Statistica Sinica **29** (2019), 1831-1849
doi:https://doi.org/10.5705/ss.202017.0466

# THE BIAS MAPPING OF THE YULE–WALKER ESTIMATOR IS A CONTRACTION

Philip A. Ernst and Paul Shaman

*Rice University and the University of Pennsylvania*

*Abstract:* This paper addresses a long-standing conjecture that order $1/T$ bias mappings arising from Yule–Walker estimation of autoregressive coefficients are contractions, and that iteration of the order $1/T$ bias mapping gives convergence to a unique set of fixed-point process coefficients. The conjecture is easily proved for processes of order 1. We provide a proof and resolve this conjecture for order 2 processes. Although it is well-known that the Yule–Walker estimator can have substantial bias, the nature of the bias has often been only partially understood, and sometimes even *misunderstood*, in the literature. We argue that Yule–Walker fixed-point processes are key to understanding the nature of the bias. These processes provide essentially maximal separation of spectral peaks, and bias pulls Yule–Walker estimated coefficients toward those of the fixed-point process for the given order of autoregression and degree of polynomial trend for the process mean. In addition, we illustrate with a simulation that, in addition to unacceptable bias, the distribution of the Yule–Walker estimator can exhibit strong skewness and excessive kurtosis. This departure from normality can occur for very large sample sizes.

*Key words and phrases:* Autoregressive process, bias mapping, contraction, fixed-point process, Yule–Walker estimation.

## 1. Introduction

Numerous estimators of the parameters of an autoregressive process have been proposed and used. These include the Yule–Walker estimator, least-squares, maximum likelihood, Burg's estimator, and an estimator proposed by Kay (1983). Yule–Walker estimation traces back to at least Yule (1927) and Walker (1931). See Katz (2002) for an interesting discussion of the contributions of Sir Gilbert Thomas Walker.

For many years, the Yule–Walker method was widely used and was perhaps the most common technique employed for estimation of autoregressive parameters. The estimating equations are simple, the Levinson–Durbin algorithm for their solution is fast and easy to program, and the resulting estimation produces a causal autoregressive process. At present, maximum likelihood (under Gaussian assumptions), Burg's estimator, and least-squares are more commonly used. As Katz (2002) observes, computational advances over the years have led to an increase in the use of maximum likelihood estimation. Among the methods cited here, only least-squares is not guaranteed to provide estimation of a causal process. The appeal of least-squares, though, is its ease of implementation. And, if one uses least-squares and determines that it does not produce a causal estimator, then one can apply one of the other methods.

The Yule–Walker estimator is defined at (2.2) and (2.3). Despite its drawbacks, though, the Yule–Walker estimator continues to be recommended for use in time series texts, with many giving it prominent attention. They describe it as a method of moments estimator, and many mention that, despite some substandard results, Yule–Walker estimation is fully efficient and has the same asymptotic properties as maximum likelihood and the other methods. Moreover, several texts cite Yule–Walker estimation with the Levinson–Durbin algorithm as a method for calculation of the sample partial autocorrelations. Some recommend that Yule–Walker estimates be employed as initial values for maximum likelihood estimation. When caution in the use of Yule–Walker estimation is recommended, the authors tend to cite sensitivity to rounding errors, especially when the zeros of the autoregressive polynomial have moduli close to 1.

In this paper we assume that the underlying time series model is an autoregressive process of known order with a polynomial time trend for the mean. We assume the mean parameters are estimated initially by least-squares and that deviations from the mean estimation are used to estimate the autoregressive parameters.

There have been many studies of the properties of autoregressive parameter estimators, beginning with Mann and Wald (1943), who proved that the least-squares estimator is consistent and asymptotically normal. All of the estimators cited here are consistent, and all have the same asymptotic normal distribution.

Bias of the estimators has been investigated extensively, and there have been two threads for this research. One is the development of analytic expressions for the order $1/T$ bias, where $T$ is the sample size. For the least-squares estimator and a first-order process with known mean, the order $1/T$ bias was given by Marriott and Pope (1954), White (1961), and Shenton and Johnson (1965), and, for an unknown mean, it is in Marriott and Pope (1954), Kendall (1954), and White (1961). For a second-order process, Tanaka (1984) and Yamamoto and Kunitomo (1984) gave the order $1/T$ bias for both known and unknown mean cases. Bhansali (1981) developed a general expression for the least-squares order $1/T$ bias, and Tanaka (1984) and Yamamoto and Kunitomo (1984) also derived general order $1/T$ bias representations. Tjøstheim and Paulsen (1983) developed an order $1/T$ bias expression for least-squares estimation for any order process, and for Yule–Walker estimation for first- and second-order processes. Shaman and Stine (1988) and Stine and Shaman (1989) gave general expressions for the order $1/T$ bias of least-squares and Yule–Walker estimators. Pham (1993) also derived general order $1/T$ bias expressions, and showed that the least-squares, maximum likelihood, Burg, and Kay estimators all have the same order $1/T$ bias. The Yule–Walker estimator stands apart from the others, in that its order $1/T$ bias includes an extra additive term, and this term can greatly increase the bias. In addition, there have been studies of the order $1/T$ bias of maximum likelihood estimation for autoregressive moving average parameters. See Cordeiro and Klein (1994) and Cheang and Reinsel (2000).

The second thread of bias research has involved the use of simulation. Tjøstheim and Paulsen (1983) used both simulation and analytical results to argue that the Yule–Walker estimator can exhibit considerable bias, and that the Burg and least-squares estimators are preferable. Shaman and Stine (1988) noted that simulations confirm reasonable accuracy for the least-squares estimator. de Hoon et al. (1996) stated that the Yule–Walker estimator should not be employed, and advocated the use of Burg's method. The authors argued that the Yule–Walker estimator performs poorly when the autocovariance matrix in the Yule–Walker equations is ill-conditioned. This occurs when the zeros of the autoregressive polynomial are close to the unit circle. Broersen (2009) further explored the Yule–Walker bias problem and advocated the use of Burg's method, and was careful to provide evidence, however, that not all autoregressive processes with polynomial zeros near the unit circle produce large Yule–Walker estimation bias.

