Int. J. Appl. Math. Comput. Sci., 2001, Vol.11, No.5, 1173–1201

A HOMOTOPY APPROACH TO RATIONAL COVARIANCE EXTENSION WITH DEGREE CONSTRAINT[†]

PER ENQVIST[*]

The solutions to the Rational Covariance Extension Problem (RCEP) are parameterized by the spectral zeros. The rational filter with a specified numerator solving the RCEP can be determined from a known convex optimization problem. However, this optimization problem may become ill-conditioned for some parameter values. A modification of the optimization problem to avoid the ill-conditioning is proposed and the modified problem is solved efficiently by a continuation method.

Keywords: stochastic realization theory, rational covariance extension problem, ARMA model design, continuation method, optimization

1. Introduction

Given a *positive covariance sequence* $r_0, r_1, \ldots, r_n$, i.e. a sequence such that the Toeplitz matrix

$$\mathbf{R} \triangleq \begin{bmatrix} r_0 & r_1 & \ldots & r_n \\ r_1 & r_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ r_n & \ldots & r_1 & r_0 \end{bmatrix} \tag{1}$$

is positive definite, the Rational Covariance Extension Problem with degree constraint (RCEP) (Georgiou, 1983; Kalman, 1981) amounts to determining a spectral density

$$\Phi(e^{i\theta}) \triangleq \tilde{r}_0 + 2\sum_{k=1}^{\infty} \tilde{r}_k \cos k\theta \tag{2}$$

such that

$$\tilde{r}_k = r_k, \quad k = 0, 1, \ldots, n, \tag{3}$$

[†] This research was supported by grants from the Swedish Research Council for Engineering Sciences (TFR).

[*] Division of Optimization and Systems Theory, Royal Institute of Technology, Lindstedtsv. 25, SE 100 44 Stockholm, Sweden, e-mail: pere@math.kth.se

and such that Φ is rational of degree at most $2n$ and positive for all z on the unit circle. Then Φ is analytic in a neighborhood of the unit circle and has a Laurent series

$$\Phi(z) \triangleq \tilde{r}_0 + \sum_{k=1}^{\infty} \tilde{r}_k (z^k + z^{-k}) \tag{4}$$

there. We say that $\tilde{r}_0, \tilde{r}_1, \tilde{r}_2, \ldots$ is a rational covariance extension of $r_0, r_1, \ldots, r_n$ with degree constraint.

One particular solution to the RCEP is given by the *Maximum Entropy* (ME) solution. The ME solution can be determined from a linear system of equations that can be solved, in a number of operations proportional to n^2, using the *Levinson algorithm* (Porat, 1994). However, there are an infinite number of solutions to the RCEP, and the other solutions require nonlinear solution methods.

Georgiou (1987) conjectured that all such extensions are completely parameterized by the zeros of the numerator of the spectral density. More precisely, for any monic stable n-th order polynomial σ, there exists a unique stable n-th order polynomial a such that

$$\Phi(z) = \frac{\sigma(z)\sigma(z^{-1})}{a(z)a(z^{-1})} \tag{5}$$

defines an extension. By a stable polynomial we mean that all the roots are inside the unit circle, and hence it is a Schur polynomial. Existence was established by Georgiou (1987). The rest of the conjecture was later proved in (Byrnes *et al.*, 1995) as a corollary of a more general result also showing that the solution depends analytically on the covariance data and the choice of the polynomial σ. Since a finite covariance sequence $r_0, r_1, \ldots, r_n$ can be estimated from a finite number of data, and the polynomial σ can be chosen according to any preference or also estimated from a finite number of data, the RCEP provides means to estimate the spectral density from a finite number of data. The ME solution corresponds to the choice $\sigma(z) = z^n$, and this method can therefore only be used to identify a small subclass of spectral densities from a finite number of data.

An independent proof of the conjecture is given in (Byrnes *et al.*, 1999). This proof is constructive and provides means for determining the unique Schur polynomial $a(z)$ which, together with the given Schur polynomial $\sigma(z)$, defines a solution to the RCEP. In fact, the Schur polynomial $a(z)$ is given by the solution to a convex optimization problem. In order for the optimization problem to be convex, it is formulated in terms of the pseudo-polynomial $Q(z) = a(z)a(z^{-1})$. It is shown here that this formulation, although the best for analysis, is not the best for calculations. In fact, it is proposed here that the optimization problem be posed in the coefficients of the polynomial $a(z)$.

There are two main reasons for formulating the optimization problem directly in the a variables. First, an ill-conditioning is introduced through the choice of the coefficients of Q as variables. Using the filter variables a, the values of the objective function and its derivatives are finite as long as a has precisely degree n, and the curvature of the function is more uniform. The second reason is that the values of the

objective function and its derivatives can be determined without a spectral factorization of $Q(z)$. By avoiding the spectral factorization, the amount of calculations can be reduced considerably and numerical problems that may occur close to the boundary are also avoided.

However, the new optimization problem has some drawbacks too. It turns out that the modified problem is locally convex but in general not globally convex. Hence the optimization procedure has to be initiated close to the optimum to ensure convergence. In order to do this in practice, a continuation method is proposed. Since the geometry of the solutions to the optimization problem for varying parameter values is well-known (Byrnes *et al.*, 1995), it follows that there is a smooth trajectory from the maximum entropy solution to any particular solution with the same n first covariances. Using a predictor-corrector path following algorithm (Allgower and Georg, 1990; 1993), the solution to the optimization problem can be found. An algorithm based on a continuation method with an adaptive step length rule is proposed, and a convergence proof for the algorithm is provided.

It should be noted that the change of variables proposed here, from pseudo-polynomials to polynomials, can be applied to other problems for similar benefits, for example, to the problem given in (Byrnes *et al.*, 2001).

The outline of this paper is as follows. In Section 2, the original optimization problem is described. In Section 3, the new formulation of the optimization problem is derived. The optimality conditions of first and second order are compared with the original problem, and expressions for the derivatives of the new objective function are determined. In Section 4, a homotopy for the new optimization problem is introduced and the concept of continuation methods is described. In Section 5, an algorithm solving the new optimization problem is proposed. In Section 6, a convergence proof for the algorithm is given.

2. The Original Optimization Problem

A convex optimization formulation for finding the polynomial $a(z)$ in (5) for an arbitrary choice of the numerator polynomial $\sigma(z)$ was presented in (Byrnes *et al.*, 1999). It will be reviewed here for clarity. Then a new formulation will be derived through a change of variables.

The numerator polynomial $\sigma(z)$ in (5) defines a symmetric pseudo-polynomial

$$P(z) = p_0 + \frac{1}{2}p_1(z + z^{-1}) + \cdots + \frac{1}{2}p_n(z^n + z^{-n}), \tag{6}$$

by $P(z) \triangleq \sigma(z)\sigma(z^{-1})$. A second pseudo-polynomial,

$$Q(z) = q_0 + \frac{1}{2}q_1(z + z^{-1}) + \cdots + \frac{1}{2}q_n(z^n + z^{-n}), \tag{7}$$

corresponding to the polynomial $a(z)$ in (5), defined by $Q(z) \triangleq a(z)a(z^{-1})$, determines the variables of the optimization problem introduced in (Byrnes *et al.*, 1999):

$$(\mathcal{P}_q) \quad \left[\begin{array}{ll} \min\limits_{\mathbf{q}} & \phi(\mathbf{q}), \\ \text{s.t.} & \mathbf{q} \in \boldsymbol{\mathcal{D}}_n \end{array} \right], \tag{8}$$

where

$$\mathbf{q} \triangleq \left[\begin{array}{cccc} q_0 & q_1 & \dots & q_n \end{array} \right]^\top, \tag{9}$$

and the feasible region will be defined below. The objective function $\phi : \mathbb{R}^{n+1} \to \mathbb{R}$ is given by

$$\phi(\mathbf{q}) \triangleq \sum_{k=0}^{n} r_k q_k - \langle \log Q, P \rangle, \tag{10}$$

where $\langle \cdot, \cdot \rangle$ denotes the L_2 inner product

$$\langle a, b \rangle \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} a(e^{i\theta}) b(e^{-i\theta}) \, \mathrm{d}\theta. \tag{11}$$

With slight abuse of notation, the same inner product notation is used for vectors and matrices, in which case the inner product is determined componentwise. Using the notation

$$\begin{array}{llllll} \mathbf{r} & \triangleq & \left[\begin{array}{cccc} r_0 & r_1 & \dots & r_n \end{array} \right]^\top, & \mathbf{z} & \triangleq & \left[\begin{array}{cccc} z^n & z^{n-1} & \dots & 1 \end{array} \right]^\top, \\ \tilde{\mathbf{z}} & \triangleq & \left[\begin{array}{cccc} 1 & z & \dots & z^n \end{array} \right]^\top, & \tilde{\mathbf{z}}^* & \triangleq & \left[\begin{array}{cccc} 1 & z^{-1} & \dots & z^{-n} \end{array} \right]^\top, \end{array}$$

explicit expressions for the derivatives of ϕ are given by the following proposition.

Proposition 1. *The gradient of ϕ is given by*

$$\nabla \phi = \mathbf{r} - \left\langle \tilde{\mathbf{z}}, \frac{P}{Q} \right\rangle.$$

The Hessian of ϕ is given by

$$\mathbf{H}_{\mathbf{q}} \triangleq \nabla^2 \phi = \left\langle \frac{1}{2}(\tilde{\mathbf{z}} + \tilde{\mathbf{z}}^*) \frac{P}{Q^2}, \frac{1}{2}(\tilde{\mathbf{z}} + \tilde{\mathbf{z}}^*)^\top \right\rangle,$$

where the elements $h_{i,j} = \dfrac{\partial^2 \phi}{\partial q_i \partial q_j}$ are given by $h_{i,j} = (d_{i+j} + d_{|i-j|})/2$, and

$$d_k \triangleq \left\langle z^k, \frac{P}{Q^2} \right\rangle, \quad k = 0, 1, \dots, 2n.$$

The feasible region $\boldsymbol{\mathcal{D}}_n$ denotes the set of vectors $\mathbf{d} \in \mathbb{R}^{n+1}$ of coefficients corresponding to symmetric positive pseudo-polynomials of order at most n, namely,

$$\boldsymbol{\mathcal{D}}_n \triangleq \Big\{ \mathbf{q} \mid Q(z) = q_0 + \frac{1}{2} q_1 (z + z^{-1}) + \cdots + \frac{1}{2} q_n (z^n + z^{-n}), \\ Q(z) > 0, \, \forall z : |z| = 1 \Big\}. \quad (12)$$

Since $P(e^{i\theta}) = |\sigma(e^{i\theta})|^2$, it is clear that $\mathbf{p} \in \boldsymbol{\mathcal{D}}_n$, and therefore the objective function is strictly convex. Since the feasible region is convex, $(\mathcal{P}_q)$ is a convex optimization problem. If, in addition, we assume that the sequence $r_0, r_1, \ldots, r_n$ is a *positive covariance sequence*, then $(\mathcal{P}_q)$ has a unique solution in the open feasible region $\boldsymbol{\mathcal{D}}_n$. This follows from, among other things, the fact that the directional derivative is infinite on the boundary. In fact, the second term acts as a barrier and pushes the solution to the interior of $\overline{\boldsymbol{\mathcal{D}}}_n$ (for details, see (Byrnes *et al.*, 1999)). The *stationarity condition*, $\nabla \phi = 0$, implies that the solution to the optimization problem $(\mathcal{P}_q)$ satisfies $\langle \tilde{\mathbf{z}}, P/Q \rangle = \mathbf{r}$, i.e.

$$\frac{P(z)}{Q(z)} = \tilde{r}_0 + \sum_{k=0}^{\infty} \tilde{r}_k (z^k + z^{-k}), \quad (13)$$

where $\tilde{r}_k = r_k$ for $k = 0, 1, \ldots, n$. Consequently, this is the unique solution to the RCEP corresponding to the numerator polynomial $\sigma(z)$, and the corresponding denominator polynomial $a(z)$ in (5) can be obtained by means of spectral factorization of $Q(z)$.

