A COVARIANCE EXTENSION APPROACH TO IDENTIFICATION OF TIME SERIES*

JORGE MARI†, ANDERS DAHLÉN†, AND ANDERS LINDQUIST†

Abstract. In this paper we consider a three-step procedure for identification of time series, based on covariance extension and model reduction, and we present a complete analysis of its statistical convergence properties. A partial covariance sequence is estimated from statistical data. Then a high-order maximum-entropy model is determined, which is finally approximated by a lower-order model by stochastically balanced model reduction. Such procedures have been studied before, in various combinations, but an overall convergence analysis comprising all three steps has been lacking. Supposing the data is generated from a true finite-dimensional system which is minimum phase, it is shown that the transfer function of the estimated system tends in $\mathcal{H}^\infty$ to the true transfer function as the data length tends to infinity, if the covariance extension and the model reduction is done properly. The proposed identification procedure, and some variations of it, are evaluated by simulations.

1. Introduction

In recent years there has been quite some interest in a certain type of procedures for identification of time series known as *subspace methods* [1, 42, 41, 28, 29]. These identification procedures are based on geometric projection methods, and they could be understood in the context of splitting geometry and partial stochastic realization theory [30, 31]. However, as pointed out in [32] and further elaborated upon in [9], these procedures are algebraically equivalent to minimal factorization of a Hankel matrix of covariance estimates, and they make no distinction between stochastic and deterministic partial realizations. Therefore they may fail because of loss of positive realness in the spectral estimation phase.

In an attempt to overcome these problems we analyze an alternative approach to time series identification proposed in [32], namely a three-step procedure consisting of estimation of a partial covariance sequence, covariance extension by the maximum-entropy method, leading to a high order autoregressive (AR) process, and finally stochastically balanced truncation. This method shares certain features with stochastic subspace identification methods, the most obvious one being that it is based on partial stochastic realization theory, but, unlike stochastic subspace methods, it guarantees positive realness. Moreover, our procedure only involves linear

* This research was supported by a grant from the Swedish Research Council for Engineering Sciences (TFR).

† Division of Optimization and Systems Theory, Royal Institute of Technology, 100 44 Stockholm, Sweden

algebra operations, and no iterations or optimization of nonconvex functions, as for maximum-likelihood (ML) methods, are needed.

The idea of approximating an autoregressive moving-average (ARMA) process by an AR process is by no means new. Its origins can be traced back to the Wold decomposition [55] where $\mathcal{L}^2$-convergence of high-order AR models to general analytic models is shown. Pioneers in the use of this concept for systems identification are Durbin [12, 13] and Whittle [54]. The convergence properties of such approximations were studied by Berk [2] and later refined in [36, 34, 33, 7]. The interesting paper [7] contains nice proofs of some of the convergence results needed in this paper, but, for the sake of completeness and insight, we provide new proofs based on some properties of fast filtering algorithms [5] and simple methods of complex analysis and Szegő polynomials. The power of the theory of Szegő polynomials and Toeplitz matrices in analyzing stochastic processes is reported in [24], but, except for elementary theory, it has not been much used in systems identification [39]. This is even more true for the newer results [16, 40, 37, 27] on orthogonal polynomials.

The idea of using model reduction for systems identification appears in the thesis by Wahlberg [50] and the subsequent paper [51], where the emphasis is on frequency weighted reduction. Instead, we use stochastically balanced truncation, for which we develop a simple computational procedure, exploiting the special structure of the AR model. We also show the advantage of this reduction procedure by theoretical analysis and simulations. In fact, a comprehensive study comprising all the steps mentioned above together with a qualitative and quantitative analysis of the entire identification strategy has been lacking, and that is what we offer in this paper.

The paper is outlined as follows. In Section 2 we formulate the problem and motivate our measure of approximation. Each of the three steps in the overall identification procedure contributes to the estimation error. In Section 3 we show that the transfer function of the maximum-entropy filter, constructed from true covariances, tends to that of the true filter in $\mathcal{H}^\infty$ norm at a geometric rate determined by the largest modulus of the zeros of the true filter as the order of the maximum-entropy filter becomes large. However the order of the approximation is too high, and therefore model reduction is performed. This is studied in Section 4. A stochastic balancing procedure, based only on linear-algebra operations so that no Riccati equations need to be solved, is provided together with the analysis of the model-reduction error. Both deterministically and stochastically balanced truncation lead to good results. However, when the covariances are estimated from statistical data, stochastic model reduction is found to be superior. In particular, variances are considerably closer to the Cramér-Rao bounds. In Section 5 we state our statistical convergence theorems, proving that the total error tends to zero as the length of the data string tends to infinity, provided the degree of the AR model tends to infinity in the proper manner. In Section 6 some simulations are presented. For comparison, a simulation using stochastic subspace identification [43] is included. For clarity of exposition, all the proofs have been deferred to two appendices, Appendix A dealing with the asymptotic properties of the maximum-entropy filter, and Appendix B devoted to the statistical error analysis. Finally, in Section 7 conclusions and open questions are discussed.

2. A covariance extension strategy for time series identification

Time series identification in the form studied here amounts to estimating the matrices (A, B, C, D) in some n-dimensional linear stochastic system

$$\begin{cases} x(t+1) & = & Ax(t) + Bw(t) \\ y(t) & = & Cx(t) + Dw(t) \end{cases} \tag{2.1}$$

driven by normalized white noise $\{w(t)\}$, from a data string of observations

$$\{y_0, y_1, y_2, \cdots, y_N\} \tag{2.2}$$

of the output process $\{y(t)\}$, which here will be taken to be scalar.

The basic idea behind our approach is very simple: given estimates of a partial sequence

$$c_0, c_1, c_2, \ldots, c_\nu \tag{2.3}$$

of the covariances $c_k = \mathrm{E}\{y(t+k)y(t)\}$, which satisfies the condition that the Toeplitz matrix

$$T_{\nu+1} := \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_\nu \\ c_1 & c_0 & c_1 & \cdots & c_{\nu-1} \\ c_2 & c_1 & c_0 & \cdots & c_{\nu-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_\nu & c_{\nu-1} & c_{\nu-2} & \cdots & c_0 \end{bmatrix} \tag{2.4}$$

is positive definite, first construct a high-order model continuing (2.3) by covariance extension. This model has all the required positivity properties, but the order is too high. Then reduce the order by means of a positivity-preserving model reduction procedure to be specified below. That this simple recipe will in fact provide a good identification method is by no means a trivial matter but is based on some rather deep results, which will be presented here.

More specifically, the approach consists of three steps, for which there are several possible variants that will be discussed below. The rigorous mathematical analysis, however, will be carried out for the following procedure, for which we shall give theoretical bounds.

(i) Estimate a partial covariance sequence

$$\hat{c}_0, \hat{c}_1, \hat{c}_2, \ldots, \hat{c}_\nu \tag{2.5}$$

from the time-series data (2.2) via the ergodic estimate

$$\hat{c}_k = \frac{1}{N+1} \sum_{t=0}^{N-k} y_{t+k} y_t \quad k = 0, 1, \ldots, \nu. \tag{2.6}$$

(ii) Use the maximum entropy extension to construct an AR model with transfer function

$$\hat{W}_\nu(z) = \frac{z^\nu}{\hat{\phi}_\nu(z)}, \tag{2.7}$$

where $\hat{\phi}_\nu(z)$ is the normalized Szegö polynomial of degree ν, to be introduced in Section 3, computed from the estimated covariance data (2.5).

(iii) Determine a reduced-degree approximation $\hat{W}(z)$ of $\hat{W}_\nu(z)$ via a stochastic model reduction procedure [11] to be described in more detail in Section 4.

In this procedure, the idea is that $\nu >> n$, the order of the system to be identified, and ideally $\hat{n} := \deg \hat{W}$ equals the degree n of the true system (2.1). However, the method will produce a valid model even if this is not the case or even if there is no "true" underlying model. This is in contrast to stochastic subspace identification models, which may fail to produce any model at all [9].

There are possibilities for variations of the procedure described above. In Step (i) we could use alternative covariance estimates or Burg's estimation of Schur parameters [3], the only requirements being that the estimated Toeplitz matrix $\hat{T}_{\nu+1}$ of (2.5) is positive definite and that $\hat{c}_k \to c_k$ a.s. as $N \to \infty$. In Step (ii) we could instead use approximate covariance extension or covariance extension with prescribed zeros, for which there is now a complete parameterization [5] and an algorithm [4]. (In the latter case a zero estimator is needed; see, e.g., [15, 35].) In Step (iii) other model reduction methods could be used. For example, an important model reduction paradigm is the one based on optimal Hankel norm approximation [21].

Before proceeding along these lines we need to decide what measure of approximation to use. Suppose that there is a true underlying system (2.1) with a stable transfer function

$$W(z) = C(zI - A)^{-1}B + D, \tag{2.8}$$

of McMillan degree n. We also assume that $W(z)$ is minimum-phase so that both zeros and poles are located in the open unit disc. Then, we need to be able to measure how the estimated model, with transfer function $\hat{W}(z)$, converges to the true one as $N \to \infty$. In this paper we have chosen to use distance between $W(z)$ and $\hat{W}(z)$ in $\mathfrak{L}^\infty$ norm as a measure of proximity between the true and estimated model. From an engineering point of view this could be called *worst case* identification. The modern literature in robust control makes extensive use of the worst case philosophy; see for example [20, 52]. There are also other reasons for using the $\mathfrak{L}_\infty$, as discussed in [35].

Returning, then, to the identification approach outlined above, the estimation error can be decomposed into three parts, one corresponding to each of the steps (i), (ii) and (iii). Hence we have the error bound

$$\|W - \hat{W}\|_\infty \leq \|W - W_\nu\|_\infty + \|W_\nu - \hat{W}_\nu\|_\infty + \|\hat{W}_\nu - \hat{W}\|_\infty, \tag{2.9}$$

where W_ν is the AR model corresponding to the true covariances (2.3) and $\hat{W}_\nu$ is the one determined from the estimated covariances (2.6). To prove convergence to zero of the estimation error (2.9), we shall need to assume that W is minimum-phase, and hence $\hat{W}$ should have the same property, which moreover is desirable in many applications. Our procedure insures this.

Estimating the first term in (2.9) is a problem in stochastic partial realization theory and function theory and will be dealt with in the next section. The third term concerns model reduction which will be studied, in the particular setting required here, in Sections 4 and 5. In Section 5, finally, we consider the second term together with the overall statistical analysis.

3. Rational modeling from a long covariance sequence

Step (ii) in the identification procedure outlined in Section 2 is based on *rational covariance extension.* To understand this, let us consider the covariance extension problem from a more general point of view. Given a partial covariance sequence

$$c_0, c_1, c_2, \dots, c_\nu, \tag{3.1}$$

covariance extension amounts to finding an infinite extension $c_{\nu+1}, c_{\nu+2}, c_{\nu+3}, \dots$ of this sequence such that the function

$$V(z) := \tfrac{1}{2}c_0 + c_1 z^{-1} + c_2 z^{-2} + \dots$$

is *strictly positive real*, i.e., it is an analytic function in the complement $\mathbb{D}^c$ of the open unit disc $\mathbb{D}$, which maps $\mathbb{D}^c$ to the open right complex half-plane. Then

$$\Phi(z) := V(z) + V(z^{-1})$$

is a spectral density for a process having $c_0, c_1, \dots, c_\nu$ as its first ν covariances and which is *coercive* in the sense that

$$\Phi(e^{i\theta}) > 0 \quad \text{for all } \theta.$$

Spectral factorization is then to find a stable transfer function $W(z)$ such that

$$|W(e^{i\theta})|^2 = \Phi(e^{i\theta}).$$

In particular, we are interested in finding covariance extensions for which $V(z)$, and hence $W(z)$, have at most degree ν.

