MA estimation in polynomial time
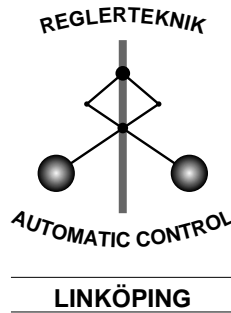
Petre Stoica[1], Tomas McKelvey, Jorge Mari[2]

Department of Electrical Engineering

Linköping University, S-581 83 Linköping, Sweden

WWW: `http://www.control.isy.liu.se`

Email: `tomas@isy.liu.se`

March 9, 1999

[1]Systems and Control Group, Uppsala University, Uppsala, Sweden
[2]Optimization and Systems Theory, Royal Institute of Technology, Stockholm, Sweden.

MA estimation in polynomial time

P. Stoica[1]   T. McKelvey[2]   J. Mari[3]

March 9, 1999

EDICS: 2-SPEC

Abstract

The parameter estimation of moving-average (MA) signals from second-order statistics was deemed for a long time to be a difficult nonlinear problem for which no computationally convenient and reliable solution was possible. In this paper we show how the problem of MA parameter estimation from sample covariances can be formulated as a semidefinite program which can be solved in polynomial time as efficiently as a linear program. Two methods are proposed which rely on two specific (over)parametrizations of the MA covariance sequence, whose use makes the minimization of the covariance fitting criterion a *convex problem*. The MA estimation algorithms proposed here are computationally fast, statistically accurate, and reliable (i.e. they "never" fail). None of the previously available algorithms for MA estimation (methods based on higher-order statistics included) shares all these desirable properties. Our methods can also be used to obtain the optimal least squares approximant of an invalid (estimated) MA spectrum (that takes on negative values at some frequencies), which was another long-standing problem in the signal processing literature awaiting a satisfactory solution.

Keywords

moving-average, estimation, time-series analysis, semidefinite programming, linear matrix inequalities, linear stochastic systems, system identification, covariance fitting.

I. Introduction

Estimation of moving-average (MA) parameters has been a research problem in the signal processing literature and elsewhere (statistics, econometrics etc.) for more than

[1] P. Stoica is with the Systems and Control Group at Uppsala University, PO Box 27, SE-751 03, Uppsala, Sweden. Email:ps@syscon.uu.se

[2] T. McKelvey is with the Division of Automatic Control at the Dep. of Electrical Eng., Linköping University, SE-581 83, Linköping, Sweden. Fax: +46 13 282622, Email:tomas@isy.liu.se. Corresponding author.

[3] J. Mari is with the Optimization and Systems Theory, Royal Institute of Technology, SE-100 44 Stockholm, Sweden. Email:jorgem@math.kth.se

50 years. Even so, all existing solutions to this problem appear to suffer from one or more drawbacks, as we explain briefly below.

• Several MA estimation methods, such as the maximum likelihood and moment fitting ones, depend on a multidimensional search; consequently they require the use of "global" search algorithms (and hence inordinately long computer time) for successful completion, or may otherwise fail to find the global optimum.

• Some methods first estimate MA covariances and then the MA parameters, for instance via spectral factorizations, and they have a hard failure each time the estimated covariances do not form a valid MA covariance sequence.

• Other methods obtain a long auto-regressive approximation of the MA signal, which is later used to estimate the MA parameters. Examples include Durbin method and the inverse covariance method. For any given sample length one can find MA signals for which this class of methods provide grossly biased MA estimates.

In summary, to our knowledge there were no methods available that:

(i) Obtain accurate MA parameter estimates in small or medium sized samples;

(ii) Require a computer time that is a polynomial function of the problem size (e.g. the MA order); and

(iii) Do not have a hard failure mode.

It should be noted that MA estimation methods based on higher-order statistics exist, which obtain MA parameter estimates by solving a linear system of equations and hence satisfy requirements (ii) and (iii) above. However accurate estimation of higher-order moments may need a rather large number of data samples, and consequently desideratum (i) may not be met by these methods.

Our main goal in this paper is to present two algorithms for MA parameter estimation from second-order statistics which possess all desiderata (i)-(iii) above. Briefly stated, these methods obtain MA estimates by minimizing a covariance fitting criterion. To do so they rely on two specific (over)parametrizations of the MA covariance sequence, whose use makes the minimization of the covariance fitting criterion a *convex problem* that can be efficiently solved in polynomial time.

We should note that the algorithms introduced here also provide a polynomial-time solution to a problem related to MA estimation that has been studied in the signal processing literature for quite some time. To describe this problem briefly, assume that the estimated covariances of an MA signal do not belong to the set of valid MA

covariance sequences. Consequently, the corresponding estimated spectrum takes on negative values at some frequencies and is hence not factorizable. To "factorize the unfactorizable" researchers have tried to correct the estimated MA covariance sequence in various ways to bring it back into the valid set see, e.g., [14] and the references therein. The methods of this paper can be used to obtain a least squares optimal solution to this problem in polynomial time.

Finally, we note that in [9], [8] we extend the ideas introduced here and combine them with subspace-based techniques to solve the more complicated problems of autoregressive moving-average (ARMA) and, respectively, vector ARMA (VARMA) parameter estimation from second order statistics. The reader who has found the present paper of interest should plan to study [9] and [8] as well.

II. Problem Formulation

Consider the following MA signal

$$y(t) = e(t) + c_1 e(t - 1) + \cdots + c_n e(t - n) = (1 + c_1 z^{-1} + \cdots + c_n z^{-n}) e(t) \overset{\Delta}{=} C(z) e(t) \tag{1}$$

where $t = 1, 2, \ldots$ denote the time index, z^{-1} is the unit delay operator, and $e(t)$ is white noise with mean zero and variance denoted by σ^2. The polynomial $C(z)$ can be assumed to have all its zeros inside the unit circle (the minimum-phase case) since such an assumption means no restriction for the second-order statistics that will be considered throughout this paper.

Let

$$r(k) = E[y(t)y(t - k)], \quad k = 0, 1, \ldots \tag{2}$$

denote the MA covariances, and let

$$\hat{r}(k) = \frac{1}{N} \sum_{t=k+1}^{N} y(t)y(t - k), \quad k = 0, 1, \ldots \tag{3}$$

denote their sample counterparts. Hereafter, N is the number of available data samples. For Gaussian signals the variance - covariances of $\{\hat{r}(k)\}$ are readily derived (see, e.g., [13]):

$$E\left\{ [\hat{r}(k) - r(k)][\hat{r}(p) - r(p)] \right\} \approx \frac{1}{N^2} \sum_{j=-n}^{n} (N - |j|)[r(j)r(j + k - p) + r(j - p)r(j + k)] \tag{4}$$

3

for $k, p = 0, 1, \ldots$ and $N \gg n$. If $\{e(t)\}$ is a sequence of random variables that are only uncorrelated but not independent (which may happen if $e(t)$ is not Gaussian distributed), then (4) no longer holds [12]. Nevertheless, in many cases (4) expresses in a reasonable manner the relative estimation errors in $\{\hat{r}(k)\}$, and we make use of it to weight or outweight the elements of our covariance fitting criterion (see below), regardless whether the data is Gaussian or not.

