On the Relationship Between the S Array and the Box-Jenkins Method of ARMA Model Identification

Author(s): Wayne A. Woodward and H. L. Gray

Source: *Journal of the American Statistical Association*, Vol. 76, No. 375 (Sep., 1981), pp. 579-587
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: https://www.jstor.org/stable/2287515
Accessed: 27-11-2018 22:56 UTC



# On The Relationship Between the S Array and the Box-Jenkins Method of ARMA Model Identification

WAYNE A. WOODWARD and H. L. GRAY*

This paper investigates an extension of the partial autocorrelation function, which we call the generalized partial autocorrelation function. These generalized partial autocorrelations, which are not true correlations except when $p = 0$, are useful in examining the relationship between the S array method of Gray, Kelley, and McIntire (1978) and the Box-Jenkins approach to ARMA model identification. Also, the generalized partial autocorrelation is shown to be a useful model identification tool to be used along with the S array. Also discussed is a reformating of the S array into a shifted S array that the authors believe is easier to use in practice than the S array. The methods of this paper are illustrated by means of examples, including an analysis of the Makridakis (1978) metals series data.

KEY WORDS: Autoregressive moving average processes; Generalized partial autocorrelation function; Partial autocorrelation function; S array; Time series.

## 1. INTRODUCTION

Gray, Kelley, and McIntire (1978), to be referred to throughout this paper as GKM, have proposed an alternative to the Box-Jenkins approach of ARMA($p$, $q$) model identification based on the S array. Their method was shown to perform well in practice, and it uniquely determines $p$ and $q$ when the true autocorrelation function is known. Using the Box-Jenkins approach, however, when the autocorrelation function is known, unique determination of $p$ and $q$ is only assured when either $p = 0$ or $q = 0$.

In this paper, the concept of a generalized partial autocorrelation function (GPAC) is discussed. It will be shown that more information concerning the order of an ARMA($p$, $q$) process is available in the generalized partial autocorrelation function than in the partial autocorrelation function, even when $q = 0$. It will also be shown that the generalized partial autocorrelation function can be obtained as a ratio of elements in the S array, and that some significant information in the S array can be lost when the ratio is taken.

## 2. METHODS OF MODEL IDENTIFICATION

Consider the ARMA($p$, $q$) process given by

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{2.1}$$

where $\phi_k$, $k = 1, \ldots, p$ and $\theta_j$, $j = 1, \ldots, q$ are real constants with $\phi_p \neq 0$ and $\theta_q \neq 0$ and where $a_t$ is white noise. Employing the backward shift operator $B$, defined by $B^k X_t = X_{t-k}$, equation (2.1) is often written in the form $\phi(B)X_t = \theta(B)a_t$, where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$. It will be assumed that the two equations $\phi(r) = 1 - \phi_1 r - \cdots - \phi_p r^p = 0$ and $\theta(r) = 1 - \theta_1 r - \cdots - \theta_q r^q = 0$ have no common roots.

Letting $\rho_k$ denote the autocorrelation function at lag $k$, it can be shown that if $X_t$ is a stationary ARMA($p$, $q$) process, then

$$\rho_i - \phi_1 \rho_{i-1} - \cdots - \phi_p \rho_{i-p} = 0, \quad i \geq q + 1. \tag{2.2}$$

In particular, if $q = 0$, then

$$\rho_i = \phi_1 \rho_{i-1} + \cdots + \phi_p \rho_{i-p}, \quad i \geq 1. \tag{2.3}$$

Successively letting $i = 1, 2, \ldots$ in (2.3) yields the well-known Yule-Walker equations.

Basic to the Box-Jenkins approach of ARMA model identification is the partial autocorrelation function given by

$$\phi_{kk} = \rho_1 \quad \text{if } k = 1$$
$$= |A(k, 0)| / |B(k, 0)| \quad \text{if } k > 1$$

where $B(s, t)$ is the $s \times s$ matrix defined by

$$B(s, t) = \begin{pmatrix} \rho_t & \rho_{t-1} & \cdots & \rho_{t-s+1} \\ \rho_{t+1} & \rho_t & \cdots & \rho_{t-s+2} \\ \vdots & & & \\ \rho_{t+s-1} & \rho_{t+s-2} & \cdots & \rho_t \end{pmatrix}$$

and where $A(s, t)$ is the matrix composed of the first $s - 1$ columns of $B(s, t)$ with the $s$th column given by $v$ where $v' = (\rho_{t+1}, \ldots, \rho_{t+s})$. That is, if it is assumed that the process actually is ARMA($k$, 0), then $\phi_{kk}$ is the Cramer's Rule solution of the first $k$ Yule-Walker equations of order $k$ for the last autoregressive coefficient.

* Wayne A. Woodward is Assistant Professor, Department of Medical Computer Science, University of Texas Health Science Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75235. H.L. Gray is C.F. Frensley Professor of Mathematical Sciences and Dean of Academic Planning, Southern Methodist University, Dallas, TX 75275. This research was partially supported under ONR Contract N00014-75-C-0439.

© Journal of the American Statistical Association
September 1981, Volume 76, Number 375
Theory and Methods Section

The Box-Jenkins procedure uses the fact that if $X_t$ actually is ARMA($p$, 0), then $\phi_{kk}$ is nonzero for $k = p$ and identically zero for $k > p$. Also used in the Box-Jenkins procedure is the fact that if the process is ARMA(0, $q$), then $\rho_k = 0$, $k > q$. That procedure involves the inspection of graphs of the autocorrelation function and partial autocorrelation functions. However, when $p$ and $q$ are both greater than zero, the autocorrelation and partial autocorrelation function do not possess graphs that uniquely determine $p$ and $q$ by simple inspection.

