Prediction-Error Approximation by Convex Optimization

Anders Lindquist

Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology. SE-10044 Stockholm, Sweden. alq@math.kth.se

This paper is dedicated to Giorgio Picci on the occasion of his 65th birthday. I have come to appreciate Giorgio not only as a great friend but also as a great scholar. When we first met at Brown University in 1973, he introduced me to his seminal paper [29] on splitting subspaces, which became the impetus for our joint work on the geometric theory of linear stochastic systems [23–26]. This led to a life-long friendship and a book project that never seemed to converge, but now is close to being finished [27].

I have learned a lot from Giorgio. The present paper grew out of a discussion in our book project, when Giorgio taught me about the connections between prediction-error identification and the Kullback-Leibler criterion. These concepts led directly into the recent theory of analytic interpolation with complexity constraint, with which I have been deeply involved in recent times. I shall try to explain these connections in the following paper.

1 Introduction

Prediction error methods for ARMA modeling play a major role in system identification [28, 30], but in general they lead to nonconvex optimization problems for which global convergence is not guaranteed. In fact, although these algorithms are computationally simple and quite reliable, as pointed out in [32, p. 103], there is so far no theoretically satisfactory algorithm for ARMA parameter estimation. Convex optimization approaches have been proposed [7, 17] for the approximation part, but it remains to verify their practical applicability and statistical accuracy.

In this paper we identify certain classes of ARMA models in which prediction error minimization leads to convex optimization. It has been shown [2, 33] that model approximation via prediction error identification leads to an optimization problem that is related to the minimization of the Kullback-Leibler divergence criterion [18, 21]. This, in turn, leads naturally to the theory of analytic interpolation and generalized moment problems with complexity constraints developed in recent years [8–14,16]. This has already been observed, at least in the context of covariance extension, in [4,6].

The paper is outlined as follows. In Section 2 we review some pertinent facts on prediction error approximation and set notations. In Section 3 we define model classes in terms of a finite number of, not necessarily rational, basis functions and show that the corresponding prediction-error minimizers can be obtained as the solution of a pair of dual convex optimization problems. In the rational case we can even compute the minimizer in closed form. The connections to the Kullback-Leibler criterion and maximum-likelihood identification is described in Section 4. In Section 5 we provide prediction-error approximants in model classes determined by interpolation conditions on the spectral density and its positve real part.

For simplicity this paper will only deal with the scalar case, but multivariable extensions are straightforward, given multivariable versions of the the theory of generalized moment problems with degree constraints [5,22].

2 Prediction-error approximation

Let $\{y(t)\}_{\mathbb{Z}}$ be a zero-mean stationary stochastic process with a spectral density $\{\Phi(e^{i\theta}); \theta \in [-\pi,\pi]\}$ that may be rational or nonrational but is zero only in isolated points θ. Let w be a normalized minimum-phase spectral of Φ; i.e.,

$$\Phi(e^{i\theta}) = \rho |w(e^{i\theta})|^2, \quad \theta \in [-\pi,\pi],$$

where $w(0) = 1$ and $\rho > 0$ is a suitable normalizing factor. Then the process y can be modeled by passing a white noise e with covariance lags $\mathbf{E}\{e(t)e(s)\} = \rho\delta_{ts}$ through a filter with a transfer function

$$w(z) = \sum_{k=0}^{\infty} w_k z^{-k}.$$

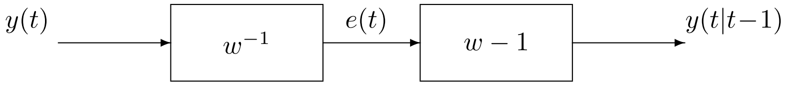
Since $w_0 = 1$,

$$y(t) = e(t) + y(t|t-1),$$

where

$$y(t|t-1) = w_1 e(t-1) + w_2 e(t-2) + \dots$$

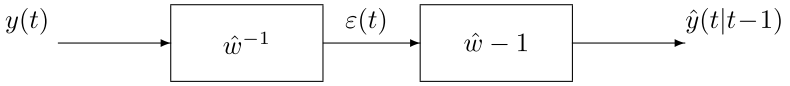
is the one-step ahead linear predictor of $y(t)$ given $\{y(s);\ s \leq t-1\}$. Hence $y(t|t$–$1)$ can be represented by passing e through a filter with transfer function $w-1$ as shown in the block diagram