Let $\hat{W}$ denote the $(n + m + 1) \times (n + m + 1)$ matrix whose (k, p)-element is the following estimate of (4):

$$\hat{W}_{kp} = \frac{1}{N^2} \sum_{j=-n}^{n} (N - |j|)[\hat{r}(j)\hat{r}(j + k - p) + \hat{r}(j - p)\hat{r}(j + k)], \quad k, p = 0, 1, \ldots, n + m \tag{5}$$

where m is an integer whose choice will be discussed shortly. We can obtain estimates of $\{c_k\}$ and σ^2 by minimizing the following weighted least-squares covariance fitting criterion:

$$f = \left[\begin{array}{c} (\hat{r} - r) \\ \gamma \end{array} \right]^T \hat{W}^{-1} \left[\begin{array}{c} (\hat{r} - r) \\ \gamma \end{array} \right] \tag{6}$$

where the inverse exists under the assumption that N is not too small, and

$$\begin{aligned} \hat{r} &= \left[\begin{array}{ccc} \hat{r}(0) & \cdots & \hat{r}(n) \end{array} \right]^T \\ r &= \left[\begin{array}{ccc} r(0) & \cdots & r(n) \end{array} \right]^T \\ \gamma &= \left[\begin{array}{ccc} \hat{r}(n + 1) & \cdots & \hat{r}(n + m) \end{array} \right]^T \end{aligned} \tag{7}$$

The reason for including the sample covariances beyond lag n in the above covariance matching criterion can be explained as follows. For MA signals, unlike auto-regressive signals, the sample covariance vector $\hat{r}$ does *not* form a "sufficient statistic" (roughly speaking this means that $\hat{r}$ does not contain all "information" about the unknown parameters that is present in the raw data $\{y(t)\}_{t=1}^N$). The m extra sample covariances in γ used in (6) correspond to true covariances equal to zero, yet they contain "information" about the unknown MA parameters via their correlation with the sample covariances in $\hat{r}$. One can in effect show that the asymptotic (for $N \gg 1$) accuracy of the parameter estimates obtained by minimizing (6) monotonically increases with increasing m, provided that m increases at a slower rate than N; and in the limit, as $m, N \to \infty$ and $\frac{m}{N} \to 0$, the estimates derived from (6) can be shown to be a realization

4

of the (Gaussian) maximum-likelihood estimator [15]. In samples of practical length (with N in the order of hundreds) we found that choosing $m = \sqrt{N}$ gives satisfactory results whenever the zeros of the MA signal are close to the unit circle. If the MA zeros are well inside the unit circle one can choose a smaller m than $\sqrt{N}$ to reduce the computational burden (without sacrificing the statistical accuracy), provided that information about the approximate zero locations is available.

The criterion in (6) can be rewritten in a more convenient form as follows. Partition $\hat{W}$ as

$$\hat{W} = \left[\begin{array}{cc} \hat{W}_{11} & \hat{W}_{12} \\ \hat{W}_{12}^T & \hat{W}_{22} \end{array} \right] \tag{8}$$

where $\hat{W}_{11}$ is of dimension $(n+1) \times (n+1)$ etc. Then, make use of a standard formula for the inverse of a partitioned matrix (see, e.g., [13]) to write f in (6) as:

$$f = (\tilde{r} - r)^T (\hat{W}_{11} - \hat{W}_{12} \hat{W}_{22}^{-1} \hat{W}_{12}^T)^{-1} (\tilde{r} - r) + \text{const.} \tag{9}$$

where

$$\tilde{r} = \hat{r} - \hat{W}_{12} \hat{W}_{22}^{-1} \gamma \tag{10}$$

The vector $\tilde{r}$ in (10) is readily checked to be a consistent estimate of r with the following large-sample covariance matrix:

$$\hat{\Gamma} = \hat{W}_{11} - \hat{W}_{12} \hat{W}_{22}^{-1} \hat{W}_{12}^T \tag{11}$$

(see, e.g., [15]). Since $\hat{W}_{11}$ is the large-sample covariance matrix of $\hat{r}$, and as $\hat{\Gamma} \leq \hat{W}_{11}$ by the definition in (11), it follows that $\tilde{r}$ is a more accurate estimate of r than $\hat{r}$.

The previous discussion implies that the optimally-weighted least-squares fitting in (6) based on the $(n + m + 1)$ sample covariances in $\hat{r}$ and γ can be reformulated as an optimally-weighted least-squares fitting involving only the $(n+1)$ *enhanced covariance estimates* in $\tilde{r}$. The first advantage of this reformulation is a dimension reduction from $(n + m + 1)$ for (6) to $(n + 1)$ for (9). To explain the second advantage, let $\tilde{\Gamma}$ denote the matrix $\hat{\Gamma}$ defined as in (11) but made from $\tilde{r}$. Since $\tilde{r}$ is a better estimate of r than $\hat{r}$ we propose to use $\tilde{\Gamma}$ instead of $\hat{\Gamma}$ in (9). Hence we will obtain estimates of the MA parameters by minimizing the following criterion

$$g = (\tilde{r} - r)^T \tilde{\Gamma}^{-1} (\tilde{r} - r) \tag{12}$$

5

Regarding the MA order, n, we assume that it is given, and write MA(n) whenever it is important to specify the order of the MA signal under discussion. Note, however, that we do not assume that $c_n \neq 0$ etc., which means that n is only an *upper bound* on the true MA order. Methods for estimating n when it is unknown are discussed, for example, in [13].

The minimization of (12) with respect to $\{c_k\}$ and σ^2 is a highly nonlinear, essentially unconstrained problem. The minimization of (12) with respect to $\{r(k)\}$ is a quadratic problem with a convex but rather complicated constraint set. Neither of these optimization problems can be conveniently solved by means of existing algorithms.

To obtain a minimization problem that can be solved in polynomial time we use a larger number of parameters than $n+1$ for describing the permissible MA covariance set. In the next section we make use of an overparametrized state-space model for the MA signal and call the resultant MA parameter estimation algorithm OS (standing for Overparametrized Signal). In Section IV we introduce another algorithm for minimizing (12) called OC (Overparametrized Covariances) which relies on an overparametrized model for the MA covariance sequence.

The use of either of the aforementioned overparametrizations makes it possible to transform (12) into a so-called Semidefinite Program (SP), the generic form of which is (see, e.g. [3], [18]):

$$
\min_x c^T x,
$$
$$
\text{s.t. } A(x) \geq 0
\tag{13}
$$

where s.t. is an abbreviation for "subject to", $A(x) = A_0 + A_1 x_1 + \cdots + A_m x_m \geq 0$ is an $\tilde{m} \times \tilde{m}$ matrix that depends affinely on the elements $\{x_k\}_{k=1}^m$ of x, $\{A_i\}$ and c are given, and $A \geq 0$ means that A is a positive semidefinite matrix. The SP (13) can be solved efficiently in polynomial time by means of interior-point methods [18], [10], [3] which are iterative in nature. The worst-case number of iterations to find a solution to (13), at a given accuracy, grows with problem size as $O(\tilde{m}^{1/2})$. In reality though, the number of iterations grows much more slowly and can often be assumed to be constant (see [10], [18]), typically 10-50 iterations. On its turn, each iteration requires the solution of a set of linear equations or a least-squares problem with a size that grows with m and $\tilde{m}$. Without utilizing any structural properties of $A(x)$ (besides symmetry), each iteration requires $O(\tilde{m}^2 m^2)$ flops. To summarize, an SP can be solved as efficiently as a linear program (which can be viewed as a special case of (13) corresponding to

diagonal $\{A_k\}$ matrices). In addition the available algorithms for solving (13) do not encounter any problems when (13) has multiple solutions.

