An efficient linear method for ARMA spectral estimation

RANDOLPH L. MOSES†, VIRGINIJA ŠIMONYTĖ‡,
PETRE STOICA§ and TORSTEN SÖDERSTRÖM‖

A three-step method for obtaining asymptotically maximum likelihood ARMA spectral estimates using only linear transformations of the data is presented. The computational efficiency of the algorithm is comparable to that of Yule–Walker algorithms, but the three-step method gives asymptotically statistically efficient estimates. The implementation of the algorithm is discussed in detail, and numerical examples are presented to illustrate its performance.

1. Introduction

Spectral estimation is a topic that continually receives a great deal of attention (Kay 1988, Ljung 1987, Marple 1987, Söderström and Stoica 1989). Of the many techniques available, parametric techniques and, in particular, the use of an autoregressive moving average (ARMA) model have become very popular (Kay 1988, Marple 1987).

Two main types of ARMA spectral estimation methods have been developed. One is the optimization-based type, including maximum likelihood (ML) methods, prediction error methods, and various nonlinear least squares methods. These procedures can sometimes be computationally intensive, and suffer from problems associated with convergence to local minima (Ljung 1987, Söderström and Stoica 1989). The other main type of estimator is the class of Yule–Walker based methods. These techniques are generally much less computationally burdensome, but can produce estimates with poor accuracy unless special steps are taken (Cadzow 1982, Chan and Langford 1982, Kay 1988, Marple 1987).

An ARMA spectral estimation procedure that combines the simplicity of Yule–Walker based methods with the accuracy of ML methods was proposed by Stoica *et al.* (1987). In this algorithm, initial covariance estimates are obtained, and an initial estimate of the ARMA spectral parameters is computed. These initial estimates are then used to correct the covariance estimates to improve their accuracy. This correction requires only linear operations, and the corrected estimates give asymptotically efficient estimates of their corresponding spectral parameters.

Received 22 August 1992. Revised 20 January 1993.

† Department of Electrical Engineering, Ohio State University, Columbus, OH 43210, U.S.A.

‡ Automatic Control and Systems Analysis Group, Department of Technology, Uppsala University, Box 27, S-751 03 Uppsala, Sweden. On leave from Institute of Mathematics and Informatics, Akademijos 4, 2600 Vilnius, Lithuania.

§ Automatic Control and Systems Analysis Group, Department of Technology Uppsala University, Box 27, S-751 03 Uppsala, Sweden. On leave from Department of Automatic Control, Polytechnic Institute of Bucharest, Splaiul Independentei 313, R-77206 Bucharest, Romania.

‖ Automatic Control and Systems Analysis Group, Department of Technology, Uppsala University Box, 27, S-751 03 Uppsala, Sweden.

0020-7179/94 $10.00

A theoretical development of the large-sample maximum likelihood spectral estimation algorithm appears in Stoica *et al.* (1987). Here we consider some practical aspects of the algorithm in Stoica *et al.* (1987). We discuss difficulties which are encountered when some theoretical (large sample) assumptions do not hold, and suggest modifications to the method. We develop a more reliable version of the algorithm in Stoica *et al.* (1987) and consider computationally efficient implementations. Finally, we present numerical examples to illustrate the performance of the algorithm.

2. The ARMA spectral model

Consider the following ARMA process

$$A(q^{-1})y(t) = C(q^{-1})e(t) \tag{2.1}$$

where q^{-1} is the unit delay operator (i.e. $q^{-1}y(t) = y(t-1)$), and

$$e(t) = \text{white noise with zero mean and variance } \lambda^2$$
$$A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_{na} q^{-na}$$
$$C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_{nc} q^{-nc}$$

The following standard assumptions are made:

(A.1) $A(z)C(z) = 0 \Rightarrow |z| > 1$

(A.2) $a_{na} c_{nc} \neq 0$

(A.3) $\{A(z), C(z)\}$ are coprime polynomials

In other words, the ARMA representation (2.1) is minimal, stable and invertible. While this is not a restrictive assumption, we note that there are cases of interest for which Assumption A1 does not hold; for example, the sinusoids-in-white-noise process can be described by an ARMA model of the form (2.1) where all the zeros of $A(z)$ and $C(z)$ lie on the unit circle. The estimation method described in this paper does not readily extend to such 'degenerate' ARMA processes; methods, developed for the sinusoids-in-noise case, can be found in, for example, Cadzow (1982), Chan and Langford (1982), Stoica *et al.* (1989 a), Stoica *et al.* (1989 b). For simplicity we also assume that the orders *na* and *nc* are given. Methods for ARMA order estimation are described in for example Kay (1988), Ljung (1987), Marple (1987), Söderström and Stoica (1989), Stoica *et al.* (1986).

Next, we introduce the following notation

$$r_k = E\{y(t)y(t-k)\} = \text{the covariance of } y(t) \text{ at lag } k \tag{2.2}$$

$$\phi(z) = \sum_{k=-\infty}^{\infty} r_k z^{-k} = \text{the spectral density of } y(t) \tag{2.3}$$

In (2.2), $E\{\cdot\}$ denotes the expectation operator, and z in (2.3) is a complex variable. It is well known (Kay 1988, Marple 1987) that

$$\phi(z) = \lambda^2 \frac{C(z)C(z^{-1})}{A(z)A(z^{-1})} \tag{2.4}$$

The problem of ARMA spectral estimation consists of first parametrizing the spectral density function, then estimating those parameters. Clearly from (2.4), $\phi(z)$ can be parametrized by $\{\lambda^2, a_1, \ldots, a_{na}, c_1, \ldots, c_{nc}\}$. However, the statistically efficient estimation of these parameters is not an easy task, and nonlinear optimization routines are generally employed. (Asymptotically efficient estimates of the $\{a_i\}$ coefficients can be obtained by using only linear operations, however, (Stoica *et al*. 1985.)

In this paper we parametrize the spectral density by

$$\theta = [r_0, \ldots, r_{na+nc}]^{\mathrm{T}} \tag{2.5}$$

This parametrization is well-defined. In order to see this, note first that $\{r_k\}$ satisfies the Yule–Walker equations

$$r_k + \sum_{i=1}^{na} a_i r_{k-i} = 0, \quad k \geqslant nc + 1 \tag{2.6}$$

The a_i coefficients are uniquely determined from θ by solving the first na linear equations of (2.6)

$$\begin{bmatrix} r_{nc} & r_{nc-1} & \cdots & r_{nc-na+1} \\ r_{nc+1} & r_{nc} & \cdots & r_{nc-na+2} \\ \vdots & \vdots & & \vdots \\ r_{nc+na-1} & r_{nc+na-2} & \cdots & r_{nc} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{na} \end{bmatrix} = - \begin{bmatrix} r_{nc+1} \\ r_{nc+2} \\ \vdots \\ r_{nc+na} \end{bmatrix} \tag{2.7}$$

or equivalently (using obvious notation)

$$Ra = -r$$

Note that under Assumptions (A.1)–(A.3), R is non-singular (Stoica 1981, Stoica *et al*. 1985). Moreover, defining

$$\begin{aligned} b_k &\triangleq E\{A(q^{-1})y(t) \cdot A(q^{-1})y(t-k)\} \\ &= \sum_{i=0}^{na}\sum_{j=0}^{na} a_i a_j r_{k+j-i}, \quad k = 0, \ldots, nc \end{aligned} \tag{2.8}$$

where $a_0 = 1$, and

$$B(z) \triangleq \sum_{k=-nc}^{nc} b_{|k|} z^{-k} = \lambda^2 C(z) C(z^{-1}) \tag{2.9}$$

it readily follows from (2.1), (2.4) and (2.9) that

$$\phi(z) = \frac{B(z)}{A(z)A(z^{-1})} \tag{2.10}$$

Since the numerator in (2.10) is uniquely defined from θ, this concludes the proof that $\phi(z)$ can be uniquely parametrized by θ.

3. An asymptotically efficient algorithm

In this section, we consider the specific problem of estimating the spectral density of the ARMA process (2.1). This problem reduces to estimating the parameter vector θ. Presented by Stoica *et al*. (1987) is an algorithm for computing an estimate of θ which asymptotically has the same statistical

properties as the ML estimate. For brevity, we present only the required results and refer the reader to Stoica *et al.* (1987) for details.

From a given set of samples $\{y(1), \ldots, y(N)\}$, θ can be estimated by using the standard unbiased sample covariances

$$\tilde{r}_k = \begin{cases} \dfrac{1}{N-k} \displaystyle\sum_{t=1}^{N-k} y(t)y(t+k) & k \geq 0 \\ \tilde{r}_{-k} & k < 0 \end{cases} \tag{3.1}$$

However, the estimates in (3.1) are, in general, not statistically efficient, and can have very bad accuracy (Porat 1987). A more accurate estimator can be obtained as follows.