In particular,

$$y(t) - y(t|t{-}1) = e(t).$$

Now, let $\hat{w}$ be a normalized ($\hat{w}(0) = 1$), stable minimum-phase function belonging to some *model class* $\mathcal{W}$ to be specified later. We shall regard $\hat{w}$ as an approximation of w, from which we can form an *approximate predictor*, denoted by $\hat{y}(t|t{-}1)$, as in the figure



Then

$$\varepsilon(t) = y(t) - \hat{y}(t|t{-}1);$$

i.e., $\varepsilon(t)$ is the *prediction error*, which is not a white noise. Indeed, it is easy to see that it has the variance

$$r := \mathbf{E}\{\varepsilon(t)^2\} = \int_{-\pi}^{\pi} |\hat{w}(e^{i\theta})|^{-2} \Phi(e^{i\theta}) \frac{d\theta}{2\pi}. \tag{1}$$

Since $\varepsilon(t) = e(t) + [y(t|t{-}1) - \hat{y}(t|t{-}1)]$ and $e(t)$ and $[y(t|t{-}1) - \hat{y}(t|t{-}1)]$ are uncorrelated,

$$r = \rho + \mathbf{E}\{|y(t|t{-}1) - \hat{y}(t|t{-}1)|^2\} \geq \rho.$$

The idea is now to find a $\hat{w} \in \mathcal{W}$ that minimizes the prediction error variance (1). To this end, define the class $\mathcal{F}$ of spectral densities

$$\hat{\Phi}(e^{i\theta}) = \hat{\rho} |\hat{w}(e^{i\theta})|^2, \tag{2}$$

where $\hat{w} \in \mathcal{W}$ and $\hat{\rho} > 0$. Then the prediction error takes the form

$$r := \hat{\rho} \int_{-\pi}^{\pi} \hat{\Phi}(e^{i\theta})^{-1} \Phi(e^{i\theta}) \frac{d\theta}{2\pi}. \tag{3}$$

The purpose of the coefficient $\hat{\rho}$ in (3) is merely to normalize $\hat{\Phi}$. Once an optimal $\hat{\Phi} \in \mathcal{F}$ has been determined, $\hat{\rho}$ and $\hat{w} \in \mathcal{W}$ are obtained by outer spectral factorization and normalzation so that $\hat{w}(0) = 1$.

3 Prediction-error approximation in restricted model classes

We begin by defining the model class $\mathcal{F}$. To this end, let

$$g_0, g_1, g_2, \dots, g_n \tag{4}$$

be a linearly independent sequence of Lipschitz continuous functions on the unit circle with zeros only in isolated points, and let the model class $\mathcal{F}$ be the set of all functions $\hat{\Phi}$ such that

$$\hat{\Phi}(e^{i\theta})^{-1} = Q(e^{i\theta}) := \mathrm{Re}\left\{\sum_{k=0}^{n} q_k g_k(e^{i\theta})\right\}, \tag{5}$$

for some $q_0, q_1, \dots, q_n \in \mathbb{C}$ such that $Q(e^{i\theta}) \geq 0$ for all $\theta \in [-\pi, \pi]$. In addition, let $\mathcal{Q}$ the class of all such functions Q.

As a simple example, consider the case $g_k = z^k$, $k = 0, 1, \dots, n$. Then the model class $\mathcal{W}$ is the family of all $AR(n)$ models. However, more general choices of rational basis functions (4) yield model classes of $ARMA$ models. Even more generally, we may choose basis functions that are not even rational.