III. Overparametrized Signal (OS) Approach

It is readily verified that the following state-space equations constitute a representation of the MA signal in (1):

$$x(t+1) = Ax(t) + ce(t)$$
$$y(t) = \zeta^T x(t) + e(t) \tag{14}$$

where

$$A = \begin{bmatrix} 0 & \cdots & & 0 \\ 1 & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & 0 \end{bmatrix}; \quad \zeta = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad c = \begin{bmatrix} c_n \\ \vdots \\ c_1 \end{bmatrix} \tag{15}$$

Defining

$$v(t) = ce(t) \tag{16}$$

we can rewrite (14) as:

$$x(t+1) = Ax(t) + v(t)$$
$$y(t) = \zeta^T x(t) + e(t) \tag{17}$$

which, in turn, yields the following equation for the output signal

$$y(t) = \zeta^T (zI - A)^{-1} v(t) + e(t) = \begin{bmatrix} z^{-n} & \cdots & z^{-1} \end{bmatrix} v(t) + e(t) \tag{18}$$

Note that we could obtain (18) directly from the signal equation (1). The use of the above state-space representation to derive (18) serves only the purpose of contrasting the approaches in this and the next section.

Let Q denote the $(n+1) \times (n+1)$ covariance matrix of the vector $\begin{bmatrix} v^T(t) & e(t) \end{bmatrix}^T$:

$$E \left\{ \begin{bmatrix} v(t) \\ e(t) \end{bmatrix} \begin{bmatrix} v^T(s) & e(s) \end{bmatrix} \right\} = Q\delta_{t,s} \tag{19}$$

where $\delta_{t,s}$ is the Kronecker delta. Also, let u_k denote the kth column of the $(n+1) \times (n+1)$ identity matrix. Using this notation and (18) we obtain the following expression

for the covariances of $\{y(t)\}$:

$$
\begin{aligned}
r(j) &= E\left\{\sum_{k=1}^{n+1}\sum_{p=1}^{n+1} u_k^T \begin{bmatrix} v(t-n-1+k) \\ e(t-n-1+k) \end{bmatrix} \begin{bmatrix} v(t-n-1+p-j) \\ e(t-n-1+p-j) \end{bmatrix}^T u_p \right\} = \\
&= \sum_{k=1}^{n+1}\sum_{p=1}^{n+1} u_k^T Q u_p \delta_{k,p-j} = \begin{cases} \sum_{k=1}^{n+1-j} u_{k+j}^T Q u_k & j = 0,\cdots,n \\ 0 & j \geq n+1 \end{cases}
\end{aligned}
\tag{20}
$$

Note from the definition (16) of $v(t)$ that:

$$
Q = \begin{bmatrix} c_n \\ \vdots \\ c_1 \\ 1 \end{bmatrix} \begin{bmatrix} c_n & \cdots & c_1 & 1 \end{bmatrix} \sigma^2
\tag{21}
$$

If we observe (21) and hence parametrize $\{r(j)\}$ via $\{c_k\}$ and σ^2 then there is no known method to solve the minimization problem (12) in polynomial time. A natural way to overcome this problem is to *(over)parametrize $\{r(j)\}$ in a linear manner via the independent elements of Q*. A similar idea has been used in [16] for estimation of power spectra from frequency measurements.

Note that the transformation from Q to $\{r(j)\}$ (and also $\{c_k\}$) is a many-to-one mapping see, e.g., [4]. In particular the covariance sequence corresponding to the rank-one matrix Q in (21) can also be described by using other Q matrices with rank larger than one. It is also important to note that *for any $Q \geq 0$ the $\{r(j)\}$ obtained by (20) form a valid MA covariance sequence of length $k \leq n$, and that one can span the whole set of permissible $MA(k)$ covariances, with $k \leq n$, by varying $Q \geq 0$ in (20).*

Let α denote the $(n+1)(n+2)/2 \times 1$ vector comprising the elements of Q on and above the main diagonal, i.e., $Q = Q(\alpha)$; also, let T denote the transformation from α to $\text{vec}(Q)$. Then, making use of (19) and a simple property of the Kronecker matrix product (denoted by the symbol $\otimes$), we obtain:

$$
\begin{aligned}
r &= \sum_{k=1}^{n+1} \begin{bmatrix} u_k^T \\ \vdots \\ u_{k+n}^T \end{bmatrix} Q u_k = \sum_{k=1}^{n+1} \left\{ u_k^T \otimes \begin{bmatrix} u_k^T \\ \vdots \\ u_{k+n}^T \end{bmatrix} \right\} T\alpha \\
&\overset{\Delta}{=} \Phi\alpha
\end{aligned}
\tag{22}
$$

where $u_k = 0$ for $k > n+1$. Inserting (22) into (12) results into the following reformu-

8

lation of the problem:

$$\min_{\alpha}(\tilde{r} - \Phi\alpha)^T\tilde{\Gamma}^{-1}(\tilde{r} - \Phi\alpha)$$

$$\text{s.t. } Q(\alpha) \geq 0 \tag{23}$$

or equivalently (see e.g., [3], [18]):

$$\min_{\rho,\alpha} \rho,$$

$$\text{s.t. } \rho - (\tilde{r} - \Phi\alpha)^T\tilde{\Gamma}^{-1}(\tilde{r} - \Phi\alpha) \geq 0 \tag{24}$$

$$Q(\alpha) \geq 0$$

where ρ is an auxiliary variable. Making use of a simple property of Schur complements we can readily rewrite (24) as a SP:

$$\min_{\rho,\alpha} \rho$$

$$\text{s.t } \left[\begin{array}{cc|c} \rho & (\tilde{r} - \Phi\alpha)^T & 0 \\ \tilde{r} - \Phi\alpha & \tilde{\Gamma} & \\ \hline & 0 & Q(\alpha) \end{array}\right] \geq 0 \tag{25}$$

Observe that both the criterion function and the *linear matrix inequality* (LMI) in (25) depend affinely on ρ and α.

Denote by $\mathcal{S}$ the set of covariances of an MA(n) signal:

$$\mathcal{S} = \{r|\{r_k\}_{k=0}^n \text{ are the covariances of an MA}(n) \text{ signal}\} \tag{26}$$

By using the convention that this set also includes the covariances of all MA signals of orders less than n we note that this set is (over)parametrized by α through (22):

$$\mathcal{S} = \{r|r = \Phi\alpha, \ Q(\alpha) \geq 0\} \tag{27}$$

The *OS algorithm* derives from (25) and consists of the following steps:

The OS Algorithm.

0. Test if $\tilde{r}$ defined by (10) belongs to the set of valid MA(n) covariance vectors $\mathcal{S}$. If yes, set $\{\hat{r}(k) = \tilde{r}(k)\}$ and go to step 3 below. If not go to step 1. There are several test procedures available for checking whether $\tilde{r} \in \mathcal{S}$ (e.g. [6]). An alternative (which we will shortly discuss in some detail) is to check the feasibility of a linear matrix inequality, i.e., to check the existence of a $P = P^T$ such that $\Pi(P, \tilde{r}) \geq 0$ where $\Pi(P, r)$ is defined by (36) below.

9

1. Obtain an estimate of Q by solving the SP (25).

2. Use the so-estimated Q to derive valid estimates of the MA(n) covariances and spectrum. Let the latter estimates be denoted by $\{\hat{\hat{r}}(k)\}$ and, respectively, $\hat{\hat{\phi}}(z) = \sum_{|k| \leq n} \hat{\hat{r}}(k) z^k$.

3. Estimate $\{c_k\}$ and σ^2 by factoring the estimated spectrum:

$$\hat{\hat{\phi}}(z) = \hat{\sigma}^2 \hat{C}(z) \hat{C}(z^{-1}) \tag{28}$$

This can be done efficiently by several procedures (see [21], [19], [2]); the resultant $\hat{C}(z)$ polynomial will be minimum phase.