A difficulty in solving the problem $(\mathcal{P}_q)$ is that the positivity constraints on Q should hold at an infinite number of points. This can be dealt with by using a Linear Matrix Inequality (LMI) formulation of the constraints, see, e.g., (Wu *et al.*, 1997). However, since we know that the constraints will not be active at the optimal point, the most computationally efficient way to solve this problem is to discretize the constraints.

Another complication is caused by the barrier-like term in the objective function. As in barrier function methods, an ill-conditioning of the Hessian may occur close to the boundary (Nash and Sofer, 1996). But, also as in barrier methods, there are ways to get around the ill-conditioning in this problem as well. The optimization problem $(\mathcal{P}_q)$ is formulated in the parameters q_k of the spectrum of the problem, while the optimization problem presented in this paper is formulated directly in the filter parameters a.

Example 1. To illustrate the sensitivity in the coefficients of the pseudo-polynomial, a polynomial $a(z)$ of degree 3 with all zeros at 0.95 is studied. Let $Q(z)$ be the corresponding pseudo-polynomial, and $Q_\epsilon(z) = Q(z) + \epsilon(Q(z) - q_0)$ a perturbed version. Then if $a_\epsilon(z)$ is a spectral factor of $Q_\epsilon(z)$, and, for example, $\epsilon = -0.01$, then the coefficients of a_ϵ change as much as 100% and the zeros shift as depicted in Fig. 1.

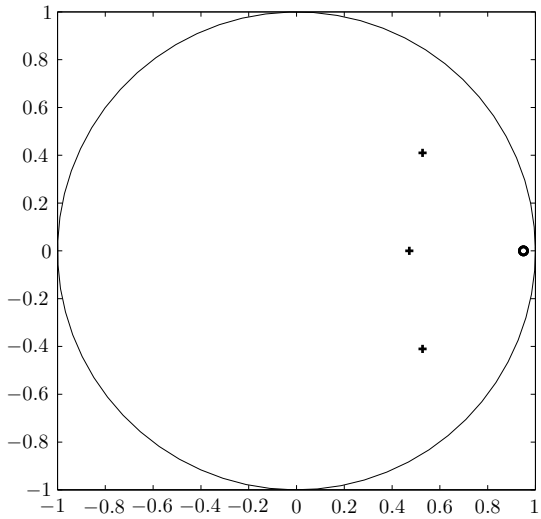

Fig. 1. Location of the zeros before (o) and after (+) a small disturbance.

For the optimization process to work properly, the condition number of the Hessian should not be too large. The problem of ill-conditioning occurs mostly close to the boundary of the feasible region. In many applications, like speech processing, the optimum will occur precisely in this region. ♦

Example 2. This example illustrates how fast the Hessian gets ill-conditioned close to the boundary of the feasible region. The *condition number* of a matrix is defined by $\kappa(\mathbf{H}) \triangleq \|\mathbf{H}\|\,\|\mathbf{H}^{-1}\|$, and if the matrix 2-norm is used, $\kappa(\mathbf{H})^2 = \max_{\|\mathbf{d}\|=1} \mathbf{d}^\top\mathbf{H}\mathbf{d} / \min_{\|\mathbf{d}\|=1} \mathbf{d}^\top\mathbf{H}\mathbf{d}$.

Assuming that $a(z) = (z - 1 + \varepsilon)\nu(z)$, we obtain

$$\begin{aligned} Q(e^{i\theta}) &= \big(\varepsilon^2 + 2(1-\varepsilon)(1-\cos\theta)\big)|\nu(e^{i\theta})|^2 \\ &\le \big(\varepsilon^2 + (1-\varepsilon)\theta^2\big)|\nu(e^{i\theta})|^2. \end{aligned} \tag{14}$$

The second derivative will now be determined in two different directions, approximating the eigenvectors of the maximum and minimum eigenvalues. Let the Hessian of $\phi(\mathbf{q})$ be denoted by $\mathbf{H}_\mathbf{q}$ as in Proposition 1, where $\mathbf{H}_\mathbf{q} = \langle \frac{1}{2}(\tilde{\mathbf{z}} + \tilde{\mathbf{z}}^*)\frac{P}{Q^2}, \frac{1}{2}(\tilde{\mathbf{z}} + \tilde{\mathbf{z}}^*)^\top \rangle$.

First, a lower bound for $\max_{\|\mathbf{d}\|=1} \mathbf{d}^\top\mathbf{H}_\mathbf{q}\mathbf{d}$ is determined. Let, e.g.,

$$\check{D}(z) = 1 + \sum_{k=1}^{n} (z^k + z^{-k}),$$

$\check{\mathbf{d}}$ being the vector of coefficients of $\check{D}$, and define $M_\delta \triangleq \min_{0\le\theta\le\delta} P(e^{i\theta})\frac{\check{D}(e^{i\theta})^2}{|\nu(e^{i\theta})|^2}$.

Then

$$\begin{aligned}\check{\mathbf{d}}^\top \mathbf{H_q} \check{\mathbf{d}} = \left\langle \frac{\check{D}}{Q^2}, P\check{D} \right\rangle &\geq \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{P(e^{i\theta})}{(\varepsilon^2+(1-\varepsilon)\theta^2)^2} \frac{\check{D}(e^{i\theta})^2}{|\nu(e^{i\theta})|^2}\,\mathrm{d}\theta \\ &\geq \frac{M_\varepsilon}{\varepsilon^4}\int_{-\varepsilon}^{\varepsilon} \frac{1}{(1+\frac{1-\varepsilon}{\varepsilon^2}\theta^2)^2}\,\mathrm{d}\theta \\ &\geq \frac{M_\varepsilon}{2\varepsilon^3}. \end{aligned} \tag{15}$$

Secondly, an upper bound for $\min_{\|\mathbf{d}\|=1} \mathbf{d}^\top \mathbf{H_q} \mathbf{d}$ is determined. Since $Q(1) = \varepsilon^2 \nu(1)^2$ tends to zero, let $\hat{D}(z) = Q(z)$ in order to cancel this term.

$$\hat{\mathbf{d}}^\top \mathbf{H_q} \hat{\mathbf{d}} = \left\langle \frac{\hat{D}}{Q^2}, P\hat{D} \right\rangle = \langle 1, P\rangle = p_0. \tag{16}$$

The condition number of the Hessian will thus increase at least as $1/\varepsilon^{3/2}$ as $\varepsilon \to 0$, since M_ε is an increasing function.

The condition number of the Hessian will thus increase close to the boundary, and if the optimal solution is located close to the boundary, it will be sensitive to small perturbations. It is important to note that the ill-conditioning mentioned here is due to the solution procedure and not to an ill-conditioning of the covariance extension problem *per se*. ⧫

3. A New Formulation of the Optimization Problem

The optimization problem $(\mathcal{P}_q)$ will now be reformulated by a change of variables. $(\mathcal{P}_q)$ was formulated in $\mathbf{q} = \begin{bmatrix} q_0 & q_1 & \dots & q_n \end{bmatrix}^\top$, the vector of coefficients of $Q(z)$. As new variables the elements of the vector

$$\mathbf{a} \triangleq \begin{bmatrix} a_0 & a_1 & \dots & a_n \end{bmatrix}^\top \tag{17}$$

are used, i.e. the coefficients of the stable spectral factor $a(z)$ corresponding to $Q(z)$. This change of variables is well defined, as will be seen next.

Let $\mathcal{S}_n$ and $\mathcal{D}_n$ define the Schur region and the region of positive pseudo-polynomials, respectively,

$$\begin{aligned}\mathcal{S}_n \triangleq \Big\{ a \mid a(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n, a_0 > 0, \\ a \text{ is a real stable polynomial} \Big\}, \end{aligned} \tag{18}$$

$$\begin{aligned}\mathcal{D}_n \triangleq \Big\{ Q \mid Q(z) = q_0 + \frac{1}{2}q_1(z+z^{-1}) + \cdots + \frac{1}{2}q_n(z^n+z^{-n}), \\ Q(z) > 0,\ \forall z : |z| = 1 \Big\}, \end{aligned} \tag{19}$$

and consider the map $T : \mathcal{S}_n \to \mathcal{D}_n$, defined by $T(a) \triangleq a(z)a(z^{-1})$. Since the coefficients of $a(z)$ are real, $(Ta)(z) = |a(z)|^2 > 0$ for all z on the unit circle. In order for the change of variables to be well defined, the map T should be one-to-one, continuously differentiable and having a nonvanishing Jacobian for all $a(z) \in \mathcal{S}_n$. This will be shown next.

Given a pseudo-polynomial $Q \in \mathcal{D}_n$, the Fejérs Theorem (Caines, 1987) ensures that there is a unique polynomial $a(z)$ such that $Q(z) = a(z)a(z^{-1})$, where $a(z) \in \mathcal{S}_n$. Therefore, there is a one-to-one correspondence between pseudo-polynomials positive on the unit circle and Schur polynomials. The inverse operation, corresponding to T^{-1}, of determining $a(z)$ from $Q(z)$ is called *spectral factorization*. There are a number of methods for spectral factorization of a positive pseudo-polynomial (Goodman *et al.*, 1997), e.g. the Bauer method (Bauer, 1955) and the method of Wilson (Wilson, 1969). The factorization usually becomes more difficult and more inaccurate numerically as the zeros of $a(z)$ get closer to the unit circle, and the mapping is actually singular on the unit circle.

It remains to show that the map T has an everywhere nonvanishing Jacobian. Defining the map $S : \mathcal{S}_n \times \mathcal{S}_n \to \mathcal{D}_n$ by

$$S(a)b \triangleq a(z)b(z^{-1}) + a(z^{-1})b(z), \tag{20}$$

we have that $T(a) = \frac{1}{2}S(a)a$. The Gâteaux differential at a in direction p is

$$\begin{aligned}\delta_p T &\triangleq \lim_{\varepsilon \to 0} \frac{\frac{1}{2}S(a+\varepsilon p)(a+\varepsilon p) - \frac{1}{2}S(a)a}{\varepsilon} \\ &= a(z)p(z^{-1}) + a(z^{-1})p(z) = S(a)p. \end{aligned} \tag{21}$$

In fact, $S(a)$ is the Fréchet derivative of the map T. It is well-known, see, e.g., (Byrnes *et al.*, 1995; Goodman *et al.*, 1997), that if a is a Schur polynomial, then $S(a)$ is nonsingular.

Corresponding to the map T, there is a map $\boldsymbol{\pi} : \boldsymbol{\mathcal{S}}_n \to \boldsymbol{\mathcal{D}}_n$, where

$$\boldsymbol{\mathcal{S}}_n \triangleq \left\{ \mathbf{a} \in \mathbb{R}^{n+1} \mid a_0 > 0, a(z) = \mathbf{a}^\top \mathbf{z} \text{ is a stable polynomial} \right\}, \tag{22}$$

defined as follows. Given an $\mathbf{a} \in \boldsymbol{\mathcal{S}}_n$, $T(\mathbf{a}^\top \mathbf{z})$ determines a pseudo-polynomial Q that can be parameterized as in (7), which in turn determines a vector $\mathbf{q}$ as in (9). The map $\boldsymbol{\pi}$ defined in this way connects the original and the new objective functions ϕ and f as depicted in Fig. 2.