Consider a random variable $\tilde{\theta}$ with unknown mean θ. Assume we are given a zero mean random vector z which is correlated with $\tilde{\theta}$

$$\operatorname{ascov}\left\{\begin{pmatrix} \tilde{\theta}-\theta \\ z \end{pmatrix}\right\} = \begin{pmatrix} W_{11} & W_{12} \\ W_{12}^{\mathrm{T}} & W_{22} \end{pmatrix} \tag{3.2}$$

where 'ascov' stands for 'asymptotic covariance'

$$\operatorname{ascov}(a) \triangleq \lim_{N\to\infty} N \operatorname{cov}(a)$$

In (3.2), it is assumed that $W_{12} \neq 0$, and W_{22} is positive definite. Then an improved estimate of θ is given by

$$\hat{\theta} = \tilde{\theta} - \hat{W}_{12}\hat{W}_{22}^{-1}z \tag{3.3}$$

where $\hat{W}_{12}$ and $\hat{W}_{22}$ are any consistent estimates of W_{12} and W_{22}, respectively (Stoica *et al.* 1987). Note that the normalized asymptotic covariance of $\tilde{\theta}$ is W_{11}. The normalized asymptotic covariance of $\hat{\theta}$ is given by

$$\operatorname{ascov}(\hat{\theta}) = W_{11} - W_{12}W_{22}^{-1}W_{12}^{\mathrm{T}} \tag{3.4}$$

If z is chosen appropriately, then the asymptotic covariance of $\hat{\theta}$ may be much smaller than that of $\tilde{\theta}$.

In the present application, the vectors $\tilde{\theta}$ and z are given by

$$\tilde{\theta} = [\tilde{r}_0, \tilde{r}_1, \ldots, \tilde{r}_{na+nc}]^{\mathrm{T}} \tag{3.5}$$

$$z = [z_1, z_2, \ldots, z_{nz}]^{\mathrm{T}} \tag{3.6}$$

$$z_k = \sum_{i=0}^{na}\sum_{j=0}^{na} \tilde{a}_i\tilde{a}_j\tilde{r}_{na+nc+k-i-j} \quad k = 1, \ldots, nz \tag{3.7}$$

In the above equations, the $\{\tilde{a}_i\}$ are consistent estimates of $\{a_i\}$ found, for example, by solving a set of Yule–Walker equations. With these definitions, the covariance matrices W_{12} and W_{22} are given by (see Stoica *et al.* 1987)

$$[W_{12}]_{ij} = \alpha_{j+i} + \alpha_{j-i}, \quad i = 0, \ldots, na + nc; \quad j = 1, \ldots, nz \tag{3.8}$$

$$\alpha_s = \text{the coefficient of } z^s \text{ in } z^{-(na+nc)}\frac{B^2(z)}{A^2(z^{-1})} \tag{3.9}$$

$$[W_{22}]_{ij} = \beta_{i-j}, \; i, j = 1, \ldots, nz \tag{3.10}$$

$$\beta_s = \text{the coefficient of } z^s \text{ in } B^2(z) \tag{3.11}$$

It is not difficult to see that

$$[W_{12}]_{ij} = 0, \; j > nc - na + i \tag{3.12}$$

This implies that $[W_{12}]_{ij} = 0$ for $j > 2nc$. Also, W_{22} is a banded, symmetric Toeplitz matrix with the band width $2nc + 1$.

Let $\{\tilde{a}_k\}$ and $\{\tilde{b}_k\}$ be consistent estimates of $\{a_k\}$ and $\{b_k\}$, and define $\hat{W}_{12}$ and $\hat{W}_{22}$ as in (3.8)–(3.12) but using the $\{\tilde{a}_k\}$ and $\{\tilde{b}_k\}$ estimates there. Then, the estimate $\hat{\theta}$ given by (3.3) along with (3.5)–(3.11) has the following properties (proved by Stoica *et al.* 1987)

(P.1) $\hat{\theta}$ is a minimum variance estimate in the class of all estimators based on the sample covariances $\{\tilde{r}_0, \ldots, \tilde{r}_{na+nc+nz}\}$. It is a large-sample approximation of the ML estimate of θ which uses these $na + nc + nz + 1$ sample covariances as a data statistic.

(P.2) Let P_{nz} denote the asymptotic (for $N \to \infty$) covariance matrix of $\hat{\theta}$ (we explicitly show the dependence of this matrix on nz). Then, $P_{nz} \geqslant P_{nz+1}$. In other words, $\{P_{nz}\}$ for $nz = 1, 2, \ldots$ forms a sequence of monotonically non-increasing positive definite matrices.

(P.3) Let P_{CR} denote the Cramér-Rao lower bound (CRLB) for the covariance matrix of any consistent estimator of θ under the gaussian hypothesis. Then

$$\lim_{nz \to \infty} P_{nz} = P_{\mathrm{CR}} \tag{3.13}$$

(P.4) The rate of convergence in (3.13) critically depends on the location of the zeros of $C(z)$. The closer these zeros are to the unit circle, the slower the convergence rate of P_{nz} to P_{CR}. The location of the zeros of $A(z)$ has only a marginal influence on the convergence rate of P_{nz}.

Explicit expressions for P_{nz} and P_{CR} may be found in Stoica *et al.* (1987).

The algorithm for determining an asymptotically efficient estimate of $\phi(z)$ can be summarized as follows.

Step 1. Compute the sample covariances $\{\tilde{r}_k\}$ in (3.1) for $k = 0, \ldots, K$, where $K = na + nc + nz$. Compute the initial AR coefficient estimate $\tilde{a}$ by solving the Overdetermined Yule–Walker (OYW) equations

$$\begin{bmatrix} \tilde{r}_{nc} & \cdots & \tilde{r}_{nc-na+1} \\ \vdots & & \vdots \\ \tilde{r}_{K-1} & \cdots & \tilde{r}_{K-na} \end{bmatrix} \tilde{a} = -\begin{bmatrix} \tilde{r}_{nc+1} \\ \vdots \\ \tilde{r}_K \end{bmatrix} \tag{3.14}$$

where $\tilde{a} = [\tilde{a}_1, \ldots, \tilde{a}_{na}]^{\mathrm{T}}$. Use $\{\tilde{a}\}$ and $\{\tilde{r}_k\}$ in (2.8) to obtain initial estimates $\tilde{b}_k$ for $k = 0, \ldots, nc$.

Step 2. Use $\tilde{r}_k$, $\tilde{b}_k$ and $\tilde{a}$ in (3.7)–(3.11) to compute z, $\hat{W}_{12}$, and $\hat{W}_{22}$. Compute improved estimates $\hat{r}_k$ for $k = 0, \ldots, na + nc$ by using (3.3).

Step 3. Compute $\hat{a}$ from $\{\hat{r}_k\}_0^{na+nc}$ using (2.7), and $\hat{b}_k$ for $k = 0, \ldots, nc$ from $\hat{a}$ and $\hat{r}_k$ using (2.8). Compute $\hat{\phi}(z)$ using $\hat{a}$ and $\hat{b}_k$ in (2.10).

In Step 1, we have used an overdetermined Yule–Walker estimate for $\tilde{a}$, where the number of equations is chosen so that sample autocovariances

$\{\tilde{r}_k\}_{k=0}^{na+nc+nz}$ are used. This particular estimate is chosen for the following reasons. First, we already need to compute $\{\tilde{r}_k\}_{k=0}^{na+nc+nz}$ to implement Step 2. If we use fewer YW equations in (3.14), then we expect the initial AR coefficients to be less accurate. This would result in poor initialization of the three-step algorithm and a correspondingly worse final accuracy of these estimates. Of course, it is not always true that an OYW produces more accurate estimates than the YW (with $K = na + nc$). However, the three-step algorithm we propose is especially useful in cases where the OYW method gives poor accuracy of the AR coefficient estimates; these cases correspond to time series whose models have zeros close to the unit circle. For time series with zeros close to the unit circle, it has been found (Bruzzone and Kaveh 1980, Cadzow 1982, Chan and Langford 1982, Porat and Friedlander 1986) that the OYW method gives (greatly) improved performance with respect to the YW method; thus, the use of the OYW method in Step 1 is justified.

Steps 2 and 3 of the above procedure may be repeated to obtain improved spectral density estimates. Specifically, the reiteration process is as follows.