Theorem 1. *Let the spectral density Φ have the property that the generalized moments*

$$c_k := \int_{-\pi}^{\pi} g_k(e^{i\theta})\Phi(e^{i\theta})\frac{d\theta}{2\pi}, \quad k = 0, 1, \dots, n, \tag{6}$$

exist, and define the functional $\mathbb{J} : \mathcal{Q} \to \overline{\mathbb{R}}$ as

$$\mathbb{J}(Q) = \int_{-\pi}^{\pi} \left[\Phi(e^{i\theta})Q(e^{i\theta}) - \log Q(e^{i\theta})\right] \frac{d\theta}{2\pi}. \tag{7}$$

Then the functional (7) *has a unique minimum Q_{opt}, which is an interior point in $\mathcal{Q}$. Moreover,*

$$\int_{-\pi}^{\pi} g_k(e^{i\theta})\frac{1}{Q_{opt}(e^{i\theta})}\frac{d\theta}{2\pi} = c_k, \quad k = 0, 1, \dots, n. \tag{8}$$

Proof. Since the functions $g_0, g_1, \dots, g_n$ are Lipschitz continuous, hypothesis $\mathbf{H1}$ in [14] is satisfied [14, Remark 1.1]. Moreover, since both $Q \in \mathcal{Q}$ and Φ are nonnegative on the unit circle with zeros only in isolated points,

$$\int_{-\pi}^{\pi} Q\Phi\frac{d\theta}{2\pi} > 0$$

for all $Q \in \mathcal{Q} \setminus \{0\}$. Hence the sequence $c = (c_1, c_2, \dots, c_n)$ is positive in the sense prescribed in [14].

The functional $\mathbb{J} : \mathcal{Q} \to \mathbb{R}$ is strictly convex on the convex set $\mathcal{Q}$, and hence, if a minimum does exist, it must be unique. However, it is shown in [14, Theorem 1.5] that $\mathbb{J}$ has a unique minimizer, Q_{opt}, which lies in the interior of $\mathcal{Q}$, provided the sequence $c = (c_1, c_2, \dots, c_n)$ is positive and hypthesis $\mathbf{H1}$ holds, which is what we have established above. Since the minimizer Q_{opt} is an interior point, the gradient of $\mathbb{J}$ must be zero there, and hence (8) follows.

Theorem 2. *Let Φ be an arbitrary spectral density such that the generalized moments* (6) *exist. Then, there is a unique spectral density $\hat{\Phi}$ in the model class $\mathcal{F}$ that minimizes the prediction error variance* (3)*, and it is given by*

$$\hat{\Phi}_{opt} := Q_{opt}^{-1}, \tag{9}$$

where Q_{opt} is the unique minimizer in Theorem 1.

Proof. By Theorem 1, $\hat{\Phi}_{opt}$ is the unique minimizer of

$$\mathbb{J}(\hat{\Phi}^{-1}) = \int_{-\pi}^{\pi} \left[\Phi(e^{i\theta})\hat{\Phi}^{-1}(e^{i\theta}) + \log \hat{\Phi}^{-1}(e^{i\theta}) \right] \frac{d\theta}{2\pi}. \tag{10}$$

However, by (3),

$$\int_{-\pi}^{\pi} \Phi(e^{i\theta})\hat{\Phi}^{-1}(e^{i\theta}) \frac{d\theta}{2\pi} = \frac{r}{\hat{\rho}}.$$

Moreover, in view of (2)

$$\begin{aligned} \int_{-\pi}^{\pi} \log \hat{\Phi} \frac{d\theta}{2\pi} &= \log \hat{\rho} + 2 \int_{-\pi}^{\pi} \log |\hat{w}| \frac{d\theta}{2\pi} \\ &= \log \hat{\rho} + 2 \log \hat{w}(0) = \log \hat{\rho}, \end{aligned}$$

where we have used Jensen's formula [1, p.184] and the facts that $\hat{w}$ is outer and $\hat{w}(0) = 1$. Consequently,

$$\mathbb{J}(\hat{\Phi}^{-1}) = \frac{r}{\hat{\rho}} + \log \hat{\rho}. \tag{11}$$

Now, for any fixed $r > 0$, (11) has a unique minimum for $\hat{\rho} = r$, and hence

$$\mathbb{J}(\hat{\Phi}_{opt}^{-1}) = 1 + \min_{r} \log r.$$

Therefore $\log r$, and hence the prediction error r, takes it unique minimum value for $\hat{\Phi} = \hat{\Phi}_{opt}$, as claimed.