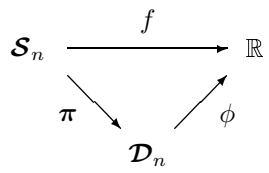

Fig. 2. Commutative diagram describing the correspondence between f and ϕ.

The objective function can then be expressed in $\mathbf{a}$ using the function $f \triangleq \phi \circ \boldsymbol{\pi}$. From the discussion above it is clear that this change of variables is well defined.

To determine an explicit expression for f, consider first the linear part of the objective function. In the L_2 inner product, we have

$$\sum_{k=0}^{n} r_k q_k = \langle R, Q \rangle, \tag{23}$$

where

$$R(z) \triangleq r_0 + r_1(z + z^{-1}) + \cdots + r_n(z^n + z^{-n}), \tag{24}$$

and the sum can now be written as $\langle R, aa^* \rangle = \langle Ra, a \rangle$, where $a^*(z) \triangleq a(z^{-1})$. A matrix representation is determined by observing that $a(z) = \mathbf{z}^\top \mathbf{a}$, then it follows that

$$\langle Ra, a \rangle = \mathbf{a}^\top \langle \mathbf{z}R, \mathbf{z}^\top \rangle \mathbf{a} = \mathbf{a}^\top \langle R, \mathbf{z}^* \mathbf{z}^\top \rangle \mathbf{a} = \mathbf{a}^\top \mathbf{R} \mathbf{a}, \tag{25}$$

where $\mathbf{R}$ is defined in (1). The second term in the objective function can be translated in a similar way. Using the same factorization, we have $\log Q(z) = \log a(z) a(z^{-1}) = \log |a(z)|^2$ on the unit circle, and the objective function can be written as

$$f(\mathbf{a}) = \mathbf{a}^\top \mathbf{R} \mathbf{a} - 2 \langle \log |a|, P \rangle. \tag{26}$$

Then the new optimization problem in $\mathbf{a}$ can be stated as

$$(\mathcal{P}_a) \quad \left[\begin{array}{ll} \min\limits_{\mathbf{a}} & f(\mathbf{a}) \\ \text{s.t.} & \mathbf{a} \in \boldsymbol{\mathcal{S}}_n \end{array} \right]. \tag{27}$$

In contrast to $(\mathcal{P}_q)$, the problem $(\mathcal{P}_a)$ is a nonconvex optimization problem. The Schur region $\boldsymbol{\mathcal{S}}_n$ is a cone with a cross section region which is nonconvex for $n \geq 3$, as depicted in Fig. 3 for the case $n = 3$. A number of the nice properties of $(\mathcal{P}_q)$ are, however, inherited by $(\mathcal{P}_a)$.

Solving the new optimization problem by an iterative method requires determining function values or derivatives of the objective function. There are two terms in the objective function (26), the first of which is just a positive definite quadratic form. Using the symmetry of the pseudo-polynomial P, the second term can be written as

$$\psi(a) \triangleq \langle \log |a|^2, P \rangle = \sum_{k=0}^{n} p_k \langle \log |a|^2, z^k \rangle = \sum_{k=0}^{n} p_k c_k, \tag{28}$$

where $c_k \triangleq \langle \log |a|^2, z^k \rangle$ are the so-called *cepstral parameters* of a Moving Average (MA) filter defined by $a(z)$. These parameters can be determined using the recursion formula

$$c_0 = 2 \log a_0, \quad c_k = \frac{a_k}{a_0} - \sum_{j=1}^{k-1} \frac{j}{k} \frac{a_{k-j}}{a_0} c_j, \quad k = 1, \ldots, n, \tag{28'}$$

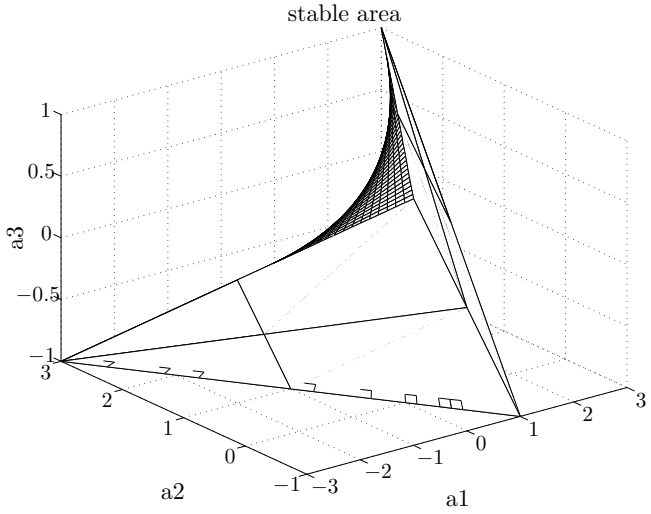

Fig. 3. The region of stable monic ($a_0 = 1$) polynomials of order 3.

which is a minor modification of the formula given in (Markel and Gray, 1976) for Auto Regressive (AR) filters.

Fixing $a_0 \neq 0$ to be constant, (28′) shows that ψ is a polynomial in $a_1, a_2, \ldots, a_n$. Therefore the gradient $\nabla\psi$ remains bounded as $\mathbf{a}$ tends to the boundary of $\boldsymbol{\mathcal{S}}_n$, in sharp contrast to the situation regarding ϕ. This is the basic reason why the optimization problem $\mathcal{P}_a$ is better behaved than $\mathcal{P}_q$, especially for minima close to the boundary.

As a simple but important example let us first consider the maximum entropy solution, i.e. the special case of the problem $(\mathcal{P}_a)$ corresponding to $\sigma(z) = z^n$. Then $P(z) = 1$, and hence $\psi(\mathbf{a}) = 2\log a_0$. Consequently, the objective function (26) becomes

$$f_0(\mathbf{a}) \stackrel{\triangle}{=} \mathbf{a}^\top \mathbf{R}\mathbf{a} - 2\log a_0. \tag{29}$$

Since the Toeplitz matrix $\mathbf{R}$ is positive definite, f_0 is strictly convex. Hence there is at most one minimum. To determine this possible minimum, set the gradient equal to zero to obtain

$$\frac{1}{2}\mathbf{g}_0(a) \stackrel{\triangle}{=} \nabla f_0 = \begin{bmatrix} r_0 & r_1 & \ldots & r_n \\ r_1 & r_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ r_n & \ldots & r_1 & r_0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} - \begin{bmatrix} 1/a_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0. \tag{30}$$

That is, defining $\varphi_{ni} \triangleq a_i/a_0$ for $i = 0, 1, \ldots, n$, we get

$$\begin{bmatrix} r_0 & r_1 & \ldots & r_{n-1} \\ r_1 & r_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ r_{n-1} & \ldots & r_1 & r_0 \end{bmatrix} \begin{bmatrix} \varphi_{n1} \\ \varphi_{n2} \\ \vdots \\ \varphi_{nn} \end{bmatrix} = - \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}, \tag{31}$$

and

$$\begin{bmatrix} r_0 & r_1 & \ldots & r_n \end{bmatrix} \begin{bmatrix} 1 & \varphi_{n1} & \ldots & \varphi_{nn} \end{bmatrix}^\top = \frac{1}{a_0^2}. \tag{32}$$

Hence, we obtain the well-known *normal equations*, which can be solved quickly using the Levinson algorithm (Porat, 1994). Since $\varphi_n(z) = z^n + \varphi_{n1} z^{n-1} + \cdots + \varphi_{nn}$ is the n-th Szegö polynomial, which is a stable polynomial, $\varphi_n \in \mathcal{S}_n$, and hence there is a unique minimum.

In the general case when P is no longer constant, the situation is more complicated. First we shall need expressions for the gradient and Hessian.

Proposition 2. *The gradient of f is given by*

$$\mathbf{g}(\mathbf{a}) \triangleq \nabla f = 2\big(\mathbf{R} - \mathbf{R}(\mathbf{a})\big)\mathbf{a}, \tag{33}$$

where $\mathbf{R}(\mathbf{a})$ is the $n \times n$ Toeplitz matrix of covariances of the spectral density $P/|a|^2$.

Proof. Noting that $a^*(z) = \mathbf{z}^{*\top}\mathbf{a}$, the gradient of f is derived as follows:

$$\nabla f = 2\mathbf{R}\mathbf{a} - 2\Big\langle \mathbf{z}\frac{1}{a}, P \Big\rangle = 2\mathbf{R}\mathbf{a} - 2\Big\langle \frac{1}{|a|^2}\mathbf{z}\mathbf{z}^{*\top}, P \Big\rangle \mathbf{a} = 2\big(\mathbf{R} - \mathbf{R}(\mathbf{a})\big)\mathbf{a}. \tag{34}$$

■

The gradient of f is thus given by the difference between the Toeplitz matrices of the desired covariances and the covariances of the filter corresponding to the current iteration point, multiplied by the denominator coefficients of this filter. This can be compared with the gradient of ϕ given in Proposition 1, which is the difference of the desired covariances and the covariances of the filter corresponding to the current iteration point.

Proposition 3. *The Hessian of f is given by*

$$\mathbf{H}_\mathbf{a} \triangleq \nabla^2 f = 2\mathbf{R} + 2\mathbf{\Xi}, \tag{35}$$

where the (k, j)-th element, $\xi_{k,j}$, of the $(n+1) \times (n+1)$ matrix $\mathbf{\Xi}$ satisfies $\xi_{k,j} = \tilde{\xi}_{k+j-2}$, and is given by

$$\tilde{\xi}_m = \begin{cases} \displaystyle\sum_{j=0}^{n}\sum_{k=0}^{n} a_j a_k \zeta_{j+k-m}, & m = 0, 1, \ldots, n, \\ 0, & m > n, \end{cases} \tag{36}$$

where ζ_k *are the covariances of the filter* σ/a^2*, i.e.*

$$\frac{P(z)}{|a(z)|^4} = \sum_{k=-\infty}^{\infty} \zeta_k z^{-k}.$$

Proof. The second-order derivative of f is given by

$$\nabla^2 f = 2\mathbf{R} + 2\Big\langle \mathbf{z}\mathbf{z}^\top \frac{1}{a^2}, P \Big\rangle. \tag{37}$$

The Hessian thus takes the form of a sum of a Toeplitz and a Hankel matrix as stated in (35). It remains to show that the elements

$$\tilde{\xi}_m = \Big\langle \frac{z^{2n-m}}{a^2}, P \Big\rangle = \Big\langle \frac{z^m}{a_*^2}, P \Big\rangle, \quad m = 0, 1, \ldots, 2n, \tag{38}$$

where $a_*(z) \triangleq z^n a(z^{-1})$, are determined by (36).

The elements $\tilde{\xi}_m$ can be determined using calculus of residues, and since $1/a_*^2$ is expanded around the origin with positive powers of z , and the only terms with inverse powers of z come from P, it follows that $\tilde{\xi}_m = 0$ for $m > n$. Therefore the Hessian is determined by only $n+1$ terms.

Another way to determine $\tilde{\xi}_m$ is to multiply the numerator and denominator in (38) by $(a^*)^2$,

$$\begin{aligned}\tilde{\xi}_m &= \Big\langle z^{2n-m} \frac{(a^*)^2}{(aa^*)^2}, P \Big\rangle = \sum_{j=0}^{n}\sum_{k=0}^{n} a_j a_k \Big\langle z^{2n-m} z^{-n+j} z^{-n+k}, \sum_{l=-\infty}^{\infty} \zeta_l z^{-l} \Big\rangle \\ &= \sum_{j=0}^{n}\sum_{k=0}^{n} a_j a_k \zeta_{j+k-m},\end{aligned}$$

and determine the covariances $\zeta_0, \zeta_1, \ldots, \zeta_{2n-m}$ of the filter σ^2/a^2. This proves the remaining part of (36). ∎

Clearly, we would like to find a stationary point for f, but it remains to demonstrate that such a point exists and that it is unique.