One should note that, in addition to estimation of autoregressive parameters, the procedures are used to estimate the partial autocorrelation coefficients, usually as an aid to the fitting of autoregressive moving average models. There can be considerable bias in estimation of these coefficients if one is using the Yule–Walker estimator.

Understanding of properties of the bias in autoregressive estimation requires consideration of the autoregressive fixed-point processes. These are processes for which the order $1/T$ bias of the parameter estimator vanishes. Their existence follows from the result that the order $1/T$ bias mappings for all of the estimation methods under discussion are contractions. That is, in each case, the distance between the order $1/T$ bias vectors for two distinct parameter vectors is less than or equal to a constant $k$, $0 \leq k < 1$, times the distance between the two parameter vectors. This implies that the order $1/T$ bias mappings have unique parameter values defining mapping fixed points, that is, parameter vector values for which application of the mapping reproduces the value of the vector. More discussion is in Section 3. There are two sets of such fixed-point processes, those for Yule–Walker estimation, and those for least-squares and the other estimation methods. In Stine and Shaman (1989), the least-squares fixed-point processes were introduced and described, and the authors also noted that numerical calculations indicate the existence of unique Yule–Walker fixed-point processes. Shaman (2010) gave coefficient vectors for least-squares and Yule–Walker fixed-point processes for autoregressive process orders 1 to 6 and polynomial trend degrees 1 and 2 for the mean. All of the coefficients of fixed-point processes are positive, and the processes are causal.

The importance of the fixed-point processes is that bias pulls the estimated autoregressive coefficients toward those of a fixed-point process, regardless of the values of the coefficients of the process generating the data. The zeros of the autoregressive polynomials for both sets of fixed-point processes are all complex in pairs, except for one negative real zero when the order is odd, and the processes have spectral peaks which are spread out on the frequency axis. That is, estimation bias moves the spectral peaks of the process being estimated, separating them, regardless of their actual positions in the observed process. Bias also alters the moduli of the autoregressive polynomial zeroes. Processes with real polynomial zeros, and those with spectral peaks that are very close together, for example, can exhibit substantial estimation bias, especially such processes with polynomial zeros close to the unit circle. Moreover, if the process structure being estimated has a relatively flat spectrum, estimation bias tends to introduce separated spectral peaks.

Stine and Shaman (1989) proved that, for least-squares estimation, the order $1/T$ bias mapping is a contraction, and observed that convergence to a fixed-point process can be achieved by iterating the order $1/T$ mapping. For least-squares estimation, the order $1/T$ bias is a linear function of the autoregressive parameters, and a fixed-point process can also be determined simply by setting the order $1/T$ bias equal to zero. Results in Pham (1993) extend these conclusions to the other autoregressive estimation methods cited above, *except* for Yule–Walker.

The first and foremost purpose of this work is to solve a conjecture from Stine and Shaman (1989) (p.1283), which states that order $1/T$ bias mappings with Yule–Walker estimation are contractions, and that iteration of the order $1/T$ bias mapping can be used to find a unique set of fixed-point process coefficients. The conjecture is easily proved for order 1 autoregressive processes. Stine and Shaman (1989) did not prove it for order 2 processes. In the current work, we successfully close this long-standing conjecture for order 2 processes. The proof is nontraditional and involves detailed calculations. After much effort, we maintain that proof of the conjecture for orders greater than 2 is analytically intractable—induction arguments fail. That is, order 2 appears to be the only nontrivial case that can be solved analytically.

The second purpose of this paper is to describe clearly the bias of the Yule–Walker estimator of the process parameters. Although it is well-known that the Yule–Walker estimator can have substantial bias, the nature of the bias has often been only partially understood, and at times misunderstood, in the literature over the past many decades.

The third purpose of this paper is to stress that, in additional to unacceptable bias, the Yule–Walker estimator can be very poorly described by its asymptotic normal distribution. In particular, the large-sample distribution of the estimator can exhibit strong skewness and excessive kurtosis. This can occur for sample sizes typically encountered in practice and, even in some cases, for very large samples. This final point is illustrated with a simulated example.

The remainder of this paper is structured as follows. Section 2 provides explicit order $1/T$ bias mapping expressions for the Yule–Walker estimator for orders 1 and 2. Section 3 constructs a fixed-point characterization for the Yule–Walker order $1/T$ bias mapping and contains the proof that this mapping is a contraction for process order 1. Section 4 has a detailed discussion of fixed-point processes. The supplement contains the proof that the Yule–Walker order $1/T$ bias mapping is a contraction for order 2 processes, as well as the proofs of Propositions 2 and 3.

## 2. Yule–Walker

### 2.1. The Yule–Walker estimator

Let $\{y_t\}$ be an autoregressive process of order $p$, henceforth denoted as

AR($p$), given by

$$\sum_{j=0}^{p} a_j(y_{t-j} - \mu_t) = \epsilon_t, \qquad a_0 = 1,$$

where $\mu_t = E(y_t)$ and the error terms $\{\epsilon_t\}$ are independently and identically distributed (i.i.d.) with mean 0 and variance $\sigma^2$. The mean is assumed to be a polynomial time trend,

$$\mu_t = \sum_{j=0}^{k-1} \beta_j t^j, \quad k \geq 0. \tag{2.1}$$

We include $k = 0$ for the case of a known mean (without loss of generality, a known mean may be taken to be 0).