For the moment disregarding this degree condition and allowing it to be meromorphic, there is a complete parameterization of all covariance extensions, which is classical and due to Schur [45]: modulo a normalization of c_0, there is a one-one correspondence between infinite covariance sequences

$$c_0, c_1, c_2, c_3, \dots \tag{3.2}$$

and a sequence of *Schur parameters*, or *reflection coefficients*,

$$\gamma_0, \gamma_1, \gamma_2, \gamma_3, \dots, \tag{3.3}$$

with the property $|\gamma_t| < 1$ for all t. In fact, fixing the value of c_0 to one, there is a one-one correspondence between the partial sequences $1, c_1, \dots, c_m$ and $\gamma_0, \gamma_1, \dots, \gamma_{m-1}$ for each m. The Schur parameters can be determined from the covariances via the Szegö polynomials

$$\varphi_t(z) = z^t + \varphi_{t1} z^{t-1} + \dots + \varphi_{tt} \quad t = 0, 1, 2 \dots,$$

computed by means of the *Szegö-Levinson recursion*

$$\begin{bmatrix} \varphi_{t+1}(z) \\ \varphi^*_{t+1}(z) \end{bmatrix} = \begin{bmatrix} z & -\gamma_t \\ -z\gamma_t & 1 \end{bmatrix} \begin{bmatrix} \varphi_t(z) \\ \varphi^*_t(z) \end{bmatrix}; \qquad \begin{bmatrix} \varphi_0(z) \\ \varphi^*_0(z) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{3.4}$$

where

$$\varphi^*_t(z) := z^t \varphi_t(z^{-1})$$

is the reciprocal polynomial of $\varphi_t(z)$, and the Schur parameters are computed via

$$\begin{aligned} \gamma_t &= \tfrac{1}{r_t} \textstyle\sum_{j=0}^{t} \varphi_{t,t-j} c_{j+1} \\ r_{t+1} &= r_t(1 - |\gamma_t|^2), \quad r_0 = c_0. \end{aligned} \tag{3.5}$$

Hence $\gamma_t = -\varphi_{t+1}(0)$, a fact that we shall use below.

In the problem to find a covariance extension for (3.1), therefore, $\gamma_0, \gamma_1, \dots, \gamma_{\nu-1}$ are fixed and the infinite continuation $\gamma_\nu, \gamma_{\nu+1}, \dots$ can be chosen freely. In particular, if we take $\gamma_t = 0$ for $t = \nu, \nu+1, \nu+2, \dots$. We obtain the *maximum entropy solution*

$$W_\nu(z) = \frac{z^\nu}{\phi_\nu(z)}, \tag{3.6}$$

where $\phi_\nu(z)$ is the normalized Szegö polynomial

$$\phi_\nu(z) := \frac{1}{\sqrt{r_\nu}}\varphi_\nu(z). \tag{3.7}$$

Thus, in this particular case, the solution to the covariance extension problem turns out to be rational of degree at most ν as required. In general, the Schur parameterization is not suitable for characterizing such rationality, but other parameterizations are needed. In fact, it has recently been shown [5] that there is *exactly* one such solution for each choice of zeros of $W_\nu(z)$, thus proving a long-standing conjecture by Georgiou [18], who had established existence. Nevertheless, as we shall see next, rationality implies that the Schur parameters tend geometrically to zero, provided $W(z)$ has no zeros on the unit circle.

In this section we shall demonstrate that the rational transfer function (2.8) can be approximated arbitrarily closely in $\mathcal{L}^\infty$ by the transfer function $W_\nu(z)$ of a maximum entropy filter for sufficiently large ν and that this ν depends on the maximum modulus of the zeros of $W(z)$. We shall first present a heuristic argument in support of this conclusion.

To this end, let (3.2) be the *infinite* covariance sequence of the output process y in (2.1), and let (3.3) be the corresponding sequence of Schur parameters determined via the Szegö-Levinson algorithm presented above. Then we have the following special case of Corollary 2.1 in [5].

Lemma 3.1. *Let the spectral density*

$$\Phi(e^{i\theta}) = |W(e^{i\theta})|^2 \tag{3.8}$$

be coercive in the sense that it is positive for all θ and let (3.3) *be the corresponding infinite sequence of Schur parameters. Moreover, let $\gamma \in (0,1)$ be greater than the maximum of the moduli of the zeros of $W(z)$. Then*

$$|\gamma_t| = O(\gamma^t), \tag{3.9}$$

i.e., $|\gamma_t| \le M\gamma^t$ for some $M \in \mathbb{R}$ and for sufficiently large t.

Remark 3.2. Since (3.9) holds for *all* γ greater than the the maximum of the moduli of the zeros of $W(z)$, we have in fact that $|\gamma_t| = o(\gamma^t)$, i.e., $\lim_{t\to\infty} |\gamma_t|\gamma^{-t} = 0$.

For ease of reference, we shall henceforth refer to this property as *geometric convergence rate* of the Schur parameters. The proof of Lemma 3.1 is based on the analysis of certain fast algorithms for Kalman filtering [6].

Remark 3.3. Coercivity is essential for the validity of Lemma 3.1. For example, the spectral density

$$\Phi(z) = -\frac{z(z-1)^2}{(z^2+z+2)(2z^2+z+1)}$$

is rational but not coercive, since it has a double zero at $z = 1$. Its Schur parameters are seen to be $-1/2, -2/3, -2/5, -2/7, -2/9, -2/11, \dots$, which tend to zero but not geometrically. On the other hand, there are coercive, analytic but nonrational models which also exhibit geometric convergence rate. A classical example [23] is obtained when $c_k = \theta^{k^2}$ for some $\theta \in (-1, 1)$. The Schur parameters in this case form an exact geometric sequence, $\gamma_k = (-\theta)^{k+1}$, $k \geq 0$.

Lemma 3.1 implies that, for a sufficiently large ν which depends on γ, the Schur parameters γ_t are close to zero for $t = \nu, \nu + 1, \nu + 2, \dots$. But, the Schur parameters of W_ν are *exactly* zero for $t = \nu, \nu + 1, \nu + 2, \dots$, and hence geometric convergence would insure that W_ν is a good approximation of $W(z)$ for sufficiently large ν. We shall prove that this is indeed the case.

Theorem 3.4. *Suppose $W(z)$ is the minimum-phase spectral factor of a coercive spectral density* (3.8)*, and let $\gamma \in (0, 1)$ be greater than the maximum of the moduli of the zeros of $W(z)$. Then*

$$\lim_{\nu\to\infty} \|W_\nu - W\|_\infty = 0, \tag{3.10}$$

and the convergence is geometric. More precisely, there is a constant M such that

$$\|W_\nu - W\|_\infty \leq M\gamma^\nu. \tag{3.11}$$

The proof of Theorem 3.4, which is given in Appendix A, amounts to first showing that

$$\lim_{\nu\to\infty} \|W_\nu^{-1} - W^{-1}\|_\infty = 0. \tag{3.12}$$

A very nice result of this type has already been given in [7]. In Appendix A we give an alternative proof of this fact based on Szegö theory, and also show that the convergence is geometric. In fact, we can choose γ arbitrarily close to the maximum modulus of the zeros of W.

However, as we shall see next, we can actually prove more. To this end, let us first observe that, since W_ν^{-1} and W^{-1} have their poles in the open unit disc $\mathbb{D}$ and thus are bounded and analytic in the complement $\mathbb{D}^c$ of $\mathbb{D}$, they belong to the Hardy space $\mathcal{H}^\infty_-$ of functions which are analytic and bounded in $\{z \in \mathbb{C} \mid |z| > 1\}$. Hence the convergence (3.12) is in $\mathcal{H}^\infty_-$, and

$$z^{-\nu}\phi_\nu(z) \to W^{-1}(z) \tag{3.13}$$

uniformly in each compact subset of $\mathbb{D}^c$. Now, W^{-1} is analytic in $\{z \in \mathbb{C} \mid |z| \geq \gamma\}$, a region that is strictly larger than $\mathbb{D}^c$. This in itself of course does not insure that the convergence (3.13) extends to this larger region. In fact, even if $z^{-\nu}\phi_\nu(z)$ did converge in $\{z \in \mathbb{C} \mid \gamma \leq |z| \leq 1\}$, it could fail to converge to $W^{-1}(z)$ there. The fact that it really does converge uniformly to this limit is another consequence of Lemma 3.1. We state this as a separate result, to be proven in Appendix A. A method for computing the maximum of the modulus of the zeros of W from statistical data, and hence an estimate of the convergence rate γ, is given in [35].

Theorem 3.5. *Suppose $W(z)$ is a minimum-phase rational function having all its poles in the open unit disc $\mathbb{D}$ and all its zeros in*

$$\mathbb{D}_\rho := \{z \in \mathbb{C} \mid |z| \leq \rho\} \subset \mathbb{D} \quad \textit{where } 0 < \rho < 1,$$

and let $\{\phi_\nu(z)\}_0^\infty$ be the normalized Szegö polynomial (3.7) determined from the covariances in the spectral density

$$|W(e^{i\theta})|^2 = c_0 + 2\sum_{k=1}^{\infty} c_k \cos k\theta.$$

Then, as $\nu \to \infty$, $z^{-\nu}\phi_\nu(z) \to W^{-1}(z)$ uniformly in every compact subset of $\mathbb{D}_\rho^c := \{z \in \mathbb{C} \mid |z| > \rho\}$, the complement of $\mathbb{D}_\rho$.

Lemma 3.1 and Theorem 3.5 give us some interesting information about the asymptotic distribution of the roots of $\phi_\nu(z)$ and hence of the poles of the high-order AR model with transfer function $W_\nu(z)$. It is known that, if the Toeplitz matrix $T_{\nu+1}$ is positive definite, all roots of $\phi_\nu(z)$ are located in the open unit disc $\mathbb{D}$, but little has been reported in the literature on their behavior as $\nu \to \infty$. This behavior is illustrated in Figure 3.1.

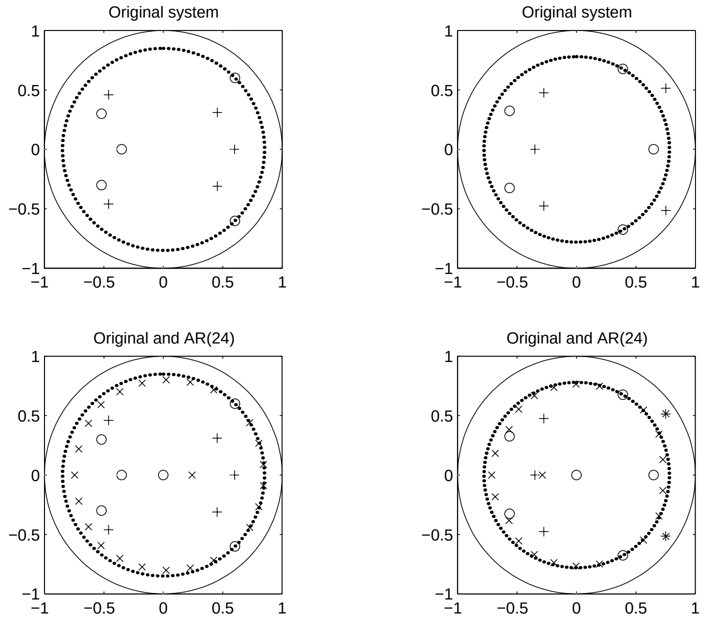

Figure 3.1: Distribution of zeros of $\phi_\nu(z)$.

The top two diagrams show the zero-pole positions, within the boundaries of the unit circle, of two minimum phase spectral factors W, both of degree five. Also indicated is a circle of radius equal to the maximum modulus of the zeros of these spectral factors. The little circles "∘" represent zeros and the "+" sign represent poles. The lower two figures show the position of zeros and poles of the original system superposed with those obtained from an AR(24) model constructed from the exact covariance sequence. The poles of the latter models are indicated with "×".

The left part of Figure 3.1 illustrates what may happen if all the poles of $W(z)$ are located in $\{z \in \mathbb{C} \mid |z| < \rho\}$, where ρ is chosen to be the maximum of the moduli of the zeros of $W(z)$. The roots of $\phi_\nu(z)$ tend to cluster inside a circle of radius ρ as $\nu \to \infty$. This phenomenon is in a sense predictable, since the constant term of the Szegö polynomials is $\varphi_{n+1}(0) = -\gamma_n$, which equals the product of the roots and, by Lemma 3.1, decays at a geometric rate, which can be chosen arbitrarily close to ρ. This does not preclude that other types of crowns may occur, because subsequences of $\{\gamma_n\}$ could decay faster than the overall rate γ, as follows from [5]. Very general statements about the distribution of zeros of orthogonal polynomials, derived with the help of potential-theoretic methods, can be found in [37, 27].