Remark 1: Note that given the estimated Q from step 1, $\{c_k\}$ can also be directly found by solving the discrete-time Riccati equation for the associated optimal filtering problem without having to form the the covariance sequence first. To see this, return to the state-space realization given in (14). It is known as the innovations form and the vector c containing the $\{c_k\}$ parameters is the Kalman gain for the associated optimal state-estimation problem. Given a matrix Q representing the covariance matrix of the driving noise in the MA model (17), solving the Riccati equation for the Kalman gain vector directly yields the MA parameters $\{c_k\}$.

The number of variables in the SP above is

$$\#\text{OS} = \frac{(n+1)(n+2)}{2} + 1 \tag{29}$$

as compared with the $n+1$ variables of the original problem. This seemingly significant overparametrization might suggest a possible degradation of the estimation accuracy. However this is *not* so. The SP in (25) is just a *reformulation* of (12). Both problems determine the valid MA(n) covariance sequence r that is minimally distanced from $\tilde{r}$, in the metric induced by $\tilde{\Gamma}$. Hence *the OS algorithm outlined above provides the same estimates of the MA parameters as the (global) minimization of* (12) *with respect to* σ^2 *and* $\{c_k\}$. Consequently the only effect of the overparametrization via α is to transform the difficult problem (12) into a tractable SP.

The discussion in the previous paragraphs also implies that *if $\tilde{r}$ does not belong to $\mathcal{S}$ (the set of valid MA(n) covariances) then the estimated covariance vector $\hat{\hat{r}}$ obtained by OS will belong to the boundary of $\mathcal{S}$.* (This follows from the equivalence of (12) and (25), and the convexity of $\mathcal{S}$). In such a case, the polynomial $\hat{C}(z)$ in (28) will have zeros on the unit circle, which is natural as the usually small estimation errors in $\tilde{r}$

may pull this vector out of $\mathcal{S}$ only if r was "close to" the boundary of $\mathcal{S}$ and hence $C(z)$ had zeros "close to" the unit circle.

IV. Overparametrized Covariance (OC) Approach

The approach in the previous section eventually relies on an overparametrized covariance model, see (27). However this model was obtained by starting from an overparametrized signal equation (17). This distinguishes the OS approach from the OC approach of this section that overparametrizes the MA covariance sequence directly.

To solve the covariance fitting problem (12) we must parametrize $\mathcal{S}$ in some way. The parametrization of $\mathcal{S}$ via r is quite complicated and, as already stated, does not lead to a tractable problem. However we can append to r a number of auxiliary variables in such a manner that *the condition on the extended vector of variables for r to belong to $\mathcal{S}$ becomes a simple LMI.* Reformulation of the problem (12) as a SP will then be straightforward. This idea, which was utilized in [20], [21] for FIR filter design, is detailed in what follows.

Let

$$\Psi(z) = \frac{1}{2}r(0) + r(1)z^{-1} + \cdots + r(n)z^{-n} \tag{30}$$

The spectrum of the MA signal in (1) can then be written as

$$\phi(z) = \Psi(z) + \Psi(z^{-1}) \tag{31}$$

and it is well known that a possible mathematical characterization of the set $\mathcal{S}$ is:

$$\mathcal{S} = \left\{ r \mid \text{Re}[\Psi(e^{i\omega})] \geq 0 \text{ for } \omega \in [-\pi, \pi] \right\} \tag{32}$$

Next let A be as defined in (15), and let

$$g = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T ; \quad h = \begin{bmatrix} r(1) & \cdots & r(n) \end{bmatrix}^T ; \quad d = r(0)/2 \tag{33}$$

It is a simple matter to check that:

$$\begin{bmatrix} z & 0 & \cdots & 0 \\ -1 & \ddots & & \ddots \\ & \ddots & \ddots & \\ 0 & \cdots & -1 & z \end{bmatrix} \begin{bmatrix} z^{-1} \\ \vdots \\ z^{-n} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

11

which directly implies that:

$$h^T (zI - A)^{-1} g + d = \Psi(z) \tag{34}$$

Hence (A, g, h, d) is a state-space realization of the FIR "transfer function" $\Psi(z)$ made from $\{r(k)\}_{k=0}^n$. Furthermore, this state-space realization is always controllable and in Appendix A we prove that it is *minimal* if and only if

$$c_n \neq 0 \tag{35}$$

In [20], [21] the "minimality condition" (35) was thought of being necessary for the formulation of (32) as an LMI. In reality (35) is not required, which allows n to be larger than the true MA order. Indeed, because (A, g) in the above state-space realization is a controllable pair it follows from the *positive real lemma* (see Appendix B) that $r \in \mathcal{S}$ *if and only if there exists an $n \times n$ symmetric matrix P such that*:

$$\Pi(P, r) = \begin{bmatrix} P - A^T P A & h - A^T P g \\ h^T - g^T P A & 2d - g^T P g \end{bmatrix} \geq 0 \tag{36}$$

Observe that (36) is an LMI in the covariances $\{r(k)\}_{k=0}^n$ and the $n(n+1)/2$ independent elements of P. Making use of this observation and a Schur complement trick similar to the one that led to (25), we can reformulate (12) as the SP:

$$\min_{\rho, P, r} \rho$$

$$\text{s.t.} \quad \left[\begin{array}{cc|c} \rho & (\tilde{r} - r)^T & 0 \\ \tilde{r} - r & \tilde{\Gamma} & \\ \hline & 0 & \Pi(P, r) \end{array} \right] \geq 0 \tag{37}$$

We can now outline the OC algorithm.

The OC Algorithm.

0. As for the OS algorithm.

1. Determine a solution of the SP (37) and obtain an estimate $\hat{r} \in \mathcal{S}$ of r from the corresponding elements of the solution vector β.

2. Same as Step 3 of the OS algorithm.

The number of variables in the SP associated with the above algorithm is

$$\#\text{OC} = \frac{(n+1)(n+2)}{2} + 1 \tag{38}$$

12

This is exactly the same as the number of variables in the OS algorithm (see (29)). Also, the size of the LMI is $2n + 3$ for both algorithms. Accordingly, one can expect that the computational complexities of the OS and OC algorithms are quite similar to one another. In the numerical example section we compare the computational burdens of the two algorithms as well as their accuracies. *Theoretically both algorithms should provide the same estimates of $\{c_k\}$ and σ^2*, as they both obtain the global minimizer of the criterion in (12). However the numerical accuracies of the two algorithms may differ from one another, and this might introduce some (small) differences between the parameter estimates obtained by using them.

V. NUMERICAL EXAMPLES

This section presents two numerical examples to illustrate the performance of the proposed algorithms and to compare them with existing techniques. The new algorithms are implemented in MATLAB and the semidefinite program is solved using the software given in [17]. The following algorithms are compared:

1. The OS algorithm with $m = 20$.

2. The OC algorithm with $m = 20$.

3. Durbin's algorithm, which can be outlined as follows. First a high order AR model $\hat{A}(z)$ is estimated by using the least-squares method (LS). The MATLAB implementation of this step (see [7]) uses an AR equation of order $4n$, which is also what we use in the following. Using the AR model, the innovations are estimated by $\hat{e}(t) = \hat{A}(z)y(t)$. The MA($n$) polynomial $C(z)$ is then estimated by a LS procedure applied to the linear regression

$$y(t) = C(z)\hat{e}(t).$$

Any zeros z_i of the estimated polynomial $\hat{C}(z)$ which are outside the unit circle are moved inside by replacing them with $\frac{1}{z_i}$ (which does not alter the power spectrum estimate).

4. The maximum-likelihood method (ML) for which the `ARMAX` command in the MATLAB System Identification Toolbox [7] is used. The solution is found by non-linear optimization using a Gauss-Newton type of algorithm. The optimization is initialized by the estimate from the Durbin algorithm. The Gauss-Newton optimization algorithm constrains the zeros of the estimate to lie strictly inside the unit circle by modifying the parameter updates which would yield zeros outside or on the unit circle.