(1) Use $\{\tilde{r}_k\}$ and the latest available estimate of $\{a_k\}$ to compute z in (3.7).
(2) Compute $\hat{W}_{12}$ and $\hat{W}_{22}$ in (3.8)–(3.11) using the latest available estimates of $\{a_k\}$ and $\{b_k\}$.
(3) Determine a new estimate of $\hat{\theta}$ using $\hat{\theta}$, z, $\hat{W}_{12}$ and $\hat{W}_{22}$ computed as above, in (3.3).

In the large-sample case, repeating Steps 2 and 3 does not improve the accuracy of $\hat{\theta}$; however, it may improve the accuracy for small data lengths.

4. Implementation of the algorithm

This section discusses the numerical and computational aspects of the estimation procedure. The integer nz is a user's variable in the algorithm. Some guidelines for choosing nz will be given in § 5.

4.1. *Conditioning of* $\hat{W}_{22}$

In this subsection we analyse a condition which was tacitly assumed to hold. It was assumed that the inverse of $\hat{W}_{22}$ (and of W_{22}) exists. To address this issue, we state the following result (see also Grenander and Szegö 1958, Söderström and Stoica 1989).

Lemma 1: *Consider the* $nz \times nz$ *matrix* $\hat{W}_{22}$ *given by (3.10), (3.11). Let* $\{\lambda_j, j = 1, \ldots, nz\}$ *denote the eigenvalues of* $\hat{W}_{22}$ *and let*

$$\lambda_{\min}^{(nz)} = \min_j \{\lambda_j\}, \quad \lambda_{\max}^{(nz)} = \max_j \{\lambda_j\} \tag{4.1}$$

Then

$$\lambda_{\min}^{(nz)} \geqslant \lambda_{\min}^{(nz+1)}, \quad \lambda_{\max}^{(nz)} \leqslant \lambda_{\max}^{(nz+1)} \tag{4.2}$$

and

$$\sigma_{\min} \triangleq \lim_{nz \to \infty} \lambda_{\min}^{(nz)} = \inf_{\omega} |\tilde{B}(e^{i\omega})|^2 \tag{4.3}$$

$$\sigma_{\max} \triangleq \lim_{nz\to\infty} \lambda_{\max}^{(nz)} = \sup_{\omega} |\tilde{B}(e^{i\omega})|^2 \tag{4.4}$$

Note that the above lemma holds for W_{22} if $\tilde{B}(e^{i\omega})$ is replaced by $B(e^{i\omega})$.

When $\tilde{B}(z)$ has no zeros on the unit circle, we conclude from Lemma 1 that $\sigma_{\min} > 0$. Thus, $\hat{W}_{22}^{-1}$ exists for any value of nz (finite or infinite). However, if $\tilde{B}(z)$ has zeros close to the unit circle, some numerical problems may be expected. Indeed, in such a case $\sigma_{\min}$ will be small and $\hat{W}_{22}$ will be ill-conditioned for large nz. These comments apply not only to the estimates $\tilde{B}(z)$ and $\hat{W}_{22}$, but to $B(z)$ and W_{22} as well.

Note that if $\tilde{B}(z)$ cannot be factored into $\tilde{C}(z)\tilde{C}(z^{-1})$ (i.e. it is not the spectral density of an MA process, owing to estimation errors), then it has zeros of odd multiplicity on the unit circle. In that case $\sigma_{\min} = 0$ and $\hat{W}_{22}$ may be very ill-conditioned for 'large' values of nz. A procedure for circumventing the difficulty induced by an ill-conditioned $\hat{W}_{22}$ is discussed in § 4.2 below.

4.2. *Efficient computation of* $\hat{W}_{22}^{-1}z$

This subsection details the steps needed to compute $\hat{W}_{22}^{-1}z$. First, the sample covariances are estimated using (3.1), and from these estimates the initial AR coefficients $\{\tilde{a}_i\}_{i=1}^{na}$ are found by solving the OYW equations (3.14). We note that the algorithm of Moses (1984) and Zohar (1979) for solving an overdetermined Toeplitz system of equations can be used, and results in fewer computations than a direct solution of (3.14) when na is large.

Once $\{\tilde{a}_i\}_{i=1}^{na}$ are found, the vector z in (3.6) must be computed. Equation (3.7) can be used for this computation, but a more computationally efficient method can be obtained by noting that $z_k = \tilde{A}^2(q^{-1})\tilde{r}_{k+na+nc}$. Thus, z_k is given by

$$z_k = \sum_{i=0}^{2na} g_i \tilde{r}_{k+na+nc-i}, \quad k = 1, \ldots, nz \tag{4.5}$$

where g_i is the coefficient of the q^{-i} term of $G(q^{-1}) \triangleq \tilde{A}^2(q^{-1})$

$$g_i = \begin{cases} \displaystyle\sum_{m=0}^{i} \tilde{a}_m \tilde{a}_{i-m} & 0 \leq i \leq na \\ \displaystyle\sum_{m=i-na}^{na} \tilde{a}_m \tilde{a}_{i-m} & na+1 \leq i \leq 2na \end{cases} \tag{4.6}$$

Computation of $\{g_i\}_{i=0}^{2na}$ using (4.6) requires about $2na^2$ flops (one flop being one multiplication and one addition), and computation of each z_k term in (4.5) requires $2na$ flops. Thus, a total of about $2(na \cdot \mathrm{nz} + na^2)$ flops are required to compute the vector z. For $nz > 2$, this compares favourably with the $na^2 \cdot nz$ flops needed to complete z directly from (3.7).

Next, the elements of $\hat{W}_{22}$ must be computed from the $\tilde{a}$ and $\tilde{r}$ sequences. This involves first computing the coefficients of $\tilde{B}(z)$ as defined in (2.8), then finding the coefficients of the polynomial $\tilde{B}^2(z)$ (cf. (3.11)). A more efficient procedure than direct computation of $\tilde{b}_k$ by using (2.8) is found by noting that

$$\tilde{b}_k = \sum_{i=0}^{na}\sum_{j=0}^{na} \tilde{a}_i \tilde{a}_j \tilde{r}_{k+j-i} = \sum_{l=-na}^{na} d_l \tilde{r}_{k+l} \tag{4.7}$$

where

$$d_l = \sum_{i=0}^{na-|l|} \tilde{a}_i \tilde{a}_{i+|l|}, \quad -na \leq l \leq na \tag{4.8}$$

Since $d_l = d_{-l}$, equation (4.7) can be written as

$$\tilde{b}_k = \tilde{r}_k d_0 + (\tilde{r}_{k+1} + \tilde{r}_{k-1}) d_1 + \cdots + (\tilde{r}_{k+na} + \tilde{r}_{k-na}) d_{na} \tag{4.9}$$

The computation of $\{\tilde{b}_k\}_{k=0}^{nc}$ using equations (4.8)–(4.9) requires about $na^2/2 + na \cdot nc$ flops, compared with $nc \cdot na^2$ flops when (2.8) is directly implemented.

Finally, in Step 2 of the algorithm, it is necessary to compute $\hat{W}_{22}^{-1} z$. We first note from (3.10)–(3.11) that

$$[\hat{W}_{22}]_{ij} = \begin{cases} \beta_{|i-j|}, & |i-j| \leq 2nc \\ 0, & \text{otherwise,} \end{cases} \tag{4.10}$$

$$\beta_k = \sum_{m=k-nc}^{nc} \tilde{b}_{|m|} \tilde{b}_{|k-m|}, \quad 0 \leq k \leq 2nc \tag{4.11}$$

As mentioned earlier, $\hat{W}_{22}$ is a banded symmetric Toeplitz matrix, a fact which enables the efficient computation of $u \triangleq \hat{W}_{22}^{-1} z$. One such method is described in Gavel (1992). This algorithm is similar to the Levinson–Durbin algorithm, but skips over multiplications by zero that are present as a result of the banded structure.

It should be noted that since the algorithm in Gavel (1992) computes $\hat{W}_{22}^{-1} z$ recursively in nz, it is possible from (3.3) to compute $\hat{\theta}$ recursively in nz. Thus, given an upper bound on nz, this proposed algorithm computes the updated estimate $\hat{\theta}(nz)$ for all values of nz up to this upper bound. This is potentially useful in determining nz adaptively if an appropriate value is not known *a priori*.