To the right in Figure 3.1 we see what happens in the case that W has poles with moduli larger than ρ. Then, for ν sufficiently large, the normalized Szegö polynomial $\phi_\nu(z)$ has roots in $\{z \in \mathbb{C} \mid \rho \leq |z| < 1\}$, but exactly as many as the poles of W in this region and approximately at the same place as these. This is of course due to the uniform convergence of $z^{-\nu}\phi_\nu(z)$ to $W^{-1}(z)$ in every compact subset of $\mathbb{D}_\rho^c$. The other roots of $\phi_\nu(z)$ behave exactly as in the previous case and tend to accumulate in a crown inside and very close to the circle $\{z \in \mathbb{C} \mid |z| = \rho\}$.

We have thus constructed an $\mathcal{H}_-^\infty$ approximation W_ν of W which can be made arbitrarily good by choosing ν sufficiently large. However, W_ν will have much larger degree and, except for the poles outside the circle $\{z \in \mathbb{C} \mid |z| = \rho\}$, a completely different zero-pole pattern. We shall rectify this situation by model reduction. In fact, for the moment considering the perfect modeling problem to identify the rational transfer function (2.8) given an exact partial covariance sequence (3.1), the last step in our procedure consists in approximating W_ν by a rational function W_{red} of smaller degree, ideally of the same degree as W.

The simplest model reduction procedure is *deterministically balanced truncation* (DBT), first introduced by Moore [38]. Though easy to implement, it may fail to yield a minimum-phase approximation, a requirement which is important in certain contexts. For this and, more importantly, for statistical reasons to be reported in Section 5, we prefer another model reduction procedure, namely *stochastically balanced truncation* (SBT), first introduced by Desai and Pal [10], which is based on a different balancing strategy to be explained in detail in Section 4.

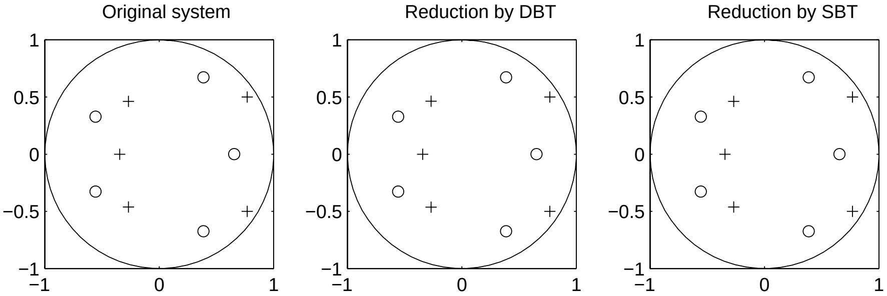

Figure 3.2: Zero-pole pattern of $W(z)$ and $W_{\text{red}}(z)$ for different model reduction methods.

Let us now return to the example depicted to the right in Figure 3.1. This fifth-order model has first been approximated by W_ν of degree $\nu = 24$, producing the pole-zero pattern in the lower right corner of Figure 3.1. Figure 3.2 illustrates what happens when the model is reduced back to order five by either deterministically balanced truncation or stochastically balanced truncation. The zeros are denoted by "$\circ$" and the poles by "+". Both reduction procedures give good approximations when applied to *exact* covariance data. However, as we shall see in Section 5, the advantages of SBT becomes apparent when applied to statistical data. Also, as explained in Remark 4.5, there are theoretical reasons to prefer stochastic model reduction.

4. Model reduction

In the present setting, model reduction amounts to replacing a stochastic system (2.1) of dimension ν by one of some dimension $r < \nu$ in such a way that most of its statistical features are retained. In particular, we want to remove the part of the system which corresponds to the weakest correlation between past and future. This idea can be formalized in the following way.

Basic concepts. In the Hilbert space generated by the random variables $\{y(t) \mid -\infty < t < \infty\}$ in the inner product $\langle u, v\rangle = \mathrm{E}\{uv\}$, let H^- be the subspace generated by the *past*, i.e., $\{y(t) \mid t < 0\}$, and H^+ that generated by the *future* $\{y(t) \mid t \geq 0\}$. Consider the Hankel operator $\mathbb{H} : H^+ \to H^-$ and its adjoint $\mathbb{H}^* : H^- \to H^+$ defined as

$$\mathbb{H} = E^{H^-}|_{H^+} \quad \text{and} \quad \mathbb{H}^* = E^{H^+}|_{H^-}, \tag{4.1}$$

where E^{H^-} denotes orthogonal projection onto the past space H^-. More precisely, $\mathbb{H}$ sends $\xi \in H^+$ to $E^{H^-}\xi \in H^-$ and $\mathbb{H}^*$ sends $\eta \in H^-$ to $E^{H^+}\eta \in H^+$. Since the process y is the output of a minimal stochastic system of dimension ν, rank $\mathbb{H} = \nu$ by Kronecker's Theorem [56], and hence $\mathbb{H}$ has exactly ν singular values, $\sigma_1, \sigma_2, \dots, \sigma_\nu$, which are positive, as usually listed so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_\nu$. These singular values are the *canonical correlation coefficients* and hence the cosines of the angles between the principal directions of the past space H^- and the future space H^+. They are therefore less than one, and the part of the stochastic system corresponding to singular values which are close to zero have a weak coupling between past and future, i.e., the corresponding subspaces are almost orthogonal. The basic idea of stochastic model reduction is to truncate the system so that this part is removed.

To each singular value σ_k there is an associated *Schmidt pair* (ξ_k, η_k) with $\xi_k \in H^+$ and $\eta_k \in H^-$ such that

$$\mathbb{H}\xi_k = \sigma_k \eta_k, \qquad \mathbb{H}^*\eta_k = \sigma_k \xi_k,$$

and such that the sequences $\xi_1, \xi_2, \xi_3, \dots$ and $\eta_1, \eta_2, \eta_3, \dots$ of *singular vectors* are orthonormal. The singular vectors corresponding to nonzero singular values span the *predictor spaces*

$$X_- := \operatorname{span}\{\eta_1, \eta_2, \dots, \eta_\nu\}, \quad X_+ := \operatorname{span}\{\xi_1, \xi_2, \dots, \xi_\nu\}.$$

Clearly, $X_- \subset H^-$ and $X_+ \subset H^+$.

The process y has one representation (2.1) for each minimal spectral factor W, having W as its transfer function. Such representations are called *minimal stochastic realizations* and the corresponding subspaces $X := \{a'x(0) \mid a \in \mathbb{R}^\nu\}$ are called *splitting subspaces* [30, 31]. In particular, X_- is the splitting subspace of the stochastic realization

$$\begin{cases} x_-(t+1) &= Ax_-(t) + B_-w_-(t) \\ y(t) &= Cx_-(t) + D_-w_-(t) \end{cases} \tag{4.2}$$

with the transfer function $W_-(z)$, the minimum-phase spectral factor; and X_+ is the splitting subspace of

$$\begin{cases} x_+(t+1) &= Ax_+(t) + B_+w_+(t) \\ y(t) &= Cx_+(t) + D_+w_+(t) \end{cases} \tag{4.3}$$

with transfer function $W_+(z)$, the *maximum-phase spectral factor*, having all its zeros in $\mathbb{D}^c$. Note that A and C are the same in both realizations (uniform choice of bases).

Each realization has a counterpart which evolves backwards in time and has the same splitting subspace. For example, the backward realization of X_+,

$$\begin{cases} \bar{x}_+(t-1) &= A'\bar{x}_+(t) + \bar{B}_+\bar{w}_+(t) \\ y(t) &= \bar{C}\bar{x}_+(t) + \bar{D}_+\bar{w}_+(t) \end{cases}, \tag{4.4}$$

has transfer function $\bar{W}_+(z)$, the *coanalytic* minimum-phase spectral factor, having all its poles and zeros in $\mathbb{D}^c$. In the present case with scalar y, we have $\bar{W}_+(z) = W_-(z^{-1})$.

Now, in order to identify the part of the system which has the weakest coupling between past and future, and hence will be removed in the model reduction, we need to balance the system in the sense of Desai and Pal, as we shall explain next. To this end, we make a coordinate transformation

$$(A, C, \bar{C}) \to (SAS^{-1}, CS^{-1}, \bar{C}S'), \tag{4.5}$$

in the minimal realization of

$$V(z) = C(zI - A)^{-1}\bar{C}' + \frac{1}{2}c_0, \tag{4.6}$$

the strictly positive real part of the spectral density of y, so that the state covariances $P_- := \mathrm{E}\{x_-(t)x_-(t)'\}$ and $\bar{P}_+ = \mathrm{E}\{\bar{x}_+(t)\bar{x}_+(t)'\}$ coincide with the diagonal $\nu \times \nu$ matrix Σ of nonzero canonical correlation coefficients, i.e.,

$$P_- = \bar{P}_+ = \Sigma := \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_\nu). \tag{4.7}$$

This is done by choosing S so that $Sx_-(0) = \Sigma^{\frac{1}{2}}\eta$, where $\eta = (\eta_1, \eta_2, \dots, \eta_\nu)'$, and $(S')^{-1}\bar{x}_+(0) = \Sigma^{\frac{1}{2}}\xi$, where $\xi := (\xi_1, \xi_2, \dots, \xi_\nu)'$.

To compute the canonical correlation coefficients, we first observe that the eigenvalues of the product $P_-\bar{P}_+$ are precisely the squares of the canonical correlation coefficients, i.e.,

$$\lambda(P_-\bar{P}_+) = \lambda(P_-P_+^{-1}) = \{\sigma_1^2, \sigma_2^2, \dots, \sigma_\nu^2\}, \tag{4.8}$$

where we have used the fact that the state covariance of (4.3) is $P_+ = \bar{P}_+^{-1}$. Therefore the canonical correlation coefficients can then be determined via (4.8) by solving the Lyapunov equations

$$P_- = AP_-A' + B_-B_-' \quad \text{and} \quad P_+ = AP_+A' + B_+B_+'. \tag{4.9}$$

The point is now to identify the canonical correlation coefficients $\sigma_1, \sigma_2, \dots, \sigma_r$ corresponding to the part of the system one wants to keep. The part corresponding to $\sigma_{r+1}, \sigma_{r+2}, \dots, \sigma_\nu$ will be disposed of. This amounts to partitioning Σ as

$$\Sigma = \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix}, \tag{4.10}$$

where Σ_1 is $r \times r$.

In order to reduce model (2.1) we make the coordinate transformation $(A, B, C) \to (SAS^{-1}, SB, CS^{-1})$, with the same balancing transformation S. Then, partition the *new* triplet (A, B, C) conformally with (4.10) as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}, \tag{4.11}$$

and perform a *principal subsystem truncation* to obtain the transfer function of a reduced-order system

$$W_{\text{red}}(z) = C_1(zI - A_{11})^{-1}B_1 + D \tag{4.12}$$

of degree r. If Σ_2 is close to zero, while Σ_1 is not, the rank of $\mathbb{H}$ is close to r, and the discarded part of the system gives a negligible contribution to y.

Stochastically balanced truncation of AR models. We now consider the problem of stochastically balanced truncation of the maximum entropy filter

$$W_-(z) := W_\nu(z) = \frac{\sqrt{r_\nu} z^\nu}{\varphi_\nu(z)} \tag{4.13}$$

of order ν, which, for the moment we denote $W_-(z)$ to emphasize its character as the minimum-phase spectral factor of the spectral density

$$\frac{r_\nu}{\varphi_\nu(z)\varphi_\nu(z^{-1})}.$$

Remark 4.1. Without loss of generality we assume that $\varphi_\nu(0) \neq 0$ so that no cancellations occur; otherwise, we may choose a smaller ν for which this condition holds. In fact, $\varphi_\nu(0) = \gamma_{\nu-1}$, and if $\gamma_{\nu-p} = \gamma_{\nu-p+1} = \dots = \gamma_{\nu-1} = 0$ and $\gamma_{\nu-p-1} \neq 0$ for some $p = 1, 2, \dots, \nu$, then $\varphi_\nu(z) = z^{\nu-p}\varphi_{\nu-p}(z)$ by (3.4), and hence (3.6) can be replaced by $W_\nu(z) = W_{\nu-p}(z)$, and for $W_{\nu-p}(z)$ the required condition holds.