Observations from this process are denoted by $\mathbf{y} = (y_1, \ldots, y_T)'$, and the vector of $p$ coefficients to be estimated is denoted by $\mathbf{a} = (a_1, \ldots, a_p)'$. We assume that the zeros of the polynomial $A_p(z) = \sum_{j=0}^{p} a_j z^{p-j}$ lie strictly inside the unit circle $|z| = 1$ so that the process is causal. The covariances of the process are given by

$$\gamma_k = E[(y_t - \mu_t)(y_{t-k} - \mu_{t-k})], \quad k = 0, \pm 1, \pm 2, \ldots,$$

and the covariance matrix of $(y_t, \ldots, y_{t-p+1})'$ is $\Gamma \equiv (\Gamma_{ij})$, where $\Gamma_{ij} = \gamma_{i-j}$. The Yule–Walker estimator of $\mathbf{a}$ is given by

$$\hat{\mathbf{a}} = -\hat{\Gamma}^{-1}\hat{\boldsymbol{\gamma}}, \tag{2.2}$$

where $\hat{\Gamma} = (\hat{\gamma}_{|i-j|})$ and $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$, with

$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=k+1}^{T} (y_t - \hat{\mu}_t)(y_{t-k} - \hat{\mu}_{t-k}), \quad k = 0, 1, \ldots, p, \tag{2.3}$$

and the parameters $\{\beta_j\}$ defining the mean function in (2.1) are estimated initially by least-squares. To ensure the validity of the approximations for the bias, we assume that the errors $\{\epsilon_t\}$ have finite moments of order 16 and that

$$E[|\hat{\Gamma}^{-1} - \Gamma^{-1}|^8] = O(1), \quad \text{as} \qquad T \to \infty,$$

where $|A|$ denotes the largest absolute eigenvalue of the matrix $A$.

### 2.2. Bias mapping expressions for the Yule–Walker estimator for $p = 1$ and $p = 2$

Following the notation of Stine and Shaman (1989), we define a $(p+1) \times (p+1)$ matrix $B$,

$$B \triangleq B_1 + B_2 + kB_3.$$

See also page 1177 of Cheang and Reinsel (2000). The matrices for $p = 1$ and $p = 2$ are $B_1 = \mathrm{diag}(0, 1, \ldots, p)$,

$$B_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad p = 1$$

$$= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \quad p = 2$$

and

$$B_3 = \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix}, \quad p = 1$$

$$= \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}, \quad p = 2.$$

Then

$$B = \begin{pmatrix} 0 & 0 \\ -k & 2+k \end{pmatrix}, \quad p = 1,$$

$$= \begin{pmatrix} 0 & 0 & 0 \\ -k & 1 & k \\ -1-k & 0 & 3+k \end{pmatrix}, \quad p = 2.$$

For $p \geq 3$, the matrices $B_j, j = 1, 2, 3$, are described in Stine and Shaman (1989). Also, define the $p \times 1$ vector $\mathbf{c}$ with elements

$$c_j = \sum_{r=0}^{p} |j - r| \gamma_{j-r} a_r, \quad j = 1, \ldots, p.$$

Then from Stine and Shaman (1989) the bias mapping for the Yule–Walker estimator is given by

$$\begin{pmatrix} 1 \\ E(\hat{\mathbf{a}}) \end{pmatrix} = \left( I - \frac{1}{T} B \right) \begin{pmatrix} 1 \\ \mathbf{a} \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{T} \Gamma^{-1} \mathbf{c} \end{pmatrix} + o\left( \frac{1}{T} \right). \qquad (2.4)$$

For $p = 1, c_1 = \gamma_1$ and, for $p = 2$, $\mathbf{c}' = (\gamma_1(1 + a_2), 2\gamma_2 + \gamma_1 a_1)$.

**Proposition 1.** *The bias mapping of the Yule–Walker estimator for $p = 1$ is given by*

$$\begin{pmatrix} 1 \\ E(\hat{a}_1) \end{pmatrix} = \begin{bmatrix} 1 \\ \dfrac{\gamma_1}{T\gamma_0} + \dfrac{k}{T} + a_1 \left( 1 - \dfrac{2+k}{T} \right) \end{bmatrix} + o\left( \frac{1}{T} \right). \qquad (2.5)$$

*Proof.* This follows directly from (2.4) and the above details.

**Proposition 2.** *The bias mapping of the Yule–Walker estimator for* $p = 2$ *is given by*

$$\begin{pmatrix} 1 \\ E(\hat{\mathbf{a}}) \end{pmatrix} = \begin{bmatrix} 1 \\ \frac{\gamma_0\gamma_1 - 2\gamma_1\gamma_2}{T(\gamma_0^2-\gamma_1^2)} + \frac{k}{T} + a_1\left(1 - \frac{\gamma_1^2}{T(\gamma_0^2-\gamma_1^2)} - \frac{1}{T}\right) + a_2\left(\frac{\gamma_0\gamma_1}{T(\gamma_0^2-\gamma_1^2)} - \frac{k}{T}\right) \\ \frac{2\gamma_0\gamma_2 - \gamma_1^2}{T(\gamma_0^2-\gamma_1^2)} + \frac{1+k}{T} + a_1\left(\frac{\gamma_0\gamma_1}{T(\gamma_0^2-\gamma_1^2)}\right) + a_2\left(1 - \frac{\gamma_1^2}{T(\gamma_0^2-\gamma_1^2)} - \frac{3+k}{T}\right) \end{bmatrix} + o\left(\frac{1}{T}\right). \tag{2.6}$$

For proof, see the supplement.

### 2.3. Covariances and coefficients

The covariances and the coefficients of the model are related through the well-known Yule–Walker equations (see Brockwell and Davis (2009)) given as

$$\sum_{i=0}^{p} a_i\gamma_{j-i} = \delta(j)\sigma^2, \quad j = 0, 1, \ldots, p,$$

where $\delta(j) = 1$ if $j = 0$ and is equal to 0 otherwise. If $p = 1$ and $j = 1$, $\gamma_1 = -a_1\gamma_0$. Substituting this into (2.5), we can rewrite the vector on the right-hand side as

$$\mathbf{h}(\tilde{\mathbf{a}}) = \begin{bmatrix} 1 \\ \frac{k}{T} + a_1\left(1 - \frac{3+k}{T}\right) \end{bmatrix} \triangleq \begin{bmatrix} 1 \\ g(a_1) \end{bmatrix}, \tag{2.7}$$

where $\tilde{\mathbf{a}} = (1, a_1)'$. For $p = 2$ and $j = 1$,

$$\gamma_1 + a_1\gamma_0 + a_2\gamma_1 = 0 \;\Leftrightarrow\; \gamma_1 = -\frac{a_1\gamma_0}{1+a_2}, \qquad a_2 \neq -1, \tag{2.8}$$

and for $j = 2$,

$$\gamma_2 + a_1\gamma_1 + a_2\gamma_0 = 0. \tag{2.9}$$

Equations (2.8) and (2.9) yield

$$\gamma_2 = -\frac{\left(a_2 + a_2^2 - a_1^2\right)\gamma_0}{1+a_2}, \qquad a_2 \neq -1. \tag{2.10}$$