As explained in § 4.1, the matrix $\hat{W}_{22}$ may be ill-conditioned if the polynomial $\tilde{B}(z)$ has zeros very close to the unit circle. Such a situation appears when $\tilde{B}(z)$ does not represent a valid MA spectral density (due to estimation errors), in which case this polynomial has zeros exactly on the unit circle. In order to avoid possible ill-conditioning of $\hat{W}_{22}$, we correct $\tilde{B}(z)$ so that the zeros of the corrected polynomial do not lie in the ring delimited by the circles of radius, say, 0·9 and 1/0·9 (the choice of these values is arbitrary and other values may work better in a given application). This operation presents no complication when no zero of $\tilde{B}(z)$ is situated on the unit circle. If $\tilde{B}(z)$ has zeros on the unit circle, first these zeros are paired and then each pair is moved away from the unit circle at moduli 0·9 and 1/0·9, respectively, and the same angular position equal to the average phase of paired zeros. Pairing the (odd-multiplicity) unit-circle zeros of $\tilde{B}(z)$ is straightforward when their number is not too large, which should be the case in most applications. If, however, in some situation this condition is not met, then pairing the unit-modulus zeros of $\tilde{B}(z)$ may be difficult. In such a case, other more elaborate methods should be used for correcting $\tilde{B}(z)$ so as to guarantee that it has no zero in a ring including the unit circle. See, for example, Stoica and Moses (1992). We note that the method in Stoica and Moses (1992) corrects $\tilde{B}(z)$ without finding its zeros; it may thus be more computationally efficient than the more direct

zero-correcting method discussed above. None the less, for small nc, the direct method is efficient and effective for most cases of interest.

4.3. *Stabilizing $\tilde{A}(z^{-1})$ and computation of $\hat{W}_{12}$*

In order to compute $\hat{W}_{12}$, it is necessary to find the first few impulse response elements of a filter whose transfer function has $\tilde{A}^2(z^{-1})$ in the denominator (cf. Step 2). It may therefore be necessary to ensure that $\tilde{A}(z^{-1})$ is a stable polynomial. Since the $\tilde{a}_k$ coefficients are consistent estimates of the a_k coefficients, $\tilde{A}(z^{-1})$ will be stable for sufficiently large N. However, for 'small' data lengths, $\tilde{A}(z^{-1})$ may be unstable. Computationally efficient procedures of stabilizing $\tilde{A}(z^{-1})$, which requires only $O(na^2)$ computations, can be found in Stoica and Moses (1992). Alternatively, one can determine the zeros of $\tilde{A}(z^{-1})$, say $\{\mu_k e^{\pm i\phi_k}\}_{k=1}^{na}$, modify them to $\{\min(\rho, \mu_k)e^{\pm i\phi_k}\}_{k=1}^{na}$ (for some $\rho < 1$, for example $\rho = 0{\cdot}95$) and then construct a stabilized $\tilde{A}(z^{-1})$ from the corrected zeros.

Once the stabilized $\tilde{A}(z^{-1})$ has been obtained, it is useful to compute $\hat{W}_{12}$ given by (3.8) and (3.9). Equation (3.9) is implemented as follows. First compute the stable polynomial $G(z^{-1}) = \tilde{A}^2(z^{-1})$ as in the preceding section (of course, if $\tilde{A}(z^{-1})$ was already stable, $G(z^{-1})$ need not be recomputed). Define

$$\Gamma(z^{-1}) = z^{-2nc}\frac{B^2(z)}{G(z^{-1})} \triangleq \sum_{i=0}^{\infty}\gamma_i z^{-i} \tag{4.12}$$

The coefficients of $\Gamma(z^{-1})$ can be recursively computed by

$$\gamma_k = \beta_{2nc-k} - \sum_{i=1}^{\min(k,2na)} \gamma_{k-i}g_i, \quad k \geq 0 \tag{4.13}$$

By comparing (4.12) and (3.9), it can be seen that $\alpha_k = \gamma_{nc-na-k}$; thus, from (3.8) we have

$$W_{12} = \begin{bmatrix} \gamma_{nc-na-1} & \gamma_{nc-na-2} & \cdots & \gamma_{nc-na-nz} \\ \gamma_{nc-na} & \gamma_{nc-na-1} & \cdots & \gamma_{nc-na-nz+1} \\ \vdots & \vdots & & \vdots \\ \gamma_{2nc-1} & \gamma_{2nc-2} & \cdots & \gamma_{2nc-nz} \end{bmatrix} + \begin{bmatrix} \gamma_{nc-na-1} & \cdots & \gamma_{nc-na-nz} \\ \gamma_{nc-na-2} & \cdots & \vdots \\ \vdots & & \gamma_0 \\ \gamma_0 & \cdots & 0 \end{bmatrix} \tag{4.14}$$

It is clear from (4.14) that γ_k must be computed for $k = 0, 1, \ldots, 2nc - 1$. Note also that since $\gamma_k = 0$ for $k < 0$, the first matrix on the right-hand side of (4.14) may have zeros in the upper-right triangle, whereas the second matrix has zeros in the lower-right triangle. Note that for $na \geq nc$, this second matrix is zero.

4.4. *Efficient computation of $\hat{a}$ and $\hat{b}$*

Note that the matrix R (or $\hat{R}$ if covariance estimates are used) in (2.7) is square and Toeplitz. This structure enables computationally efficient Levinson-type methods to be used to compute the $\hat{a}$ vector. Such algorithms can be found

in Kay (1988), Marple (1987), Söderström and Stoica (1989) and Zohar (1979). The number of computations required is $O(na^2)$.

Once $\hat{a}$ is found, the $\hat{b}$ coefficients can be efficiently computed using (4.8)-(4.9), with $\hat{a}$ and $\hat{r}$ replacing $\tilde{a}$ and $\tilde{r}$ there.

4.5. *Ensuring the non-negativity of the spectral estimate*

It is clear from (2.10) that the estimated spectrum $\hat{\phi}(e^{i\omega})$ is non-negative if and only if $\hat{B}(e^{i\omega})$ is non-negative for all frequencies. However, $\hat{B}(e^{i\omega})$ provided by the proposed algorithm is not guaranteed to be so. If a non-negative spectral estimate is needed, then $\hat{B}(e^{i\omega})$ must be modified.

One simple way to modify $\hat{B}(z)$ to make it non-negative was described in § 4.2. As already stated, that procedure works well if the set of unit-circle zeros of $\hat{B}(z)$ is not too dense. Another procedure that can be used to ensure a non-negative MA spectral density is the window method of Moses (1984), Moses and Beex (1986). This procedure, however, may introduce an unduly large bias in the corresponding modified spectral density estimate. Other more elaborate procedures for enforcing the non-negativity condition on an estimated MA spectral density can be found in Stoica and Moses (1992) and Moses and Liu (1991). In the numerical applications considered in this paper (see § 5), we found it unneccessary to recourse to these more complicated procedures.

The results of the previous discussions are summarized in Table 1 as a detailed description of the proposed algorithm.

Detailed implementation of the efficient linear spectral ARMA algorithm

Given: $y(1), \ldots, y(N)$; na, nc, and nz.

Step 1(a). Generate $\{\tilde{r}_k\}_{k=0}^{nz+na+nc}$ using (3.1).

Step 1(b). Compute $\tilde{a}$ from (3.14) using a non-symmetric Toeplitz equation solution procedure.

Step 1(c). Compute $\{\tilde{b}_k\}_{k=0}^{nc}$ using (4.7)–(4.9).

Step 2(a). Compute the coefficients $\{g_k\}_{k=0}^{2na}$ of $G(z) = \tilde{A}^2(z)$ using (4.6).

Step 2(b). Compute $z_k = \sum_{i=0}^{2na} g_i \tilde{r}_{k+na+nc-i}$, $k = 1, \ldots, nz$.

Step 2(c). Compute $\{\beta_k\} = \sum_{s=k-nc}^{nc} \tilde{b}_{|s|}\tilde{b}_{|k-s|}$, $k = 0, 1, \ldots, 2nc$.

Step 2(d). Set $[\hat{W}_{22}]_{ij} = \begin{cases} \beta_{|i-j|}, & |i-j| \leq 2nc \\ 0, & \text{otherwise.} \end{cases}$

Solve for $u = \hat{W}_{22}^{-1} z$ using a fast Toeplitz algorithm. Optionally, if $\hat{W}_{22}$ is ill-conditioned, correct $\{\tilde{b}_k\}$ as described in § 4.2.