Proposition 4. *The optimization problem* $\mathcal{P}_a$ *has a unique stationary point* $\hat{\mathbf{a}}$*. Moreover, if* $\hat{\mathbf{q}}$ *is the unique stationary point of* $\mathcal{P}_q$*, then* $\boldsymbol{\pi}(\hat{\mathbf{a}}) = \hat{\mathbf{q}}$*.*

To prove this, we can use the fact that the original optimization problem has a unique optimum, as proven in (Byrnes *et al.*, 1999). To this end, we first establish the relation between the gradients and Hessians of the two optimization problems $\mathcal{P}_q$ and $\mathcal{P}_a$.

Proposition 5. *The gradient* ∇f *is related to* $\nabla\phi$ *as*

$$\nabla f = (\nabla \boldsymbol{\pi})^\top \nabla \phi, \tag{39}$$

where $\nabla\boldsymbol{\pi}$ *is given by*

$$\nabla\boldsymbol{\pi} = \begin{bmatrix} \dfrac{\partial q_0}{\partial a_0} & \dfrac{\partial q_0}{\partial a_1} & \cdots & \dfrac{\partial q_0}{\partial a_n} \\ \dfrac{\partial q_1}{\partial a_0} & \dfrac{\partial q_1}{\partial a_1} & \cdots & \dfrac{\partial q_1}{\partial a_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial q_n}{\partial a_0} & \dfrac{\partial q_n}{\partial a_1} & \cdots & \dfrac{\partial q_n}{\partial a_n} \end{bmatrix}$$

$$= \begin{bmatrix} a_0 & a_1 & \dots & a_n \\ 2a_1 & \dots & 2a_n & 0 \\ \vdots & \iddots & \iddots & \vdots \\ 2a_n & 0 & \dots & 0 \end{bmatrix} + \begin{bmatrix} a_0 & a_1 & \dots & a_n \\ 0 & 2a_0 & \dots & 2a_{n-1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 2a_0 \end{bmatrix}. \quad (40)$$

Proof. In order to use the chain rule, the derivatives of $\boldsymbol{\pi}$ are needed. First, we determine the elements $\boldsymbol{\pi}_k$. Let $a(z)$ be an n-th order real polynomial

$$a(z) \stackrel{\triangle}{=} a_0 z^n + a_1 z^{n-1} + \cdots + a_n, \quad a_0 > 0, \quad (41)$$

and $Q(z)$ the pseudo-polynomial determined from $a(z)$ by

$$Q(z) = Ta = a(z)a(z^{-1}) = \sum_{m=-n}^{n} z^m \sum_{k=0}^{n-|m|} a_k a_{k+|m|}. \quad (42)$$

By identifying coefficients in (7) and (42), we get

$$q_0 = \boldsymbol{\pi}_0(\mathbf{a}) = \sum_{k=0}^{n} a_k^2, \quad q_m = \boldsymbol{\pi}_m(\mathbf{a}) = 2\sum_{k=0}^{n-m} a_k a_{k+m}, \quad m = 1, \dots, n. \quad (43)$$

Differentiating $\phi(\boldsymbol{\pi}(\mathbf{a}))$ using the chain rule, we obtain

$$\frac{\partial}{\partial a_k}(\phi \circ \boldsymbol{\pi}) = \sum_{j=0}^{n} \frac{\partial \phi}{\partial q_j} \frac{\partial \boldsymbol{\pi}_j}{\partial a_k}, \quad k = 0, \dots, n, \quad (44)$$

where differentiation of (43), using the convention $a_j = 0$ if $j < 0$ or $j > n$, gives

$$\frac{\partial \boldsymbol{\pi}_0}{\partial a_l} = 2a_l, \quad \frac{\partial \boldsymbol{\pi}_k}{\partial a_l} = 2(a_{l+k} + a_{l-k}), \quad k = 1, \dots, n. \quad (45)$$

Equations (39) and (40) are just the matrix form of (44) and (45). ■

Proposition 6. *The Hessians of* f *and* ϕ *are related as*

$$\mathbf{H_a} = (\nabla\boldsymbol{\pi})^\top \mathbf{H_q} (\nabla\boldsymbol{\pi}) + 2\mathbf{T}, \quad (46)$$

where $\mathbf{T}$ *is the Toeplitz matrix of* $\nabla\phi$*. Moreover,*

$$\mathbf{T} = \mathbf{R} - \mathbf{R}(\mathbf{a}), \tag{47}$$

where $\mathbf{R}(\mathbf{a})$ *is defined as in Proposition 2.*

Proof. Repeated use of the chain rule determines the second-order derivatives. Differentiate (44) again to get

$$\begin{aligned}\frac{\partial^2 f}{\partial a_l \partial a_k} &= \sum_{j=0}^{n} \left(\frac{\partial}{\partial a_l} \left(\frac{\partial \phi}{\partial q_j} \right) \frac{\partial \boldsymbol{\pi}_j}{\partial a_k} + \frac{\partial \phi}{\partial q_j} \frac{\partial^2 \boldsymbol{\pi}_j}{\partial a_l \partial a_k} \right) \\ &= \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{\partial \boldsymbol{\pi}_i}{\partial a_l} \frac{\partial^2 \phi}{\partial q_i \partial q_j} \frac{\partial \boldsymbol{\pi}_j}{\partial a_k} + \sum_{j=0}^{n} \frac{\partial \phi}{\partial q_j} \frac{\partial^2 \boldsymbol{\pi}_j}{\partial a_l \partial a_k},\end{aligned} \tag{48}$$

for all $k = 0, \dots, n$ and $l = 0, \dots, n$. Using (43) again and the Kronecker delta function, we have

$$\frac{\partial^2 \boldsymbol{\pi}_0}{\partial a_l \partial a_k} = 2\delta_{k-l}, \quad \frac{\partial^2 \boldsymbol{\pi}_m}{\partial a_l \partial a_k} = 2(\delta_{k-l+m} + \delta_{k-l-m}) \tag{49}$$

for $m = 1, \dots, n,$ and hence the second term in (48) becomes

$$\sum_{j=0}^{n} \frac{\partial \phi}{\partial q_j} \frac{\partial^2 \boldsymbol{\pi}_j}{\partial a_l \partial a_k} = 2 \left(\frac{\partial \phi}{\partial q_{l-k}} + \frac{\partial \phi}{\partial q_{k-l}} \right) = 2 \frac{\partial \phi}{\partial q_{|k-l|}}, \tag{50}$$

if we set $\partial\phi/\partial q_j = 0$ whenever $j < 0$ or $j > n$. The Toeplitz matrix $\mathbf{T}$ is determined using (50) and Proposition 1. ■

We are now in a position to prove Proposition 4.

Proof. We will show that f has a unique stationary point. From (39) it is clear that $\nabla f = 0$ if and only if $\nabla\phi = 0$ as long as $\nabla\boldsymbol{\pi}$ is nonsingular. It was shown in (21) that the Fréchet derivative of T is given by $S(a)$. Since $\nabla\boldsymbol{\pi}$ is a matrix representation of $S(a)$, $\nabla\boldsymbol{\pi}$ must be nonsingular.

It was shown in (Byrnes *et al.*, 1999) that $\mathcal{P}_q$ has exactly one stationary point $\hat{\mathbf{q}}$. Since $\boldsymbol{\pi}$ is bijective, there is exactly one $\hat{\mathbf{a}} \in \boldsymbol{\mathcal{S}}_n$ such that $\boldsymbol{\pi}(\hat{\mathbf{a}}) = \hat{\mathbf{q}}$ and $\nabla f(\hat{\mathbf{a}}) = 0$. Consequently, the stable spectral factor $\hat{\mathbf{a}}$ of $\hat{\mathbf{q}}$ is the only stationary point in $\boldsymbol{\mathcal{S}}_n$. ■

Unlike ϕ, the objective function $f(\mathbf{a})$ is not convex on the whole feasible region $\boldsymbol{\mathcal{S}}_n$, but we will show here that it is convex in a neighborhood of the optimum. This is suggested by the following example.

Example 3. Let $a(z) = a_0 z^2 + a_1 z + a_2$ be a stable polynomial and $P(z) = p_0 + p_1(z + z^{-1}) + p_2(z^2 + z^{-2})$ be a pseudo-polynomial strictly positive on the unit circle.

The second term ψ of the objective function f is determined here, using (28) and (28′), as

$$\psi(a) = 2 \left(p_0 \log a_0 + p_1 \frac{a_1}{a_0} + p_2 \left(\frac{a_2}{a_0} - \frac{1}{2} \left(\frac{a_1}{a_0} \right)^2 \right) \right). \tag{51}$$

Then

$$f(\mathbf{a}) = \mathbf{a}^\top \mathbf{R}\mathbf{a} - \psi(a)$$
$$= \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}^\top \begin{bmatrix} r_0 & r_1 & r_2 \\ r_1 & r_0 & r_1 \\ r_2 & r_1 & r_0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} - \begin{bmatrix} p_0 \\ p_1 \\ p_2 \end{bmatrix}^\top \begin{bmatrix} 2\log a_0 \\ 2a_1/a_0 \\ 2a_2/a_0 \end{bmatrix} + p_2 \frac{a_1^2}{a_0^2}. \quad (52)$$

It is clear that this function is strictly convex for large a_0, and it is also clear that the function is only locally convex. ♦

Proposition 7. *The function f is locally strictly convex in a neighborhood of the stationary point $\hat{\mathbf{a}}$.*

Proof. According to Proposition 6, the Hessians of f and ϕ are related as $\mathbf{H_a} = (\nabla\boldsymbol{\pi})^\top \mathbf{H_q} (\nabla\boldsymbol{\pi}) + 2\mathbf{T}$. The first term $(\nabla\boldsymbol{\pi})^\top \mathbf{H_q} (\nabla\boldsymbol{\pi})$ is positive definite, since $\mathbf{H_q}$ is positive definite and the map $\nabla\boldsymbol{\pi}$ is nonsingular. Therefore, $\mathbf{H_a}$ is positive definite as long as any negative eigenvalue of $\mathbf{T}$ is sufficiently small.

We know that $\nabla\phi = 0$ at the optimum, so at the optimum we have $\mathbf{T} = 0$ in (46). Then the Hessian of ϕ is positive definite. Since the elements of T are continuous functions of $\mathbf{a}$, it also follows that $\mathbf{H_a}$ is positive definite in a neighborhood of the optimum. The objective function f is thus locally convex around the optimum. ■

To further analyze the region where f is convex, we study the matrix $\mathbf{T}$. According to Proposition 5, $\mathbf{T} = \mathbf{R} - \mathbf{R}(\mathbf{a})$. From this we can also see that the Hessian is positive definite if $\|\mathbf{a}\|$ is large, since $\mathbf{R}(k\mathbf{a}) = (1/k^2)\mathbf{R}(\mathbf{a})$, and this implies that the components of $\mathbf{R}(k\mathbf{a})$ are small if k is large. Practically, this means that an iterative optimization procedure applied to $(\mathcal{P}_a)$ should start with an $a^{(0)}$ of large norm and approach the optimum from the locally convex region.

Most of the problems with the optimization problem $(\mathcal{P}_q)$ occur close to the boundary of the feasible region. In (Byrnes *et al.*, 1999) it was shown that the directional derivative of $\phi(\mathbf{q})$ pointing out of the feasible region tends to infinity as $\mathbf{q}$ tends to the boundary. This property repels the optimum of the problem $(\mathcal{P}_q)$ from the boundary, which is really important for that problem. Thanks to this property, the solution will correspond to a strictly stable filter, and the interpolation conditions will be satisfied at the optimum.