Using (2.8) and (2.10), we can rewrite the vector on the right-hand side of (2.6) as a function of $\tilde{\mathbf{a}} = (1, a_1, a_2)'$,

$$\mathbf{h}(\tilde{\mathbf{a}}) = \begin{pmatrix} 1 \\ a_1 + \frac{1}{T}\left[k(1-a_2) - a_1\right] + \frac{a_1^3 - 4a_1a_2 - 3a_1a_2^2 - a_1}{T\left[(1+a_2)^2 - a_1^2\right]} \\ a_2 + \frac{1}{T}[k(1-a_2) + (1-3a_2)] - \frac{2a_2(1+a_2)^2}{T\left[(1+a_2)^2 - a_1^2\right]} \end{pmatrix} \triangleq \begin{bmatrix} 1 \\ \mathbf{g}(\mathbf{a}) \end{bmatrix}, \tag{2.11}$$

with $(1+a_2)^2 - a_1^2 \neq 0$. Note that the vector (2.11) is a nonlinear function of the autoregressive coefficients.

## 3. A Fixed-Point Characterization for the Yule–Walker Bias Mapping

We now provide a fixed-point characterization for the terms up to order $1/T$ in the bias mappings (2.5) and (2.6). The expression for a general autoregressive order $p$ is given in (2.4). See Stine and Shaman (1989) for details. The terms up to order $1/T$ of the mappings for autoregressive orders 1 and 2 are denoted by $\mathbf{h(a)}$ and are shown in (2.7) and (2.11), respectively. We argue that each of these functions forms a contraction mapping, thereby providing a unique fixed point. And numerical calculations confirm that the order $1/T$ bias mapping of the Yule–Walker estimator is a contraction for autoregressive orders greater than 2.

The function $\mathbf{h(a)}$ is a contraction mapping if, for a metric distance $d$ and some constant $0 \leq k < 1$, $d(\mathbf{h}(\mathbf{a}_1), \mathbf{h}(\mathbf{a}_2)) \leq kd(\mathbf{a}_1, \mathbf{a}_2)$, for any pair of parameter values $(\mathbf{a}_1, \mathbf{a}_2)$. The Banach fixed-point theorem states that a contraction mapping has a unique fixed point and that iteration of the mapping provides convergence to this fixed point. The implication of this in the present context is that for each autoregressive order and degree of the polynomial time trend for the process mean, there is a unique Yule–Walker fixed point parameter value, and the order $1/T$ bias at this fixed point is $\mathbf{0}$. As we have noted, such a fixed-point process has separation of spectral peaks.

Simulation studies show that the order $1/T$ bias term accurately represents the bias of Yule–Walker estimation for parameter values close to a fixed point, but it can be very inaccurate for parameter values distant from a fixed point. Simulation 1 in Section 4 is for a process very close to the fixed point. The empirical estimation bias for it is very small, and the order $1/T$ bias is extremely close to the empirical bias. In Simulation 2 in Section 4, the parameter values are far from the fixed point. The spectrum has two peaks which are close together and the zeroes of the autoregressive polynomial have amplitude 0.9, close to the unit circle. In the simulation the empirical estimation bias is substantial, and the order $1/T$ bias is grossly inaccurate. The significance of the order $1/T$ bias, though, is that it leads to calculation of a fixed point process, thereby clarifying the nature of the bias in Yule–Walker autoregressive estimation.

We specify a condition under which the order $1/T$ bias mappings have a fixed point $\tilde{\mathbf{a}}$ with first coordinate equal to 1 satisfying $\mathbf{h}(\tilde{\mathbf{a}}) = \tilde{\mathbf{a}}$. By Theorem 2.11 of Olver (2015), this holds if, for $p = 1$, the mapping $g(a_1)$ is a contraction at the fixed point $a_1^*$, and if, for $p = 2$, the mapping $\mathbf{g}(\mathbf{a})$ is a contraction at the fixed point $\mathbf{a}^* = (a_1^*, a_2^*)$. By Theorem 2.12 of Olver (2015), this reduces to showing that $|g'(a_1^*)| < 1$ for $p = 1$, and, for $p = 2$, it involves showing that the Jacobian $\mathbf{g}'(\mathbf{a}^*)$ satisfies $|\mathbf{g}'(\mathbf{a}^*)| < 1$. As in Olver (2015), we shall show that the eigenvalues of $\mathbf{g}'(\mathbf{a}^*)$ are less than one in absolute value.

### 3.1. The first-order process

It is easy to show that the bias mapping is a contraction for $p = 1$. The fixed points are obtained by solving $g(a_1^*) = a_1^*$, which yields $a_1^* = k/(3 + k)$. The derivative $g'(a_1^*)$ is $1 - (3 + k)/T$, and this is less than 1 in magnitude if $T > (3 + k)/2$.