Step 2(e). Optionally stabilize $G(z)$ as discussed in § 4.3, if necessary. Compute

$$\gamma_k = \beta_{2nc-k} - \sum_{i=1}^{\min(k,2na)} \gamma_{k-i} g_i \quad \text{for } k = 0, \ldots, 2nc-1$$

($\gamma_k = 0$ for $k < 0$). Form

$$\hat{W}_{12} = \begin{bmatrix} \gamma_{nc-na-1} & \gamma_{nc-na-2} & \cdots & \gamma_{nc-na-nz} \\ \gamma_{nc-na} & \gamma_{nc-na-1} & \cdots & \gamma_{nc-na-nz+1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{2nc-1} & \gamma_{2nc-2} & \cdots & \gamma_{2nc-nz} \end{bmatrix}$$

$$+ \begin{bmatrix} \gamma_{nc-na-1} & \cdots & \gamma_{nc-na-nz} \\ \gamma_{nc-na-2} & \cdots & \gamma_{nc-na-nz-1} \\ \vdots & \ddots & \vdots \\ \gamma_{-2na-1} & \cdots & \gamma_{-2na-nz} \end{bmatrix}$$

Step 2(f). Compute $\hat{r} = \tilde{r} - \hat{W}_{12}u$.

Step 3(a). Compute $\hat{a}$ from (2.7) with $\{\hat{r}_k\}$ replacing $\{r_k\}$, by using a fast Levinson solver; and $\{\hat{b}_k\}$ from (4.10)–(4.12) with $\{\hat{a}_k\}$ and $\{\hat{r}_k\}$ replacing $\{\tilde{a}_k\}$ and $\{\tilde{r}_k\}$ there.

Step 3(b). Compute $\hat{\phi}(z)$ using $\hat{a}$ and $\hat{b}_k$ in (2.10). Optionally, if $\hat{\phi}(z)$ is not non-negative definite, correct $\{\hat{b}_k\}$ as described in § 4.5.

5. Numerical examples

In this section, we present some numerical examples which illustrate the performance of the three-step spectral ARMA algorithm. The ARMA processes in these examples are similar to the ones considered in Friedlander and Porat (1984). We also compare the performance of the proposed algorithm with the Cramér-Rao bound (CRB). The CRB is computed using the expressions in Friedlander and Porat (1984). In each example, we present the estimates of the spectral density for short data realizations ($N = 200$ or $N = 300$) and/or for long data realizations ($N = 1500$ or $N = 2000$). In each case considered we plot $mc = 10$ realizations of the estimated spectral density in a superimposed fashion. The plots of the estimates of the spectral density are shown in pairs. The first plot of a pair corresponds to the initial estimates $\{\tilde{a}_k\}$ and $\{\tilde{b}_k\}$ obtained from the first step of the proposed algorithm. We call the estimates of the spectral density shown in this plot the 'tilde estimates'. The second plot of a pair corresponds to the final estimates $\{\hat{a}_k\}$ and $\{\hat{b}_k\}$. These estimates of the spectral density are referred to as the 'hat estimates'. For each example, we present several plots of the estimates of the spectral density, corresponding to different values of nz. In our examples, nz is taken to vary in the range $\{2, 3, \ldots, 10\}$.

Concerning the choice of the user's variable nz in applications, we have the following suggestion based on empirical experience with the ARMA spectral method under consideration. In our experiments, increasing nz always leads to a spectral density estimate which was non-negative for all values of the frequency. Moreover, the first value of nz, say $\hat{nz}$, which produced a non-negative spectral density estimate also gave a satisfactory accuracy that could not be significantly improved by further increasing nz. We recommend choosing $nz = \hat{nz}$ in any application in which one does not dispose of *a priori* information on the data under study which might permit a more judicious choice of this parameter.

Example 1: Consider the ARMA(4, 4) model with

$$A(q^{-1}) = 1 + 0{\cdot}1q^{-1} + 1{\cdot}66q^{-2} + 0{\cdot}093q^{-3} + 0{\cdot}8649q^{-4} \tag{5.1}$$

$$C(q^{-1}) = 1 + 0{\cdot}0226q^{-1} + 0{\cdot}8175q^{-2} + 0{\cdot}0595q^{-3} + 0{\cdot}0764q^{-4} \tag{5.2}$$

The poles of the process are at $0{\cdot}9644\exp(\pm j0{\cdot}4335\pi)$ and $0{\cdot}9644\exp(\pm j0{\cdot}5835\pi)$, the zeros are at $0{\cdot}8471\exp(\pm j0{\cdot}4867\pi)$ and $0{\cdot}3263\exp(\pm j0{\cdot}5457\pi)$. Figures 1 and 2 show the plots of the true spectral

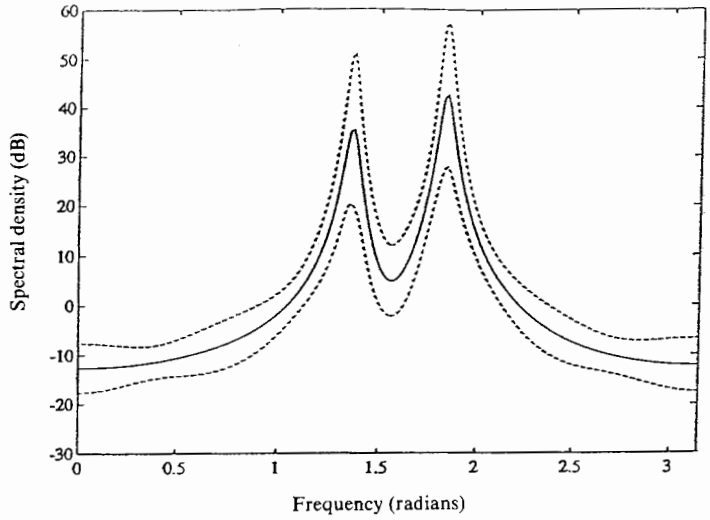

Figure 1. Spectral density $\pm 2\sigma$ bounds for Example 1. The data length is $N = 200$. Solid curve = spectral density; dashed curves = bounds.

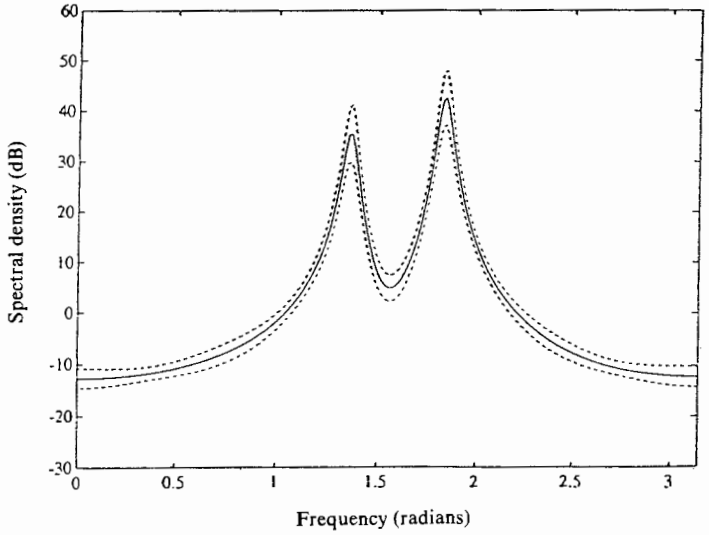

Figure 2. Spectral density $\pm 2\sigma$ bounds for Example 1. The data length is $N = 1500$. Solid curve = spectral density; dashed curves = bounds.

density and the spectral density $\pm 2\sigma$, where σ is the CRB, for the data lengths of $N = 200$ and $N = 1500$. In both cases, we have computed the tilde and hat estimates for $nz = 2, \ldots, 10$. Typical results so obtained are shown in Figs 3 and 4.

From Figs 3 and 4, one can see that negative spectral density estimates always appear among the tilde estimates. Increasing nz does not necessarily improve the performance of the tilde estimator.

The hat estimator attains higher accuracy of the estimates. For $N = 200$, increasing nz not only improves the accuracy of the hat estimates, but also eliminates the occurrence of negative spectral estimates there. In the case of $N = 1500$, increasing nz does not affect the accuracy of the hat estimates significantly. As Fig. 4 shows, the hat estimates have a tendency to converge as nz increases.