The functional $f(\mathbf{a})$ does not have the same property, which is illustrated by the example below.

Example 4. Consider the filter

$$w(z) = \frac{z+\sigma}{z+a}$$

and the derivatives

$$\nabla f = 2\mathbf{R}\begin{bmatrix} 1 \\ a \end{bmatrix} - 2\begin{bmatrix} 1+\sigma^2-\sigma a \\ \sigma \end{bmatrix}, \qquad \nabla\phi = \mathbf{r} - \begin{bmatrix} 1+\dfrac{(\sigma-a)^2}{(1-a)^2} \\ \dfrac{(\sigma-a)(1-a\sigma)}{1-a^2} \end{bmatrix}.$$

It is clear that ∇f is bounded as a tends to $+1$ or -1, whereas $\nabla\phi$ is unbounded if $|\sigma| \neq 1$ and a tends to $+1$ or -1.

This also follows from observing (28′) and noting that, for a constant $a_0 \neq 0$, $\psi(a)$ is a polynomial in $a_1, \ldots, a_n$. Therefore, the partial derivatives exist and are bounded for all $\mathbf{a} \in \boldsymbol{\mathcal{S}}_n$. However, since the two optimization problems are intimately connected, the problem $(\mathcal{P}_a)$ inherits the property that the solution will be at an interior point by Proposition 4. By Proposition 5, the gradient $\nabla\phi$ and ∇f are related by $\nabla\boldsymbol{\pi}$. Since for bounded a_0 the derivative of f is bounded at the boundary, the infinite derivative of ϕ at the boundary is cancelled by the singularity of $\nabla\boldsymbol{\pi}$ at the boundary.

Similarly, the ill-conditioning of the Hessian is also related to the singularity of $\nabla\boldsymbol{\pi}$ at the boundary. As seen in Example 2, the condition number of $\mathbf{H_q}$ increases rapidly as $\mathbf{q}$ tends to the boundary. By Proposition 6, $\nabla\boldsymbol{\pi}$ connects the Hessians of ϕ and f, and since $\nabla\boldsymbol{\pi}$ is singular at the boundary, the condition number changes drastically close to the boundary. Therefore, the optimization formulated directly in the $a(z)$ parameters is preferred, although it is not convex. ♦

To illustrate that there is a difference in the condition numbers of the Hessians $\mathbf{H_q}$ and $\mathbf{H_a}$, a low dimensional example is considered next.

Example 5. Let $r_0 = 1, r_1 = 0.99$ and $r_2 = 0.99$. Also, let $\sigma(z) = (z-0.8)(z+0.8) = z^2-0.64$. With these parameters, the objective function is depicted in Fig. 4 for monic polynomials $a(z)$. At the optimum, where the roots of a are 0.9996 and -0.3930, we have the condition number $\kappa(\mathbf{H_a}) = 64.6$. If we determine the Hessian corresponding to the objective function formulated in the q parameters, we have $\kappa(\mathbf{H_q}) = 2.5 \cdot 10^9$.

Next, the singularity of $\nabla\boldsymbol{\pi}$ at the boundary of the feasible region is studied. We noted above that $\nabla T = S(a)$, and it is well-known that $S(a)$ is singular if $a(z)$ has roots on the unit circle. A real polynomial $a(z)$ with roots on the unit circle can be written in at least one of the following three forms: $a^{(1)}(z) = (z+1)\nu(z)$, $a^{(2)}(z) = (z-1)\nu(z)$ and $a^{(3)}(z) = (z^2+\alpha z+1)\nu(z)$. First, consider the case $a^{(1)}(z) = (z+1)\nu(z)$. The singular direction of $S(a^{(1)})$ is given by $b^{(1)}(z) = (z-1)\nu(z)$, i.e. $S(a^{(1)})b^{(1)} = 0$. Further, the singular direction $b^{(1)}$ is orthogonal to $a^{(1)}$,

$$\left\langle a^{(1)}, b^{(1)} \right\rangle = \left\langle (z+1)\nu, (z-1)\nu \right\rangle = \left\langle -(z-z^{-1}), |\nu|^2 \right\rangle = 0.$$

By symmetry, the same relations hold for $a^{(2)}$ and $b^{(2)} = (z+1)\nu(z)$. Finally, $S(a^{(3)})b^{(3)} = 0$ for $b^{(3)}(z) = (z^2-1)\nu(z)$, and

$$\left\langle a^{(3)}, b^{(3)} \right\rangle = \left\langle (z^2+\alpha z+1)\nu, (z^2-1)\nu \right\rangle = -\left\langle (z^2-z^{-2}) + \alpha(z-z^{-1}), |\nu|^2 \right\rangle = 0$$

by symmetry of $|\nu(z)|^2$.

In Fig. 5 the singular directions (the kernels) of $\nabla\boldsymbol{\pi}(\mathbf{a})$ are depicted for a series of $\mathbf{a}$'s with $a_0 = 1$. We note that these are roughly orthogonal to the feasible region, similarly as the directions of the unbounded derivative of the function $\phi(\mathbf{q})$. ⧫

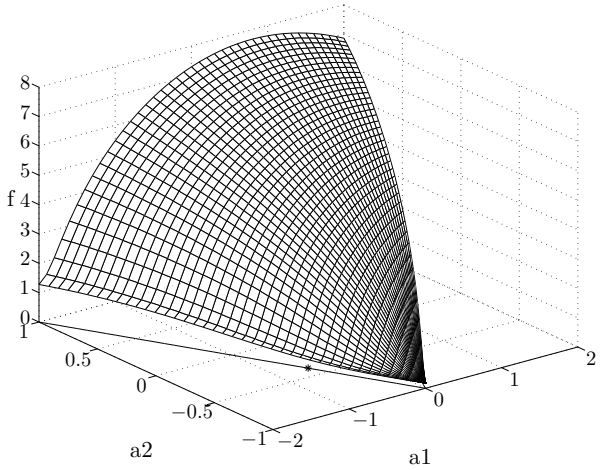

Fig. 4. Function values for monic a.

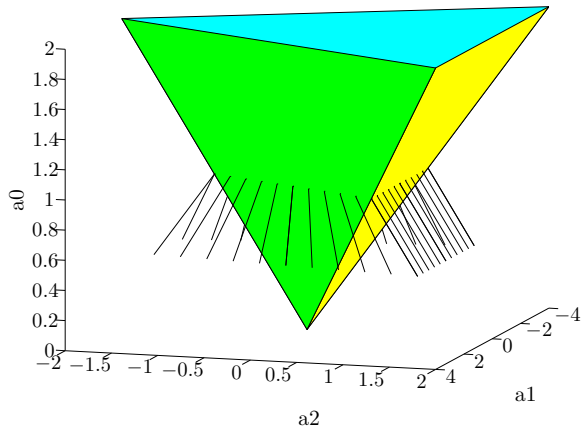

Fig. 5. Singular-directions for $\nabla\boldsymbol{\pi}$ of monic $a(z)$ of degree 2.

4. Homotopy Approach

The solution to the optimization problem $(\mathcal{P}_a)$ is characterized by

$$\mathbf{g}(\mathbf{a}) \stackrel{\triangle}{=} \nabla f(\mathbf{a}) = 0. \tag{53}$$

This is a parameterized family of nonlinear equations, with the parameters $r_0, r_1, \ldots, r_n$ and $p_0, p_1, \ldots, p_n$. In this section, a continuation method (Allgower and Georg, 1990) to solve this system of equations will be proposed. The concept of continuation methods will be explained in the context of the optimization problem $(\mathcal{P}_a)$.

For a fixed positive covariance sequence $r_0, r_1, \ldots, r_n$, the stationary points of $(\mathcal{P}_a)$ form a parameterized solution manifold

$$\mathcal{P}(\mathbf{r}) \stackrel{\triangle}{=} \{\mathbf{a} \mid \mathbf{g}(\mathbf{a}) = 0 \text{ for some } P \in \mathcal{D}_n \text{ s.t. } p_0 = 1\},$$

where a small variation of the parameters results in a small change of the stationary point. Considering parameter combinations with one degree of freedom, any two points on the manifold can be connected with a smooth trajectory. A so-called homotopy is used to deform the first-order optimality conditions of the optimization problem, and then a predictor-corrector method is used for tracking the trajectory from a known starting point on the solution manifold to the desired one.

Definition 1. A continuous function $G : \mathcal{U} \times [0, 1] \to \mathcal{V}$, where $\mathcal{U}$ and $\mathcal{V}$ are topological spaces, is a *homotopy* between the functions $g_0 : \mathcal{U} \to \mathcal{V}$ and $g_1 : \mathcal{U} \to \mathcal{V}$, if $G(\cdot, 0) = g_0(\cdot)$ and $G(\cdot, 1) = g_1(\cdot)$.

A number of different homotopy deformations for the problem $(\mathcal{P}_a)$ exists. Three examples, each of which is of the same type as the original problem, are

- Deformation of the covariances,
- Deformation of the zero-polynomial $\sigma(z)$,
- Deformation of the covariances and the zero-polynomial $\sigma(z)$.

We shall adopt a deformation of the zero-polynomial here, which generates a solution trajectory in the solution manifold $\mathcal{P}(\mathbf{r})$. Such a homotopy is proposed to have $g_0 = f_0$, the objective function defined in (29) corresponding to $\sigma(z) = z^n$, and $g_1 = f$, the objective function corresponding to the desired $\sigma(z)$. Then the Maximum Entropy solution can be used to determine the starting point of the trajectory defined by the homotopy. Since the objective function $\phi(q)$ of $(\mathcal{P}_q)$ can be thought of as a linear problem plus a barrier function, it is most natural to use a homotopy that acts on the barrier term, as the popular interior point methods used for Linear Programming (Nash and Sofer, 1996). Also numerical examples indicate that the deformation of the zero-polynomial $\sigma(z)$ gives the best results. An intuitive explanation of why this gives the best results is that the dominant poles of the Maximum Entropy solution are almost invariant under the deformation of $\sigma(z)$. Maximization of the frequency weighted Maximum Entropy measure will still concentrate on the matching of the spectrum at the peaks, at least in regions where P is large, and the peaks are, of course, a result of the position of the poles. There is another important reason to choose the deformation of the zero-polynomial $\sigma(z)$, namely, because the intermediate results provide covariance interpolating filters. So if the optimization procedure is interrupted before it has converged, relevant partial results exist.

The zero-polynomial $\sigma(z)$ can also be deformed in several ways. This deformation should be such that

$$\sigma_0(z) = z^n, \quad \sigma_1(z) = \sigma(z), \tag{54}$$

and $\sigma_\rho(z)$ should be stable for all $\rho \in [0,1]$. The choice of deformation made here is implicitly defined by a homotopy of the pseudo-polynomial

$$P_\rho(z) \stackrel{\triangle}{=} p_0 + \rho\frac{1}{2}p_1(z+z^{-1}) + \cdots + \rho\frac{1}{2}p_n(z^n+z^{-n}). \tag{55}$$

Thus the following set of problems is considered:

$$(\mathcal{P}_a^\rho) \quad \left[\begin{array}{ll} \min\limits_{\mathbf{a}} & f_\rho(\mathbf{a}), \\ \text{s.t.} & \mathbf{a} \in \boldsymbol{\mathcal{S}}_n, \end{array}\right], \tag{56}$$

where $\rho \in [0,1]$ and

$$f_\rho(\mathbf{a}) \stackrel{\triangle}{=} \mathbf{a}^\top \mathbf{R}\mathbf{a} - 2\langle \log|a|, P_\rho\rangle. \tag{57}$$

For a fixed value of ρ in $[0,1]$, this problem is of the same form as $(\mathcal{P}_a)$. For $\rho = 1$ we have precisely $(\mathcal{P}_a)$, and for $\rho = 0$ we have $P_0(z) = p_0$, so $(\mathcal{P}_a^0)$ is the maximum entropy problem. For intermediate values we have the following lemma.