### 3.2. The second-order process

The fixed points of $\mathbf{g}$ are obtained by solving the equation $\mathbf{g}(\mathbf{a}^*) = \mathbf{a}^*$, which gives the system

$$\begin{aligned} a_1^* + \frac{1}{T}\left[k(1 - a_2^*) - a_1^*\right] + \frac{(a_1^*)^3 - 4a_1^*a_2^* - 3a_1^*(a_2^*)^2 - a_1^*}{T\left[(1 + a_2^*)^2 - (a_1^*)^2\right]} &= a_1^*, \\ a_2^* + \frac{1}{T}[k(1 - a_2^*) + (1 - 3a_2^*)] - \frac{2a_2^*(1 + a_2^*)^2}{T\left[(1 + a_2^*)^2 - (a_1^*)^2\right]} &= a_2^*. \end{aligned}$$

The second equation of the system gives the relationship

$$(a_1^*)^2 = \frac{k + 1 - (k + 5)a_2^*}{k + 1 - (k + 3)a_2^*}(1 + a_2^*)^2, \quad a_2^* \neq \frac{k + 1}{k + 3}. \tag{3.1}$$

Substituting $(a_1^*)^2$ into the first equation of the system, we obtain:

$$a_1^* = \frac{k(1 - a_2^*)(1 + a_2^*)}{(k + 1)(1 - a_2^*) + 2}, \qquad a_2^* \neq \frac{k + 3}{k + 1}\,. \tag{3.2}$$

Finally, since from (2.11), $(1 + a_2)^2 \neq a_1^2$, the points $a_1^* = \pm 1, a_2^* = 0$ and

$a_1^* = 0, a_2^* = -1$ cannot be fixed points.

**Proposition 3.** *The matrix* $\mathbf{g}'(a_1, a_2)$ *is given by*

$$\mathbf{g}'(a_1, a_2) =$$

$$\begin{bmatrix} 1-\frac{4}{T}+\frac{2(1+a_2-a_1^2)}{T[(1+a_2)^2-a_1^2]}-\frac{4a_1^2a_2(1+a_2)}{T[(1+a_2)^2-a_1^2]^2} & -\frac{k}{T}-\frac{2a_1}{T[(1+a_2)^2-a_1^2]}+\frac{4a_1^3a_2}{T[(1+a_2)^2-a_1^2]^2} \\ -\frac{4a_1a_2(1+a_2)^2}{T[(1+a_2)^2-a_1^2]^2} & 1-\frac{k+3}{T}-\frac{2(1+a_2)^2}{T[(1+a_2)^2-a_1^2]}+\frac{4a_1^2a_2(1+a_2)}{T[(1+a_2)^2-a_1^2]^2} \end{bmatrix}. \tag{3.3}$$

For proof, see the supplement.

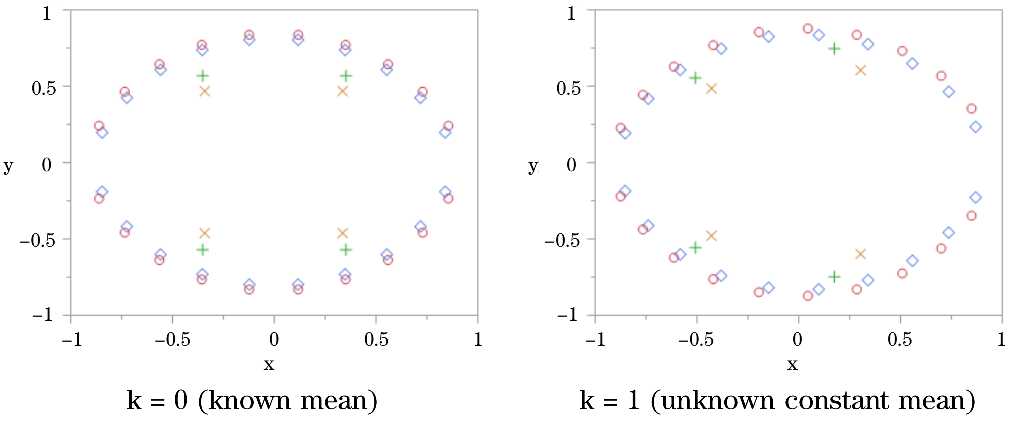
The proof for $p = 2$ that the order $1/T$ bias mapping is a contraction is given in the supplement.

## 4. Fixed-Point Processes and Discussion

We have noted that there are two sets of fixed-point processes associated with autoregressive parameter estimation, those arising from Yule–Walker estimation and those arising from estimation via least-squares and the other methods. The fixed-point processes exist for each autoregressive order $p$ and each order $k$ of the polynomial trend modeling the mean.

There are similarities and differences between the two sets of fixed-point processes. The least-squares fixed-point process coefficients can be determined analytically and are rational functions of $p$ and $k$. The Yule–Walker fixed-point process coefficients are determined numerically by iterating the order $1/T$ bias mapping. For each $k$, as $p$ varies, the least-squares fixed-point processes form a Levinson–Durbin sequence (see Shaman (2010)), but the Yule–Walker fixed-point processes do not have this property. Given this Levinson–Durbin property, the least-squares fixed-point processes are finite order projections of an infinite-dimensional parent process. The correlation function and spectral density for this infinite-dimensional process are given in Stine and Shaman (1989) for $k =$ 0 and 1, and they can also be determined for $k > 1$.

It is instructive to compare the two types of fixed-point processes by comparing the zeros of the autoregressive polynomials which the processes determine. Figure 1 compares the Yule–Walker and least-squares zeros for $p = 4$ and 20, and for $k = 0$ and 1. The arguments of the Yule–Walker and least-squares zeros have similar values, and the least-squares zeros have somewhat larger magnitudes. In addition, the least-squares fixed-point process parameter values are larger in value than the corresponding Yule–Walker parameter values. For both sets of



k = 0 (known mean)