Note that the plots in Figs 3 and 4 were obtained without any correction. For this example, we have also tried to correct the polynomials $\tilde{A}(z)$ and $\tilde{B}(z)$ as described in § 4. The corrections removed the negative estimates of the spectral density but at the expense of an inflated bias. Interestingly enough, the bias was nearly constant for all frequencies, and the variance has not been affected in a visible way. In the interest of brevity, we omit the plots showing corrected spectral density estimates. □

Example 2: Consider the ARMA(4,3) process with

$$A(q^{-1}) = 1 - 1{\cdot}3136q^{-1} + 1{\cdot}4401q^{-2} - 1{\cdot}0919q^{-3} + 0{\cdot}83527q^{-4} \quad (5.3)$$

$$C(q^{-1}) = 0{\cdot}13137 + 0{\cdot}023543q^{-1} + 0{\cdot}10775q^{-2} + 0{\cdot}03516q^{-3} \quad (5.4)$$

The poles of the process are at $0{\cdot}9245\exp(\pm j0{\cdot}5433\pi)$ and $0{\cdot}9886\exp(\pm j0{\cdot}2095\pi)$; the zeros are at $-0{\cdot}3110$ and $0{\cdot}9283\exp(\pm j0{\cdot}4773\pi)$. The plots of the true spectral density and the spectral density $\pm 2\sigma$ for a data length $N = 2000$ are presented in Fig. 5. Figure 7 shows how the tilde and hat estimates of the spectral density change when increasing nz. From Fig. 7 one can see the improvement of estimation accuracy offered by the hat estimator over the tilde estimator. Increasing nz results in better estimates both in the case of the tilde and the hat estimators. As Fig. 7 shows, the peak is estimated rather accurately. The more complicated problem is to estimate accurately the 'valley' of the spectrum. □

Example 3: Consider the ARMA(4,4) process with

$$A(q^{-1}) = 1 - 2{\cdot}7607q^{-1} + 3{\cdot}8106q^{-2} - 2{\cdot}6535q^{-3} + 0{\cdot}9238q^{-4} \quad (5.5)$$

$$C(q^{-1}) = 1 - 2{\cdot}1398q^{-1} + 2{\cdot}3672q^{-2} - 1{\cdot}3729q^{-3} + 0{\cdot}3930q^{-4} \quad (5.6)$$

The poles of the process are at $0{\cdot}9805\exp(\pm j0{\cdot}2801\pi)$ and $0{\cdot}9803\exp(\pm j0{\cdot}2199\pi)$; the zeros are at $0{\cdot}8328\exp(\pm j0{\cdot}3238\pi)$ and $0{\cdot}7528\exp(\pm j0{\cdot}1828\pi)$. Figure 6 shows the spectral density $\pm 2\sigma$ of the ARMA process above. The data length is $N = 2000$. The tilde and hat spectral density estimates are presented in Fig. 8. Figure 8 shows quite poor initial tilde estimates which are significantly improved by using the three-step algorithm. The large number of data considered was necessary to obtain initial estimates with reasonable accuracy. For smaller values of N, the initial estimates are very erratic and their use to initialize the three-step algorithm is not appropriate. □

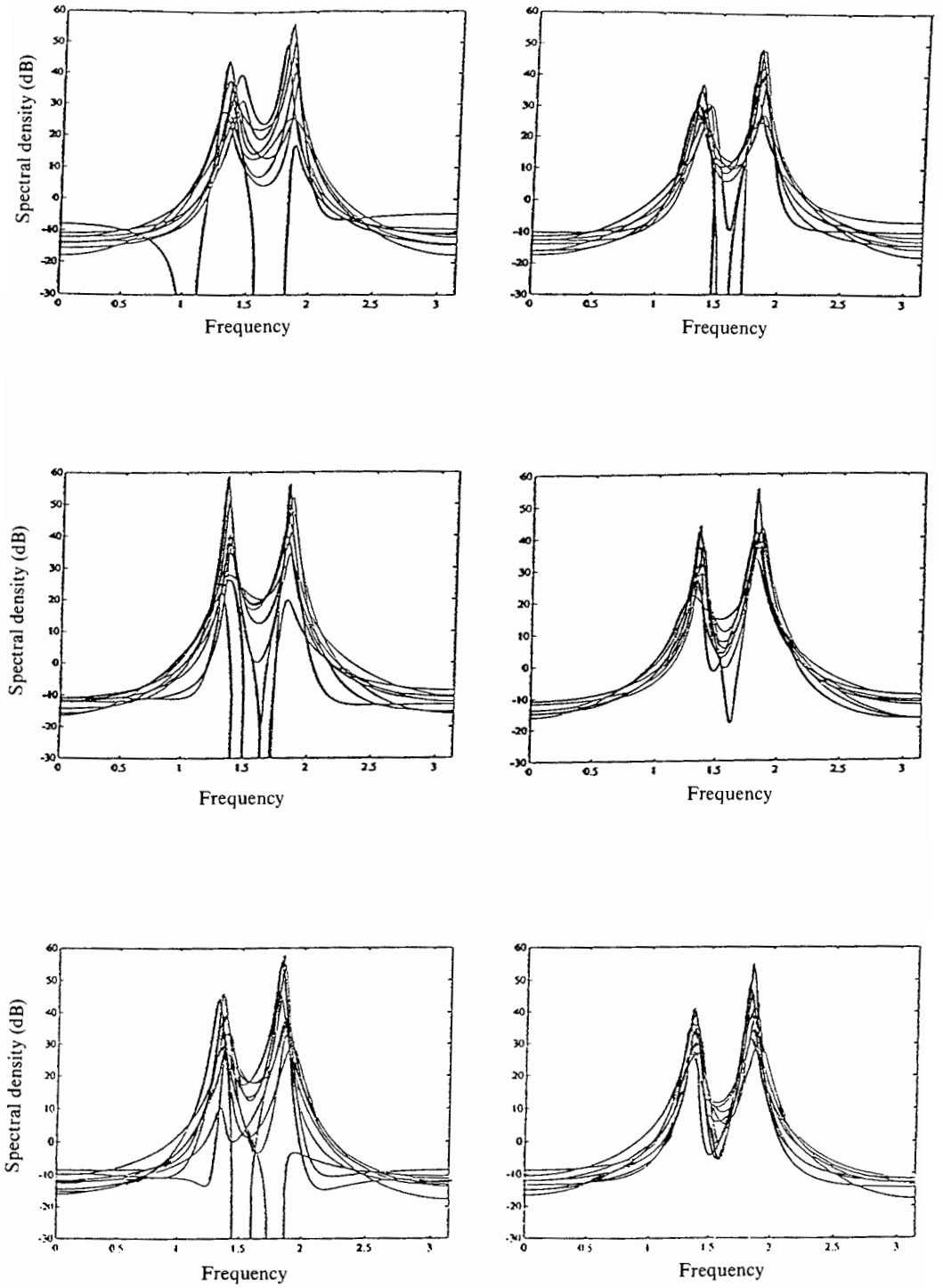

Figure 3. Estimates of the spectral density for Example 1. Left-hand side: the tilde estimates, right-hand side: the hat estimates. The data length is $N = 200$. Upper plots: $nz = 2$; middle plots: $nz = 6$; lower plots: $nz = 10$.