Upon examination of the Yule-Walker equations it can be seen that if $X_t$ is ARMA($p$, $q$), then the autocorrelation function does not satisfy the first $q$ Yule-Walker equations but does satisfy equations $q + 1$ and following. Using this observation, we can define the GPAC function as

$$\begin{aligned} \phi_{kk}^{(j)} &= \rho_{j+1}/\rho_j && \text{if } k = 1 \\ &= |A(k,j)|/|B(k,j)| && \text{if } k > 1. \end{aligned} \tag{2.4}$$

Thus, $\phi_{kk}^{(j)}$ is the Cramer's Rule solution for the last autoregressive coefficient using the $(j + 1)$st through $(j + k)$th Yule Walker equations of order $k$. That is, $\phi_{kk}^{(j)}$ is the last autoregressive coefficient if it is assumed that the process is ARMA($k$, $j$). In particular, if $X_t$ is actually ARMA($p$, $q$), then $\phi_{pp}^{(q)} = \phi_p$. Also, if $X_t$ is ARMA($p$, $q$), then $\phi_{kk}^{(q)}$ is nonzero for $k = p$ and identically zero for $k > p$. In addition, if the process is ARMA($p$, $q$), one could use equations $q + 1$ through $q + p$ to solve for the autoregressive parameters. However, one could use the $p$ equations $q + i$ through $q + i + p - 1$ for $i = 1, 2, \ldots$ and obtain the same solution for the parameters if the true autocorrelation function is known. Thus, if $X_t$ is ARMA($p$, $q$), then $\phi_{pp}^{(q+i)} = \phi_p$, $i = 0, 1, \ldots$. These properties make it possible to uniquely identify $p$ and $q$ of a mixed process by simple inspection if the true autocorrelation function is known. Granger and Newbold (1977) note that Jenkins has also suggested the use of $\phi_{pp}^{(q)}$ for aiding in model identification.

For each $q$, $q = 0, 1, \ldots$, we will consider the GPAC function $\phi_{pp}^{(q)}$, $p = 1, 2, \ldots$. Of course, $\phi_{pp}^{(0)}$, $p = 1, 2, \ldots$ is the usual partial autocorrelation function. Two methods of summarizing this collection of GPAC functions for purposes of identifying $p$ and $q$ will be discussed. The first method is an array in which the $(i, j)$ element is given by $\phi_{jj}^{(i)}$ for $i \geq 0$ and $j \geq 1$. If $X_t$ is an ARMA ($p$, $q$) process, then the associated array will have the pattern shown in Table 1. Thus the procedure is to search for a column $p$ in which constant behavior occurs and a row $q$ in which the elements are zero for columns $k$, $k > p$. An alternative to this approach is a graphical procedure in which each row of Table 1 is graphed on the same set of axes and the patterns of Table 1 are viewed graphically. This approach is similar to the Box-Jenkins graphical approach. It should be mentioned that for $q > 0$ the generalized partial autocorrelations are not correlation coefficients of any sort as they are when $q = 0$.

GKM approach the problem of identifying $p$ and $q$ by

*Table 1. GPAC Array for an ARMA (p,q) Process*

| Moving Average Order | Autoregressive Order | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | | $p-1$ | $p$ | $p+1$ | $p+2$ | |
| 0 | $\phi_{11}^{(0)}$ | ... | $\phi_{p-1,p-1}^{(0)}$ | $\phi_{pp}^{(0)}$ | $\phi_{p+1,p+1}^{(0)}$ | $\phi_{p+2,p+2}^{(0)}$ | ... |
| ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ | |
| $q-1$ | $\phi_{11}^{(q-1)}$ | | $\phi_{p-1,p-1}^{(q-1)}$ | $\phi_{pp}^{(q-1)}$ | $\phi_{p+1,p+1}^{(q-1)}$ | $\phi_{p+2,p+2}^{(q-1)}$ | ... |
| $q$ | $\phi_{11}^{(q)}$ | | $\phi_{p-1,p-1}^{(q)}$ | $\phi_p$ | 0 | 0 | ... |
| $q+1$ | $\phi_{11}^{(q+1)}$ | | $\phi_{p-1,p-1}^{(q+1)}$ | $\phi_p$ | u* | u | ... |
| ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ | |

* u = undefined.

defining S array elements as the following ratios: $S_n(f_m) = H_{n+1}(1; f_m)/H_n(f_m)$, where $H_n(f_m)$ is the determinant of the $n \times n$ matrix with $(i, j)$ element given by $f_{m+i+j-2}$, and $H_{n+1}(1; f_m)$ is the determinant of the $(n + 1) \times (n + 1)$ matrix with $(1, j)$ element equal to 1 and $(i, j)$ element for $i \geq 2$ given by $f_{m+i+j-3}$. In their work, $f_m = \rho_m$ or $f_m = (-1)^m\rho_m$, where $\rho_m$ is the autocorrelation function of the ARMA($p$, $q$) process at lag $m$. A simple iterative method of calculating the S array elements is also available (see GKM).

The properties of the S array on which the GKM procedure depend are summarized in Theorem 1.

*Theorem 1.* Let $\{X_t\}$ be a stationary ARMA($p$, $q$) process with autocorrelation $\rho_m$. Suppose $S_n(f_m)$ is defined, $p > 0$ and $S_n(f_m) \neq 0$ and let $f_m = \rho_m$ or $f_m = (-1)^m\rho_m$.

(a) For some integer $M$ and some constant $C_1 \neq 0$,

$$S_n(f_m) = C_1, \quad m \geq M,$$
$$S_n(f_{M-1}) \neq C_1$$

if and only if $n = p$ and $M = q - p + 1$. Also

$$S_n(f_m) = C_2, \quad m \leq N$$
$$S_n(f_{N+1}) \neq C_2$$

if and only if $n = p$ and $N = -q - p$. Moreover,

$$\left.\begin{aligned} C_1 &= (-1)^p\left(1 - \sum_{k=1}^{p}\phi_k\right) \\ C_2 &= -C_1/\phi_p \end{aligned}\right\} \text{if } f_m = \rho_m$$

and

$$\left.\begin{aligned} C_1 &= (-1)^p\left(1 - \sum_{k=1}^{p}\phi_k\right) \\ C_2 &= (-1)^{p+1}C_1/\phi_p \end{aligned}\right\} \text{if } f_m = (-1)^m\rho_m.$$

(b) For $k > n$, $S_k(f_{-k-m}) = \pm\infty$ and $S_k(f_{-k+m+1}) = (-1)^{k-n}S_n(f_{-n+m+1})$ iff $n = p$ and $m = q$.

*Proof.* For the proof of (a) see Gray, Kelley, and McIntyre (1978). To prove (b) we first let $f_m = \rho_m$ and suppose $n = p$ and $m = q$, i.e. $X_t$ is ARMA($n$, $m$) with

autoregressive coefficients $\phi_k$, $k = 1, \ldots, n$. Since

$$\rho_j = \sum_{k=1}^{n} \phi_k \rho_{j-k}, \quad j \geq m + 1,$$

it can be shown that the first column in the matrix associated with $H_{n+i}(\rho_{-n-m-i})$, $i > 0$, can be expressed as a linear combination of columns two through $n + 1$ and thus that $H_{n+i}(\rho_{-n-m-i}) = 0$.