Lemma 1. *If $P \in \mathcal{D}_n$, then $P_\rho \in \mathcal{D}_n$ for all $\rho \in [0,1]$.*

Proof. For each $\theta \in [-\pi, \pi]$, and each $\rho \in [0,1]$,

$$P_\rho(e^{i\theta}) = \rho P(e^{i\theta}) + (1-\rho)p_0 > 0,$$

since $P(e^{i\theta}) > 0$ and $p_0 > 0$. ■

From the Fejérs Theorem and Lemma 1 we know that there exists a unique stable polynomial $\sigma_\rho(z)$ such that $\sigma_\rho(z)\sigma_\rho(z^{-1}) = P_\rho(z)$, and then for each ρ in $[0,1]$ there is one and only one $\mathbf{a} \in \boldsymbol{\mathcal{S}}_n$ such that

$$\boldsymbol{\gamma}(\mathbf{a},\rho) \stackrel{\triangle}{=} \nabla f_\rho(\mathbf{a}) = 2\mathbf{R}\mathbf{a} - 2\Big\langle \mathbf{z}\frac{1}{a}, P_\rho \Big\rangle = 0. \tag{58}$$

We denote this $\mathbf{a}$ by $\hat{\mathbf{a}}(\rho)$. The function $\boldsymbol{\gamma} : \boldsymbol{\mathcal{S}}_n \times [0,1] \to \mathbb{R}^n$ is a homotopy between $\mathbf{g}_0$ defined in (30) and $\mathbf{g}$ defined in (53).

4.1. Initial Value Problem Formulation

In Proposition 7 it was shown that $\nabla^2 f(\mathbf{a})$ is positive definite at the optimum, from which it is clear that $\nabla_{\mathbf{a}}\boldsymbol{\gamma}(\mathbf{a},\rho)$ is positive definite at the optimum for all $\rho \in [0,1]$. The Implicit Function Theorem (Rudin, 1976) implies that there is a differentiable function $\hat{\mathbf{a}} : \mathbb{R} \to \boldsymbol{\mathcal{S}}_n$ such that $\boldsymbol{\gamma}(\hat{\mathbf{a}}(\rho),\rho) = 0$ for all ρ in $[0,1]$. The function $\hat{\mathbf{a}}$ inherits the C^∞ property from $\boldsymbol{\gamma}$, and defines a smooth trajectory in the space $\boldsymbol{\mathcal{S}}_n$. We will show below that this function satisfies an Initial Value Problem (IVP).

Differentiation of the identity $\boldsymbol{\gamma}(\hat{\mathbf{a}}(\rho),\rho)\equiv 0$ with respect to ρ gives rise to the so-called *Davidenko equation* (Allgower and Georg, 1990; Davidenko, 1953)

$$\nabla_{\mathbf{a}}\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho),\rho\big)\frac{\mathrm{d}\hat{\mathbf{a}}(\rho)}{\mathrm{d}\rho}+\frac{\partial}{\partial\rho}\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho),\rho\big)=0,\quad \rho\in[0,1]. \tag{59}$$

Since $\nabla_{\mathbf{a}}\boldsymbol{\gamma}(\hat{\mathbf{a}}(\rho),\rho)=\nabla^2 f_\rho(\hat{\mathbf{a}}(\rho))$ is positive definite, we know that $\hat{\mathbf{a}}$ must satisfy the well-defined initial value problem

$$\text{(IVP)}\quad\begin{cases}\dfrac{\mathrm{d}\hat{\mathbf{a}}(\rho)}{\mathrm{d}\rho}=-\big[\nabla_{\mathbf{a}}\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho),\rho\big)\big]^{-1}\dfrac{\partial}{\partial\rho}\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho),\rho\big),\\ \hat{\mathbf{a}}(0)=\mathbf{a}_{\mathrm{ME}},\end{cases} \tag{60}$$

where $\mathbf{a}_{\mathrm{ME}}=a_0\sqrt{p_0}\begin{bmatrix}1 & \varphi_{n1} & \cdots & \varphi_{nn}\end{bmatrix}^\top$ is easily obtained from (31) and (32).

4.2. Predictor-Corrector Method

A Predictor-Corrector method is a trajectory following method. It is an iterative method that alternates between two steps. The first step is the predictor step based on numerical integration of the IVP; here an Euler predictor step will be used. The second step is the corrector step where we will apply Newton steps to the current $(\mathcal{P}_a^\rho)$. An important property of the IVP that makes it well-suited for an embedding method (Allgower and Georg, 1993; Arnold, 1983) is that the trajectory $\hat{\mathbf{a}}(\rho)$ has no bifurcations. This follows from the Implicit Function Theorem and the discussion in Section 4.1.

A predictor step is now defined as an Euler step for the IVP, namely,

$$\delta\hat{\mathbf{a}}(\rho)=-\big[\nabla_{\mathbf{a}}\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho),\rho\big)\big]^{-1}\frac{\partial}{\partial\rho}\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho),\rho\big)\delta\rho, \tag{61}$$

for some step size $\delta\rho$.

The corrector step is defined as an iterative optimization method applied to the optimization problem $(\mathcal{P}_a^{\rho+\delta\rho})$ using the predictor point $\hat{\mathbf{a}}(\rho)+\delta\hat{\mathbf{a}}(\rho)$ as the starting point. Here, Newton's method is used as the optimization procedure. A modified version using an inaccurate line search, such as Armijo's rule (Luenberger, 1984), increases the robustness.

5. The Algorithm

One of the most important parts of the Predictor-Corrector method is the choice of the step size $\delta\rho$. Since too large a step may cause convergence problems for the corrector algorithm, and too small a step will lead to long computation times, great care has to be taken at this point.

With minor modifications the convergence theorem in (Allgower and Georg, 1990) can be used to show convergence for the Predictor-Corrector method applied on $(\mathcal{P}_a^\rho)$

if a sufficiently small uniform step length is used. But since it is hard to get a bound on how small the steps have to be, it gives no aid in designing a numerical algorithm.

For a computationally efficient method one should use another approach, with an adaptive procedure for choosing the step length.

5.1. Adaptive Step Length Procedure

One way to determine the step length is to use the convergence proofs available for the corrector step method. In our case Newton's method is used, and the well-known Newton-Kantorovich (NK) Theorem can be applied. The NK Theorem was used by Den Heijer and Rheinboldt (1981) for determining error models for general problems, but it will be used here for guaranteeing convergence in the specific problem considered here.

The following formulation of the NK Theorem can be found in (Ortega and Rheinboldt, 1970).

Theorem 1. (Newton-Kantorovich) *Assume that* $\mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ *is Fréchet-differentiable on a convex set* $D_0 \subset D$ *and that*

$$\|\nabla \mathbf{g}(\mathbf{x}) - \nabla \mathbf{g}(\mathbf{y})\|_2 \leq \gamma \|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in D_0,$$

for some $\gamma > 0$. *Suppose that there exists an* $\mathbf{x}^0 \in D_0$ *such that* $\alpha = \beta\gamma\eta \leq 1/2$, *for some* $\beta, \eta > 0$ *meeting*

$$\beta \geq \left\|\nabla \mathbf{g}(\mathbf{x}^0)^{-1}\right\|_2, \quad \eta \geq \left\|\nabla \mathbf{g}(\mathbf{x}^0)^{-1}\mathbf{g}(\mathbf{x}^0)\right\|_2. \tag{62}$$

Set

$$t^* = (\beta\gamma)^{-1}\left[1 - (1-2\alpha)^{1/2}\right],$$

$$t^{**} = (\beta\gamma)^{-1}\left[1 + (1-2\alpha)^{1/2}\right],$$

and assume that the closed ball $\bar{B}(\mathbf{x}^0, t^*)$ *is contained in* D_0.

Then the Newton iterates

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \nabla \mathbf{g}(\mathbf{x}^k)^{-1}\mathbf{g}(\mathbf{x}^k), \quad k = 0, 1, \ldots$$

are well-defined, remain in $\bar{B}(\mathbf{x}^0, t^*)$, *and converge to a solution* $\mathbf{x}^*$ *of* $\mathbf{g}(\mathbf{x}) = 0$ *which is unique in* $\bar{B}(\mathbf{x}^0, t^{**}) \cap D_0$.

In order to apply this theorem to the function $\mathbf{g}$ defined by (53), the parameters η, β and γ need to be determined. First, let a^0 be the polynomial corresponding to the predictor point $\hat{\mathbf{a}}(\rho) + \delta\hat{\mathbf{a}}$, and define the constants η and β so that (62) hold with $\mathbf{x}^0 = \hat{\mathbf{a}}(\rho) + \delta\hat{\mathbf{a}}$.

Next, we prove the Lipschitz continuity of the derivative $\nabla \mathbf{g}(\mathbf{a})$, and determine a feasible value of the Lipschitz constant γ, so that $D_0 = \bar{B}(a^0, t^*)$. To this end, note that

$$\begin{aligned}\left\|\nabla \mathbf{g}(\mathbf{a}) - \nabla \mathbf{g}(\mathbf{b})\right\|_2 &= 2\left\|\mathbf{R} + \left\langle \frac{1}{a^2}\mathbf{z}\mathbf{z}^\top, P\right\rangle - \mathbf{R} - \left\langle \frac{1}{b^2}\mathbf{z}\mathbf{z}^\top, P\right\rangle\right\|_2 \\ &= 2\left\|\left\langle \left(\frac{1}{a^2} - \frac{1}{b^2}\right)\mathbf{z}\mathbf{z}^\top, P\right\rangle\right\|_2.\end{aligned}$$

Let us define

$$\chi(z) = P(z)\left(\frac{1}{a(z)^2} - \frac{1}{b(z)^2}\right) = \sum_{k=-n}^{\infty} \chi_k z^{-k}.$$

Then

$$\left\langle \left(\frac{1}{a^2} - \frac{1}{b^2}\right)\mathbf{z}\mathbf{z}^\top, P \right\rangle = \begin{bmatrix} \chi_0 & \chi_1 & \iddots & \chi_n \\ \chi_1 & \chi_2 & \iddots & \chi_{n+1} \\ \iddots & \iddots & \iddots & \vdots \\ \chi_n & \chi_{n+1} & \cdots & \chi_{2n} \end{bmatrix} \tag{63}$$

is a Hankel matrix with symbol χ. The Nehari Theorem (Chui and Chen, 1992) implies $\|\mathbf{H}_\chi\|_2 \leq \|\chi\|_\infty$, where $\mathbf{H}_\chi$ is the infinite Hankel matrix, and it is clear that the following inequality holds for the $n \times n$ submatrix in (63):

$$\begin{aligned} 2\left\|\left\langle \left(\frac{1}{a^2} - \frac{1}{b^2}\right)\mathbf{z}\mathbf{z}^\top, P \right\rangle\right\|_2 &\leq 2\left\|\left(\frac{1}{a^2} - \frac{1}{b^2}\right)P\right\|_{L^\infty} \\ &\leq 2\left\|\left(\frac{1}{a^2} - \frac{1}{b^2}\right)\right\|_{L^\infty} \|P\|_{L^\infty}. \end{aligned} \tag{64}$$

A bound on the infinity norm of the pseudo-polynomial P is easy to determine. In fact, it is clear that

$$\|P\|_{L^\infty} \leq \|\mathbf{p}\|_1, \quad \text{where } \mathbf{p} \triangleq \begin{bmatrix} p_0 & p_1 & \dots & p_n \end{bmatrix}. \tag{65}$$

In order to determine the infinity norm of the first factor in (64), the following lemma is useful.