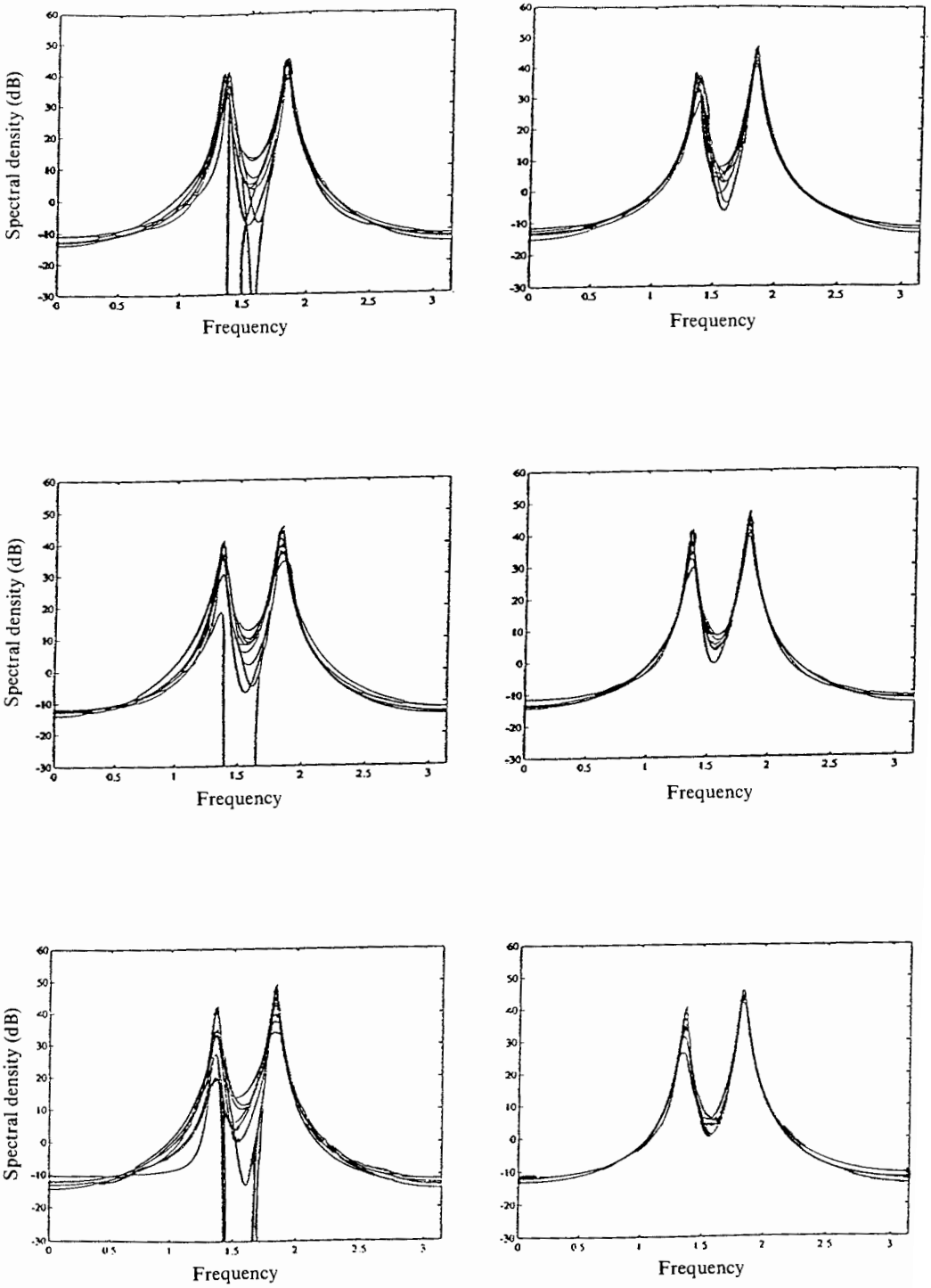

Figure 4. Estimates of the spectral density for Example 1. Left-hand side: the tilde estimates, right-hand side: the hat estimates. The data length is $N = 1500$. Upper plots: $nz = 2$; middle plots: $nz = 6$; lower plots: $nz = 8$.

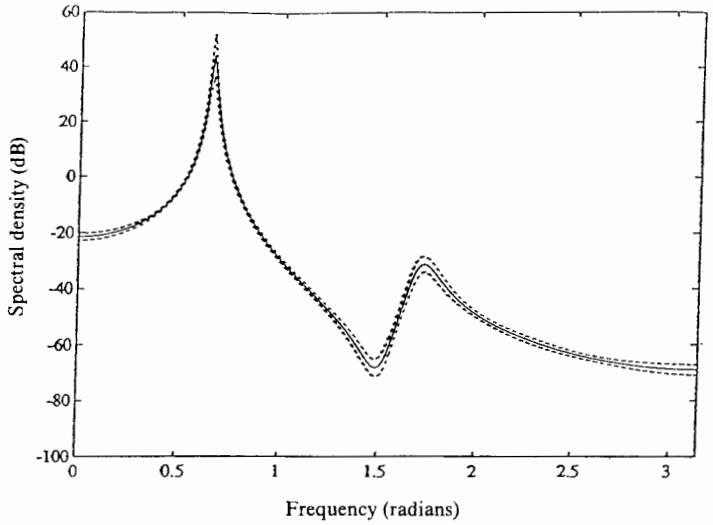

Figure 5. Spectral density $\pm 2\sigma$ bounds for Example 2. The data length is $N = 2000$. Solid curve = spectral density, dashed curves = bounds.

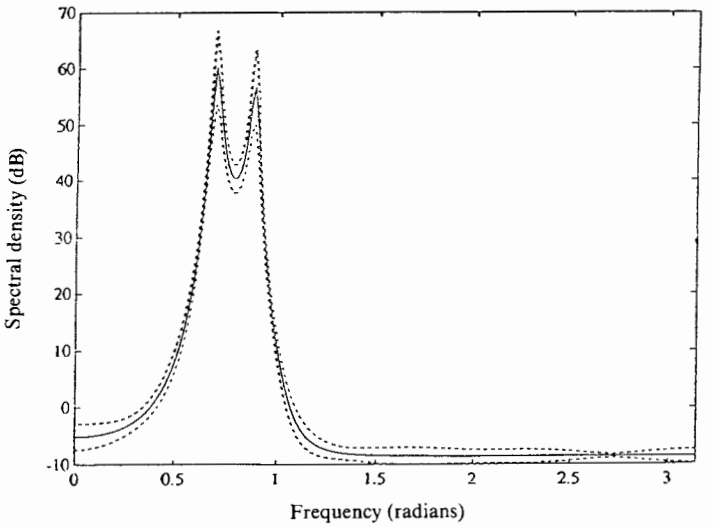

Figure 6. The spectral density $\pm 2\sigma$ bounds for Example 3. The data length is $N = 2000$. Solid curve = spectral density, dashed curves = bounds.

6. Conclusions

We have presented a three-step algorithm for obtaining asymptotically statistically efficient estimates of ARMA spectral densities. The algorithm includes an initial estimation step which is a Yule–Walker based estimation method. Then, the parameter estimates are updated to obtain estimates with asymptotically minimum variance. The algorithm is computationally efficient. In addition, the algorithm readily incorporates steps which give stable denominator

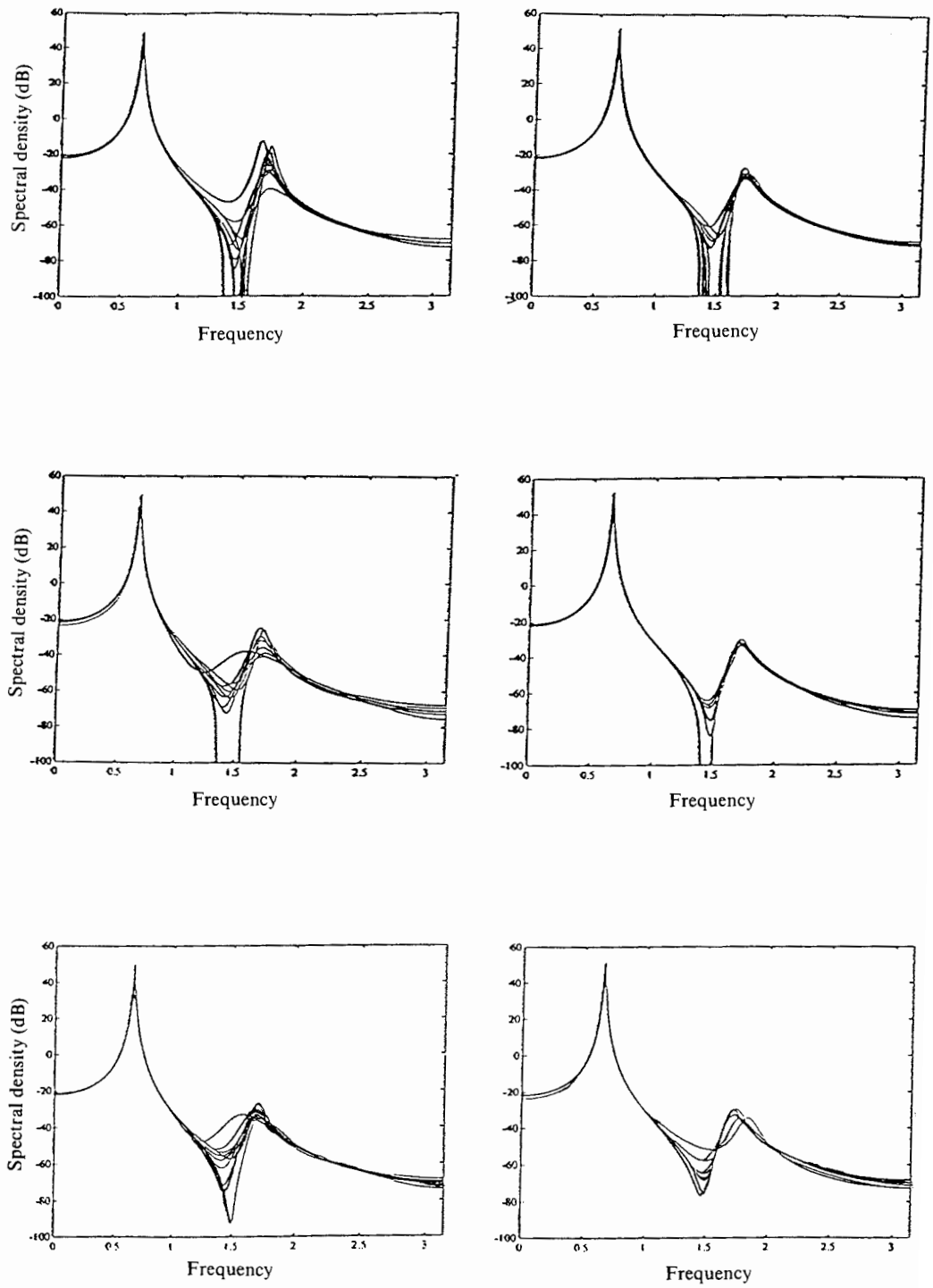

Figure 7. Estimates of the spectral density for Example 2. Left-hand side: the tilde estimates, right-hand side: the hat estimates. The data length is $N = 2000$. Upper plots: $nz = 2$; middle plots: $nz = 4$; lower plots: $nz = 8$.