Since $X_t$ is ARMA($n$, $m$), it follows that $\rho_j = \sum_{i=1}^{n} \phi_i \rho_{j-i}$, $j \geq m + 1$ but $\rho_m \neq \sum_{i=1}^{n} \phi_i \rho_{m-i}$. Using these properties, we can show that $H_{n+i+1}(1; \rho_{-n-m-i}) \neq 0$. Thus, for $i > 0$, $S_{n+i}(\rho_{-n-i-m}) = \pm\infty$. Again, similar manipulations reveal that for $i > 0$,

$$H_{n+i+1}(1; \rho_{-n+m-i+1}) = (-1)^{n+i+2}\left(1 - \sum_{j=1}^{n} \phi_j\right) H_{n+i}(\rho_{-n+m-i+1})$$

and that $H_{n+i}(\rho_{-n+m-i+1}) \neq 0$. Thus, it follows that

$$S_{n+i}(\rho_{-n+m-i+1}) = (-1)^i\left[(-1)^n\left(1 - \sum_{j=1}^{n} \phi_j\right)\right] = (-1)^i S_n(\rho_{-n+m+1}).$$

Similarly, one can show $S_{n+i}(\rho_{-n-i-m+1})$ is indeed defined.

To prove that the conditions in (b) are sufficient to imply that $X_t$ is ARMA($n$, $m$), we first note that since $X_t$ is ARMA($p$, $q$) for some $p$ and $q$, the proof of necessity shows that $n \leq p$. It can be shown that the conditions in Theorem 1(b) are sufficient to show that $\rho_k$ satisfies an $n$th order difference equation for $k \geq m + 1$ but not for $k = m$ and thus the result follows. A similar argument applies with $f_m = (-1)^m \rho_m$.

The $S$ values will be placed in an S array as in Table 2 where we have employed the shortened notation $S_n(m) = S_n(f_m)$. It should be noted that the S array given in Table 2 is actually a modification of the format of the S array given by GKM in that element $S_k(j)$ of Table 2 is located in the row corresponding to lag $j + k - 1$ instead of lag $j$ as previously defined. In the remainder of this paper we will refer to this shifted S array as simply the S array.

In Table 3 we present the behavior of the S array when $X_t$ is ARMA($p$, $q$). Again, the reader familiar with the earlier work of GKM should note the modifications in

*Table 2. S Array (Shifted)*

| $m$ | 1 | 2 | ... | $k$ |
|---|---|---|---|---|
| $-i$ | $S_1(-i)$ | $S_2(-i-1)$ | | $S_k(-i-k+1)$ |
| ⋮ | ⋮ | ⋮ | | ⋮ |
| $-1$ | $S_1(-1)$ | $S_2(-2)$ | | $S_k(-k)$ |
| 0 | $S_1(0)$ | $S_2(-1)$ | | $S_k(-k+1)$ |
| 1 | $S_1(1)$ | $S_2(0)$ | | $S_k(-k+2)$ |
| ⋮ | ⋮ | ⋮ | | ⋮ |
| $j$ | $S_1(j)$ | $S_2(j-1)$ | | $S_k(j-k+1)$ |

*Table 3. S Array Where $X_t$ Is ARMA (p,q)*

| $m$ \ $n$ | 1 | ... | $p$ | $p+1$ | ... | $p+i$ | ... |
|---|---|---|---|---|---|---|---|
| ⋮ | | | ⋮ | ⋮ | | ⋮ | |
| $-q-2$ | | | $C_2$ | u | | u | |
| $-q-1$ | | | $C_2$ | $\pm\infty$ | ... | $\pm\infty$ | |
| $-q$ | ⋮ | | NC* | NC | | NC | |
| ⋮ | | | ⋮ | ⋮ | | ⋮ | |
| $-1$ | $S_1(-1)$ | | NC | NC | | NC | |
| 0 | $S_1(0)$ | | NC | NC | | NC | |
| 1 | $S_1(+1)$ | | NC | NC | | NC | |
| ⋮ | | | ⋮ | ⋮ | | ⋮ | |
| $q-1$ | | | NC | NC | | NC | |
| $q$ | ⋮ | | $C_1$ | $-C_1$ | ... | $(-1)^i C_1$ | |
| $q+1$ | | | $C_1$ | u | | u | |
| $q+2$ | | | $C_1$ | u | | u | |
| ⋮ | | | ⋮ | ⋮ | | ⋮ | |

* NC = Nonconstants.

patterns in the S array due to the shifting. We have used this new format because we believe the patterns are simplified. For example, given a stationary ARMA($p$, $q$) process the $2q$ nonconstant terms in column $p$ are the elements from lags $-q$ through $q - 1$, regardless of $p$. That is, the nonconstant behavior will always be centered around lags $-1$ and 0, thus eliminating the need for the "starred quantities" used by GKM to assist in locating the correct position for inspection of a particular column.

The following theorem establishes the relationship between the S array and the generalized partial autocorrelation function.

*Theorem 2.* Let $\rho_k$ be the autocorrelation function of a stationary time series.

(a) If $f_m = \rho_m$, then

$$\phi_{kk}^{(j)} = -S_k(f_{-k+j+1})/S_k(f_{-k-j}).$$

(b) If $f_m = (-1)^m \rho_m$, then

$$\phi_{kk}^{(j)} = (-1)^{k+1} S_k(f_{-k+j+1})/S_k(f_{-k-j}).$$

*Proof.* (a) By definition,

$$\frac{S_k(\rho_{-k+j+1})}{S_k(\rho_{-k-j})} = \frac{H_{k+1}(1; \rho_{-k+j+1}) H_k(\rho_{-k-j})}{H_k(\rho_{-k+j+1}) H_{k+1}(1; \rho_{-k-j})}.$$