Now, in view of (8) and (9),

$$\int_{-\pi}^{\pi} g_k(e^{i\theta})\hat{\Phi}_{opt}(e^{i\theta}) \frac{d\theta}{2\pi} = c_k, \quad k = 0, 1, \ldots, n. \tag{12}$$

However, $\hat{\Phi}_{opt}$ is not the only spectral density that satisfies these moment conditions. In fact, following [12, 14], we can prove that, among all such solutions, $\hat{\Phi}_{opt}$ is the one maximizing the entropy gain.

Theorem 3. *The optimal prediction-error approximation $\hat{\Phi}_{opt}$ of Theorem 2 is the unique maximizer of the entropy gain*

$$\mathbb{I}(\hat{\Phi}) := \int_{-\pi}^{\pi} \log \hat{\Phi}(e^{i\theta}) \frac{d\theta}{2\pi} \tag{13}$$

subject to the moment constraints

$$\int_{-\pi}^{\pi} g_k(e^{i\theta})\hat{\Phi}(e^{i\theta}) \frac{d\theta}{2\pi} = c_k, \quad k = 0, 1, \ldots, n. \tag{14}$$

Let us stress again that the basis functions $g_0, g_1, \ldots, g_n$ need not be rational. Although, in general, we want the model class $\mathcal{W}$ to consist of rational functions of low degree, there may be situations when it is desirable to include nonrational components, such as, for example, exponentials.

Identification in terms of orthogonal basis functions is a well studied topic [19, 34, 35]. The most general choice is

$$g_k(z) = \frac{\sqrt{1-|\xi_k|^2}}{z-\xi_k} \prod_{j=0}^{k-1} \frac{1-\xi_j^* z}{z-\xi_j},$$

where $\xi_0, \xi_1, \xi_2, \ldots$ are poles to be selected by the user. The functions $g_0, g_1, g_2, \ldots$ form a complete sequence in the Hardy space $H^2(\mathbb{D}^c)$ over the complement of the unit disc $\mathbb{D}$ provided $\sum_{k=0}^{\infty}(1-|\xi_k|) = \infty$. In [19] the problem to determine a minimum-degree rational function of the form

$$\hat{F}(z) = \frac{1}{2}c_0 g_0(z) + \sum_{k=1}^{\infty} c_k g_k(z),$$

where $c_0, c_1, \ldots, c_n$ are prescribed, was considered.

In our present setting, in order for $\hat{\Phi} := \mathrm{Re}\{\hat{F}\}$ to be a spectral density, $\hat{F}$ needs to be positive real, leading to a problem left open in [19]. Let $c_0, c_1, \ldots, c_n$ be given by (6). Then, by Theorem 3, the problem of determining the minimum prediction-error approximant of Φ in the model class defined by $g_0, g_1, \ldots, g_n$ amounts to finding the function $\hat{\Phi}$ that maximizes the entropy gain

$$\int_{-\pi}^{\pi} \log \hat{\Phi} \frac{d\theta}{2\pi},$$

subject to

$$\int_{-\pi}^{\pi} g_k \hat{\Phi} \frac{d\theta}{2\pi} = c_k, \quad k = 0, 1, \ldots, n.$$

Alternatively, we may solve the convex optimization problem of Theorem 1.

Theorem 1 enables us to determine, under general conditions, the minimum prediction-error in closed form. Here, following [16], we state such a result under the assumption that the basis functions are rational.