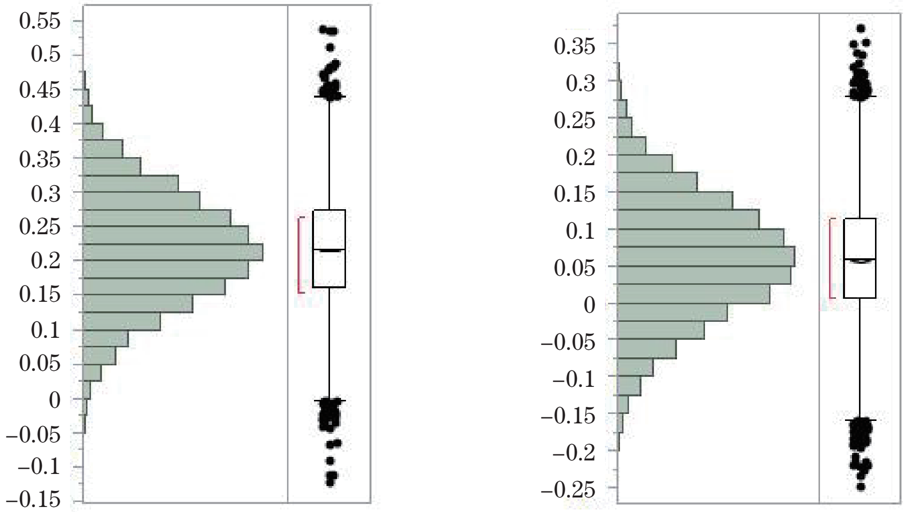
k = 1 (unknown constant mean)

Figure 1. Zeros of polynomials for least-squares and Yule–Walker fixed point processes: green $+$, $p = 4$, least-squares; orange $\times$, $p = 4$, Yule–Walker; red $\circ$, $p = 20$, least-squares; blue $\diamond$, $p = 20$, Yule–Walker.

fixed-point processes, the zeros are approximately uniformly spread around the unit circle when $k = 0$, and the angles all the zeros make with the positive horizontal axis move slowly toward 180 degrees as $k$ increases. In addition, as $p$ increases, the zeros move toward the circumference of the unit circle.

All of the estimation methods experience some finite sample bias. Bias works to separate spectral peaks and introduce complex-valued autoregressive polynomial zeros which do not exist in the data generating process. In doing so, bias moves the estimators toward the fixed-point model for the method of estimation, the order of the autoregression, and the degree of polynomial in time for the mean. If one iterates the order $1/T$ bias mapping, convergence to a fixed-point model will occur, but the number of iterations required can be large, and the path taken can be circuitous. In addition, as noted in the Introduction, for sample sizes typically encountered in practice, the distribution of the estimator can deviate substantially from the asymptotic normal distribution. These features occur when the data generating process deviates substantially from a fixed-point process. We explore these issues with several simulations.

In the first two simulations, Yule–Walker estimation is considered with 10,000 samples of length $T = 150$ for an AR(4) process with $k = 1$ (unknown constant mean). The fixed-point process coefficient vector is (0.2386, 0.3474, 0.1350, 0.1888), and the fixed-point process polynomial zeros are $(0.6755\exp(\pm i1.0995)$, $0.6432\exp(\pm i2.2947))$. The empirical bias vector is taken to be the average of the 10,000 coefficient vector estimates minus the simulated process parameter



(a) Histogram for estimate of $a_1$. Skewness = −0.00541. Kurtosis = 0.02458.

(b) Histogram for estimate of $a_3$. Skewness = −0.03796. Kurtosis = 0.03667.

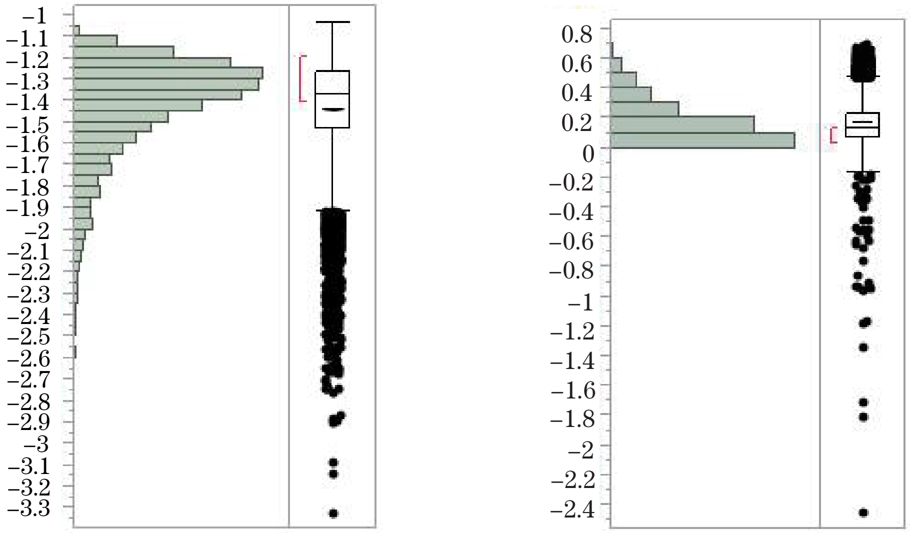
Figure 2. Histograms and summary statistics for two coefficient estimates from simulation 1.

vector. The process disturbance variance is taken to be 1.

*Simulation 1.* The simulated process has zeros $(0.5\exp(\pm i\,1.1),\ 0.5\exp(\pm i\,2.3))$, with separated and nonprominent spectral peaks. Thus, the zeros of the simulated process have arguments essentially equal to the arguments of the fixed-point process zeros, but have smaller amplitudes. The coefficient vector is (0.2127, 0.1978, 0.0532, 0.0625). The empirical bias vector is (0.0044, 0.0078, 0.0058, 0.0091), and the order $1/T$ bias vector is (0.0027, 0.0077, 0.0059, 0.0090). The average of the 10,000 estimates gives a process for which the zeroes are $(0.5180\exp(\pm i\,1.0937),\ 0.5166\exp(\pm i\,2.3058))$. The bias adjustment shows very small movement toward the fixed-point process.

The bias is very modest, and the empirical distribution of the parameter vector from this simulation approximates very well the theoretical asymptotic normal distribution for the estimation. The 4×4 inverse covariance matrix for the simulated process, multiplied by 1/150, is

$$\frac{1}{150}\Gamma^{-1} = \begin{bmatrix} 0.00664 & 0.00140 & 0.00124 & 0.00027 \\ 0.00140 & 0.00692 & 0.00161 & 0.00124 \\ 0.00124 & 0.00161 & 0.00692 & 0.00140 \\ 0.00027 & 0.00124 & 0.00140 & 0.00664 \end{bmatrix}$$



(a) Histogram for estimate of $a_1$. Skewness = −1.70850. Kurtosis = 4.11105.

