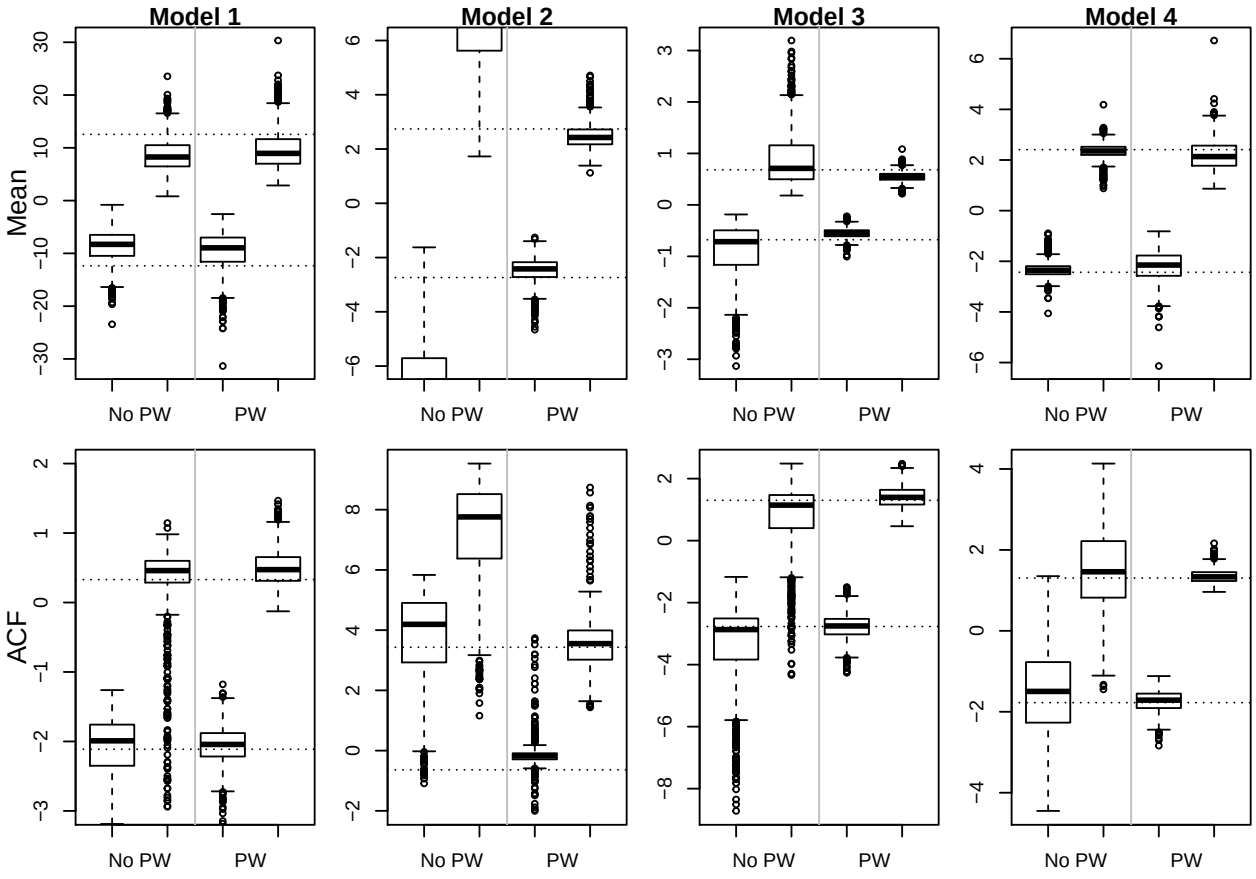

Figure 1: Bootstrap estimates of the 10'th and 90'th percentiles of $\sqrt{n}(\bar{X} - \mu)$ (top row) and $\sqrt{n}(\hat{\rho}(2) - \rho(2))$ (bottom row) with $n = 128$. The left side of each panel shows the 10'th and 90'th percentiles estimated by the pure MA bootstrap without pre-whitening, while the right shows the results of the pre-whitened bootstrap. The dashed horizontal lines indicate the theoretical quantiles.