The maximum-phase spectral factor $W_+(z)$ has all its zeros at infinity, and hence

$$W_+(z) = h'(zI - F)^{-1}b = \frac{\sqrt{r_\nu}}{\varphi_\nu(z)}, \tag{4.14}$$

where (F, b, g) is the (observable) canonical form

$$F = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\varphi_{\nu\nu} & -\varphi_{\nu,\nu-1} & \cdots & -\varphi_{\nu 1} \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sqrt{r_\nu} \end{bmatrix}, \quad h = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{4.15}$$

$\varphi_{\nu 1}, \varphi_{\nu 2}, \dots, \varphi_{\nu\nu}$ being the coefficients of the Szegö polynomial $\varphi_\nu(z)$. In this basis, it follows from (4.9) that

$$\begin{aligned}[P_+]_{jk} &= \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}(e^{i\theta}I - A)^{-1}bb'(e^{-i\theta}I - A')^{-1}d\theta\right]_{jk}\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}e^{-(j-k)i\theta}\frac{r_\nu}{\varphi_\nu(e^{i\theta})\varphi_\nu(e^{-i\theta})}d\theta = c_{j-k},\end{aligned}$$

and hence $P_+ = T_\nu$. It is well-known and easy to prove that $\Phi_\nu' T_\nu \Phi_\nu = R_\nu$, where

$$\Phi_{\nu+1} = \begin{bmatrix} \varphi_{\nu\nu} & \varphi_{\nu-1,\nu-1} & \varphi_{\nu-2,\nu-2} & \cdots & 1\\ \vdots & \vdots & \vdots & & \\ \varphi_{\nu 2} & \varphi_{\nu-1,1} & 1 & & \\ \varphi_{\nu 1} & 1 & & & \\ 1 & & & & \end{bmatrix} \quad \text{and} \quad R_\nu = \begin{bmatrix} r_{\nu-1} & & & \\ & r_{\nu-2} & & \\ & & \ddots & \\ & & & r_0\end{bmatrix}, \tag{4.16}$$

and consequently

$$\bar{P}_+ = T_\nu^{-1} = \Phi_\nu R_\nu^{-1}\Phi_\nu'. \tag{4.17}$$

It remains to determine P_-. From (4.13) is easy to see that

$$W_-(z) = -\boldsymbol{\varphi}_\nu'(zI - F)^{-1}b + \sqrt{r_\nu}, \tag{4.18}$$

where

$$\boldsymbol{\varphi}_\nu' := \begin{bmatrix}\varphi_{\nu\nu} & \varphi_{\nu,\nu-1} & \cdots & \varphi_{\nu 1}\end{bmatrix}, \tag{4.19}$$

but, in order to determine P_-, this realization needs to be transformed so that the A and C matrices are the same as in (4.14) (uniform choice of bases). More precisely, we need to perform a transformation

$$(F, b, -\boldsymbol{\varphi}_\nu') \to (QFQ^{-1}, Qb, -\boldsymbol{\varphi}_\nu' Q^{-1}) =: (F, Qb, h').$$

Then P_- is the solution of the Lyapunov equation $P_- = FP_-F' + Qbb'Q'$, and therefore, since $T_\nu = FT_\nu F' + bb'$ and $QF = FQ$ and consequently

$$QT_\nu Q' = FQT_\nu Q'F' + Qbb'Q',$$

we have

$$P_- = QT_\nu Q'. \tag{4.20}$$

To determine Q, notice that $-\boldsymbol{\varphi}_\nu' = h'Q$ and $QF = FQ$ to form

$$\begin{bmatrix} -\boldsymbol{\varphi}_\nu' \\ -\boldsymbol{\varphi}_\nu' F \\ \vdots \\ -\boldsymbol{\varphi}_\nu' F^{\nu-1}\end{bmatrix} = \begin{bmatrix} h' \\ h'F \\ \vdots \\ h'F^{\nu-1}\end{bmatrix} Q = Q. \tag{4.21}$$

Next, define the symmetric matrix

$$M := R_\nu^{-1/2}\Phi_\nu' QT_\nu Q'\Phi_\nu R_\nu^{-1/2}. \tag{4.22}$$

In view of (4.20) and (4.17), $\det(zI - M) = \det(zI - P_-\bar{P}_+)$, and hence, by (4.8), M has the eigenvalues $\sigma_1^2, \sigma_2^2, \dots, \sigma_\nu^2$, and the singular-value decomposition

$$M = U\Sigma^2 U', \tag{4.23}$$

where $U'U = UU' = I$. It is then well-known and simple to check that

$$S := \Sigma^{-1/2}U'R_\nu^{-1/2}\Phi_\nu' \tag{4.24}$$

is the required balancing transformation (4.5) such that $SP_-S' = (S')^{-1}\bar{P}_+S^{-1} = \Sigma$.

Proposition 4.2. *Given the partial covariance sequence*

$$c_k = E\{y(t+k)y(t)\}, \quad k = 0, 1, \dots, \nu,$$

let $\varphi_1(z), \varphi_2(z), \dots, \varphi_\nu(z)$ *and* $r_0, r_1, \dots, r_\nu$ *be the corresponding Szegö polynomials and error variances. Supposing that* $\gamma_{\nu-1} = -\varphi_\nu(0) \neq 0$, *let* (F, b, h) *be given by* (4.15), R_ν *and* Φ_ν *by* (4.16) *and* Q *by* (4.21). *Moreover, let* U *and* Σ *be defined by the singular value decomposition* (4.23) *of* (4.22). *Then, the canonical correlation coefficients* $\sigma_1, \sigma_2, \dots, \sigma_\nu$ *are the diagonal elements of* Σ, *as described in* (4.7), *and the stochastically balanced realization of* W_ν *is given by*

$$(A, B, C, D) = (SFS^{-1}, SQb, h'S^{-1}, \sqrt{r_\nu}), \tag{4.25}$$

where S *is defined by* (4.24).

Stochastic balanced truncation (SBT) is then performed as described above. Principal subsystem truncation (4.11) is executed on the balanced realization (4.25) to yield a transfer function (4.12) of degree r. Bounds can be derived for the approximation error, and the procedure can be designed so that it preserves the minimum-phase property. In fact, we have the following result, the proof of which is given in Section A.

Theorem 4.3. *Let* W_{red} *be the SBT approximation of degree* r *of* W_ν, *and set*

$$\epsilon := 2\sum_{k=r+1}^{\nu} \frac{\sigma_k}{1-\sigma_k} \quad \text{and} \quad \kappa := \sqrt{c_0}\prod_{k=0}^{\nu-1}\sqrt{\frac{1+|\gamma_k|}{1-|\gamma_k|}}, \tag{4.26}$$

where $\gamma_0, \gamma_1, \dots, \gamma_{\nu-1}$ *are the Schur parameters of* $c_0, c_1, c_2, \dots, c_\nu$. *Then*

$$c_0(1-\epsilon)\kappa^{-1} \leq |W_{red}(e^{i\theta})| \leq (1+\epsilon)\kappa \quad \text{for all } \theta, \tag{4.27}$$

and, if $\epsilon < 1$, W_{red} *is minimum phase. Finally, the approximation error has the bound*

$$\|W_\nu - W_{red}\|_\infty \leq \epsilon\kappa. \tag{4.28}$$

A properly executed SBT procedure should imply that the canonical correlation coefficients $\sigma_{r+1}, \dots, \sigma_\nu$, and hence ϵ, are close to zero, insuring the minimum-phase condition.

Remark 4.4. Stochastic model reduction can also be carried out by instead performing principal subsystem truncation on $(A, C, \bar{C})$ in $V_\nu(z) = C(zI - A)^{-1}\bar{C} + \frac{1}{2}c_0$, where A and C are given by (4.25) and $\bar{C}' = S(c_1, c_2, \dots, c_n)$. It was shown in [32] that this preserves the necessary positivity, i.e., V_{red} is positive real. Finally, the spectral density $\Phi_{\text{red}}(z) := V_{\text{red}}(z) + V_{\text{red}}(z^{-1})$ is factorized to yield a minimum-phase spectral factor $\tilde{W}$. This is in a sense a more natural procedure, but we do not know of any error bound for it. Statistically it behaves essentially as SBT, and for small Σ_2 it yields almost the same result. In fact, it is shown in [53], that $|\tilde{W}(e^{i\theta})|^2 = |W_{\text{red}}(e^{i\theta})|^2 + H(e^{i\theta})\Sigma_2 H(e^{-i\theta})$, where $H(z) = C_1(zI - A_{11})^{-1}A_{12}$.

Remark 4.5. There are good reasons to prefer stochastic over deterministic model reduction, as seen from the following heuristics. In fact, it can be seen that

$$V_\nu(z) = \frac{c_0}{2}\frac{\psi_\nu(z)}{\varphi_\nu(z)}, \tag{4.29}$$

where $\psi_\nu(z)$ is the Szegö polynomial of the *second kind* (obtained by exchanging $-\gamma_t$ for γ_t in the recursion (3.4)). Now, the matrix representation of the Hankel operator $\mathbb{H}$ in the innovation bases of the past and the future, provided by w_- and $\bar{w}_+$ respectively, is given by $L^{-1}\mathcal{H}(L^{-1})'$, where $\mathcal{H}$ is the infinite Hankel matrix of the sequence $c_1, c_2, c_3, \dots$ and L is the lower triangular Cholesky factor of the Toeplitz matrix T_∞; see, e.g., [32, p. 714]. It is easy to see that $\psi_\nu(z)$ has the same asymptotic behavior as $\varphi_\nu(z)$, i.e., the roots tend to cluster uniformly inside the circle $z = \rho$ as $\nu \to \infty$, and hence these roots are close to canceling in (4.29). Consequently, the corresponding Hankel matrix $\mathcal{H}$ is close to having low rank. This massive "almost cancellation" does not occur in $W_\nu(z)$, and hence the corresponding infinite Hankel matrix, constructed from the Laurent coefficients of $W_\nu(z)$, may have a less distinct separation between Σ_1 and Σ_2. On the other hand, since the Schur parameters tend geometrically to zero, the lower part of L tends to the identity, and hence the asymptotic behavior of the canonical correlation coefficients is very much like that of the singular values of $\mathcal{H}$. Therefore we may expect SBT to have better statistical behavior than DBT. In Section 6 we shall see that this is the case.

5. Identification from statistical data

We now return to our original problem of time series identification: Given a data string (2.2) of observations of the output process y of some n-dimensional linear stochastic system (2.1) with minimum-phase transfer function $W(z)$, given by (2.8), find an estimate $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ of the matrices (A, B, C, D).

The identification method proceeds as follows. Given the covariance estimates (2.5), we compute the corresponding maximum entropy filter (2.7), a balanced realization (4.25), and the canonical correlation coefficients

$$\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3, \dots, \hat{\sigma}_\nu, \tag{5.1}$$

determined as in Proposition 4.2 from the covariance estimates $\hat{c}_0, \hat{c}_1, \dots, \hat{c}_\nu$.

Based on (5.1), choose an integer $\hat{n}$ such that $\hat{\sigma}_{\hat{n}+1}, \hat{\sigma}_{\hat{n}+2}, \dots, \hat{\sigma}_\nu$ are close to zero or at least distinctively smaller than $\hat{\sigma}_1, \hat{\sigma}_2, \dots, \hat{\sigma}_{\hat{n}}$. Then, the balanced realization (4.25) is truncated accordingly as in (4.11) to yield a $\hat{n}$-dimensional triplet (A_{11}, B_1, C_1) and a transfer function

$$\hat{W}(z) = C_1(zI - A_{11})^{-1}B_1 + D. \tag{5.2}$$

Then, (A_{11}, B_1, C_1, D) is the required estimate $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$.