Proposition 1. *Suppose that the basis functions $g_0, g_1, \ldots, g_n$ are rational and analytic in the unit disc $\mathbb{D}$. Then,*

$$Q_{opt}(z) = \frac{|g^*(z)P^{-1}g(0)|^2}{g^*(0)P^{-1}g(0)}, \tag{15}$$

where

$$g(z) := \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix}, \qquad P := \int_{-\pi}^{\pi} g(e^{i\theta})\Phi(e^{i\theta})g(e^{i\theta})^* \frac{d\theta}{2\pi}.$$

Proof. Clearly the basis functions $g_0, g_1, \dots, g_n$ belong to the Hardy space $H^2(\mathbb{D})$, and $g := (g_0, g_1, \dots, g_n)'$ has a representation

$$g(z) = (I - zA)^{-1}B,$$

where (A, B) is a reachable pair. Then

$$\varphi(z) = \frac{\det(zI - A^*)}{\det(I - zA)}$$

is an inner function, and it can be shown that the basis functions $g_0, g_1, \dots, g_n$ span the coinvariant subspace $\mathcal{K} := H^2 \ominus \varphi H^2$. Moreover, for any $Q \in \mathcal{Q}$, there is an outer function in $a \in \mathcal{K}$ such that $Q = a^* a$ ([13, Proposition 9]). Consequently (7) can be written

$$J(a) = \int_{-\pi}^{\pi} a^* \Phi a \frac{d\theta}{2\pi} - \int_{-\pi}^{\pi} 2\log|a| \frac{d\theta}{2\pi}.$$

Here the second term can be written $2\log|a(0)|$ by Jensen's formula [1, p.184], and since $a \in \mathcal{K}$, there is a vector $\mathbf{a} \in \mathbb{C}^{n+1}$ such that $a(z) = g^*(z)\mathbf{a}$, so the second term be written $\mathbf{a}^* P \mathbf{a}$. Hence the optimization problem is reduced to determining the $\mathbf{a}$ that minimizes

$$\tilde{J}(\mathbf{a}) = \mathbf{a}^* P \mathbf{a} - 2\log|\mathbf{a}^* g(0)|.$$

Setting the gradient equal to zero, we obtain $\mathbf{a} = P^{-1}g(0)/|a(0)|$ and hence $a(z) = g^*(z)P^{-1}g(0)/|a(0)|$. Then $|a(0)|^2 = g^*(0)P^{-1}g(0)$, and therefore the optimal $\mathbf{a}$ becomes

$$a(z) = \frac{g^*(z)P^{-1}g(0)}{\sqrt{g^*(0)P^{-1}g(0)}},$$

from which (15) follows.

Remark 1. The pair of dual optimization problems in Theorems 1-3 are special cases of a more general formulation [8–14, 16] where (7) is replaced by

$$\mathbb{J}_\Psi(Q) = \int_{-\pi}^{\pi} \left[\Phi(e^{i\theta})Q(e^{i\theta}) - \Psi(e^{i\theta}) \log Q(e^{i\theta})\right] \frac{d\theta}{2\pi}, \tag{16}$$

with Ψ is a parahermitian function that is positive on the unit circle and available for tuning; and (13) is replaced by

$$\mathbb{I}_\Psi(\hat{\Phi}) := \int_{-\pi}^{\pi} \Psi(e^{i\theta}) \log \hat{\Phi}(e^{i\theta}) \frac{d\theta}{2\pi}. \tag{17}$$

The particular choice $\Psi = I$, corresponding to the minimum prediction-error approximation, is called the *central* or *maximum entropy* solution. As suggested by Blomqvist and Wahlberg [4, 6] in the context of covariance extension, a nontrivial Ψ corresponds to a particular choice of prefiltering that may lead to better results; cf, page 274.

4 The Kullback-Leibler criterion and maximum-likelihood identification

The optimization problem of Theorem 1 is intimately connected to the Kullback-Leibler divergence [18, 21]

$$D(y\|z) := \limsup_{N\to\infty} \frac{1}{N} D(p_y^N \mid p_z^N)$$

from one stationary, Gaussian stochastic processes z to another y, where p_y^N and p_z^N are the N-dimensional density functions of y and z respectively, and where

$$D(p_1 \mid p_2) := \int_{\mathbb{R}^n} p_1(x) \log \frac{p_1(x)}{p_2(x)}\, dx.$$