(b) Histogram for estimate of $a_3$. Skewness = −0.93000. Kurtosis = 23.08779.

Figure 3. Histograms and summary statistics for two coefficient estimates from simulation 2.

and the corresponding inverse sample matrix from the simulations is

$$\frac{1}{150}\hat{\Gamma}^{-1} = \begin{bmatrix} 0.00676 & 0.00152 & 0.00138 & 0.00032 \\ 0.00152 & 0.00696 & 0.00158 & 0.00120 \\ 0.00138 & 0.00158 & 0.00679 & 0.00131 \\ 0.00032 & 0.00120 & 0.00131 & 0.00629 \end{bmatrix}.$$

Figure 2 shows the empirical distribution of the 10,000 parameter estimates for the first and third coefficients. There is good agreement with normality.

*Simulation 2.* The zeros of the simulated process are $(0.9\exp(\pm i\,0.1)$, $0.9\exp(\pm i\,0.3))$, with prominent spectral peaks that are close together, and far removed from the peaks of the fixed-point process. Moreover, the amplitudes of the zeros are much closer to 1 than are those for the fixed-point process. The coefficient vector is (−3.5106, 4.6998, −2.8436, 0.6561). The empirical bias vector is (2.0759, −4.4620, 3.0112, −0.5826), but the order $1/T$ bias vector is (1,175.905, −3,441.679, 3,426.500, −1,160.420). The average of the 10,000 estimates yields a process for which the zeroes are $(0.9209\exp(\pm i\,0.1608)$, $0.2944\exp(\pm i\,2.2799))$. There is substantial bias, especially in separation of the spectral peaks.

The empirical distribution of the parameter vector from this simulation is very poorly approximated by the theoretical asymptotic normal distribution. The 4×4 inverse covariance matrix of the simulated process, multiplied by 1/150, is

$$\frac{1}{150}\Gamma^{-1} = \begin{bmatrix} 0.00380 & -0.01097 & 0.01078 & -0.00360 \\ -0.01097 & 0.03205 & -0.03186 & 0.01078 \\ 0.01078 & -0.03186 & 0.03205 & -0.01097 \\ -0.00360 & 0.01078 & -0.01097 & 0.00380 \end{bmatrix}$$

and the corresponding inverse sample matrix from the simulations is

$$\frac{1}{150}\hat{\Gamma}^{-1} = \begin{bmatrix} 0.06126 & -0.06584 & -0.02200 & 0.02979 \\ -0.06584 & 0.07761 & 0.01568 & -0.03055 \\ -0.02200 & 0.01568 & 0.02079 & -0.01582 \\ 0.02979 & -0.03055 & -0.01582 & 0.01820 \end{bmatrix}.$$

Figure 3 gives the empirical distribution of the 10,000 parameter estimates for the first and third coefficients. The distributions are heavily skewed and there is substantial excess kurtosis. The estimation is so bad that all 10,000 estimates exceed the target values for each coefficient.

*Simulation 3.* This is the same as simulation 2, except that the sample length $T$ is taken to be 1,000. The empirical bias vector is (1.8074, −4.2561, 3.2320, −0.7617), and the order $1/T$ bias vector is (176.386, −516.252, 513.975, −174.063). The average of the 10,000 estimates gives a process with zeros (0.2600, −0.4459, $0.9544\exp(\pm i\, 0.1437)$), notably including two real values. There continues to be enormous bias, and the empirical distribution of the parameter vector is still very poorly approximated by the asymptotic normal distribution. The empirical distributions of the 10,000 estimates for the first and third coefficients are very outlier prone, with skewness values −33.14 and 21.20 and excessive kurtosis values 1,305.0 and 631.2.

*Simulation 4.* This is the same as simulation 2, except that least-squares estimation is employed. The empirical bias vector is (0.00306, −0.00413, −0.00035, 0.00174), and the order $1/T$ bias vector is (0.02570, −0.06253, 0.05027, 0.01291). The average of the 10,000 estimates gives a process with zeros ($0.9081\exp(\pm i\, 0.1213)$, $0.8931\exp(\pm i\, 0.3035)$). Despite the relatively modest bias from this simulation, the empirical distribution of the parameter vector is poorly approximated by the asymptotic normal distribution. The empirical distributions of the estimates of the first and third coefficients have skewness values 32.43 and 36.79 and excess kurtosis values 1,469.1 and 1,773.2. When the sample size $T$ is increased to 1,000, the corresponding skewness and excess kurtosis values are 7.30, 7.39 and 64.69, 64.96.

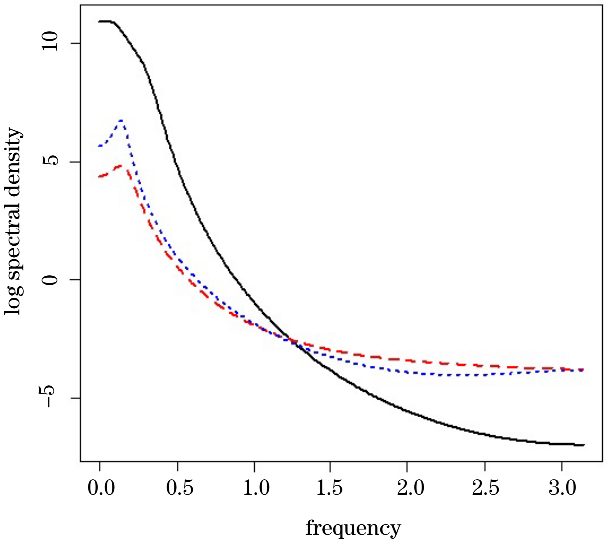


Figure 4. Log spectral densities for simulations 2 and 3: solid black line, generating process; dashed red line, process for average of 10,000 estimates, $T = 150$; dotted blue line, process for average of 10,000 estimates, $T = 1{,}000$.