For $2 \leq r \leq k + 1$ and $1 \leq c \leq k + 1$, both the $(r, c)$ element of $H_{k+1}(1; \rho_{-k-j})$ and the $(k + 3 - r, k + 2 - c)$ element of $H_{k+1}(1; \rho_{-k+j+1})$ are $\rho_{-k-j+r+c-3}$. Thus by interchanging columns $c$ and $k + 2 - c$, $c = 1, \ldots, [(k + 1)/2]$, where $[\ ]$ denotes the greatest integer function, of the matrix associated with $H_{k+1}(1; \rho_{-k+j+1})$ and then interchanging rows $r$ and $k + 3 - r$, $r = 2, \ldots, 1 + [k/2]$, it can be seen that $H_{k+1}(1; \rho_{-k-j}) = (-1)^k H_{k+1}(1; \rho_{-k+j+1})$. Similar relationships exist among the elements of $H_k(\rho_{-k+j+1})$ and $|B(k, j)|$ and among the elements of $H_k(\rho_{-k-j})$ and $|A(k, j)|$, from which it can be shown that

$$H_k(\rho_{-k+j+1}) = (-1)^{[k/2]} |B(k, j)|$$

and

$$H_k(\rho_{-k-j}) = (-1)^{[(k-1)/2]} \mid A(k, j) \mid.$$

Thus,

$$\frac{-S_k(\rho_{-k+j+1})}{S_k(\rho_{-k-j})} = \frac{(-1)(-1)^k(-1)^{[(k-1)/2]} \mid A(k, j) \mid}{(-1)^{[k/2]} \mid B(k, j) \mid}$$

$$= \frac{\mid A(k, j) \mid}{\mid B(k, j) \mid}$$

$$= \phi_{kk}{}^{(j)}.$$

(b) The proof is similar to that of (a).

As a practical point, it should be noted that although the elements of the GPAC array could be calculated by their defining relation (2.4), their calculation is facilitated using Theorem 2 and the iterative procedure for calculating S array elements.

GKM recommend the inspection of the S arrays both with $f_m = \rho_m$ and with $f_m = (-1)^m\rho_m$ since the model identification behavior in the presence of noise is often clearer in one of the arrays than in the other, yet neither is uniformly better than the other. It does appear that for low-frequency data the S array appears more stable for $f_m = (-1)^m\hat{\rho}_m$ and for high-frequency data S arrays are usually better at $f_m = \hat{\rho}_m$ ($\hat{\rho}_m$ will be defined in Sec. 3).

## 3. USE OF THE GPAC ARRAY IN CONJUNCTION WITH THE S ARRAY FOR MODELING

In this section we will illustrate the use of the GPAC array along with the S array through three examples. In these examples the autocorrelation function will be estimated based on a realization of length $T$ by

$$\hat{\rho}_m = \frac{\sum_{t=1}^{T-m} (X_t - \bar{X})(X_{t+m} - \bar{X})}{\sum_{t=1}^{T} (X_t - \bar{X})^2}.$$

The patterns in the GPAC array and S array, although somewhat disturbed, are still very useful in identifying the order of the model. Example 3 shows that the patterns in the GPAC array should always be checked via the S array.

Finally, before proceeding to the examples, it should be mentioned that if $p = 0$, the Box-Jenkins procedure, the S array procedure, and the GPAC procedure all use primarily the autocorrelation function with its property that $\rho_k = 0$, $k \geq q + 1$ (see GKM).

*Table 4. True Generalized Partial Autocorrelation Function Array for Series (3.1)*

| Moving Average Order | Autoregressive Order 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 0 | .812 | −.650 | .000 | .000 | .000 | .000 | .000 | .000 |
| 1 | .540 | −.650 | u* | u | u | u | u | u |
| 2 | .135 | −.650 | u | u | u | u | u | u |
| 3 | −3.458 | −.650 | u | u | u | u | u | u |
| 4 | 1.528 | −.650 | u | u | u | u | u | u |
| 5 | .915 | −.650 | u | u | u | u | u | u |

* u = undefined.

*Example 1.* Consider the process

$$X_t - 1.34X_{t-1} + .65X_{t-2} = a_t. \tag{3.1}$$

In Table 4 the GPAC array using the true autocorrelation function $\rho_m$ is presented. The identification as an ARMA(2, 0) process is clear, because $\phi_{2+i,2+i}{}^{(0)} = 0$, $i = 1, 2, \ldots$ and $\phi_{22}{}^{(i)} = -.65$, $i = 0, 1, 2, \ldots$. In Table 5 the sample generalized partial autocorrelation function array is given based on a realization from (3.1) of length 100. The patterns that occurred exactly in Table 4 are still visible in Table 5; that is, $\mid \hat{\phi}_{2+i,2+i}{}^{(0)} \mid \leq .127$, $i = 1, \ldots, 8$ and $\hat{\phi}_{22}{}^{(i)} \approx -.65$, $i = 0, \ldots, 5$. Of course, the process is strictly autoregressive and it is clear that one could have identified the process as ARMA(2, 0) strictly on the basis of the first row of Table 5, which is the partial autocorrelation function. However, as will usually be the case, the constant behavior that occurred in the second column in this example is the most visible pattern. In any event, the combination of both is clearly more informative than is either alone. In Example 3 we will deal more completely with the comparison of the model identification capabilities of the constant column behavior and the zero row behavior in the GPAC array.

*Example 2.* Consider the process

$$X_t - 1.5X_{t-1} + 1.21X_{t-2} - .455X_{t-3} = a_t + .2a_{t-1} + .9a_{t-2}. \tag{3.2}$$

Tables 6 and 7 give the GPAC arrays based on theo-

*Table 5. Sample Generalized Partial Autocorrelation Function Array for a Realization of Length 100 From Series (3.1)*

| Moving Average Order | Autoregressive Order 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 0 | .821 | −.650 | .029 | −.114 | −.075 | −.127 | −.034 | .091 |
| 1 | .562 | −.628 | −2.487 | −.133 | .116 | −.107 | −.370 | .074 |
| 2 | .209 | −.702 | .392 | .232 | −.196 | −.170 | −.160 | .012 |
| 3 | −1.894 | −.715 | .873 | 1.143 | −.738 | −.278 | −.167 | −2.244 |
| 4 | 1.901 | −.670 | −.276 | −1.067 | .450 | −.561 | −.343 | .481 |
| 5 | 1.195 | −.720 | 3.137 | −1.292 | −1.692 | −.281 | .480 | −.052 |

*Table 6. True Generalized Partial Autocorrelation Function Array for Series (3.2)*

| Moving Average Order | Autoregressive Order 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 0 | .845 | −.706 | .414 | .299 | −.304 | −.145 | .245 | .062 |
| 1 | .606 | −.458 | .836 | .683 | −.434 | −.646 | .279 | .848 |
| 2 | .391 | −.070 | .455 | .000 | .000 | .000 | .000 | .000 |
| 3 | .328 | 2.073 | .455 | u* | u | u | u | u |
| 4 | 1.356 | −.119 | .455 | u | u | u | u | u |
| 5 | 1.632 | 5.367 | .455 | u | u | u | u | u |

* u = undefined.

retical autocorrelations and sample autocorrelations from a realization of size 300 respectively. In Table 6 the fact that $\phi_{3+i,3+i}^{(2)} = 0$, $i = 1, 2, \ldots$ and that $\phi_{33}^{(2)} = \phi_{33}^{(2+i)} = .455$, $i = 1, 2, \ldots$, indicates that $p = 3$ and $q = 2$. These patterns again are also quite clear in Table 7.