In fact, it was shown in [33] that, if y and z have spectral densities Φ and $\hat{\Phi}$, respectively, then

$$D(y\|z) = \frac{1}{2}\int_{-\pi}^{\pi} \left[(\Phi - \hat{\Phi})\hat{\Phi}^{-1} - \log(\Phi\hat{\Phi}^{-1}) \right] \frac{d\theta}{2\pi}. \tag{18}$$

Consequently,

$$D(y\|z) = \frac{1}{2}\mathbb{J}(\hat{\Phi}^{-1}) - \frac{1}{2}\left[1 + \int_{-\pi}^{\pi} \log \Phi \frac{d\theta}{2\pi} \right], \tag{19}$$

where the last integral is constant.

Given the process y, consider the problem to find the minimum divergence $D(y\|z)$ over all z with a spectral density $\hat{\Phi} \in \mathcal{F}$. Then we have established that this minimum is attained precisely when $\hat{\Phi}^{-1}$ is the unique minimizer of $\mathbb{J}$ in Theorem 1, which in turn is the minimum prediction-error estimate in the model class $\mathcal{F}$.

Next, suppose that we have a finite sample record

$$\{y_0, y_1, \dots, y_N\} \tag{20}$$

of the process y and an estimate Φ_N of Φ based on (20) that is consistent in the sense that $\lim_{N\to\infty} \Phi_N(e^{i\theta}) = \Phi(e^{i\theta})$ with probability one for almost all $\theta \in [-\pi, \pi]$. The periodogram

$$\Phi_N(e^{i\theta}) = \frac{1}{N}\left| \sum_{t=0}^{N} e^{-i\theta t} y_t \right|^2.$$

is one such estimate of Φ. Then, under some mild technical assumptions,

$$J_N(\hat{\Phi}) := \frac{1}{2}\int_{-\pi}^{\pi} \left[\Phi_N(e^{i\theta})\hat{\Phi}(e^{i\theta})^{-1} + \log \hat{\Phi}(e^{i\theta}) \right] \frac{d\theta}{2\pi} \tag{21}$$

tends to $\mathbb{J}(\hat{\Phi}^{-1})$ as $N \to \infty$. The functional $J_N(\Psi)$ is known as the *Whittle log-likelihood*, and it is a widely used approximation of $-\log L_N$, where $L_N(\hat{\Phi})$ is the *likelihood function*. In fact, $J_N(\hat{\Phi})$ and $L_N(\hat{\Phi})$ tend to the same limit $\mathbb{J}(\hat{\Phi}^{-1})$ as $N \to \infty$.

5 Prediction-error approximation by analytic interpolation

Let Φ be the given (or estimated) spectral density defined as above. Then, by the Herglotz formula,

$$F(z) = \int_{-\pi}^{\pi} \frac{e^{i\theta}+z}{e^{i\theta}-z} \Phi(e^{i\theta}) \frac{d\theta}{2\pi} \tag{22}$$

is *the positive real part* of Φ. More precisely, F is the unique function in $H(\mathbb{D})$ such that $F(0)$ is real and

$$\Phi(e^{i\theta}) = \text{Re}\{F(e^{i\theta})\}. \tag{23}$$

Now, let us select a number of points

$$z_0, z_1, \dots, z_n \tag{24}$$

in the unit disc $\mathbb{D}$. Then, in view of (22),

$$F(z_k) = \int_{-\pi}^{\pi} g_k(z) \Phi(e^{i\theta}) \frac{d\theta}{2\pi},$$

where

$$g_k(z) = \frac{z+z_k}{z-z_k}. \tag{25}$$

Therefore, if the points (24) are distinct, we may choose $g_0, g_1, \dots, g_n$ as our basis functions, and then

$$F(z_k) = c_k, \quad k = 0, 1, \dots, n,$$

where

$$c_k := \int_{-\pi}^{\pi} g_k(e^{i\theta}) \Phi(e^{i\theta}) \frac{d\theta}{2\pi}, \quad k = 0, 1, \dots, n. \tag{26}$$