Lemma 2. *If $x, y \in \mathbb{C}$ and $x, y \in B(a, |a|/4)$, then*

$$\left|\frac{1}{x^2} - \frac{1}{y^2}\right| \leq \frac{8}{|a|^3}|x - y|. \tag{66}$$

Proof. If $x, y \in B(a, |a|/4)$, then we can write $x = a + a_x$ and $y = a + a_y$, where $|a_x| < |a|/4$ and $|a_y| < |a|/4$. Then

$$\left|\frac{1}{(a + a_x)^2} - \frac{1}{(a + a_y)^2}\right| = \left|\frac{(a_y - a_x)(2a + a_x + a_y)}{(a + a_x)^2(a + a_y)^2}\right|,$$

and since

$$|2a + a_x + a_y| = |a|\left|2 + \frac{a_x}{a} + \frac{a_y}{a}\right| \leq 5/2|a|$$

and

$$|a + a_x| = |a|\left|1 + \frac{a_x}{a}\right| \geq 3/4|a|,$$

we have

$$\left|\frac{1}{x^2}-\frac{1}{y^2}\right| \le \frac{1}{|a|^3}\frac{5/2}{(3/4)^2(3/4)^2}|a_x-a_y| \le \frac{8}{|a|^3}|x-y|.$$

■

So assuming that $a, b \in \bar{B}(a^0, 1/4\|1/a^0\|_{L^\infty}^{-1})$, it follows that

$$\left\|\nabla \mathbf{g}(\mathbf{a})-\nabla \mathbf{g}(\mathbf{b})\right\|_2 \le 2\cdot 8\left\|\frac{1}{a^0}\right\|_{L^\infty}^3 \|a-b\|_{L^\infty}\|\mathbf{p}\|_1.$$

Finally, using the equivalence of the vector norms to get back to the 2-norm $\|a-b\|_{L^\infty} \le \|\mathbf{a}-\mathbf{b}\|_1 \le \sqrt{n}\|\mathbf{a}-\mathbf{b}\|_2$ shows that

$$\left\|\nabla \mathbf{g}(\mathbf{a})-\nabla \mathbf{g}(\mathbf{b})\right\|_2 \le \bar{\gamma}\|\mathbf{a}-\mathbf{b}\|_2, \quad \bar{\gamma} = 16\sqrt{n}\|\mathbf{p}\|_1\left\|1/a^0\right\|_{L^\infty}^3, \tag{67}$$

an upper bound on the Lipschitz constant is $\bar{\gamma}$. This Lipschitz constant holds in $\bar{B}(a^0, 1/4\|1/a^0\|_{L^\infty}^{-1})$, so take $t^* < \|1/a^0\|_{L^\infty}^{-1}/4$. Note that if P is replaced by P_ρ and $\rho \in [0,1]$, the bound on the infinity norm in (65) still holds and the same $\bar{\gamma}$ is valid.

In order to use the NK Theorem to prove convergence, the step length has to be chosen so that $\alpha = \beta\gamma\eta \le 1/2$ and $t^* < \|1/((\hat{\mathbf{a}}(\rho)+\delta\hat{\mathbf{a}})^\top \mathbf{z})\|_{L^\infty}^{-1}/4$. This can be achieved since η approaches zero as the step length approaches zero.

An algorithm based on this step length procedure and on predictor corrector steps as discussed in Section 4.2 is proposed next.

Algorithm 1. *(Predictor-corrector with adaptive step length)*

1. *Let* $k=0$, $\rho=0$ *and* $\mathbf{a}^{(0)} = \mathbf{a}_{ME} = a_0\sqrt{p_0}\begin{bmatrix} 1 & \varphi_{n1} & \cdots & \varphi_{nn}\end{bmatrix}^\top$.
2. *Determine an initial step length* $\delta\rho$.
3. *If necessary, reduce the step length until* $\alpha = \beta\gamma\eta \le 1/2$ *and* $t^* < \|1/a^{(k)}\|_{L^\infty}^{-1}/4$ *is satisfied.*
4. *Let* $\rho := \min\{1, \rho+\delta\rho\}$, $\quad k := k+1$.
5. *Predictor step: Let* $\mathbf{b}^{(k)} = \mathbf{a}^{(k-1)} + \delta\hat{\mathbf{a}}(\rho)$ *be the estimate of* $\hat{\mathbf{a}}(\rho)$ *defined by (61).*
6. *Corrector step: Solve* $\mathcal{P}_a^\rho$ *for* $\mathbf{a}^{(k)}$ *initiated at* $\mathbf{b}^{(k)}$*, using Newton's method.*
7. *If* $\rho = 1$ *then* $\mathbf{a}^{(k)}$ *is the solution, otherwise go to 2.*

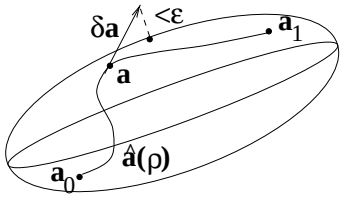

Fig. 6. $\hat{\mathbf{a}}(\rho)$ and predictor step.

5.2. How to Choose the Initial Step Size $\delta\rho$

In this section a procedure to determine an initial step length is proposed. We know that $\boldsymbol{\gamma}(\hat{\mathbf{a}}(\rho), \rho) = 0$, so

$$\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho), \rho\big)^{\top} \hat{\mathbf{a}}(\rho) = 2\big\langle \hat{\mathbf{a}}(\rho), \mathbf{R}\hat{\mathbf{a}}(\rho)\big\rangle - 2\langle 1, P_\rho\rangle = 0 \tag{68}$$

follows from (58). Now, since $\langle 1, P_\rho\rangle = p_0$ independently of ρ, we have

$$\big\langle \hat{\mathbf{a}}(\rho), \mathbf{R}\hat{\mathbf{a}}(\rho)\big\rangle \equiv p_0, \tag{69}$$

so $\hat{\mathbf{a}}(\rho)$ lies on an ellipse defined by the covariances as depicted in Fig. 6.

From (61), the predictor step can be written as $\delta\hat{\mathbf{a}} = \mathbf{v}\,\delta\rho$, where

$$\mathbf{v} = -\big[\nabla_{\mathbf{a}}\boldsymbol{\gamma}(\hat{\mathbf{a}}(\rho), \rho)\big]^{-1} \frac{\partial}{\partial\rho}\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho), \rho\big). \tag{70}$$

The step size $\delta\rho$ can now be chosen so that $\hat{\mathbf{a}} + \delta\hat{\mathbf{a}}$ does not deviate from the ellipse more than some given ε in the metric defined by $\langle \cdot\, \mathbf{R}, \cdot\rangle$. Since $\hat{\mathbf{a}}(\rho)$ is confined to the ellipsoid, the tangent $\delta\hat{\mathbf{a}}$ is orthogonal to the normal $\hat{\mathbf{n}} \triangleq 2\mathbf{R}\hat{\mathbf{a}}$ of the ellipsoid at $\hat{\mathbf{a}}$. Then

$$(\hat{\mathbf{a}} + \delta\hat{\mathbf{a}})^{\top}\mathbf{R}(\hat{\mathbf{a}} + \delta\hat{\mathbf{a}}) = \underbrace{\hat{\mathbf{a}}^{\top}\mathbf{R}\mathbf{a}}_{=p_0} + 2\underbrace{\delta\hat{\mathbf{a}}^{\top}\mathbf{R}\hat{\mathbf{a}}}_{=0} + \underbrace{\delta\hat{\mathbf{a}}^{\top}\mathbf{R}\delta\hat{\mathbf{a}}}_{\geq 0} \geq p_0,$$

and $\hat{\mathbf{a}} + \delta\hat{\mathbf{a}}$ will lie outside the ellipsoid. The initial step length $\delta\rho$ is chosen so that

$$(\hat{\mathbf{a}} + \delta\rho\mathbf{v})^{\top}\mathbf{R}(\hat{\mathbf{a}} + \delta\rho\mathbf{v}) = \hat{\mathbf{a}}^{\top}\mathbf{R}\hat{\mathbf{a}} + \varepsilon$$

for some $\varepsilon > 0$. This criterion leads to the step size

$$\delta\rho = \sqrt{\frac{\varepsilon}{\mathbf{v}^{\top}\mathbf{R}\mathbf{v}}}. \tag{71}$$

Note that the distance between $\hat{\mathbf{a}} + \delta\hat{\mathbf{a}}$ and $\hat{\mathbf{a}}(\rho + \delta\rho)$ is in general larger than ε. Note also that, since $\hat{\mathbf{a}} + \delta\hat{\mathbf{a}}$ is "larger" than $\hat{\mathbf{a}}(\rho + \delta\rho)$, the Hessian $\nabla\boldsymbol{\gamma}(\hat{\mathbf{a}} + \delta\hat{\mathbf{a}}, \rho + \delta\rho)$ is more likely to be positive definite, as argued in the discussion following Proposition 6.

5.3. A Practical Algorithm

The step length determined by Algorithm 1 is in general too small to be useful in practice. Therefore, less stringent conditions need to be used in Step 3. If Newton's method is used with an inaccurate line search in the corrector step, the domain of convergence for the corrector step is enlarged and it pays off to take longer predictor steps. The initial step length determined in Section 5.2 is almost always satisfactory, and gives only a few predictor steps with "rule of thumb" choices of ε.

6. Convergence of the Proposed Algorithm

Theorem 2. *If the constant $\varepsilon > 0$ is chosen sufficiently small, Algorithm 1 will converge in a finite number of steps.*

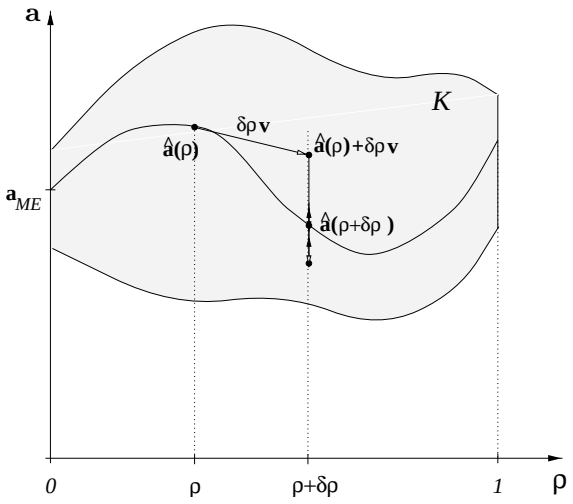

Fig. 7. $\hat{\mathbf{a}}(\rho)$, predictor and corrector step.

Proof. It follows from the NK Theorem that the corrector steps will converge. What remains to be proven is that there will only be a finite number of predictor steps. Since the trajectory $\hat{\mathbf{a}}(\rho)$ is in the interior of $\mathcal{S}_n$ and on an ellisoid (69), there exists a compact neighborhood $\mathcal{K} \subset \mathcal{S}_n$ of the trajectory and, especially $0 \notin \mathcal{K}$.