Table 8 presents the theoretical S array and Table 9 presents the S array for the realization of length 300 on which the GPAC array in Table 7 is based. From Table 8, the $\pm\infty$ behavior occurs at lag $-q - 1 = -3$ beginning in column 4. In Table 9 this behavior is manifested in the "large" numbers 268.053, −42.435, and −319.257. At lag $q = 2$ in columns 4, 5, and 6, the quantities 4.494, −4.300, and 4.454, respectively, correspond to the $(-1)^i C_1$ behavior of Theorem 1(b). These observations along with the two sets of near constants in column 3 of Table 9 identify the process as an ARMA(3, 2). Of course the nonconstant behavior is centered at the line drawn between lags 0 and −1.

*Example 3.* Table 10 presents the GPAC array using the true autocorrelation function for the process

$$X_t - .5X_{t-1} + .5X_{t-2} = a_t - a_{t-1}. \qquad (3.3)$$

In that array it should be noted that $\phi_{22}^{(j)} = -.5, j \geq 0$ and $\phi_{kk}^{(1)} = 0, k \geq 2$. Since the process is ARMA(2, 1) we would have expected $\phi_{22}^{(j)} = -.5, j \geq 1$, and $\phi_{22}^{(0)} \neq -.5$. This example, however, points out that the constant behavior in the $p$th column of the GPAC array of a stationary ARMA($p$, $q$) process may begin before row $q$. In other words, if the process actually is ARMA($p$, $q$), with $\phi_p$ the $p$th autoregressive coefficient, and it is treated as an ARMA($p$, $q - 1$), for example, the estimate of the $p$th autoregressive coefficient in this case may also be $\phi_p$. Thus, the user must exercise caution when using the GPAC array to determine $q$. The constant behavior in a column of the GPAC array should be accompanied by the appropriate zero behavior. It is clear that for a stationary ARMA process, $\phi_{kk}^{(q)} = 0$ for $k \geq p + 1$ and $\phi_{pp}^{(q)} \neq 0$ if and only if the process is ARMA($p$, $q$). Thus, in the array in Table 10, although the constant behavior is misleading, the zero behavior correctly identifies $p$ and $q$. As we will shortly comment, the zero behavior can also be misleading. We formally state the results of the preceding discussion in the following theorem.

*Table 8. S Array for Model (3.2) ($f_m = (-1)^m \rho_m$)*

| Lag | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|---|---|---|---|---|---|---|
| −8 | −2.872 | | | | | |
| −7 | −2.037 | 3.552 | | | | |
| −6 | −1.613 | 1.247 | −9.154 | | | |
| −5 | −1.737 | 23.860 | −9.154 | u* | | |
| −4 | −4.052 | 2.915 | −9.154 | u | u | |
| −3 | −3.556 | 22.437 | −9.154 | ±∞ | ±∞ | ±∞ |
| −2 | −2.651 | 5.686 | −7.573 | 1.470 | 6.806 | −7.911 |
| −1 | −2.184 | 4.456 | −10.750 | −10.450 | 7.143 | −65.674 |
| 0 | −1.845 | 3.148 | −4.452 | 3.122 | −2.173 | 2.489 |
| 1 | −1.606 | 2.606 | −6.334 | −1.004 | −3.527 | −1.090 |
| 2 | −1.391 | 1.578 | −4.165 | 4.165 | −4.165 | 4.165 |
| 3 | −1.328 | −6.044 | −4.165 | u | u | |
| 4 | −2.356 | 2.838 | −4.165 | u | | |
| 5 | −2.632 | −6.691 | −4.165 | | | |
| 6 | −1.964 | 3.598 | | | | |
| 7 | −1.534 | | | | | |

* u = undefined.

*Theorem 3.* Let $X_t$ be a stationary ARMA($p$, $q$) process ($p > 0$).

(a) $\phi_{pp}^{(j)} = \phi_p, j > q$.

(b) Suppose that the $n$th order Yule-Walker equations are nonsingular. For some constant $m$, $\phi_{kk}^{(m)} = 0, k \geq n + 1$, and $\phi_{nn}^{(m)} \neq 0$, with all these quantities well defined, iff $n = p$ and $m = q$.

Referring back to Theorem 1, we can see that the constant behavior in the S array is necessary and sufficient. Table 11 presents the second column of the S array with $f_k = \rho_k$ from which the GPAC array in Table 10 was calculated. We see that, as would be expected from Theo-

*Table 7. Sample Generalized Partial Autocorrelation Function Array for a Realization of Length 300 From Series (3.2)*

| Moving Average Order | Autoregressive Order 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 0 | .819 | −.725 | .344 | .315 | −.164 | −.177 | .145 | .105 |
| 1 | .528 | −.540 | .930 | .477 | −.494 | −.307 | .270 | .271 |
| 2 | .165 | −.309 | .458 | −.017 | .101 | .014 | .004 | .037 |
| 3 | −1.124 | −.204 | .452 | 2.850 | .104 | −.017 | −.123 | .045 |
| 4 | .296 | −1.048 | .468 | .024 | .092 | −1.702 | −.316 | .072 |
| 5 | −4.174 | −1.596 | .467 | −1.771 | .145 | .005 | .294 | .393 |