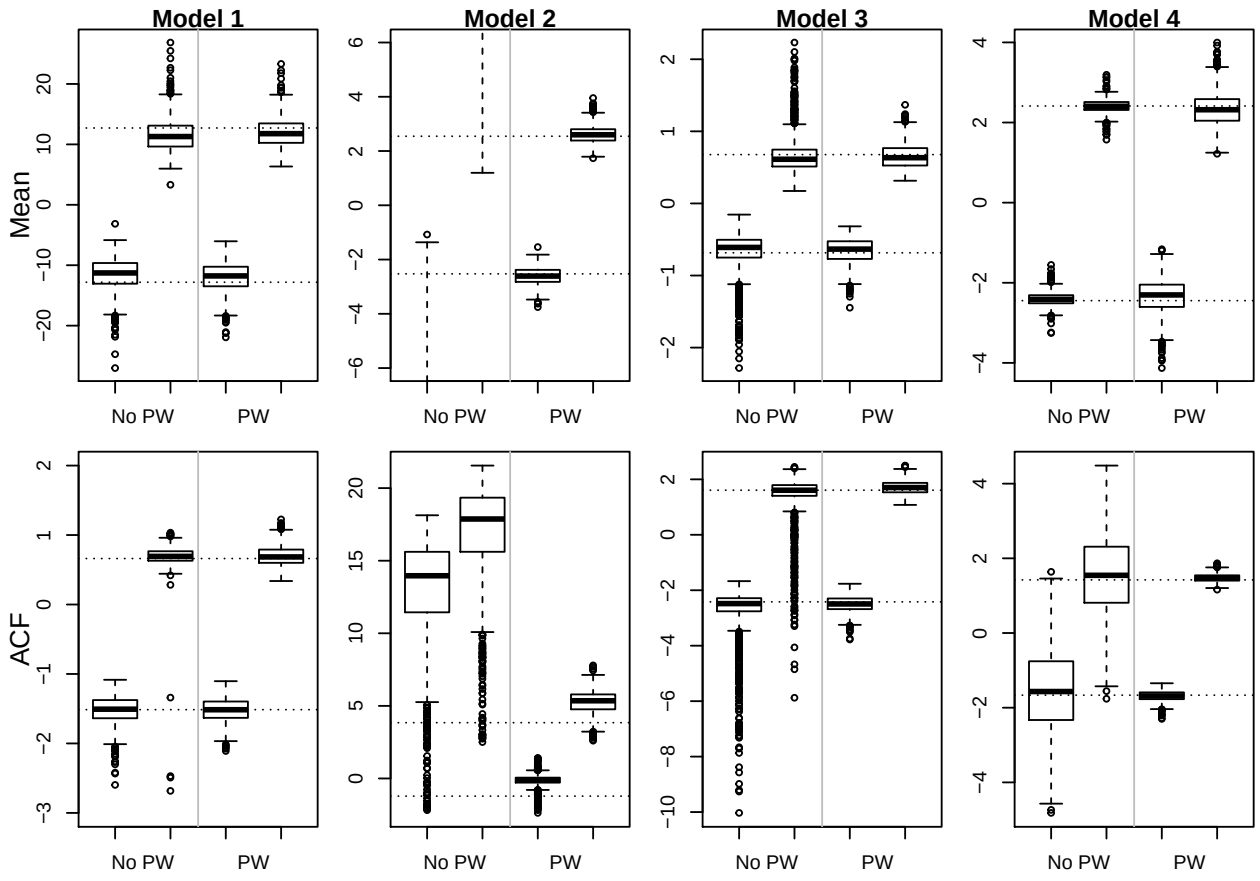

Figure 2: Bootstrap estimates of the 10'th and 90'th percentiles of $\sqrt{n}(\bar{X} - \mu)$ (top row) and $\sqrt{n}(\hat{\rho}(2) - \rho(2))$ (bottom row) with $n = 512$. The left side of each panel shows the 10'th and 90'th percentiles estimated by the pure MA bootstrap without pre-whitening, while the right shows the results of the pre-whitened bootstrap. The dashed horizontal lines indicate the theoretical quantiles.

$$|\hat{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m|_2 = O_p(r_n), \tag{18}$$

and

$$|\hat{\boldsymbol{\phi}}_m - \boldsymbol{\phi}_m|_2 = O_p(r_n), \tag{19}$$

where $\phi_m = \Gamma_m^{-1}\boldsymbol{\gamma}_m$ *is the vector of optimal prediction coefficients,* $\hat{\phi}_m = \hat{\Gamma}_m^{-1}\hat{\boldsymbol{\gamma}}_m$ *is their finite sample estimate, and where* $r_n = ln^{-1/2} + \sum_{i=l}^{\infty} |\gamma_i|$ *independent of* m*.*