13

A. Example 1

Consider an MA(3) signal $y(t) = C(z)e(t)$ where

$$C(z) = (1 - 0.95z^{-1})(1 - 0.98e^{i\frac{\pi}{4}}z^{-1})(1 - 0.98e^{-i\frac{\pi}{4}}z^{-1}) \tag{39}$$

and $e(t)$ is a zero-mean Gaussian white noise with unit variance. The sample size is $N = 400$. A Monte-Carlo simulation is performed based on 20 independent data sets $\{y(t)\}$. For each data set the four aforementioned estimation algorithms were applied. The results of the simulation are summarized in Figures 1–3.
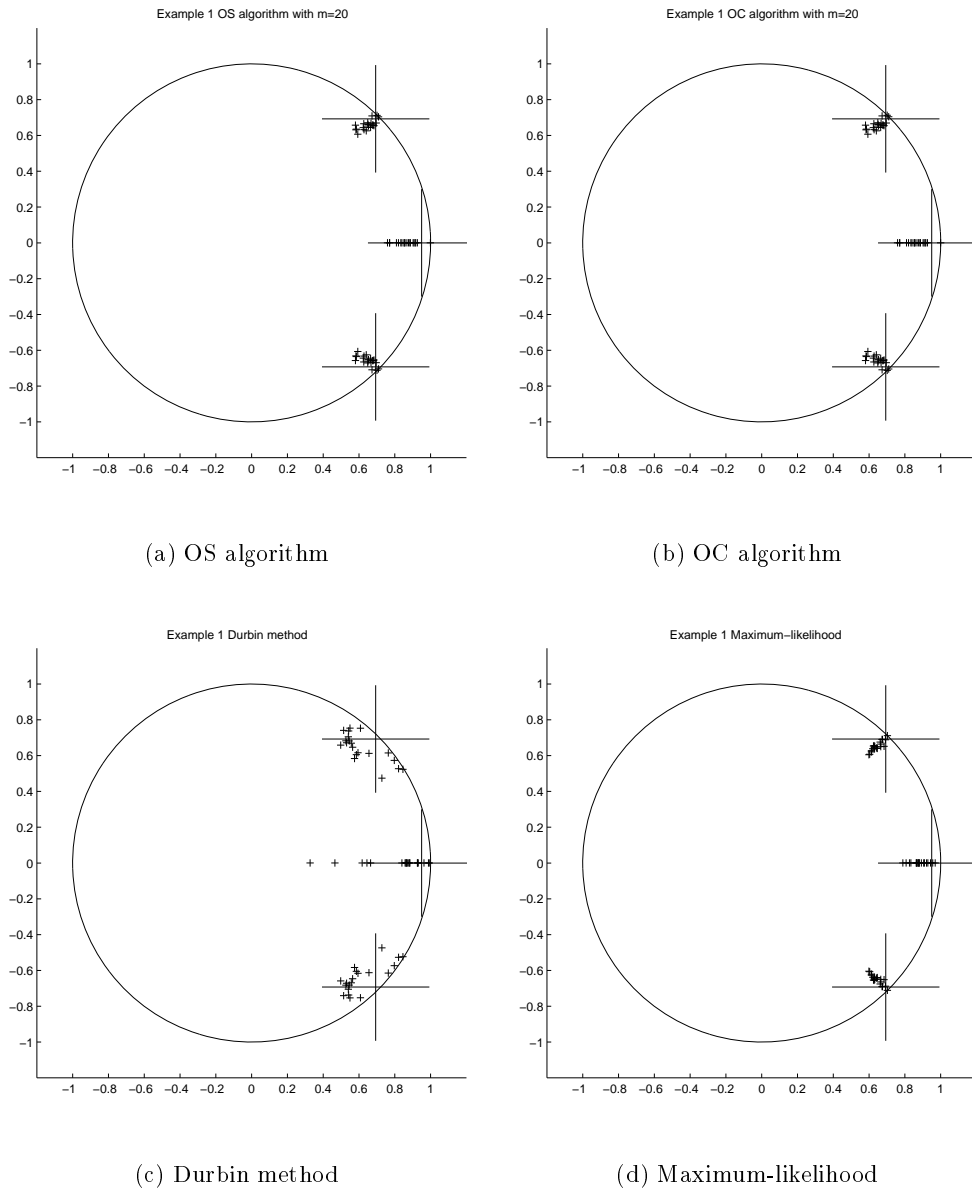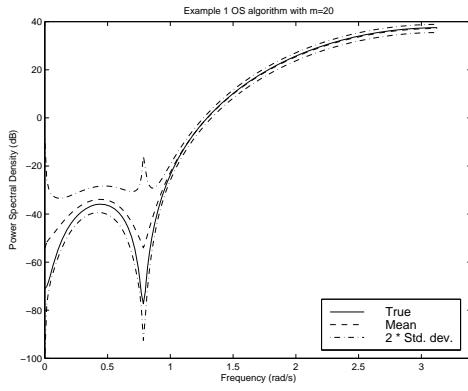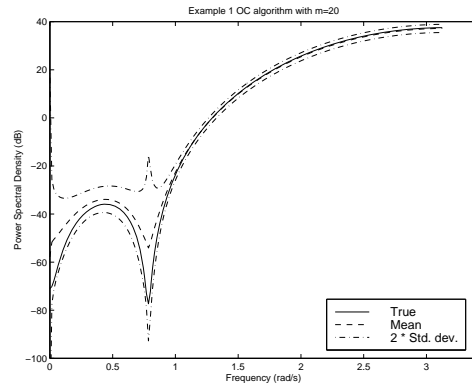

(a) OS algorithm

(b) OC algorithm

(c) Durbin method

(d) Maximum-likelihood

Fig. 1. Example 1: Superimposed estimated zero locations for 20 independent data sets. The large crosses indicate the true locations.

(a) OS algorithm

(b) OC algorithm
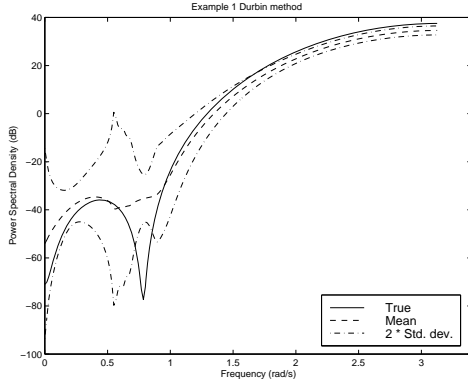
(c) Durbin method
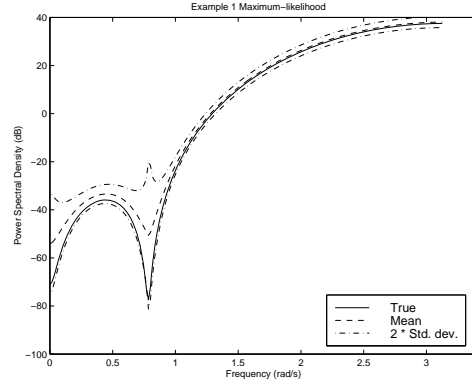
(d) Maximum-likelihood

Fig. 2. Example 1: Estimated power spectra. Sample statistics based on 20 independent realizations. Solid line: true spectrum. Dashed line: estimated mean. Dash-dotted lines: $\pm$ twice the standard deviation.

The two new methods (OS and OC) start from an initial estimate $\tilde{r}$ of the covariance vector. If this estimate is not a valid covariance sequence ($\tilde{r} \notin \mathcal{S}$) then the SP optimization is initiated. In Table I the percentage of invalid initial estimates is reported together with the average execution time.