*Table 9. S Array for Realization of Length 300 From Model (3.2) ($f_m = (-1)^m\hat{\rho}_m$)*

| Lag | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|---|---|---|---|---|---|---|
| −8 | −2.312 | | | | | |
| −7 | −1.628 | 3.467 | | | | |
| −6 | −.760 | 2.386 | −9.505 | | | |
| −5 | −4.375 | 1.347 | −9.472 | −177.715 | | |
| −4 | −.110 | 6.459 | −9.579 | 8.022 | −48.460 | |
| −3 | −7.064 | 6.664 | −9.543 | 268.053 | −42.435 | −319.257 |
| −2 | −2.893 | 5.049 | −7.264 | −1.905 | 10.998 | 2.990 |
| −1 | −2.221 | 4.329 | −12.256 | −9.160 | 14.689 | 16.074 |
| 0 | −1.819 | 3.138 | −4.218 | 2.888 | −2.414 | 2.840 |
| 1 | −1.528 | 2.728 | −6.752 | .908 | −5.435 | .919 |
| 2 | −1.165 | 2.059 | −4.372 | 4.494 | −4.300 | 4.454 |
| 3 | .124 | 1.319 | −4.334 | −22.865 | −5.046 | |
| 4 | −1.296 | 1.412 | −4.438 | 4.211 | | |
| 5 | 3.174 | 3.807 | −4.435 | | | |
| 6 | −2.593 | 3.322 | | | | |
| 7 | −1.762 | | | | | |

rem 1, the ambiguity seen in Table 10 is not present in Table 11. The additional constant term in Table 10 was due to the fact that 1.0/2.0 = 1.5/3.0. It is suggested that sequences of constants in the GPAC array due to constant ratios with numerators and denominators that vary can in fact occur in any column of the GPAC array. Thus, the S array should always be checked when considering constant behavior in the GPAC array. This is true even though the zero behavior does characterize $q$; in Section 4 we will demonstrate that the zero behavior can seem to appear prematurely in the GPAC array because of missing coefficients. This problem is somewhat alleviated via the S array, as we will see.

## 4. A COMPREHENSIVE EXAMPLE

In this section we employ the model identification techniques mentioned in this paper, namely the GPAC array and the S array, to model the Makridakis (1978) metal series data. We hope this example will demonstrate the use of these techniques in modeling rather easily this series that has been difficult to handle with previous techniques.

The metal series data consist of 144 monthly values of carbon steel monthly shipments from 1961 to 1972. Makridakis, using the first 108 values of the series, fits the model

$$(1 - B)X_t = (1 - \theta_1 B)(1 - \theta_{12}B^{12})a_t, \qquad (4.1)$$

*Table 10. Theoretical GPAC Array for Series (3.3)*

| Moving Average Order | Autoregressive Order 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | .000 | −.500 | −.333 | −.250 | −.200 | −.167 |
| 1 | −∞* | −.500 | −.000 | .000 | .000 | .000 |
| 2 | .500 | −.500 | u** | u | u | u |
| 3 | −.500 | −.500 | u | u | u | u |
| 4 | 1.500 | −.500 | u | u | u | u |

* It is not necessary for prior elements to be finite for the element to follow to be finite.
** u = undefined.

where $X_t = Y_t - \bar{Y}$ and $Y_t$ are the data. He reports that this model does not forecast as well as the random walk model, an apparent paradox since he believes this to be the optimum model. Parzen (1979) also analyzes the data and suggests the solution to the paradox is that the model of Makridakis is not satisfactory. Parzen identifies the AR(2) process

$$X_t - .44X_{t-1} - .35X_{t-2} = a_t \qquad (4.2)$$

as the best choice according to Parzen's (1974) criterion autoregressive transfer (CAT). The same result is obtained by using Akaike's (1974) FPE criterion. However, Parzen reports that the sample spectrum does not agree well with the spectrum of the AR(2) model (4.2) and that the second choice according to CAT is an AR(13) (the second choice by FPE is AR(3)). Moreover, the spectrum obtained from the fitted AR(13) is compatible with the sample spectrum. For this reason Parzen prefers the AR(13) model, even though it is not selected as best by either CAT or FPE.

A plot of the metals data is given in Figure 1 and the sample autocorrelations are shown in Figure 2. Table 12 shows portions of the S array for the data with $f_m = (-1)^m\hat{\rho}_m$. The array is evaluated with $f_m = (-1)^m\hat{\rho}_m$ because the data are clearly predominantly low frequency. The patterns in the S array suggest an ARMA(1, 1),

*Table 11. Column 2 From Theoretical S Array With $f_k = \rho_k$ for Series (3.3)*

| Lag | Column 2 |
|---|---|
| −5 | 2.0 |
| −4 | 2.0 |
| −3 | 2.0 |
| −2 | 2.0 |
| −1 | 3.0 |
| 0 | 1.5 |
| 1 | 1.0 |
| 2 | 1.0 |
| 3 | 1.0 |
| 4 | 1.0 |

*Figure 1. Makridakis Metals Data*

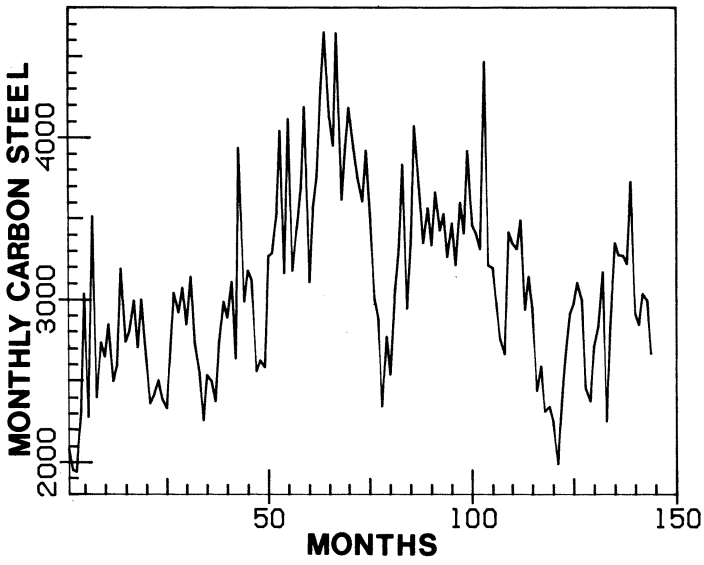
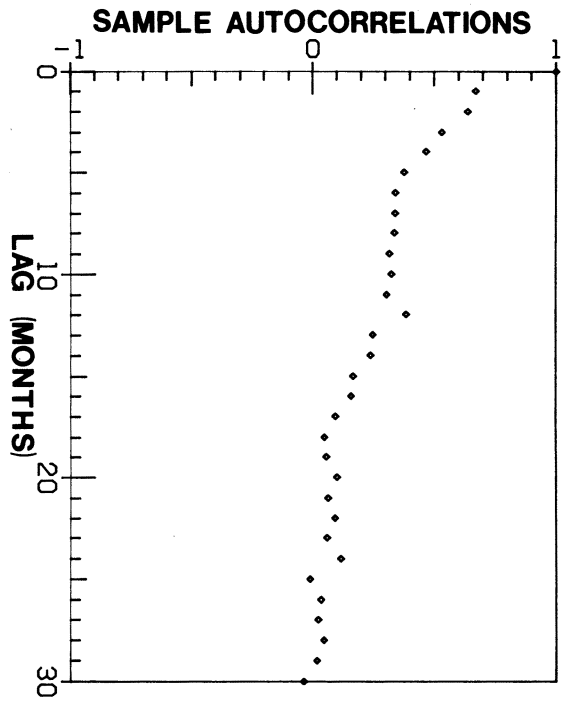