Let $\mathbf{v}$ be the predictor step direction as defined in (70). Choose $\varepsilon > 0$ small enough such that $\hat{\mathbf{a}}(\rho) + \delta\rho\,\mathbf{v} \in \mathcal{K}$ for all $\rho \in [0, 1]$. Then

$$\mathbf{d}(\mathbf{a}) \stackrel{\triangle}{=} -\frac{\partial}{\partial\rho}\boldsymbol{\gamma}(\mathbf{a}, \rho) = 2\Big\langle \mathbf{z}\frac{1}{a}, P - p_0 \Big\rangle, \tag{72}$$

which follows from (58) and the definition of P_ρ in (55). Next, define

$$\begin{aligned}
C_1 &\triangleq \max\big\{\|1/a\|_{L^\infty} \mid \mathbf{a} \in \mathcal{K}\big\},\\
C_2 &\triangleq \max\big\{\|(\nabla_{\mathbf{a}}\boldsymbol{\gamma}(\mathbf{a},\rho))^{-1}\| \mid \mathbf{a} \in \mathcal{K}, \rho \in [0,1]\big\},\\
C_3 &\triangleq \max\big\{\|\nabla^2_{\mathbf{a}}\boldsymbol{\gamma}(\mathbf{a},\rho)\| \mid \mathbf{a} \in \mathcal{K}, \rho \in [0,1]\big\},\\
C_4 &\triangleq \max\big\{\|\mathbf{d}(\mathbf{a})\| \mid \mathbf{a} \in \mathcal{K}\big\},\\
C_5 &\triangleq \max\big\{\|\nabla_{\mathbf{a}}\mathbf{d}(\mathbf{a})\| \mid \mathbf{a} \in \mathcal{K}\big\},
\end{aligned} \tag{73}$$

and note that these maxima exist since the arguments define continuous functions.

We will prove that the step length can uniformly be bounded from below. Thus it is clear that ρ will converge to one in a finite number of steps.

Consider the step length condition $\alpha = \beta\gamma\eta \leq 1/2$ in Step 3 of the algorithm. It is clear that $\beta \leq C_2$, and from (67) it follows that $\gamma \leq 16\sqrt{n}\|\mathbf{p}\|_1 C_1^3$. Since $\|\nabla\boldsymbol{\gamma}(\mathbf{a},\rho)^{-1}\boldsymbol{\gamma}(\mathbf{a},\rho)\| \leq \|\nabla\boldsymbol{\gamma}(\mathbf{a},\rho)^{-1}\|\|\boldsymbol{\gamma}(\mathbf{a},\rho)\|$, it follows that the step length criterion is implied by

$$\big\|\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho) + \delta\rho\,\mathbf{v}, \rho + \delta\rho\big)\big\| \leq \frac{1/2}{16\sqrt{n}\|\mathbf{p}\|_1 C_1^3 C_2^2}. \tag{74}$$

The predictor step is now considered. Using the Taylor expansion of $\boldsymbol{\gamma}$ in the first variable around $(\hat{\mathbf{a}}(\rho), \rho + \delta\rho)$, we have

$$\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho) + \delta\rho\,\mathbf{v}, \rho + \delta\rho\big) \;=\; \boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho), \rho + \delta\rho\big) + \delta\rho\nabla_{\mathbf{a}}\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho), \rho + \delta\rho\big)\mathbf{v} + \frac{\delta\rho^2}{2}M_1[\mathbf{v},\mathbf{v}].$$

Since $\boldsymbol{\gamma}$ is affine in the second argument,

$$\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho), \rho + \delta\rho\big) = \boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho), \rho\big) - \delta\rho\,\mathbf{d}\big(\hat{\mathbf{a}}(\rho)\big) = -\delta\rho\,\mathbf{d}\big(\hat{\mathbf{a}}(\rho)\big),$$

as seen from (72), it follows that

$$\begin{aligned}
\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho) + \delta\rho\,\mathbf{v}, \rho + \delta\rho\big) \;&=\; -\delta\rho\,\mathbf{d}\big(\hat{\mathbf{a}}(\rho)\big) + \frac{\delta\rho^2}{2}M_1[\mathbf{v},\mathbf{v}]\\
&\quad + \delta\rho\left(\nabla_{\mathbf{a}}\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho), \rho\big) - \delta\rho\nabla_{\mathbf{a}}\mathbf{d}\big(\hat{\mathbf{a}}(\rho)\big)\right)\mathbf{v}\\
&=\; -\delta\rho^2\,\nabla_{\mathbf{a}}\mathbf{d}\big(\hat{\mathbf{a}}(\rho)\big)\mathbf{v} + \frac{\delta\rho^2}{2}M_1[\mathbf{v},\mathbf{v}].
\end{aligned} \tag{75}$$

Consequently,

$$\big\|\boldsymbol{\gamma}\big(\hat{\mathbf{a}}(\rho) + \delta\rho\,\mathbf{v}, \rho + \delta\rho\big)\big\| \leq C_5\delta\rho^2\|\mathbf{v}\| + \frac{1}{2}C_3\delta\rho^2\|\mathbf{v}\|^2, \tag{76}$$

where

$$\|\mathbf{v}\| \leq \left\|\left(\nabla_{\mathbf{a}}\boldsymbol{\gamma}(\hat{\mathbf{a}}(\rho),\rho)\right)^{-1}\right\| \left\|\mathbf{d}\left(\hat{\mathbf{a}}(\rho)\right)\right\| \leq C_2C_4. \tag{77}$$

So the bound for η is of order $\delta\rho^2$:

$$\left\|\boldsymbol{\gamma}\left(\hat{\mathbf{a}}(\rho)+\delta\rho\,\mathbf{v},\rho+\delta\rho\right)\right\| \leq \left(C_5C_2C_4+\frac{1}{2}C_3(C_2C_4)^2\right)\delta\rho^2, \tag{78}$$

and the condition on the step length (for any ρ) is satisfied if

$$\delta\rho \leq \sqrt{\frac{1/2}{16\sqrt{n}\|\mathbf{p}\|_1C_1^3C_2^2}\,\frac{1}{C_5C_2C_4+\frac{1}{2}C_3(C_2C_4)^2}}. \tag{79}$$

A similar argument can be used to show a uniform bound for the condition $t^* < \|1/((\hat{\mathbf{a}}(\rho)+\delta\hat{\mathbf{a}})^\top\mathbf{z})\|_{L^\infty}^{-1}/4$. Using (73), it is clear that the condition $t^* < 1/(4C_1)$ is a stricter version, and this is the one used below.

The definition $t^* = (\beta\gamma)^{-1}(1-\sqrt{1-2\alpha})$, and the expression (67) for γ shows that the condition on the step length is satisfied if

$$\frac{1-\sqrt{1-2\alpha}}{\beta} \leq \frac{16\sqrt{n}\|\mathbf{p}\|_1C_1^3}{4C_1}.$$

Define a sixth constant

$$C_6 \stackrel{\triangle}{=} \min\left\{\left\|(\nabla_{\mathbf{a}}\boldsymbol{\gamma}(\mathbf{a},\rho))^{-1}\right\| \mid \mathbf{a}\in\mathcal{K}, \rho\in[0,1]\right\},$$

in order to bound β. Then

$$1-\sqrt{1-2\alpha} \leq 4\sqrt{n}\|\mathbf{p}\|_1C_1^2C_6,$$

and assuming that (79) holds, $\alpha < 1/2$ and

$$\alpha \leq \left(1-(1-4\sqrt{n}\|\mathbf{p}\|_1C_1^2C_6)^2\right)/2,$$

determines another bound on α. Now, $\alpha = \beta\gamma\eta$, and $\beta \leq C_2$, $\gamma = 16\sqrt{n}\|\mathbf{p}\|_1C_1^3$ are bounded by a constant and η is bounded by (78), which is quadratic in $\delta\rho$. This leads to the following bound on $\delta\rho$:

$$\delta\rho \leq \sqrt{\frac{1-(1-4\sqrt{n}\|\mathbf{p}\|_1C_1^2C_6)^2}{2C_216\sqrt{n}\|\mathbf{p}\|_1C_1^3}\,\frac{1}{C_5C_2C_4+\frac{1}{2}C_3(C_2C_4)^2}}. \tag{80}$$

■

Acknowledgments

The author wishes to express his thanks to Prof. T.T. Georgiou for suggesting the continuation method approach. He would also like to thank his collegues R. Nagamune and Dr. U. Jönsson for the help with some early versions of this paper, and finally to thank his advisor Prof. A. Lindquist for helping to make this paper as interesting for the reader as the research behind it.

References

Allgower E.L. and Georg K. (1990): *Numerical Continuation Methods.* — Berlin, New York: Springer.

Allgower E.L. and Georg K. (1993): *Continuation and path following.* — Acta Numerica, Vol.2, pp.1–64.

Arnold V.I. (1983): *Geometrical Methods in the Theory of Ordinary Differential Equations.* — New York, Berlin: Springer.

Bauer F.L. (1955): *Ein direktes iterationverfahren zur Hurwitz-zerlegung eines polynoms.* — Arch. Elek. Ubertragung, Vol.9, pp.285–290.

Byrnes C.I., Enqvist P. and Lindquist A. (2001): *Cepstral coefficients, covariance lags and pole-zero models for finite data strings.* — IEEE Trans. Sign. Process, Vol.49, No.4.

Byrnes C.I., Gusev S.V. and Lindquist A. (1999): *A convex optimization approach to the rational covariance extension problem.* — SIAM J. Contr. Optim., Vol.37, No.1, pp.211–229.

Byrnes C.I., Lindquist A., Gusev S.V. and Matveev A.S. (1995): *A complete parametrization of all positive rational extensions of a covariance sequence.* — IEEE Trans. Automat. Contr., Vol.40, No.11, pp.1841–1857.

Caines P.E. (1987): *Linear Stochastic Systems.* — New York: Wiley.

Chui C.K. and Chen G. (1992): *Signal Processing and Systems Theory.* — Berlin: Springer.

Davidenko D. (1953): *On a new method of numerically integrating a system of nonlinear equations.* — Dokl. Akad. Nauk SSSR, Vol.88, pp.601–604 (in Russian).

Den Heijer C. and Rheinboldt W.C. (1981): *On steplength algorithms for a class of continuation methods.* — SIAM J. Numer. Anal., Vol.18, No.5, pp.925–948.

Georgiou T.T. (1983): *Partial Realization of Covariance Sequences.* — Ph.D. Thesis, University of Florida.

Georgiou T.T. (1987): *Realization of power spectra from partial covariance sequences.* — IEEE Trans. Acoust. Speech Sign. Process., Vol.ASSP–35, No.4, pp.438–449.

Goodman T., Michelli C., Rodriguez G. and Seatzu S. (1997): *Spectral factorization of Laurent polynomials.* — Adv. Comp. Math., Vol.7, No.4, pp.429–454.

Kalman R.E. (1981): *Realization of covariance sequences.* — Toeplitz Memorial Conference, Tel Aviv, Israel, pp.331–342.

Luenberger D.G. (1984): *Linear and Nonlinear Programming.* — Reading, Mass.: Addison Wesley.

Markel J.D. and Gray Jr. A.H. (1976): *Linear Prediction of Speech*. — New York: Springer.

Nash S.G. and Sofer A. (1996): *Linear and Nonlinear Programming*. — New York: McGraw-Hill.

Ortega J.M. and Rheinboldt W.C. (1970): *Iterative Solution of Nonlinear Equations in Several Variables*. — New York: Academic Press.

Porat B. (1994): *Digital Processing of Random Signals, Theory & Methods*. — Englewood Cliffs. NJ.: Prentice Hall.

Rudin W. (1976): *Principles of Mathematical Analysis*. — New York: McGraw Hill.

Wilson G. (1969): *Factorization of the covariance generating function of a pure moving average process*. — SIAM J. Numer. Anal., Vol.6, pp.1–7.

Wu S-P., Boyd S. and Vandenberghe L. (1997): *FIR filter design via spectral factorization and convex optimization*, In: Applied Computational Control, Signal and Communications (Biswa Datta, Ed.) — Boston: Birkhäuser, pp.215–245.