B. Example 2

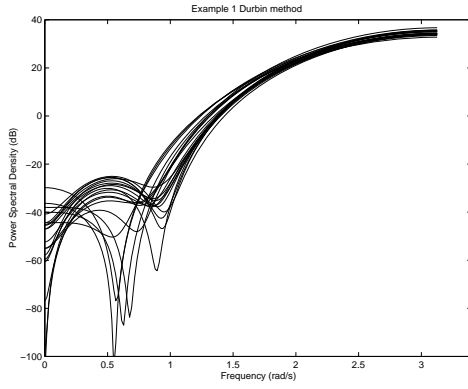The second example investigates parameter estimation performance for an MA system of order 6 with zeros located at $0.99\,e^{\pm i\frac{\pi}{4}}, 0.99\,e^{\pm i\frac{3\pi}{4}}$ and $0.999\,e^{\pm i\frac{\pi}{2}}$. The data are generated using the same setup as in Example 1. Figures 4–6 and Table II present the results of the Monte-Carlo simulation.
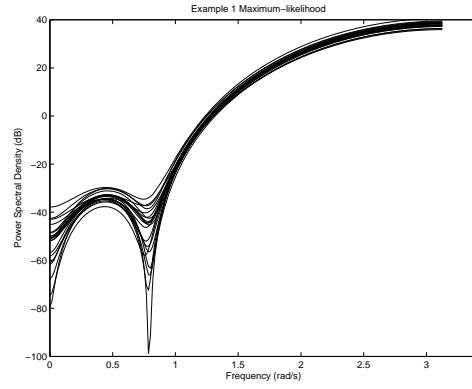
15
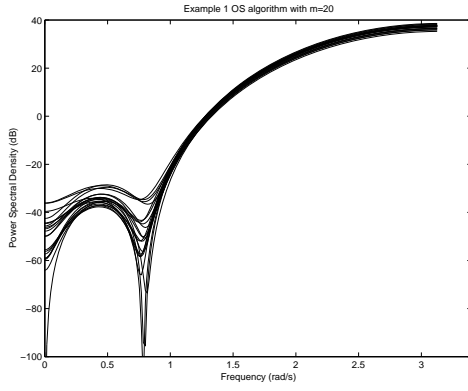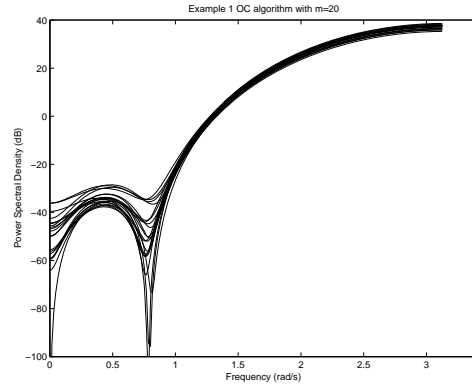
(a) OS algorithm

(b) OC algorithm

(c) Durbin method

(d) Maximum-likelihood

Fig. 3. Example 1: Superimposed estimated spectra for 20 independent data sets.

TABLE I

Method	Aver. exec. time (sec)	Invalid initial estim.(%)
OS	0.249	15
OC	0.240	15
Durbin	0.043	n.a.
Maximum-likelihood	0.245	n.a.

(a) OS algorithm

(b) OC algorithm
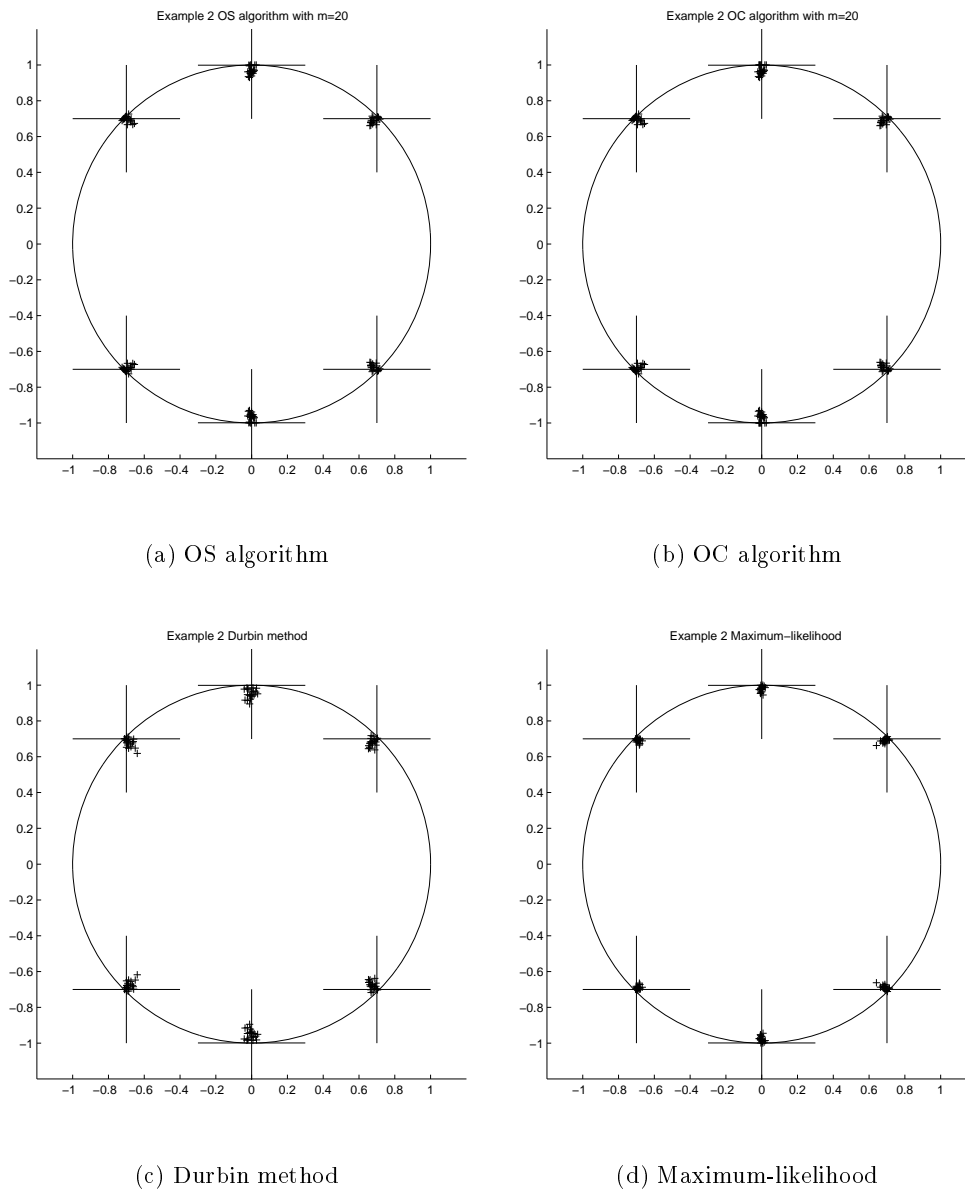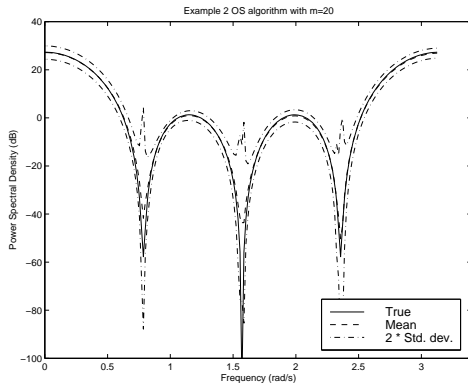
(c) Durbin method

(d) Maximum-likelihood

Fig. 4. Example 2: Superimposed estimated zero locations for 20 independent data sets. The large crosses indicate the true locations.