As pointed out in Section 2, we have a bound

$$\|W - \hat{W}\|_\infty \leq \|W - W_\nu\|_\infty + \|W_\nu - \hat{W}_\nu\|_\infty + \|\hat{W}_\nu - \hat{W}\|_\infty, \tag{5.3}$$

for the estimation error. As seen from Theorem 3.4, the first term $\|W - W_\nu\|_\infty$, which does not depend on the statistical data (2.2) but only on the underlying system (2.1), tends to zero geometrically with a rate $\gamma \in (0, 1)$ as $\nu \to \infty$. The other two terms depend on the data (2.2), and here N must grow at a faster rate than ν. In fact, we shall assume that

$$\nu = \nu(N) = O(\log N), \tag{5.4}$$

which in particular requires that $\lim_{N\to\infty} \frac{\nu}{N} = 0$. We also need to assume that the white noise process in (2.1) satisfies a mild technical condition, namely

$$\mathrm{E}\{w(t)^4\} < \infty. \tag{5.5}$$

This condition is, of course, satisfied if w is Gaussian.

Next, we present our main convergence theorem.

Theorem 5.1. *Suppose y is the output process of a system* (2.1)*, having a minimum-phase transfer function W, and driven by a white noise with the property* (5.5)*. Then, to each length N of the data string* (2.2)*, there is a $\nu(N)$, tending to infinity with N at the rate* (5.4)*, such that any sequence of estimated transfer functions $\hat{W}$ of fixed degree $\hat{n} \geq n$, determined, for each N and corresponding $\nu = \nu(N)$, by the procedure described above, satisfies*

$$\|W - \hat{W}\|_\infty \to 0$$

almost surely as $\nu(N) \to \infty$. For sufficiently large $\nu(N)$, the transfer function $\hat{W}$ has minimum phase.

We have already proven that the first term in (5.3) tends to zero, so Theorem 5.1 follows from the next two theorems, each corresponding to one of the remaining terms in (5.3). As for the second term, we have the following result, the proof of which is deferred to Appendix B.

Theorem 5.2. *Suppose the system* (2.1) *satisfies the conditions of Theorem* 5.1*. Let W_ν be the maximum-entropy filter* (3.6) *determined from the partial covariance sequence* (3.1) *of y and let $\hat{W}_\nu$ be the corresponding function determined from the ergodic estimates* (2.5)*. Then, if $\nu(N)$ is defined as in Theorem* 5.1*,*

$$\|W_{\nu(N)} - \hat{W}_{\nu(N)}\|_\infty \to 0$$

almost surely as $\nu(N) \to \infty$.

There are several results of this type in the literature [2, 36, 7, 33]. In particular, Berk [2] proved that, provided $\frac{\nu^3}{N} \to 0$ as $N \to \infty$ and Φ is coercive (i.e. positive on the unit circle), the estimated AR spectral density $\widehat{\Phi}(e^{i\theta}) \to \Phi(e^{i\theta})$ in probability. Under the same hypotheses, Caines and Baykal-Gürsoy [7] showed that if $N \geq \nu^{5+\eta}$ for some $\eta > 0$, then $\|\hat{W}_\nu^{-1} - W^{-1}\|_\infty \to 0$ almost surely as $\nu \to \infty$. However, in both cases, ergodic estimates are used which are not quite the same as (2.5).

Finally, we consider the last term in (5.3). The proof of the following theorem is given in Appendix B.

Theorem 5.3. *Suppose the system* (2.1) *and the function $\nu(N)$ are defined as in Theorem* 5.1. *Moreover, for each N, let $\hat{W}_{\nu(N)}$ be defined as in Theorem* 5.2 *and $\hat{W}$ as in Theorem* 5.1. *Then, for sufficiently large $\nu(N)$, $\hat{W}$ has minimum phase, and*

$$\|\hat{W}_{\nu(N)} - \hat{W}\|_\infty \to 0$$

almost surely as $\nu(N) \to \infty$.

6. Simulations

Performing model reduction on $\hat{W}_\nu$, rather than on the maximum-entropy filter of exact covariance data as in Section 3, the advantage of stochastically balanced truncation becomes apparent. First stochastically balanced truncation allows for easier and more accurate order determination, as the heuristics of Remark 4.5 suggest. There are also alternative order determination statistical tests based on the canonical correlation coefficients [17, 26, 46]. But, even more importantly, there is less bias, and the error variances are closer to the Cramér-Rao bound.

Since we are approximating rational models with AR models the method will be biased for finite amount of data, unless the model generating the data really is an AR model. The consistency result given in Theorem 5.1 implies that the method is asymptotically unbiased and therefore we consider the Cramér-Rao bound for unbiased methods; see [44, pp. 137–138]. The Cramér-Rao bound for biased estimation requires knowledge about the bias as a function of the parameter to be estimated. As already mentioned, the method will be unbiased and even statistically efficient for Gaussian AR processes if the model reduction step is omitted. Despite the fact that an algorithm based on covariance estimates (2.6) is not asymptotically efficient for general ARMA models [44, p. 144], our method can be used to provide a starting guess for other algorithms, for example the maximum likelihood method.

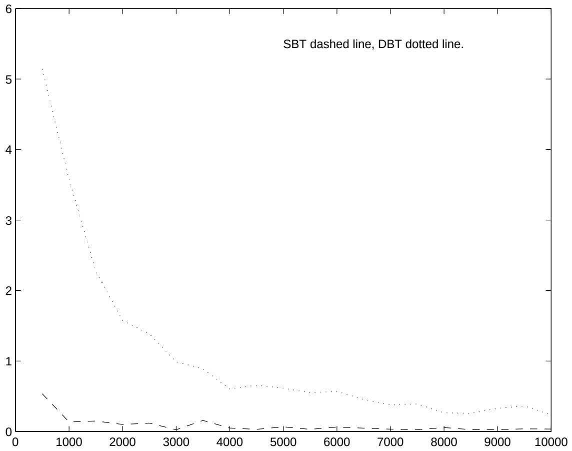

Figure 6.1: Biases of the identification estimates.

To illustrate our procedure, let us consider data generated by passing white noise through a "true system" with transfer function

$$W(z) = \frac{z^5 - 0.0550z^4 - 0.1497z^3 - 0.2159z^2 + 0.1717z - 0.0495}{z^5 - 0.7031z^4 + 0.3029z^3 + 0.1103z^2 - 0.1461z + 0.2845}.$$

Model reduction is performed, both with DBT and SBT, on a maximum-entropy model $\hat{W}_\nu$ of degree $\nu = 24$ determined from estimated covariances. Based on 100 test runs, the empirical means and standard deviations are determined. Figure 6.1 illustrates the statistical bias as a function of the length N of the data string when using stochastic (dashed curve) and deterministic (dotted curve) model reduction respectively.

For the same test runs, Figure 6.2 illustrates the corresponding standard deviations together with the Cramér-Rao bound (solid curve). More precisely, the figures depict the sums of the moduli of the biases and standard deviations respectively for the coefficients of the numerator and denominator polynomials of $\hat{W}(z)$.

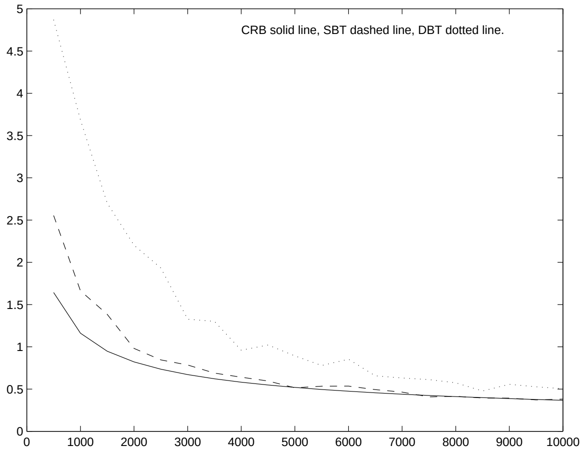

Figure 6.2: Standard deviations of estimates and the Cramér-Rao bound.

The same pattern can be observed in the next experiment, where a model $W(z)$ with poles and zeros closer to the unit circle is considered. The poles and zeros of $\hat{W}(z)$ are determined for 100 runs and a data length $N = 500$. As before, $\nu = 24$. Figure 6.3 depicts these poles and zeros in the case that SBT is used, together with the poles and zeros of $W(z)$, which are denoted by "$\circ$".

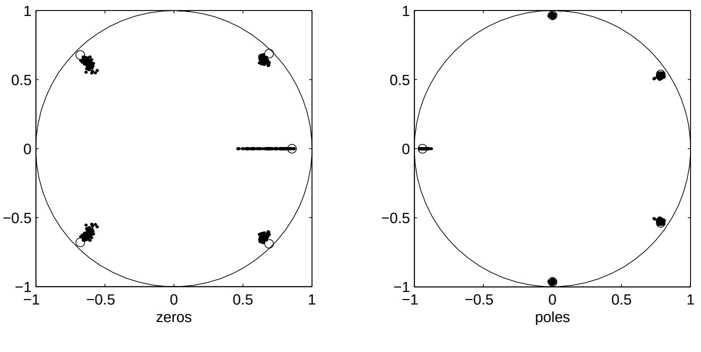

Figure 6.3: SBT estimates of zeros (left) and poles (right) for $N = 500$ and $\nu = 24$.

The approximation obtained from the alternative stochastic model reduction procedure discussed in Remark 4.4 shows the same pattern as SBT.

To further motivate the reason for preferring stochastic model reduction, the corresponding experiment with deterministically balanced truncation is illustrated in Figure 6.4. As expected, the spread is greater, and some of the solutions are nonminimum phase.

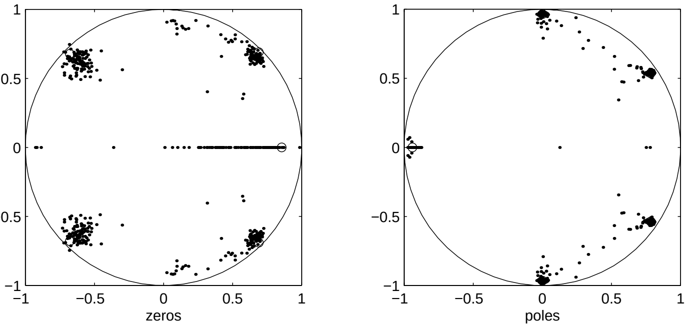

Figure 6.4: DBT estimates of zeros (left) and poles (right) for $N = 500$ and $\nu = 24$.

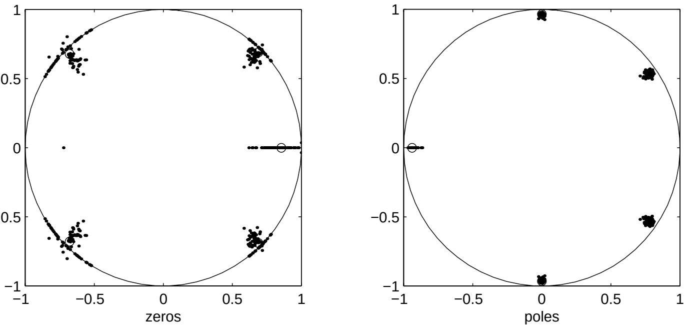

Figure 6.5: Estimates of zeros (left) and poles (right) when using subspace identification.

Figure 6.5 describes the result obtained when applying stochastic subspace identification to the same data. More precisely, Algorithm # 2 in [43] is used. In order to make the experiments comparable, we have chosen a Hankel matrix of dimension 13×13, which corresponds to $\nu = 25$ in our procedure.

Note that the estimates are much less focused, and many zeros tend to cluster on the unit circle, implying that coercivity becomes critical. This is related to the positivity issues discussed in [9]. Also for the model illustrated in Figure 6.1 and Figure 6.2 the subspace identification method performs worse than our SBT identification method, yielding larger biases and standard deviations, but performs better than when DBT is used.

Although the method considered here yields very focused pole-zero estimates, as illustrated in Figure 6.3, there is a noticeable bias in the zero estimates. It will disappear as ν and N are increased. In Figure 6.6 we show the same experiment for $\nu = 64$ and $N = 2000$.