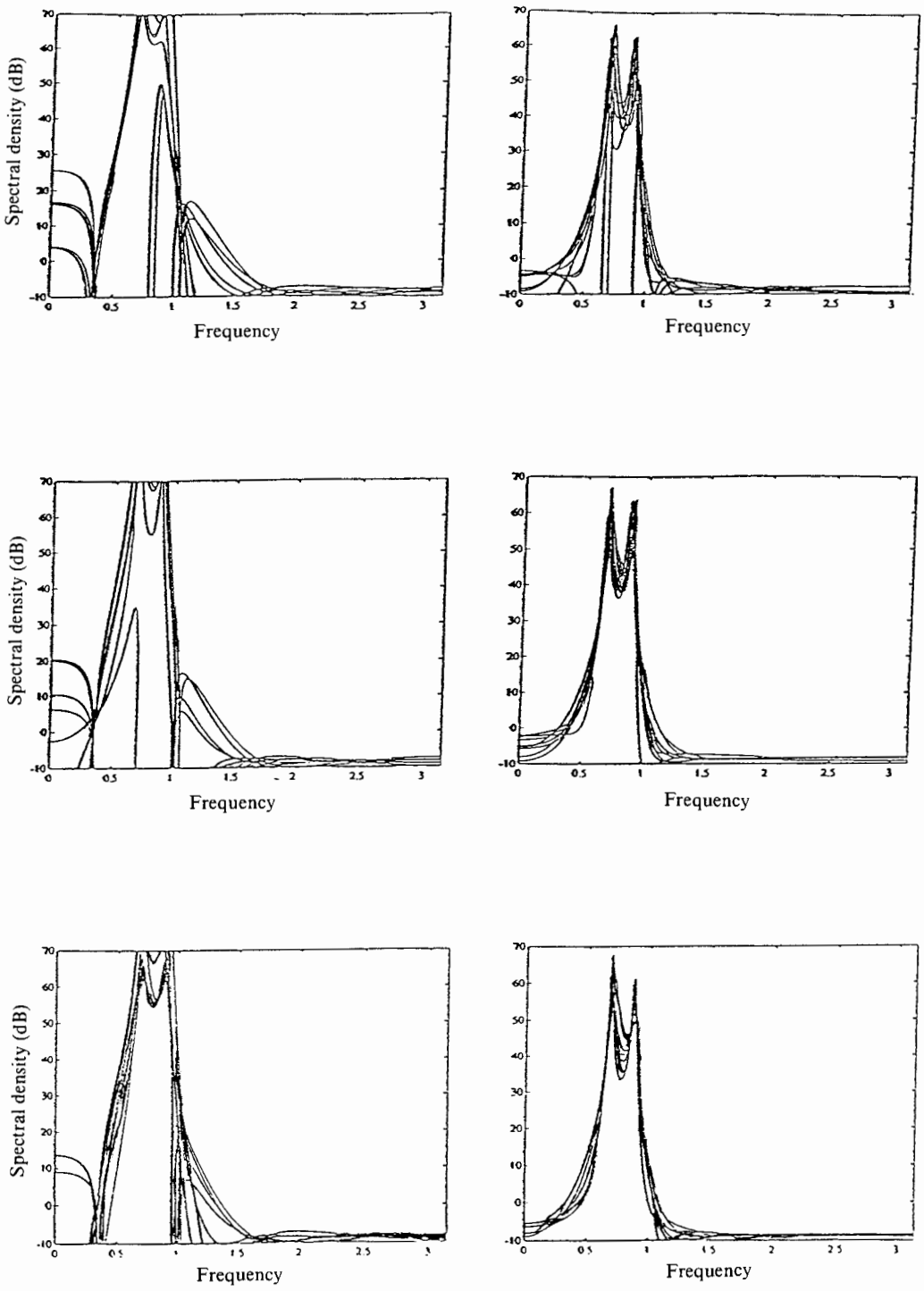

Figure 8. Estimates of the spectral density for Example 3. Left-hand side: the tilde estimates, right-hand side: the hat estimates. The data length is $N = 2000$. Upper plots: $nz = 4$; middle plots: $nz = 6$; lower plots: $nz = 8$.

polynomial estimates and non-negative spectral estimates. It is shown that the algorithm provides improved accuracy of ARMA spectral parameter estimates. In addition, the algorithm reduced the occurrence of negative spectral estimates even when no explicit step is used to ensure non-negativity.

ACKNOWLEDGMENTS

The work of V. Šimonytė was supported by the Swedish Institute. The work of P. Stoica was supported in part by the Swedish Research Council for Engineering Sciences under Contract 91-676.

REFERENCES

BRUZZONE, S., and KAVEH, M., 1980, On some suboptimum ARMA spectral estimators. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **28**, 753–755.

CADZOW, J. A., 1982, Spectrum estimation: an overdetermined rational model equation approach. *Proceedings of the Institute of Electrical and Electronics Engineers*, **70**, 907–939.

CHAN, Y. T., and LANGFORD, R. P., 1982, Spectral estimation via the high-order Yule–Walker equations. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **30**, 689–698.

FRIEDLANDER, B., and PORAT, B., 1984, A general lower bound for parametric spectrum estimation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **32**, 728–732.

GAVEL, D. T., 1992, Solution to the problem of instability in banded Toeplitz solvers. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **40**, 464–466.

GRENANDER, U., and SZEGÖ, G., 1958, *Toeplitz Forms and their Applications* (Berkeley, California: University of California Press).

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

LJUNG, L., 1987, *System Identification. Theory for the User* (Englewood Cliffs, NJ: Prentice Hall).

MARPLE, L., 1987, *Digital Spectral Analysis with Applications* (Englewood Cliffs, NJ: Prentice Hall).

MOSES, R., 1984, Design and analysis of fast recursive ARMA spectral estimators. Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

MOSES, R. L., and BEEX, A. A., 1986, A comparison of numerator estimators for ARMA spectra. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **34**, 1668–1671.

MOSES, R. L., and LIU, D., 1991, Optimal nonnegative definite approximations of estimated moving average covariance sequences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **39**, 2007–2015.

PORAT, B., 1987, Some asymptotic properties of the sample covariances of Gaussian auto-regressive moving average processes. *Journal of Time Series Analysis*, **8**, 205–220.

PORAT, B., and FRIEDLANDER, B., 1986, Bounds on the accuracy of Gaussian ARMA parameter estimation methods based on sample covariances. *IEEE Transactions on Automatic Control*, **31**, 579–582.

SÖDERSTRÖM, T., and STOICA, P., 1989, *System Identification* (Hemel Hempstead, U.K.: Prentice Hall International).

STOICA, P., 1981, On a procedure for testing the order of time-series. *IEEE Transactions on Automatic Control*, **26**, 572–573.

STOICA, P., EYKHOFF, P., JANSSEN, P., SÖDERSTRÖM, T., 1986, Model-structure selection by cross-validation. *International Journal of Control*, **43**, 1841–1878.

STOICA, P., FRIEDLANDER, B., and SÖDERSTRÖM, T., 1987, Approximate maximum-likelihood approach to ARMA spectral estimation. *International Journal of Control*, **45**, 1281–1310.

STOICA, P., and MOSES, R. L., 1992, On the unit circle problem: the Schur–Cohn procedure revisited. *Signal Processing*, **26**, 95–118.

STOICA, P., SÖDERSTRÖM, T., and FRIEDLANDER, B., 1985, Optimal instrumental variable estimates of the AR parameters of an ARMA process. *IEEE Transactions on Automatic Control*, **30**, 106–114.

STOICA, P., SÖDERSTRÖM, T., and TI, F.-N., 1989 a, Asymptotic properties of the high-order Yule-Walker estimates of sinusoidal frequencies. *IEEE Transactions on Acoustics, Speech and Signal Processing,* **37**, 1721–1734; 1989 b, Overdetermined Yule–Walker estimation of the frequencies of multiple sinusoids: Accuracy aspects. *Signal Processing*, **16**, 155–174.

ZOHAR, S., 1979, Fortran subroutines for the solution of Toeplitz sets of linear equations. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **27**, 656–658.