TABLE II
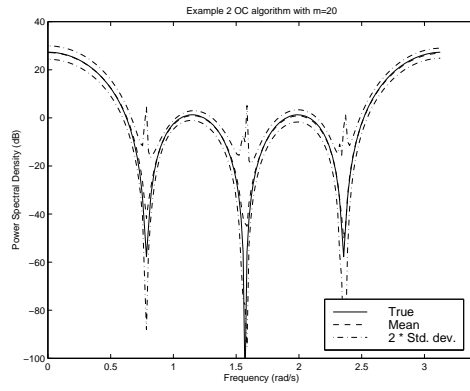
RESULTS FROM THE MONTE-CARLO SIMULATION FOR EXAMPLE 2. (N.A. = NOT APPLICABLE)
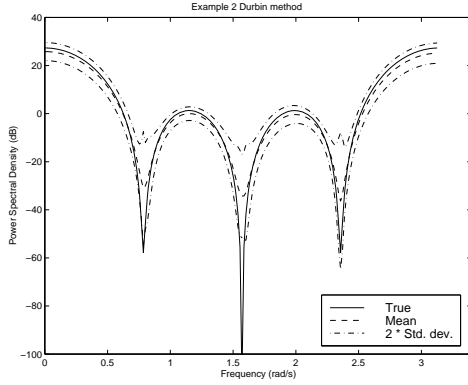
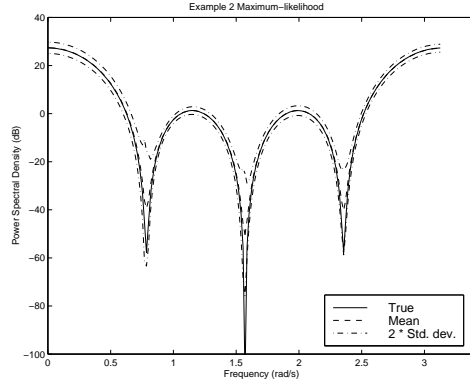| Method	Aver. exec. time (sec)	Invalid initial estim.(%)
OS	1.08	80
OC	1.10	80
Durbin	0.078	n.a.
Maximum-likelihood	0.322	n.a.

17

(a) OS algorithm

(b) OC algorithm

(c) Durbin method

(d) Maximum-likelihood

Fig. 5. Example 2: Estimated power spectra. Sample statistics based on 20 independent realizations. Solid line: true spectrum. Dashed line: estimated mean. Dash-dotted lines: ± twice the standard deviation.

C. Summary

As already stated the OS and OC algorithms should theoretically yield identical estimates since they only differ by the way the constraint is implemented. The simulation results confirm this, as only small numerical differences can be detected between the OS and OC estimates which are likely due to the choice of stopping criteria in the SP solver.

Note that the MA signals considered in both examples have zeros near the unit circle. Such signals have a large spectral range and estimation of their parameters is more challenging and hence more interesting. We have also considered MA signals with zeros well inside the unit circle. While we omit the results obtained in such cases,

(a) OS algorithm

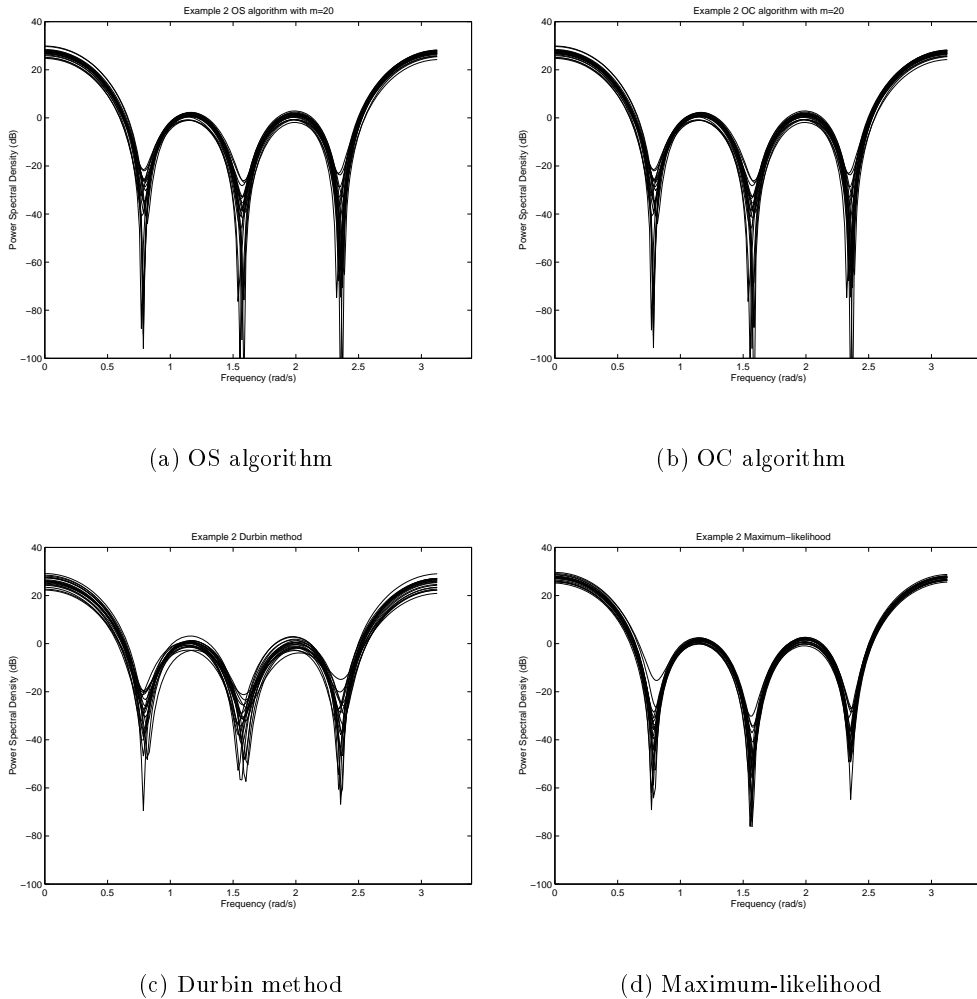(b) OC algorithm

(c) Durbin method

(d) Maximum-likelihood

Fig. 6. Example 2: Superimposed estimated spectra for 20 independent data sets.

let us mention that for this type of MA signals the four methods above behave quite similarly and that the OS and OC algorithms usually terminate after step 0.

The performance of OS and OC algorithms is clearly better than that of the Durbin method and is very close to the maximum-likelihood performance. Although the performance of the ML method in the two examples was quite good it should be noted that the ML method relies on a non-linear optimization without any guarantee of global convergence. The risk of the ML algorithm being trapped in a false maximum of the likelihood function is higher for real data that do not exactly satisfy any MA equation than for synthetic MA data. In contrast the OC and OS methods are based on a convex optimization formulation and convergence to the optimum is guaranteed in polynomial time.

The execution time for the OS and OC methods seems to increase faster with in-

19

creasing model order n as compared with the other methods. The reason is that for the new methods the number of estimated "parameters" grows as $O(n^2)$, see (29) and (38). However, we should note that the SP solver used does not exploit any structure of the problem. Significant execution time reductions are expected if the sparseness of the problem is properly utilized. Furthermore, our code for the OS and OC algorithms is far from optimal; in particular it contains, unlike the code for the ML algorithm, a relatively large number of external loops that are notoriously slow in MATLAB. Optimization of future versions of our code should lead to further reductions of the execution times associated with the OS and OC algorithms. The reader can download the latest available version of the code for the OS and OC algorithms from the www address `http://www.control.isy.liu.se/~tomas/maestim`

VI. Conclusions

Two novel methods for estimation of the parameters of a moving average signal have been introduced. The methods share the following important features:

(i) They obtain accurate MA parameter estimates in small or medium sized samples;

(ii) They require a computer time which, in worst case, is a polynomial function of the problem size (e.g. the MA order); and

(iii) They do not have a hard failure mode.