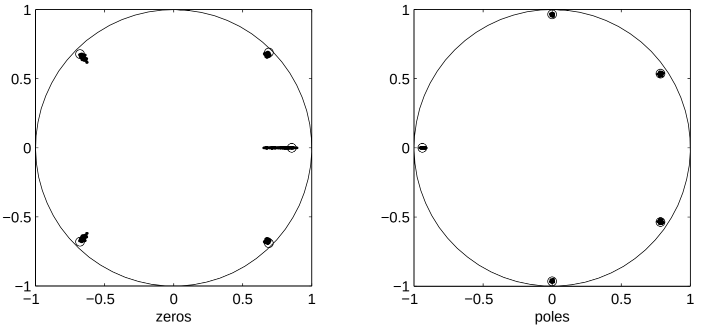

Figure 6.6: SBT estimates of zeros (left) and poles (right) for $N = 2000$ and $\nu = 64$.

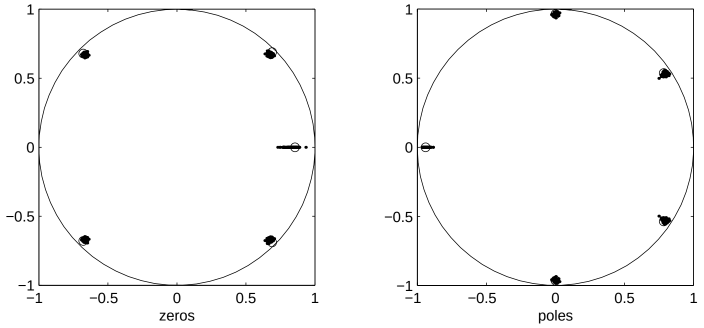

Figure 6.7: SBT estimates of zeros (left) and poles (right) for $N = 500$ and $\nu = 40$ using Burg's method.

In practice, there is a trade-off between the quality of the ergodic estimates, which roughly speaking depend on $|\lambda_{\max}(A)|$, the $\mathfrak{L}^\infty$-error tolerance, which is a function of $|\lambda_{\max}(A - BD^{-1}C)|$, and the numerical accuracy of the computations. For example, if the zeros of $W(z)$ are far from the unit circle and ν is chosen very large, the error may increase.

In the present example, it turns out that using Burg's method [3] in lieu of the ergodic estimate (2.6) yields better estimates for smaller ν and N, as illustrated in Figure 6.7 which shows the case $N = 500$ and $\nu = 40$.

A more detailed picture of the same experiment is given In Table 6.1 and 6.2. There we give the empirical bias and standard deviation for the coefficients of the numerator and the denominator, respectively, of the estimated transfer functions together with the Cramér-Rao bound. It is the authors experience that Burg's method gives at least as good results as when using the ergodic covariance estimate (2.6), unless the intermediate AR model used has a very high model order.

Parameter		σ_w^2	b_1	b_2	b_3	b_4	b_5
True value		1.0000	-0.8762	0.0184	0.0197	0.8591	-0.7491
Bias:	CE:	0.5458	0.1434	0.0100	0.0573	-0.2193	0.2895
	Burg:	0.0971	0.0383	-0.0147	-0.0117	-0.0544	0.0734
Std.dev.:	CE:	0.2332	0.1314	0.0508	0.0611	0.0722	0.0802
	Burg:	0.0712	0.0411	0.0381	0.0339	0.0339	0.0356
CRB:		0.0632	0.0313	0.0312	0.0313	0.0312	0.0309

Table 6.1. Bias and standard deviation of estimated numerator polynomials for $N = 500$ and $\nu = 40$ using covariance estimation (CE) or Burg estimation and , in both cases, followed by SBT.

Parameter		a_1	a_2	a_3	a_4	a_5
True value		-0.6281	0.3597	0.2634	-0.5322	0.7900
Bias:	CE:	0.0087	-0.0044	-0.0003	0.0066	-0.0152
	Burg:	-0.0293	-0.0014	-0.0047	-0.0138	0.0125
Std.dev.:	CE:	0.0274	0.0304	0.0371	0.0305	0.0304
	Burg:	0.0336	0.0307	0.0358	0.0324	0.0306
CRB:		0.0293	0.0321	0.0342	0.0322	0.0290

Table 6.2. Bias and standard deviation of estimated denominator polynomials for $N = 500$ and $\nu = 40$ using covariance estimation (CE) or Burg estimation and, in both cases, followed by SBT.

7. Conclusions

We have presented a three-step procedure for identification of time series, which is easy to understand and implement. Just like for subspace identification methods, robust linear-algebra algorithms can be used and no nonconvex optimization computations are required. Moreover, it has a sound theoretical basis and is computationally competitive to stochastic subspace identification, as our extensive simulations indicate. In particular, its good performance has been confirmed by Monte Carlo simulations. The paper only covers the scalar case, but the multivariate case is presently being worked out.

The three steps, covariance estimation, covariance extension and model reduction have each been studied separately before. This is an advantage which should make the method easy to grasp. However, a comprehensive study of the entire identification strategy, giving appropriate bounds, has been missing and this is what we offered here.

The observation that the Schur parameters converge geometrically simplifies our application of Szegö theory and allows us to give a complete account of the asymptotic behavior of maximum entropy models of growing order. This analysis provides us with a clear indication as to when the identification strategy is good and when it might face difficulties, based purely on the closeness of the maximum modulus zero to the unit circle. The parsimony permeating other system identification methods should not be a reason for refraining from high-order modeling as an intermediate step. In fact, such a strategy might be desirable, since we have shown that the poles of the "true" system which lie outside a circle in the complex plane containing all of its zeros are directly inherited by the high order models. The rest of the poles cluster inside the perimeter of this circle, providing a justification for choosing stochastically balanced model reduction, rather than deterministically balanced truncation, in the last step. With this reduction procedure, we have confirmed better statistical properties with variances closer to the Cramér Rao bound. The procedure could also be modified by exchanging exact covariance extension for approximate one, as outlined in [35].

Even though, in general, stochastic balancing would require the solution of a pair of Riccati equations, this is not the case for the particular maximum entropy models used here. In fact, the balancing procedure only requires linear algebra, and hence an intelligent use of the Levinson algorithm may substantially reduce the number of arithmetic operations.

Finally, by decomposing the total error as a sum of three terms, each one independently adjustable by choosing the integers N, ν and $\hat{n}$, we gave worst-case guaranteed bounds, which complement the nice statistical properties of the method. The error analysis is based on the assumption that the data comes from a rational coercive stochastic system, but the method returns a valid model also for generic data. In fact, in contrast to many stochastic subspace identification [9], all steps of the procedure preserve the positive real property.

Appendix A. Asymptotic behavior of the maximum entropy filter

Theorem 3.4 is actually a modification to the rational setting of a theorem due to Szegö [47], and the proof is modeled after [19], which in turn includes aspects already present in the work of Schur [45]. See also [48], [49] and [16] for more facts on orthogonal polynomials. However, rationality and coercivity allows us to present a simplified and self-contained proof of a version of Szegö's classical theorem, to which we also are able to add geometric convergence. The derivation of Caines and Baykal-Gürsoy [7] is shorter, but we feel that our approach is more systematic and gives additional insight into the mechanism of identification.

To prove Theorems 3.4, 3.5 and 4.3 we need the following lemmas.

Lemma A.1. *Let $\{\phi_\nu(z)\}_0^\infty$ be the normalized Szegö polynomials (3.7). Then $|z^{-\nu}\phi_\nu(z)|$ is uniformly bounded from above and away from zero in the complement $\mathbb{D}^c$ of the open unit disc, i.e., there are positive numbers $\alpha, \beta \in \mathbb{R}$ such that*

$$\alpha \leq |z^{-\nu}\phi_\nu(z)| \leq \beta$$

for all ν and all $z \in \mathbb{D}^c$.

Proof. In view of the Szegö-Levinson recursion (3.4),

$$\varphi_{t+1}(z) = \varphi_t(z)\left[z - \bar{\gamma}_t \frac{\varphi_t^*(z)}{\varphi_t(z)}\right],$$

and hence

$$z^{-\nu}\varphi_\nu(z) = \prod_{k=0}^{\nu-1}\left[1 - z^{-1}\bar{\gamma}_k \frac{\varphi_k^*(z)}{\varphi_k(z)}\right].$$

Now, if $z_1, z_2, \dots, z_\nu$ are the roots of $\varphi_\nu(z)$, it is immediately seen that

$$\frac{\varphi_\nu^*(z)}{\varphi_\nu(z)} = \prod_{k=1}^{\nu} \frac{1 - z\bar{z}_k}{z - z_k},$$

which is a Blaschke product, analytic in $\mathbb{D}^c$ and having modulus one on the unit circle, and thus modulus less than or equal to one in $\mathbb{D}^c$. Hence, since $|z^{-1}| \leq 1$ in $\mathbb{D}^c$,

$$\prod_{k=0}^{\nu-1}(1 - |\gamma_k|) \leq |z^{-\nu}\varphi_\nu(z)| \leq \prod_{k=0}^{\nu-1}(1 + |\gamma_k|) \tag{A.1}$$

for all $z \in \mathbb{D}^c$. But, these products converge to positive numbers as $\nu \to \infty$. This follows from the absolute convergence of the infinite sum $\sum_{k=0}^{\infty} |\gamma_k|$, a fact that, in the present context, stems from Lemma 3.1. From (3.5) we also have $0 < r_\infty \leq r_\nu \leq r_0$, and consequently the lemma follows. □

Remark A.2. An equivalent statement of this lemma is that the maximum entropy solution $W_\nu(z)$, defined by (3.6), is uniformly bounded from above and away from zero for all ν and $z \in \mathbb{D}^c$.

Lemma A.3. *Suppose W is rational and minimum-phase. Then, the sequence of functions*

$$f_\nu(z) := z^{-\nu}\phi_\nu(z)$$

converges uniformly to an analytic function f_∞ in $\mathbb{D}^c_\rho$, where $\mathbb{D}^c_\rho$ is defined in the statement of Theorem 3.5.

Proof. Recall from the theory of polynomials orthogonal on the unit circle [49] the purely algebraic relation

$$\sum_{k=0}^{\nu} \phi_k(z)\overline{\phi_k(w)} = \frac{\phi_\nu^*(z)\overline{\phi_\nu^*(w)} - z\bar{w}\phi_\nu(z)\overline{\phi_\nu(w)}}{1 - z\bar{w}}, \tag{A.2}$$

which is called the Christoffel-Darboux-Szegö formula. In particular, setting $w = 0$ and exchanging z for z^{-1} in (A.2), (3.5) and (3.7) yield

$$\frac{f_\nu(z)}{\sqrt{r_\nu}} = \frac{1}{c_0} - \sum_{k=1}^{\nu} \phi_k(z^{-1})\frac{\gamma_{k-1}}{\sqrt{r_k}}. \tag{A.3}$$

Observe that $\phi_k(z^{-1})$ is analytic and bounded in $\mathbb{D}^c_\rho$, and hence in $\mathbb{D}^c$, and therefore it belongs to $\mathcal{H}^\infty_-$. Moreover, by the maximum modulus principle, it attains its maximum value in $\mathbb{D}^c$ on the unit circle where, by Lemma A.1, it is bounded by β. Hence

$$|\phi_k(z^{-1})| \leq \beta \quad \text{for } z \in \mathbb{D}^c \text{ and for all } k. \tag{A.4}$$

Therefore, in view of (A.3) and the fact that $r_k \geq r_\infty$, we have

$$\left|\frac{f_\nu(z)}{\sqrt{r_\nu}} - \frac{f_\mu(z)}{\sqrt{r_\mu}}\right| \leq \frac{\beta}{\sqrt{r_\infty}} \sum_{k=\mu+1}^{\nu} |\gamma_{k-1}|, \tag{A.5}$$

which, by Lemma 3.1, can be made arbitrarily small for sufficiently large ν and μ. This establishes (A.3) as a Cauchy sequence in $\mathcal{H}_-^\infty$. The same holds for $f_\nu(z)$. In fact, since $r_\nu \leq c_0$, for all $z \in \mathbb{D}^c$

$$\begin{aligned}|f_\nu(z) - f_\mu(z)| &\leq \sqrt{c_0}\left|\frac{f_\nu(z)}{\sqrt{r_\nu}} - \frac{f_\mu(z)}{\sqrt{r_\nu}}\right| \\ &\leq \sqrt{c_0}\left|\frac{f_\nu(z)}{\sqrt{r_\nu}} - \frac{f_\mu(z)}{\sqrt{r_\mu}}\right| + |f_\mu(z)|\left|\frac{1}{\sqrt{r_\nu}} - \frac{1}{\sqrt{r_\mu}}\right|. \qquad \text{(A.6)}\end{aligned}$$

But, by Lemma A.1, $|f_\mu(z)| \leq \beta$ for all ν and $z \in \mathbb{D}^c$, and therefore, in view of (A.5), we obtain

$$|f_\nu(z) - f_\mu(z)| \leq \beta\sqrt{\frac{c_0}{r_\infty}}\sum_{k=\mu+1}^{\nu}|\gamma_{k-1}| + \beta\left|\frac{1}{\sqrt{r_\nu}} - \frac{1}{\sqrt{r_\mu}}\right| \quad \text{for all } z \in \mathbb{D}^c. \quad \text{(A.7)}$$

Since $r_\nu \to r_\infty$ as $\nu \to \infty$, we see that, for each $\epsilon > 0$, $|f_\nu(z) - f_\mu(z)| < \epsilon$ for sufficiently large ν and μ. Consequently, f_ν tends uniformly in $\mathbb{D}^c$ to a function $f_\infty \in \mathcal{H}_-^\infty$.