Figure 4 shows the log spectral densities for the generating process in simulation 2 and for the processes corresponding to the average of the 10,000 YuleWalker estimates for T = 150 and 1,000 in simulations 2 and 3.

In summary, when the process structure being estimated is far removed from the fixed-point process, many problems ensue for Yule–Walker estimation. There is substantial estimation bias, and the theoretical asymptotic distribution does not accurately describe the distribution of the estimate for the range of sample sizes usually encountered in practice. The positions of the arguments of the zeros of the process being estimated usually play a greater role in the behavior of the estimates than do the amplitudes of the zeros. Difficulties with estimation can occur when the amplitudes of the zeros are close to 1. All of these problems are mitigated if the sample size is increased, but unreasonably large sample size can be required to obtain reliable results.

One can lessen the problems with Yule–Walker estimation by applying a data taper at the outset (see, e.g., (Zhang (1991))). Alternatively, one can employ other methods of estimation, such as Burg, maximum likelihood, Kay, or even least-squares.

In describing the theoretical bias and determining fixed-point processes for Yule–Walker estimation, we have used only the order $1/T$ component of the bias, and we have argued that this component is a contraction mapping. Simulation indicates clearly that for some generating processes this component does not accurately capture the true bias, and that in fact it can be extremely inaccurate, as simulation 2 illustrates. The purpose of the focus on the order $1/T$ component of the bias is to develop the fixed-point processes and to emphasize that these processes give guidance on how bias operates and which types of generating processes are subject to severe bias.

A key aim of this paper has been to expose and clearly describe the severe bias and distributional problems which can result from use of the Yule–Walker estimator. In addition to its use for autoregressive process parameter estimation, the Yule–Walker procedure is often employed to calculate estimates of the partial autocorrelations to aid in ARIMA model fitting. It is advisable to avoid such usage.

Additional work is needed to consider vector autoregressive process estimation bias via Yule–Walker. Also, the modified Yule–Walker procedure described by Tjøstheim and Paulsen (1983) can be studied.

## Supplementary Materials

In the online supplement we provide proofs for Propositions 2 and 3. We then provide the proof that the Yule–Walker bias mapping is a contraction for $p = 2$.

## Acknowledgments

We are grateful to the Associate Editor and to an anonymous referee, whose excellent suggestions greatly improved the quality of this manuscript.

# References

Bhansali, R. J. (1981). Effects of not knowing the order of an autoregressive process on the mean squared error of prediction–I. *Journal of the American Statistical Association* **76**, 588–597.

Brockwell, P. J. and Davis, R. A. (2009). *Time Series: Theory and Methods* (2nd Edition). New York: Springer.

Broersen, P. M. T. (2009). Finite-sample bias propagation in autogressive estimation with the Yule–Walker method. *IEEE Transactions on Instrumentation and Measurement* **58**, 1354–1360.

Cheang, W.-K. and Reinsel, G. C. (2000). Bias reduction of autoregressive estimates in time series regression models through restricted maximum likelihood. *Journal of the American*

*Statistical Association* **95**, 1173–1184.

Cordeiro, G. M. and Klein, R. (1994). Bias correction in ARMA Models. *Statistics and Probability Letters* **19**, 169–176.

de Hoon, M. J. L., van der Hagen, T. H. J. J., Schoonewelle, H. and van Dam, H. (1996). Why Yule–Walker should not be used for autoregressive modelling. *Annals of Nuclear Energy* **23**, 1219–1228.

Katz, R. W. (2002). Sir Gilbert Walker and a connection between El Niño and statistics. *Statistical Science* **17**, 97–117.

Kay, S. M. (1983). Recursive maximum likelihood estimation of autoregressive processes. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **31**, 56–65.

Kendall, M. G. (1954). Note on bias in the estimation of autocorrelation. *Biometrika* **41**, 403–404.

Mann, H. B. and Wald, A. (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica* **11**, 173–220.

Marriott, F. H. C. and Pope, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika* **41**, 390–402.

Olver, P. J. (2015). *Nonlinear Systems*. Lecture Notes, University of Minnesota.

Pham, D. T. (1993). On the asymptotic expansions for the bias and covariance matrix of autoregressive estimators. In *Developments in Time Series Analysis*, (Edited by T. Subba Rao), 80–100. London: Chapman and Hall.

Shaman, P. (2010). Generalized Levinson–Durbin sequences, binomial coefficients and autoregressive estimation. *Journal of Multivariate Analysis* **101**, 1263–1273.

Shaman, P. and Stine, R. A. (1988). The bias of autoregressive coefficient estimators. *Journal of the American Statistical Association* **83**, 842–848.

Shenton, L. R. and Johnson, W. L. (1965). Moments of a serial correlation coefficient. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **27**, 308–320.

Stine, R. A. and Shaman, P. (1989). A fixed point characterization for bias of autoregressive estimators. *The Annals of Statistics* **17**, 1275–1284.

Tanaka, K. (1984). An asymptotic expansion associated with the maximum likelihood estimators in ARMA models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **46**, 58–67.

Tjøstheim, D. and Paulsen, J. (1983). Bias of some commonly-used time series estimates. *Biometrika* **70**, 389–399.

Walker, G. (1931). On periodicity in series of related terms. *Proceedings of the Royal Society London, Series A* **131**, 518–532.

White, J. S. (1961). Asymptotic expansions for the mean and variance of the serial correlation coefficient. *Biometrika* **48**, 85–94.

Yamamoto, T. and Kunitomo, N. (1984). Asymptotic bias of the least-squares estimator for multivariate autoregressive models. *Annals of the Institute of Statistical Mathematics* **36**, 419–430.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society A* **226**, 267–298.

Zhang, H.-C. (1991). Reduction of the asymptotic bias of autoregressive and spectral estimators by tapering. *Journal of Time Series Analysis* **13**, 451–469.

Department of Statistics, Department of Rice University, Houston, TX 77005, USA.
E-mail: philip.ernst@rice.edu

Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104, USA.
E-mail: shaman@wharton.upenn.edu

(Received February 2017; accepted January 2018)