*Figure 2. Sample Autocorrelations From the Makridakis Data*




ARMA(1, 2), ARMA(1, 6) or an ARMA(13, 1), or AR(13). A statistic, referred to as the D statistic by GKM has been suggested for helping to identify possible patterns in the S array. Closer examination of the S array suggests from columns 1 and 13 that the process is ARMA(13, 1) with a strong low frequency component. In such a case GKM recommended that a simple prefiltering may be of value in determining the proper underlying model. As an ad hoc method of determining an initial estimate of such a first order filter (as is indicated here), they suggest averaging appropriate values of the S array. In keeping with the procedure of GKM, the data were transformed by $1 - \hat{\phi}B$, where

$$\hat{\phi} = \frac{S_1(1) + S_1(2) + S_1(3)}{S_1(-2) + S_1(-3) + S_1(-4)} \approx .9.$$

It is, of course, conventional to difference the data in such circumstances as this. However, the data clearly do not indicate a strong tendency toward nonstationarity but do suggest the possibility of a root near one in the characteristic equation. Although in most problems it makes little difference at this stage of the identification process whether one transforms the data by the difference operator $1 - B$ or the operator $1 - .9B$, these authors have found some cases where this seemingly small change in the operators can have a large effect. The S array of $(1 - .9B)X_t$ is shown in Table 13. The suggested model there is clearly an ARMA(12, 1) and the moving average effect is obviously not large. This is consistent with our initial identification of an ARMA(13, 1). Thus, the identified model is

$$\phi(B)X_t = (1 - \theta_1 B)a_t \tag{4.3}$$

where

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_{13} B^{13}.$$

Yule-Walker estimation of the coefficients yields the model

$$(1 - .84B - .14B^2 + .08B^3 + .01B^4 + .10B^5 - .02B^6 - .10B^7 - .01B^8 + .04B^9 + .01B^{10} + .01B^{11} - .33B^{12} + .29B^{13})X_t = (1 - .40B)a_t. \tag{4.4}$$

*Table 12. Portion of S Array for Makridakis Metals Series $f_m = (-1)^m \hat{\rho}_m$)*

| Lag | $S_1$ | $S_2$ | $S_3$ | | $S_{12}$ | $S_{13}$ | $S_{14}$ | $S_{15}$ |
|---|---|---|---|---|---|---|---|---|
| −5 | −2.240 | .091 | −.399 | | −3.079 | 2.143 | −5.921 | 4.662 |
| −4 | −2.134 | .406 | −1.313 | | 1.780 | 2.755 | −.400 | 2.006 |
| −3 | −2.206 | −3.458 | 3.731 | | −4.595 | 3.020 | −2.152 | 1.174 |
| −2 | −2.048 | −4.468 | 18.637 | | .969 | 3.263 | −17.066 | 1.676 |
| −1 | −2.493 | −3.168 | −30.088 | ... | −3.951 | 3.008 | 8.151 | 10.565 |
| 0 | −1.670 | 1.093 | −1.135 | | .815 | −.641 | .696 | −.653 |
| 1 | −1.954 | 1.254 | −1.511 | ... | −.201 | −.953 | 1.156 | −.197 |
| 2 | −1.829 | 1.149 | 5.814 | | 1.236 | −.935 | −1.811 | −.134 |
| 3 | −1.882 | −.424 | .117 | | −.488 | −.731 | −.111 | −1.535 |
| 4 | −1.807 | −.081 | −4.636 | | 1.203 | −.657 | 1.229 | −1.344 |

*Table 13. Portion of S Array for $Z_t = (1 - .9B)X_t$ Where $X_t$ Is the Makridakis Metals Series ($f_m = (-1)^m\hat{\rho}_m$)*

| Lag | $S_1$ | $S_2$ | $S_3$ | | $S_{11}$ | $S_{12}$ | $S_{13}$ | $S_{14}$ | $S_{15}$ |
|---|---|---|---|---|---|---|---|---|---|
| −5 | −.187 | .252 | .657 | | 2.007 | −1.051 | 3.166 | −1.475 | 1.989 |
| −4 | −.264 | 3.037 | 1.534 | | −3.219 | −1.239 | −.300 | −1.746 | −2.014 |
| −3 | 1.298 | −1.789 | .410 | | 4.449 | −1.437 | 1.504 | −1.848 | −6.442 |
| −2 | 2.560 | −7.586 | −.570 | | −1.381 | −1.422 | −43.835 | .258 | 4.163 |
| −1 | 1.188 | 5.875 | 11.809 | ... | 2.056 | −1.181 | −2.534 | −4.810 | 4.054 |
| 0 | −.543 | .598 | −.569 | | −.397 | .297 | −.337 | .315 | −.292 |
| 1 | −.719 | .853 | .092 | ... | .220 | .494 | −.480 | −.054 | −.350 |
| 2 | −.565 | −1.311 | −.075 | | −.973 | .496 | −9.085 | .372 | −.387 |
| 3 | .359 | −.252 | −1.514 | | 1.040 | .401 | −.069 | .498 | −.260 |
| 4 | .230 | 4.512 | −.802 | | −1.444 | .356 | −.584 | .485 | −3.975 |