If (24) are not distinct, we modify $g_0, g_1, \dots, g_n$ in the following way to make them linearly independent. If $z_k = z_{k+1} = \dots = z_{k+m-1}$, then $g_k, \dots, g_{k+m-1}$ are replaced by

$$g_k(z) = \frac{z+z_k}{z-z_k}, \quad g_{k+1}(z) = \frac{2z}{(z-z_k)^2}, \quad \dots, \quad g_{k+m-1}(z) = \frac{2z}{(z-z_k)^m}. \tag{27}$$

Then, differentiating (22), we have the modified interpolation conditions

$$F(z_k) = c_k, \quad \frac{dF}{dz}(z_k) = c_{k+1}, \quad \dots, \quad \frac{1}{(m-1)!} \frac{d^{(m-1)}F}{dz^{(m-1)}}(z_k) = c_{k+m-1}.$$

Now, given the points (24), let $\mathcal{F}(z_0, z_1, \dots, z_n)$ be the class of all spectral densities $\hat{\Phi}$ with positive real part $\hat{F}$ of degree at most n and satisfying the interpolation conditions

$$\hat{F}(z_k) = c_k \tag{28a}$$

for distinct points and

$$\hat{F}(z_k) = c_k, \quad \frac{d\hat{F}}{dz}(z_k) = c_{k+1}, \quad \dots, \quad \frac{1}{(m-1)!}\frac{d^{(m-1)}\hat{F}}{dz^{(m-1)}}(z_k) = c_{k+m-1}, \tag{28b}$$

if $z_k = z_{k+1} = \cdots = z_{k+m-1}$, where $c_0, c_1, \dots, c_n$ are given by (26). In particular,

$$\hat{\Phi}(z_k) = \Phi(z_k), \quad k = 0, 1, \dots, n \tag{29}$$

for all $\hat{\Phi} \in \mathcal{F}(z_0, z_1, \dots, z_n)$, where some of the conditions (29) may be repeated (in case of multiple points).

With the basis (4) chosen as above, the minimum prediction-error approximation in the model class $\mathcal{F}(z_0, z_1, \dots, z_n)$ defined by these functions is as described in the following theorem, which now is a direct consequence of Theorems 1 and 2.

Theorem 4. *The minimum prediction-error approximation of Φ in the class $\mathcal{F}(z_0, z_1, \dots, z_n)$ is the unique $\hat{\Phi} \in \mathcal{F}(z_0, z_1, \dots, z_n)$ that minimizes the entropy gain*

$$\int_{-\pi}^{\pi} \log \hat{\Phi} \frac{d\theta}{2\pi},$$

or, dually, the $\hat{\Phi}$ that minimizes (7)*, where $g_0, g_1, \dots, g_n$ are given by* (25)*, or* (27) *for multiple points.*

It follows from Theorem 1 that all $\hat{\Phi} \in \mathcal{F}(z_0, z_1, \dots, z_n)$, and in particular the optimal one, has (spectral) zeros that coincide with $z_0, z_1, \dots, z_n$ and hence with the interpolation points. Recently, Sorensen [31] has developed an efficient algorithm for solving large problems of this type. In [15] we point out the connection between this approach, initiated by Antoulas [3], and our theory for analytic interpolation with degree constraints [10–14, 16]. We show that a better spectral fit can often be obtained by choosing a nontrivial weight P in the objective function (16). This corresponds to prefiltering; see Remark 1.

An important question in regard to the application of Theorem 4 to system identification is how to choose the interpolation points $z_0, z_1, \dots, z_n$. Here (29) could serve as an initial guide. However, a more sofisticated procedure is proposed in [20].

6 Conclusion

In this paper we have shown that in large model classes of ARMA models, as well as in some model classes of nonrational functions, prediction-error approximation leads to convex optimization. The connections to Kullback-Leibler and maximum-likelihood criteria have been described. Model classes defined in terms of interpolation conditions have also been considered, connecting to literature in numerical linear algebra. Generalizations to the multivarable case should be straight-forward relying on mutivarable versions [5, 22] of the theory of analytic interpolation and generalized moment problems with complexity constraints.