Proof of Lemma 1. Following the proof of Theorem 1 in McMurry and Politis (2010).

$$\begin{aligned}
\left|\left|\hat{\Gamma}_m - \Gamma_m\right|\right|_2 &\leq \max_{1\leq j\leq n} \sum_{i=1}^{n} \left|\hat{\gamma}_{|i-j|}\kappa(|i-j|/l) - \gamma_{|i-j|}\right| \\
&\leq 2\sum_{i=0}^{l\wedge m} |\hat{\gamma}_i - \gamma_i| + 2\sum_{i=l+1}^{\lfloor c_\kappa l\rfloor \wedge m} |\hat{\gamma}_i\kappa(i/l) - \gamma_i| + 2\sum_{\lfloor c_\kappa l\rfloor+1}^{m} |\gamma_i| \\
&= T_1' + T_2' + T_3'
\end{aligned} \tag{20}$$

To prove (17), we note that because the first two sums in (20) are potentially truncated, we have $T_1' \leq T_1$ and $T_2' \leq T_2$, where T_1 and T_2 are as described in the proof of Theorem 1 in McMurry and Politis (2010). T_3' is potentially larger than T_3 when $m > n$, but both terms are bounded by $\sum_{i=l}^{\infty} |\gamma_i|$.

To establish (18), by the proof of Lemma 1 in McMurry and Politis (2015), bound (20) also holds for $|\hat{\boldsymbol{\gamma}}_p - \boldsymbol{\gamma}_p|_2$ without the factors of 2; therefore rate r_n holds.

Bound (19) results from replacing the dimension n quantities of Theorem 2 in McMurry and Politis (2015) with the matched m dimensional quantities described here, and then making use of bounds (17) and (18) instead of their n dimensional counterparts. □

Lemma 2. *Under Assumptions 1–4,*

$$\left|\left|\hat{L}_m - L_m\right|\right|_2 = O_p\left(r_n \log m\right),$$

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

Proof of Lemma 2. The proof uses a matrix factorization to take advantage of a bound in Edelman and Mascarenhas (1995). They consider the Cholesky factorization of a perturbation of the identity matrix $I + \Delta A = (I + \Delta\mathcal{L})(I + \Delta\mathcal{L})'$, where $\Delta\mathcal{L}$ is a lower triangular matrix which represents the resulting perturbation of the Cholesky factor. They show $||\Delta\mathcal{L}||_2 \leq (2\log_2 m + 4)\,||\Delta A||_2$, where the matrices are $m \times m$; they also establish a lower bound with the same rate of growth (see their Bound (2)). Using this approach, we consider the perturbation to Γ_m when it is approximated by $\hat{\Gamma}_m$:

$$\begin{aligned}
\hat{\Gamma}_m = \Gamma_m + (\hat{\Gamma}_m - \Gamma_m) &= L_m(I_m + [L_m^{-1}\hat{\Gamma}_m(L_m')^{-1} - I_m])L_m' \\
&= L_m(I_m + \Delta\mathcal{L}_m)(I_m + \Delta\mathcal{L}_m)'L_m'.
\end{aligned}$$

Therefore

$$L_m - \hat{L}_m = L_m - L_m(I_m + \Delta\mathcal{L}_m) = -L_m\Delta\mathcal{L}_m,$$

and

$$\begin{aligned}\left|\left|L_m - \hat{L}_m\right|\right|_2 &= ||L_m \Delta \mathcal{L}_m||_2 \\ &\leq (2\log_2 m + 4)\,||L_m||_2 \left|\left|L_m^{-1}\hat{\Gamma}_m (L_m')^{-1} - I_m\right|\right|_2 \\ &\leq (2\log_2 m + 4)\,||L_m||_2 \left|\left|L_m^{-1}\right|\right|_2 \left|\left|\hat{\Gamma}_m - \Gamma_m\right|\right|_2 \left|\left|(L_m')^{-1}\right|\right|_2 \\ &= O(r_n \log m),\end{aligned}$$

where the final equality follows by Lemma 1 and because the eigenvalues of L_m are bounded from above and away from 0. □