This suggests a parsimonious model of the form

$$(1 - \phi B)(1 - \psi B^{12})X_t = (1 - \theta B)a_t. \qquad (4.5)$$

Thus far we have not considered the generalized partial autocorrelation in this example. In this example, the GPAC presents the model identification nicely and at the same time shows why the zero behavior in the rows should not be used alone for identification, even for AR($p$) models only. Table 14 shows the GPAC for the metals data and Table 15 for $(1 - .9B)X_t$. Row 1 of Table 14 clearly demonstrates why CAT and FPE chose an AR(2) model. That is, since $\phi_{3,3}^{(0)}$ through $\phi_{11,11}^{(0)}$ are all approximately zero it is not surprising that CAT and FPE pick an AR(2). On the other hand, note that column 2 in the GPAC is not even approximately constant and hence the long string of zeros in row 1 is only suggestive of a number of zero coefficients prior to $\phi_p$. That is, the row and column behavior together suggest the process is *not* AR(2) but does have a number of zero or near zero coefficients following $\phi_2$. Note this is also clearly seen in the S array where column 2 shows no constancy behavior. In that array the zero behavior, seen in row 1 of the GPAC array, is obtained from the ratio of the elements on each side of the center line. Thus the S array and GPAC use both the row and column behavior for identifying a process. The information gained by this procedure is clearly demonstrated in this example where one can easily see that the constant column behavior is influenced by all the coefficients, even if a number of zeros lie between the first and the last autoregressive coefficients. However, the zero behavior of rows gives no information about the values of $\phi_k$ for larger values of $k$, that is, those farther out than the zeros. The ARMA(13, 1) identified by the S array is also vividly displayed by the GPAC array. Note that the ARMA(13, 1) is the only choice that shows both the proper row and column behavior. Nevertheless, the large values of .95, .83, .88, and so on, in column 1 suggest, for reasons already mentioned, that it is a good idea to transform the data to more stationary behavior (as we did) before making the final identification.

In concluding this example we make two comments. The D statistic measures jointly both the constant column behavior of the S array and the zero row behavior cur-

*Table 14. GPAC Array for Makridakis Data*

| Moving Average Order | Autoregressive Order | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 5 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 0 | .67 | .35 | .04 | .01 | −.05 | .02 | .12 | .09 | .01 | .04 | .00 | .21 | −.21 | −.09 | −.06 |
| 1 | .95 | .28 | −.08 | .15 | −.04 | .25 | .10 | .07 | −.29 | .04 | −10.27 | .21 | −.29 | .07 | −.12 |
| 2 | .83 | .33 | 1.56 | .38 | .09 | .10 | .07 | .15 | −.04 | −.35 | −.02 | .27 | −.31 | −.84 | .11 |
| 3 | .88 | 1.05 | −.09 | .94 | −1.31 | .07 | .00 | .13 | −1.32 | −.34 | −6.17 | .27 | −.27 | −.28 | −.77 |
| 4 | .81 | .88 | 11.62 | 1.12 | −.56 | .10 | −1.81 | .14 | −.11 | .61 | −.15 | .39 | −.31 | .21 | −.29 |

*Table 15. GPAC Array for $Y_t = (1 = .9B)X_t$ Where $X_t$ Is the Makridakis Metals Data*

| Moving Average Order | Autoregressive Order | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 0 | −.46 | −.10 | −.05 | −.05 | −.04 | −.10 | −.10 | −.02 | −.03 | .03 | −.19 | .25 | .13 | .07 | −.07 |
| 1 | −.28 | .11 | −.16 | .01 | −.17 | −.06 | −.08 | .10 | −.05 | −.16 | −.16 | .35 | .01 | .21 | −.08 |
| 2 | −.44 | −.73 | −.18 | −1.47 | −.15 | .06 | −.10 | −.12 | .25 | .18 | −.22 | .35 | −6.04 | .20 | .06 |
| 3 | −1.36 | .08 | −.99 | .63 | −.19 | −.48 | −.09 | −.55 | .80 | .83 | −.32 | .32 | .23 | .28 | .13 |
| 4 | −1.23 | −17.94 | −1.22 | .20 | .21 | −.15 | .28 | −.13 | −.48 | −.31 | −.72 | .34 | −.19 | .33 | −2.00 |

rently measured by such statistics as CAT and FPE. It is probably overly sensitive to pairwise constant behavior and the very first "zero" in the zero behavior. Some modifications of this are now under way and it is hoped that it may eventually evolve not just as a guide to inspection of the S array but as a dependable estimator of $p$ and $q$ for the ARMA($p$, $q$) process. Finally, as a word of caution, the GPAC is a useful measure; however, Example 3, which demonstrates that its column behavior is not sufficient to identify $q$, is not pathological and these authors have encountered a number of real data sets in which the behavior demonstrated in Example 3 was observed. When these observations are coupled with the present example demonstrating that zero behavior alone can also be misleading, it is clear that the S array should be consulted, being the only measure that uniquely characterized the ARMA($p$, $q$) by both its column and row properties.

## 5. CONCLUDING REMARKS

In this paper we have demonstrated the use of the generalized partial autocorrelation for identifying an ARMA($p$, $q$) model. In the process we have shown that it is a natural extension of the Box-Jenkins method. In addition we have demonstrated the relationship between the S array approach of GKM and the generalized partial autocorrelation. That is, we have shown that the generalized partial autocorrelation, much like the D statistic, represents a condensation of the information in the S array. We recommend the use of the GPAC array along with the S array and the D statistic. Actually, we recommend the use of a slight modification of the S array, which presents the information in the S array in a more easily interpretable form.

[*Received August 1979. Revised December 1980.*]

## REFERENCES

AKAIKE, H. (1970), "Statistical Predictor Identification," *Annals of the Institute of Statistical Mathematics*, 22, 203–217.

BOX, G.E.P., and JENKINS, G.M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden Day.

GRANGER, C.W.J., and NEWBOLD, P. (1977), *Forecasting Economic Time Series*, New York: Academic Press.

GRAY, H.L., KELLEY, G.D., and McINTIRE, D.D. (1978), "A New Approach to ARMA Modeling," *Communications in Statistics*, B7, 1–77.

MAKRIDAKIS, S. (1978), "Time Series Analysis and Forecasting: An Update and Evaluation," *International Statistical Review*, 46, 255–278.

McINTIRE, D.D. (1977), "A New Approach to ARMA Modeling," Ph.D. dissertation, Southern Methodist University.

PARZEN, E. (1974), "Some Recent Advances in Time Series Modeling," *IEEE Transactions on Automatic Control*, AC-19, 723–730.

——— (1977), "Multiple Time Series: Determining the Order of Approximating Autoregressive Schemes," in *Multivariate Analysis–IV*, ed. P. Krishnaiah, Amsterdam: North Holland, 283–295.

——— (1979), "Time Series Modeling, Spectral Analysis, and Forecasting," ONR Technical Report No. N-7, Texas A&M Research Foundation, Project No. 3838.