The uniform convergence and the analyticity can be extended to any compact subset of $\mathbb{D}_\rho^c$. To see this, first note that $z \in \mathbb{D}$ if and only if $z^{-1} \in \mathbb{D}^c$. Therefore, by Lemma A.1,

$$|\phi_k(z^{-1})| \leq \beta|z|^{-k} \quad \text{for } z \in \mathbb{D},$$

and consequently, since $r_\nu \leq r_k$, (A.3) yields

$$|f_\nu(z)| \leq \frac{1}{\sqrt{c_0}} + \beta|z|^{-1}\sum_{k=0}^{\nu-1}|\gamma_k||z|^{-k}. \quad \text{(A.8)}$$

Similarly, instead of (A.5) we have

$$\left|\frac{f_\nu(z)}{\sqrt{r_\nu}} - \frac{f_\mu(z)}{\sqrt{r_\mu}}\right| \leq \frac{\beta}{\sqrt{r_\infty}}|z|^{-1}\sum_{k=\mu}^{\nu-1}|\gamma_k||z|^{-k}. \quad \text{(A.9)}$$

Now, for any compact subset $K \in \mathbb{D}_\rho^c$, there is a $\gamma \in (\rho, 1)$ and an $\epsilon > 0$ such that $|z| > \gamma + \epsilon$ for all $z \in K$. Hence, by Lemma 3.1, $|\gamma_k||z|^{-k} \leq M\hat{\gamma}^k$ where $\hat{\gamma} := \gamma(\gamma+\epsilon)^{-1} < 1$. Consequently, by (A.8), $f_\nu(z)$ is uniformly bounded in K, and (A.9) can be made arbitrarily small for sufficiently large ν and μ. Therefore, by (A.6), f_ν tends uniformly in K to the analytic function f_∞. □

Lemma A.4. *Let γ be a real number such that $\rho < \gamma < 1$. Then $\|f_\nu - f_\infty\|_\infty = O(\gamma^\nu)$.*

Proof. It follows from (A.7) that

$$|f_\nu(z) - f_\infty(z)| \leq \frac{\beta}{\sqrt{r_\infty}}\left[\sqrt{c_0}\sum_{k=\nu+1}^{\infty}|\gamma_{k-1}| + \left|1 - \sqrt{\frac{r_\infty}{r_\nu}}\right|\right] \quad \text{for all } z \in \mathbb{D}^c. \quad \text{(A.10)}$$

By Lemma 3.1, the first term is $O(\gamma^\nu)$. It remains to show that the same holds for the second term. To this end, first note that, by (3.5),

$$1 - \sqrt{\frac{r_\infty}{r_\nu}} = 1 - \prod_{k=\nu}^{\infty}\sqrt{1-\gamma_k^2}.$$

But, by Lemma 3.1, $|\gamma_k| \leq M\gamma^k$ for some M. Therefore, since $\sqrt{1-x^2} \geq 1-x$ for each $x \in [0,1]$,

$$1-\sqrt{\frac{r_\infty}{r_t}} \leq 1-\prod_{k=t}^{\infty}(1-M\gamma^k) = O(\gamma^t)$$

for t large enough. This concludes the proof. □

Recalling the definition (3.6) of W_ν, we note that Lemma A.4 may be written

$$\|W_\nu^{-1} - f_\infty\|_\infty = O(\gamma^\nu).$$

As it turns out, by coercivity, this implies that $W_\nu \to f_\infty^{-1}$ in the same manner.

Lemma A.5. *Let W_ν be the transfer function* (3.6) *of the maximum entropy filter. Then*

$$\|W_\nu - f_\infty^{-1}\|_\infty = O(\gamma^\nu),$$

where f_∞ is the limit function of Lemma A.3.

Proof. Note that the limit function f_∞ has the same uniform bounds as f_ν in Lemma A.1. In particular, $|f_\infty(z)| \geq \alpha$, $|f_\infty(z)|^{-1} \leq \alpha^{-1}$, and $|W_\nu(z)| \leq \alpha^{-1}$ for all $z \in \mathbb{D}^c$. Consequently,

$$\|W_\nu - f_\infty^{-1}\|_\infty \leq \|W_\nu\|_\infty \|f_\infty^{-1}\|_\infty \|W_\nu^{-1} - f_\infty\|_\infty \leq \alpha^{-2}\|W_\nu^{-1} - f_\infty\|_\infty,$$

so the required result follows from Lemma A.4. □

Lemma A.6. *Let W be the rational minimum-phase function defined above, and let f_∞ be the limit function in Lemma* A.3. *Then $W(z) = f_\infty(z)^{-1}$ for all $z \in \mathbb{D}_\rho^c$.*

Proof. Let $\Phi_\nu(e^{i\theta}) := |W_\nu(e^{i\theta})|^2$ be the spectral density of the maximum entropy process. Then, in view of the interpolation condition,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta})d\theta = c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\Phi_\nu(e^{i\theta})d\theta \quad \text{for } k = 0,1,\dots,\nu, \tag{A.11}$$

from which we have pointwise convergence of the Fourier coefficients of $\Phi_\nu(e^{i\theta})$ to those of $\Phi(e^{i\theta})$ as $\nu \to \infty$, and hence $\Phi_\nu(e^{i\theta}) \to \Phi(e^{i\theta})$ in the $\mathfrak{L}^2$ sense. However, by Lemma A.5, $\Phi_\nu(e^{i\theta}) \to |f_\infty(e^{i\theta})|^{-2}$ in $\mathfrak{L}^\infty$ norm, and hence *a fortiori* in $\mathfrak{L}^2$ norm, as $\nu \to \infty$. Since, in addition, not only $\Phi(e^{i\theta})$ but also f_∞ is analytic in a neighborhood of the unit circle (Lemma A.3), we have

$$\Phi(e^{i\theta}) = |f_\infty(e^{i\theta})|^{-2}. \tag{A.12}$$

In the language of Hardy space theory [14], a minimum-phase spectral factor is *outer*. In particular, W_ν is an outer spectral factor of $\Phi_\nu(e^{i\theta})$ satisfying

$$W_\nu(z) = \exp\left[\frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{e^{it}+z}{e^{it}-z}\log|W_\nu(e^{it})|^2 dt\right].$$

But Lemma A.5, Equation (A.12) and the fact that $\Phi(e^{i\theta}) = |W(e^{i\theta})|^2$,

$$W_\nu(z) \to \exp\left[\frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{e^{it}+z}{e^{it}-z}\log|W(e^{it})|^2 dt\right] = W(z),$$

the outer spectral factor of Φ. But, by Lemma A.3, $W_\nu(z) \to f_\infty(z)^{-1}$ in $\mathbb{D}_\rho^c$, and therefore $f_\infty(z) = W^{-1}(z)$ as claimed. □

Proof of Theorem 3.4. The theorem is a direct consequence of Lemmas A.5 and A.6. □

Proof of Theorem 3.5. The theorem follows from Lemmas A.3 and A.6. □

Proof of Theorem 4.3. Following [53] we see that

$$\|W_\nu^{-1}(W_\nu - W_{\text{red}})\|_\infty \leq \epsilon, \tag{A.13}$$

and consequently

$$|W_\nu(e^{i\theta}) - W_{\text{red}}(e^{i\theta})| \leq \epsilon |W_\nu(e^{i\theta})|$$

holds for all θ, from which we have

$$(1-\epsilon)|W_\nu(e^{i\theta})| \leq |W_{\text{red}}(e^{i\theta})| \leq (1+\epsilon)|W_\nu(e^{i\theta})|.$$

However, in view of (3.6) and (3.7), it follows from (A.1) that

$$\frac{\sqrt{r_\nu}}{\prod_{k=0}^{\nu-1}(1+|\gamma_k|)} \leq |W_\nu(e^{i\theta})| \leq \frac{\sqrt{r_\nu}}{\prod_{k=0}^{\nu-1}(1-|\gamma_k|)},$$

which together with (3.5) yields

$$\frac{c_0}{\kappa} \leq |W_\nu(e^{i\theta})| \leq \kappa \tag{A.14}$$

for all θ. This establishes (4.27). To see that W_{red} is minimum phase if $\epsilon < 1$, note that, by (4.27), W_{red} cannot have a zero on the unit circle. Moreover, by Rouché's Theorem, W_{red} has the same number of zeros in $\mathbb{D}^c$ (including ∞) as W_ν. Hence, since W_ν is minimum phase, so is W_{red}.

To establish the bound (4.28) note that

$$\|W_\nu - W_{\text{red}}\|_\infty \leq \|W_\nu\|_\infty \|W_\nu^{-1}(W_\nu - W_{\text{red}})\|_\infty.$$

From (A.14) we have $\|W_\nu\|_\infty \leq \kappa$, and hence (4.28) follows from (A.13). □

Appendix B. Statistical convergence proofs

Proof of Theorem 5.2. Given the covariance estimates (2.6) we determine the corresponding Szegö polynomial $\hat{\varphi}_\nu(z)$ and predictor error variance $\hat{r}_\nu$ from (3.4) and (3.5), and form the maximum-entropy filter

$$\hat{W}_\nu(z) = \frac{\sqrt{\hat{r}_\nu} z^\nu}{\hat{\varphi}_\nu(z)}.$$

To determine $\|W_\nu - \hat{W}_\nu\|_\infty$ let $z \in \mathbb{D}^c$ and form

$$\begin{aligned} W_\nu(z) - \hat{W}_\nu(z) &= \frac{\sqrt{r_\nu} z^\nu}{\varphi_\nu(z)} - \frac{\sqrt{\hat{r}_\nu} z^\nu}{\hat{\varphi}_\nu(z)} \\ &= \frac{(\sqrt{r_\nu} - \sqrt{\hat{r}_\nu}) z^{-\nu}\varphi_\nu(z) - \sqrt{r_\nu} z^{-\nu}(\varphi_\nu(z) - \hat{\varphi}_\nu(z))}{z^{-\nu}\varphi_\nu(z) z^{-\nu}\hat{\varphi}_\nu(z)}. \end{aligned}$$