References

1. Ahlfors LV (1953) Complex Analysis. McGraw-Hill,
2. Anderson BDO, Moore JB, Hawkes RM (1978) Automatica 14: 615–622
3. Antoulas AC (2005) Systems and Control Letters 54: 361–374
4. Blomqvist, A (2005) A Convex Optimization Approach to Complexity Constrained Analytic Interpolation with Applications to ARMA Estimation and Robust Control. PhD Thesis, Royal Institute of Technology, Stockholm, Sweden
5. Blomqvist A, Lindquist A, Nagamune R (2003) IEEE Trans Autom Control 48: 2172–2190
6. Blomqvist A, Wahlberg B (2007) IEEE Trans Autom Control 55: 384–389
7. Byrnes CI, Enqvist P, Lindquist A (2002) SIAM J. Control and Optimization 41: 23–59
8. Byrnes CI, Gusev SV, Lindquist A (1998) SIAM J. Contr. and Optimiz. 37: 211–229
9. Byrnes CI, Gusev SV, Lindquist A (2001) SIAM Review 43: 645–675
10. Byrnes CI, Georgiou TT, Lindquist A (2001) IEEE Trans Autom Control 46: 822–839
11. Byrnes CI, Georgiou TT, Lindquist A (2000) IEEE Trans. on Signal Processing 49: 3189–3205
12. Byrnes CI, Lindquist A (2003) A convex optimization approach to generalized moment problems. In: Hashimoto K, Oishi Y, Yamamoto Y (eds) Control and Modeling of Complex Systems: Cybernetics in the 21st Century. Birkhäuser, Boston Basel Berlin
13. Byrnes CI, Georgiou TT, Lindquist A, Megretski (2006) Trans American Mathematical Society 358: 965–987
14. Byrnes CI, Lindquist A (2006) Integral Equations and Operator Theory 56: 163–180
15. Fanizza G, Karlsson J, Lindquist A, Nagamune R (2007) Linear Algebra and Applications. To be published
16. Georgiou TT, Lindquist A (2003) IEEE Trans. on Information Theory 49: 2910–2917
17. Georgiou TT, Lindquist A (2007) IEEE Trans Autom Control. Submitted
18. Good, IJ (1963) Annals Math. Stat. 34: 911–934
19. Heuberger PSC, Van den Hof PMJ, Szabó Z (2001) Proc. 40th IEEE Conf. Decision and Control, Orlando, Florida, USA: 3673–3678
20. Karlsson J, Lindquist A (2007) To appear
21. Kullback S (1959) Information Theory and Statistics. John Wiley, New York
22. Kuroiwa Y, Lindquist A (2007) Proc 2007 Decision and Control Conference. Submitted for publication
23. Lindquist A, Picci G (1985) SIAM J Control Optim 23: 809–857
24. Lindquist A, Picci G (1995) Stochastics 15: 1–50
25. Lindquist A, Picci G (1991) J Math Systems Estim Control 1: 241–333
26. Lindquist A, Picci G (1996) Automatica 32:709–733
27. Lindquist A, Picci G (2007) Linear Stochastic Systems: A Geometric Approach to Modeling, Estimation and Identification. To appear
28. Ljung L (1987) System Identification: Theory for the User. Prentice Hall, Englewood Cliffs
29. Picci G (1976) Proc. IEEE 64: 112–122
30. Söderström T, Stoica P (1989) System Identification. Prentice Hall, New York
31. Sorensen DC (2005) Systems and Control Letters 54: 347-360
32. Stoica P, Moses R (1997) Introduction to Spectral Analysis. Prentice Hall, Upper Saddle River, NJ
33. Stoorvogel AA, van Schuppen JH (1996) System identification with information theoretic criteria. In: Bittanti S, Picci G (eds) Identification, Adaptation, Learning: The Science of learning Models from Data. Springer, Berlin Heidelberg
34. Wahlberg B (1991) IEEE Trans Autom Control 36: 551–562
35. Wahlberg B (1994) IEEE Trans Autom Control 39: 1276–1282