Lemma 3. *Assumptions 1–4,*

$$\left|\left|\hat{\Sigma}_m - \Sigma_m\right|\right|_2 = O_p\left(r_n\right),$$

Proof of Lemma 3. Since $\hat{\Sigma}_m$ and Σ_m are diagonal matrices, it is sufficient to examine converge of their diagonal entries.

$$\begin{aligned}\hat{\sigma}_k^2 - \sigma_k^2 &= \hat{\gamma}_0 - \hat{\boldsymbol{\gamma}}_k' \hat{\Gamma}_k^{-1} \hat{\boldsymbol{\gamma}}_k - \left(\gamma_0 - \boldsymbol{\gamma}_k' \Gamma_k^{-1} \boldsymbol{\gamma}_k\right) \\ &= (\hat{\gamma}_0 - \gamma_0) - \left(\hat{\boldsymbol{\gamma}}_k' \hat{\Gamma}_k^{-1} \hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k' \Gamma_k^{-1} \boldsymbol{\gamma}_k\right) \\ &= (\hat{\gamma}_0 - \gamma_0) + (\boldsymbol{\gamma}_k - \hat{\boldsymbol{\gamma}}_k)' \Gamma_k^{-1} \boldsymbol{\gamma}_k + (\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k)' \left(\Gamma_k^{-1} \boldsymbol{\gamma}_k - \hat{\Gamma}_k^{-1} \hat{\boldsymbol{\gamma}}_k\right) + \boldsymbol{\gamma}_k' \left(\Gamma_k^{-1} \boldsymbol{\gamma}_k - \hat{\Gamma}_k^{-1} \hat{\boldsymbol{\gamma}}_k\right) \\ &= S_1 + S_2 + S_3 + S_4\end{aligned}$$

By Lemma 1 in Wu and Pourahmadi (2009), $S_1 = O_p(n^{-1/2})$. Since the eigenvalues of Γ_k^{-1} are bounded by $(2\pi c_1)^{-1}$, $|\Gamma_k^{-1}\boldsymbol{\gamma}_k|_2 \leq (2\pi c_1)^{-1}|\boldsymbol{\gamma}_k|_2 \leq (2\pi c_1)^{-1}\sum_{i=1}^{\infty}|\gamma_i|$. Therefore, by Lemma 1 and the Cauchy-Schwarz inequality, $S_2 = O_p(r_n)$. Since there exists a finite bound for $|\boldsymbol{\gamma}_k|_2$ independent of k, bound (19) gives $S_4 = O_p(r_n)$. Finally, bounds (18) and (19) combined with the Cauchy-Schwarz inequality give $S_3 = O_p(r_n^2)$. □

Proof of Theorem 2.

$$\begin{aligned}\left[\sum_{k=0}^{\infty}(\hat{\beta}_{k,m} - \beta_k)^2\right]^{1/2} \leq &\left[\sum_{k=0}^{\lfloor c_\kappa l \rfloor}(\hat{\beta}_{k,m} - b_{m,m-k})^2\right]^{1/2} + \left[\sum_{k=0}^{\lfloor c_\kappa l \rfloor}(b_{m,m-k} - \beta_k)^2\right]^{1/2} \\ &+ \left[\sum_{k=\lfloor c_\kappa l \rfloor + 1}^{\infty}\beta_k^2\right]^{1/2} \qquad (21)\end{aligned}$$

The first term in (21) is $O_p(r_n \log m)$ by Theorem 1. The second term quantifies the difference between the finite and infinite predictors. By inequality (2.2) in Brockwell and Davis (1988) it is bounded by $\gamma_0 \sigma^{-2} l^{1/2}\left(\sum_{j=m-\lfloor c_\kappa l \rfloor + 1}^{\infty}|\phi_j|\right)$. □

Proof of Theorem 3. Under the assumed conditions, Theorem 3 of Wu (2005) implies that the large-sample distribution of $\sqrt{n}(\bar{Y} - \mu)$ is $N(0, \tau^2)$ where

$$\tau^2 = 2\pi f(0) = \sum_{k=-\infty}^{\infty}\gamma_k = \sigma^2\left|\sum_{k=0}^{\infty}\beta_k\right|^2$$

and $\sigma^2 = \lim_{k\to\infty} \sigma_k^2$ is the innovations variance in the Wold decomposition.