The two methods obtain MA estimates by minimizing an optimally weighted covariance fitting criterion. To do so they rely on two specific (over)parametrizations of the MA covariance sequence, whose use makes the minimization of the covariance fitting criterion a *convex problem* that can be efficiently solved in polynomial time.

Acknowledgment

I. Proof of the minimality property

A state-space realization (A, g, h, d) is minimal if and only if its observability matrix

$$\mathcal{O} = \begin{bmatrix} h^T \\ h^T A \\ \vdots \\ h^T A^{n-1} \end{bmatrix}$$

and its controllability matrix

$$\mathcal{C} = \begin{bmatrix} g & Ag & \cdots & A^{n-1}g \end{bmatrix}$$

are non-singular. A straightforward calculation shows that

$$\mathcal{O} = \begin{bmatrix} r(1) & r(2) & \cdots & r(n) \\ r(2) & & \cdot & 0 \\ \vdots & \cdot & & \vdots \\ r(n) & 0 & \cdots & 0 \end{bmatrix}$$

and

$$\mathcal{C} = I$$

Hence $\mathcal{C}$ is always nonsingular, whereas $\mathcal{O}$ is so if and only if $r(n) \neq 0$, or equivalently $c_n = r(n)/\sigma^2 \neq 0$.

II. The positive real lemma

First we state the discrete-time version of the celebrated Kalman-Yacubovich-Popov (KYP) lemma. We present the result for the multivariable case for the sake of generality.

Lemma 1 (KYP) Given $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $M \in \mathbb{R}^{(n+m) \times (n+m)}$, with $\det(e^{i\omega}I - A) \neq 0$ for $\omega \in \mathbb{R}$ and (A, B) controllable, the following two statements are equivalent.
(i) For $\omega \in \mathbb{R}$

$$\begin{bmatrix} (e^{i\omega}I - A)^{-1}B \\ I \end{bmatrix}^* M \begin{bmatrix} (e^{i\omega}I - A)^{-1}B \\ I \end{bmatrix} \geq 0$$

where superscript * denotes the conjugate transpose.

(ii) There exists a matrix $P \in \mathbb{R}^{n \times n}$ such that $P = P^T$ and

$$M + \begin{bmatrix} P - A^T P A & -A^T P B \\ -B^T P A & -B^T P B \end{bmatrix} \geq 0$$

The corresponding equivalence for strict inequalities holds even if (A, B) is not controllable.

    Proof: See e.g. [11]. ∎

Based on the KYP-lemma, the following positive real lemma is easily derived, (see also [1], [5]).

    Lemma 2 (PR) Given $A \in \mathbb{R}^{n \times n}$, $B, C^T \in \mathbb{R}^{n \times m}$, $D \in \mathbb{R}^{m \times m}$, with $\det(e^{i\omega} I - A) \neq 0$ for $\omega \in \mathbb{R}$, (A, B) controllable and $\Psi(\omega) = D + C(e^{i\omega} I - A)^{-1} B$, the following two statements are equivalent.

(i) For $\omega \in \mathbb{R}$

$$\Psi^*(\omega) + \Psi(\omega) \geq 0$$

(ii) There exists a $P \in \mathbb{R}^{n \times n}$ such that $P = P^T$ and

$$\begin{bmatrix} P - A^T P A & C^T - A^T P B \\ C - B^T P A & D + D^T - B^T P B \end{bmatrix} \geq 0 \tag{40}$$

The corresponding equivalence for strict inequalities holds even if (A, B) is not controllable.

    Proof: The result follows directly from the KYP-lemma by letting

$$M = \begin{bmatrix} 0 & C^T \\ C & D + D^T \end{bmatrix}$$

∎

Note that for scalar systems we have

$$\Psi^*(\omega) + \Psi(\omega) = 2 \operatorname{Re}[\Psi(\omega)]$$

which clarifies the name of the lemma.

    Consider the state-space formulation $\Psi(z)$ as defined by (15), (33) and (34). The pair (A, g) is controllable as shown in Appendix A. Consequently the counterpart of (40) in (36) is the necessary and sufficient condition for the positive realness of $\Psi(z)$ in (30), or equivalently for the vector r to be a valid covariance sequence ($r \in \mathcal{S}$).

References

[1] B. D. O. Anderson. A system theory criterion for positive real matrices. *SIAM Journal of Control*, 5:171–182, 1967.

[2] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, New Jersey, 1979.

[3] S. Boyd, L. El Ghaoui, E. Feron, and V.Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, 1994.

[4] P. Faurre. Stochastic realization algorithms. In R. Mehra and D. Lainiotis, editors, *System Identification: Advances and Case Studies*, volume 126 of *Mathematics in Science and Engineering*. Academic Press, NY, 1976.

[5] P. Faurre, M. Clerget, and F. Germain. *Opérateurs Rationnels Postitifs*. Dunod, 1979.

[6] E. Jury. *Inners and Stability of Dynamic Systems*. Wiley, New York, 1974.

[7] L. Ljung. *System Identification Toolbox*. The Mathworks, Inc, 1998.

[8] J. Mari, P. Stoica, and T. McKelvey. Vector ARMA Estimation: a complete story. In preparation.

[9] T. McKelvey, P. Stoica, and J. Mari. ARMA estimation made simple. In preparation.

[10] Y. Nesterov and A. Nemirovsky. *Interior-Point Polynomial Methods in Convex Programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, 1994.

[11] A. Rantzer. On the Kalman-Yakubovich-Popov lemma. *Systems Control Lett.*, 28(1):7–10, 1996.

[12] J. Romano and L. Thombs. Inference for autocorrelations under weak assumptions. Technical Report 436, Dept. of Statistics, Stanford University, CA, Aug 1993.

[13] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall International, Hemel Hempstead, Hertfordshire, 1989.

[14] A. Stoica, R. Moses, and P. Stoica. Enforcing positiveness on estimated spectral densities. *Electronics Lett.*, 29:2009–2011, 1993.

[15] P. Stoica, B. Friedlander, and T. Söderström. Approximate maximum-likelihood approach to ARMA spectral estimation. *Int. J. of Control*, 45(4):1281–1310, 1987.

[16] P. Van Overschee, B. De Moor, W. Dehandschutter, and J. Swevers. A subspace algorithm for the identification of discrete time frequency domain power spectra. *Automatica*, 33(12):2147–2157, Dec 1997.

[17] L. Vandenberghe and S. Boyd. *SP: Software for Semidefinite Programming. User's Guide*. K.U. Leuven and Stanford University, http://www-isl.stanford.edu/~boyd/SP.html, Nov 1994.

[18] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Rev.*, 38(1):49–95, 1996.

[19] G. Wilson. Factorization of the covariance generating function of a pure moving average process. *SIAM J. Numerical Analysis*, 6:1–8, 1969.

[20] S. P. Wu, S. Boyd, and L. Vandenberghe. FIR Filter design via semidefinite programming and spectral factorization. In *Proc. 35th IEEE Conference on Decision and Control, Kobe, Japan*, Dec 1996.

[21] S. P. Wu, S. Boyd, and L. Vandenberghe. FIR Filter design via spectral factorization and convex optimization. In Biswa Datta, editor, *Applied Computational Control, Signal and Communications*. Birkhauser, 1997.