The proof of asymptotic normality of $\bar{Y}^*$ in the bootstrap world follows similar arguments as in the proof of Theorem 3.2 of Krampe et al. (2016). Thus, we only focus on showing eq. (16). By construction, $Var^*(\sqrt{n}\bar{Y}^*) = \hat{\sigma}^2_{m-1}|\sum_{k=0}^{\lfloor c_\kappa l \rfloor} \hat{\beta}_k|^2$ where $\hat{\sigma}^2_{m-1}$ satisfies eq. (10) with $k = m-1$. But under the assumed conditions, $\hat{\boldsymbol{\gamma}}_{m-1}$ and $\hat{\Gamma}^*_{m-1}$ are consistent for $\boldsymbol{\gamma}_{m-1}$ and Γ_{m-1} respectively. Hence, $\hat{\sigma}^2_{m-1} \xrightarrow{P} \sigma^2$ by eq. (5).

As already mentioned, Assumption 4 implies that the Wold coefficients are absolutely summable. Hence, eq. (14) yields that $|\hat{\beta} - \beta|_1 \xrightarrow{P} 0$ which implies $\sum_{k=0}^{\lfloor c_\kappa l \rfloor} \hat{\beta}_k \xrightarrow{P} \sum_{k=0}^{\infty} \beta_k$, and eq. (16) follows. □

7 Acknowledgements

Many thanks are due to Jonas Krampe, Jens-Peter Kreiss, and Stathis Paparoditis for several helpful discussions and for sharing their 2016 paper preprint as well as the draft of the Kreiss and Paparoditis (2017) textbook. Research of the second author was partially supported by NSF grants DMS 13-08319 and DMS 16-13026

References

K. I. Beltrão and P. Bloomfield. Determining the bandwidth of a kernel spectrum estimate. *Journal of time series analysis*, 8(1):21–38, 1987.

P. J. Brockwell and R. A. Davis. Simple consistent estimation of the coefficients of a linear filter. *Stochastic Processes and their Applications*, 28(1):47–59, 1988.

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

A. Edelman and W. F. Mascarenhas. On Parlett's matrix norm inequality for the Cholesky decomposition. *Numerical Linear Algebra with Applications*, 2(3):243–250, 1995.

C. Jentsch and D. N. Politis. Covariance matrix estimation and linear process bootstrap for multivariate time series of possibly increasing dimension. *The Annals of Statistics*, 43(3):1117–1140, 2015.

J. Krampe, J.-P. Kreiss, and E. Paparoditis. Estimated Wold representation and spectral density driven bootstrap for time series. Preprint, 2016.

J.-P. Kreiss and E. Paparoditis. Bootstrap for time series: Theory and applications. Textbook preprint., 2017.

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

T. L. McMurry and D. N. Politis. Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, 31(6):471–482, 2010. Corrigendum, *J. Time Ser. Anal.* **33** 2012.

T. L. McMurry and D. N. Politis. High-dimensional autocovariance matrices and optimal linear prediction. *Electronic Journal of Statistics*, 9(1):753–788, 2015.

T. L. McMurry and D. N. Politis. *iosmooth: Functions for smoothing with infinite order flat-top kernels*, 2017. R package version 0.94.

E. Parzen. On consistent estimates of the spectrum of a stationary time series. *The Annals of Mathematical Statistics*, pages 329–348, 1957.

E. Parzen. On asymptotically efficient consistent estimates of the spectral density function of a stationary time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 303–322, 1958.

E. Parzen. Mathematical considerations in the estimation of spectra. *Technometrics*, 3(2):167–190, 1961.

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

D. N. Politis. Adaptive bandwidth choice. *Journal of Nonparametric Statistics*, 15(4-5):517–533, 2003.

D. N. Politis. Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. *Econometric Theory*, 27(04):703–744, 2011.

M. Pourahmadi. Exact factorization of the spectral density ann its application to wrf, castiilg and time series analysis. *Communications in Statistics-Theory and Methods*, 12(18):2085–2094, 1983.

M. Pourahmadi. *Foundations of Time Series Analysis and Prediction Theory*, volume 379 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, New York, 2001.

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

W. B. Wu and M. Pourahmadi. Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, 19(4):1755–1768, 2009.

H. Xiao and W. B. Wu. Covariance matrix estimation for stationary time series. *The Annals of Statistics*, 40(1):466–493, 2012.