Since $r_\infty > 0$, by (3.7) and Lemma A.1,

$$0 < \mu := \sqrt{r_\infty}\alpha \leq |z^{-\nu}\varphi_\nu(z)| \leq \sqrt{c_0}\beta =: M,$$

and, by (A.1),

$$|z^{-\nu}\hat{\varphi}_\nu(z)| \geq \hat{\mu}_\nu := \prod_{k=0}^{\nu-1}(1-|\hat{\gamma}_k|),$$

where $\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_{\nu-1}$ are the Schur parameters corresponding to the estimated covariances (2.6). Therefore, by the maximum-modulus principle,

$$|W_\nu(z) - \hat{W}_\nu(z)| \leq \max_{|z|=1} \frac{1}{\mu\hat{\mu}_\nu}\{M|\sqrt{r_\nu} - \sqrt{\hat{r}_\nu}| + \sqrt{c_0}|\varphi_\nu(z) - \hat{\varphi}_\nu(z)|\},$$

where we have also used the fact that $r_\nu \leq c_0$. But, for $|z| = 1$,

$$|\varphi_\nu(z) - \hat{\varphi}_\nu(z)| \leq \|\boldsymbol{\varphi}_\nu - \hat{\boldsymbol{\varphi}}_\nu\|_1,$$

where $\boldsymbol{\varphi}_\nu$ and $\hat{\boldsymbol{\varphi}}_\nu$ are the ν-vectors formed as in (4.19) and $\|\cdot\|_1$ is the ℓ_1 norm. Recall that $\boldsymbol{\varphi}_\nu$ is the unique solution of the normal equations

$$T_\nu\boldsymbol{\varphi}_\nu = -\mathbf{c}_\nu \quad \text{where } \mathbf{c}_\nu := \begin{bmatrix} c_\nu & c_{\nu-1} & \dots & c_1 \end{bmatrix}', \tag{B.1}$$

where T_ν is the Toeplitz matrix defined by (2.4), and that

$$r_\nu = c_0 + \mathbf{c}_\nu'\boldsymbol{\varphi}_\nu. \tag{B.2}$$

Also, the analogous relations hold for $\hat{\varphi}_\nu$ and $\hat{r}_\nu$. Then,

$$r_\nu - \hat{r}_\nu = (c_0 - \hat{c}_0) + (\mathbf{c}_\nu - \hat{\mathbf{c}}_\nu)'\boldsymbol{\varphi}_\nu + \hat{\mathbf{c}}_\nu'(\boldsymbol{\varphi}_\nu - \hat{\boldsymbol{\varphi}}_\nu)$$

and hence

$$|r_\nu - \hat{r}_\nu| \leq |c_0 - \hat{c}_0| + \|\mathbf{c}_\nu - \hat{\mathbf{c}}_\nu\|_1\|\boldsymbol{\varphi}_\nu\|_\infty + \|\hat{\mathbf{c}}_\nu\|_\infty\|\boldsymbol{\varphi}_\nu - \hat{\boldsymbol{\varphi}}_\nu\|_1.$$

Finally,

$$|\sqrt{r_\nu} - \sqrt{\hat{r}_\nu}| \leq \frac{|r_\nu - \hat{r}_\nu|}{\sqrt{r_\nu} + \sqrt{\hat{r}_\nu}} \leq \frac{|r_\nu - \hat{r}_\nu|}{\sqrt{r_\infty}},$$

and consequently, since $\|x\|_1 \leq \nu\|x\|_\infty$ for any $x \in \mathbb{R}^\nu$,

$$\begin{aligned} \|W_\nu - \hat{W}_\nu\|_\infty \leq \; & \frac{M}{\mu\hat{\mu}_\nu\sqrt{r_\infty}}\{|c_0 - \hat{c}_0| + \|\boldsymbol{\varphi}_\nu\|_\infty \nu \|\mathbf{c}_\nu - \hat{\mathbf{c}}_\nu\|_\infty\} \\ & + \frac{1}{\mu\hat{\mu}_\nu}\left[\sqrt{c_0} + \frac{M\|\hat{\mathbf{c}}_\nu\|_\infty}{\sqrt{r_\infty}}\right]\nu\|\boldsymbol{\varphi}_\nu - \hat{\boldsymbol{\varphi}}_\nu\|_\infty. \end{aligned} \tag{B.3}$$

Recall now that $\boldsymbol{\varphi}_\nu$ and $\hat{\boldsymbol{\varphi}}_\nu$ are each solutions of a normal equation (B.1). More precisely, $T_\nu\boldsymbol{\varphi}_\nu = -\mathbf{c}_\nu$ and $\hat{T}_\nu\hat{\boldsymbol{\varphi}}_\nu = -\hat{\mathbf{c}}_\nu$. Since $c_k = CA^{k-1}\bar{C}'$ for $k > 0$, where all eigenvalues of A are less than one in modulus, $c_k \to 0$ exponentially, we have

$$\|\mathbf{c}_\nu\|_\infty \leq K_1 \quad \text{and} \quad \|T_\nu\|_\infty \leq c_0 + 2\sum_{k=1}^{\nu-1}|c_k| \leq K_2$$

for some constants K_1 and K_2. Moreover, from [8] we have

$$\|T_\nu^{-1}\|_\infty \leq \frac{1}{c_0}\prod_{k=0}^{\nu-1}\frac{1+|\gamma_k|}{1-|\gamma_k|} \leq K_3$$

for some constant K_3; see the proof of Lemma A.1. Hence

$$\|\boldsymbol{\varphi}_\nu\|_\infty \leq \|T_\nu^{-1}\|_\infty \|\mathbf{c}_\nu\|_\infty \leq K_1 K_3$$

and the condition number

$$\kappa(T_\nu) := \|T_\nu\|_\infty \|T_\nu^{-1}\|_\infty \leq K := K_2 K_3$$

is bounded for all ν.

Now, it is known [25] that for each data length N in (2.2), there is a $\nu(N)$ of order $O(\log N)$ such that

$$\max_{0\leq k\leq \nu(N)} |c_k - \hat{c}_k| = O\left(\sqrt{\frac{\log\log N}{N}}\right), \tag{B.4}$$

and therefore, for any $a \in \mathbb{R}$,

$$\nu^a |c_0 - \hat{c}_0| \to 0 \quad \text{and} \quad \nu^a \|\mathbf{c}_\nu - \hat{\mathbf{c}}_\nu\|_\infty \to 0 \quad \text{as } \nu = \nu(N) \to \infty. \tag{B.5}$$

Consequently the first term in the bound (B.3) tends to zero as $N \to \infty$ and $\nu(N) \to \infty$ provided it is done at the specified relative rates and provided $\hat{\mu}_\nu$ is bounded away from zero. However, the estimate (2.6) has the property that the corresponding Toeplitz matrix $\hat{T}_\nu$ is positive definite for each finite ν, and this in turn is equivalent to $|\hat{\gamma}_k| < 1$ for $k = 0, 1, \dots, \nu - 1$ so that $\hat{\mu}_\nu > 0$. Since, in addition $\hat{\mu}_\nu \to \mu > 0$ as $\nu(N) \to \infty$ by (B.4) and continuity, the second requirement is also fulfilled. To simplify notations, we have suppressed the index N in the quantities marked with a hat, which of course depend on the data (2.2) and hence also on N.

Next we show that also the second term in (B.3) tends to zero. Since $\|\hat{\mathbf{c}}_\nu\|_\infty \leq \|\mathbf{c}_\nu\|_\infty + \|\mathbf{c}_\nu - \hat{\mathbf{c}}_\nu\|_\infty$ is bounded, it thus remains to demonstrate that

$$\nu \|\boldsymbol{\varphi}_{\nu(N)} - \hat{\boldsymbol{\varphi}}_{\nu(N)}\|_\infty \to 0 \quad \text{as } \nu(N) \to \infty.$$

This follows from the more general fact, needed for the proof of Corollary B.1, that

$$\nu^a \|\boldsymbol{\varphi}_{\nu(N)} - \hat{\boldsymbol{\varphi}}_{\nu(N)}\|_\infty \to 0 \quad \text{as } \nu(N) \to \infty \tag{B.6}$$

for any $a \in \mathbb{R}$. To prove this, first note that

$$\|T_\nu - \hat{T}_\nu\|_\infty \leq |c_0 - \hat{c}_0| + 2\nu \|\mathbf{c}_\nu - \hat{\mathbf{c}}_\nu\|_\infty,$$

and hence $\|T_\nu - \hat{T}_\nu\|_\infty \to 0$. Therefore $\rho_\nu := \|T_\nu - \hat{T}_\nu\|_\infty \|T_\nu^{-1}\|_\infty < 1$ for $\nu := \nu(N)$ sufficiently large, and, provided $\mathbf{c}_\nu \neq 0$, the standard perturbation estimate [22] yields

$$\frac{\|\boldsymbol{\varphi}_\nu - \hat{\boldsymbol{\varphi}}_\nu\|_\infty}{\|\boldsymbol{\varphi}_\nu\|_\infty} \leq \frac{1}{1-\rho_\nu} \kappa(T_\nu) \left[\frac{\|T_\nu - \hat{T}_\nu\|_\infty}{\|T_\nu\|_\infty} + \frac{\|\mathbf{c}_\nu - \hat{\mathbf{c}}_\nu\|_\infty}{\|\mathbf{c}_\nu\|_\infty} \right], \tag{B.7}$$

and consequently, since $\|T_\nu\|_\infty \geq c_0 > 0$, it follows from (B.5) that (B.6) tends to zero in the required manner.

If $\mathbf{c}_\nu = 0$, $\boldsymbol{\varphi}_\nu = 0$, and hence

$$\|\boldsymbol{\varphi}_\nu - \hat{\boldsymbol{\varphi}}_\nu\|_\infty = \|\hat{\boldsymbol{\varphi}}_\nu\|_\infty \leq \|\hat{T}_\nu^{-1}\|_\infty \|\hat{\mathbf{c}}_\nu\|_\infty = \|\hat{T}_\nu^{-1}\|_\infty \|\mathbf{c}_\nu - \hat{\mathbf{c}}_\nu\|_\infty,$$

which shows that (B.6) tends to zero also in the case $\mathbf{c}_\nu = 0$. In fact, since $\hat{\mu}$ is bounded away from zero, by continuity, for each $\epsilon > 0$, there is a N_0 such that

$$\|\hat{T}_\nu^{-1}\|_\infty \leq \frac{1}{\hat{c}_0}\prod_{k=0}^{\nu-1}\frac{1+|\hat{\gamma}_k|}{1-|\hat{\gamma}_k|} \leq \frac{1}{c_0}\prod_{k=0}^{\nu-1}\frac{1+|\gamma_k|}{1-|\gamma_k|} + \epsilon \leq K_3 + \epsilon$$

for $\nu \geq N_0$. □

Corollary B.1. *If $\nu(N)$ is defined as in Theorem* 5.1, *then, for any $a \in \mathbb{R}$,*

$$\nu^a\|W_\nu - \hat{W}_\nu\|_\infty \to 0$$

almost surely as $\nu := \nu(N) \to \infty$.

To prove Theorem 5.3, we first note that the Hankel operator $\mathbb{H}$, defined by (4.1), has a nice representation in the space $\mathcal{L}^2$ of square-integrable functions. In fact, let $\mathcal{H}_+^2$ be the subspace in $\mathcal{L}^2$ of functions with vanishing negative Fourier coefficients, hence being analytic in the unit disc $\mathbb{D}$. In this setting, $\mathbb{H}$ has the representation $H_\Theta : \mathcal{H}_+^2 \to \mathcal{L}^2 \ominus \mathcal{H}_+^2$ given by

$$H_\Theta f = P^\perp \Theta f, \tag{B.8}$$

where $P^\perp$ is the orthogonal projection onto the orthogonal complement $\mathcal{L}^2 \ominus \mathcal{H}_+^2$ of $\mathcal{H}_+^2$ in $\mathcal{L}^2$, and where Θ is the $\mathcal{L}^\infty$-function

$$\Theta(z) = W_-(z)\bar{W}_+(z)^{-1}. \tag{B.9}$$

Here $W_-(z)$ and $\bar{W}_+(z)$ are the analytic and coanalytic minimum-phase spectral factors defined in Section 4. (See, e.g., [30, 31].) In the present scalar case, $\bar{W}_+(z) = W_-(z^{-1})$. In fact, the *phase function* Θ is the transfer function of an all-pass filter transforming the white noise w_- in (4.2) to the white noise w_+ in (4.3) [30, p. 834]. Let $d\hat{w}_-$ and $d\hat{\bar{w}}_+$ be the stochastic measures such that

$$w_-(t) = \int_{-\pi}^{\pi} e^{i\theta t} d\hat{w}_- \quad \text{and} \quad \bar{w}_+(t) = \int_{-\pi}^{\pi} e^{i\theta t} d\hat{\bar{w}}_+$$

Then

$$\begin{aligned} H^+ &= \int_{-\pi}^{\pi} \mathcal{H}_+^2 d\hat{\bar{w}}_+ = \int_{-\pi}^{\pi} \mathcal{H}_+^2 \Theta(e^{i\theta t}) d\hat{w}_- \\ H^- &= \int_{-\pi}^{\pi} \mathcal{L}^2 \ominus \mathcal{H}_+^2 d\hat{w}_- \end{aligned}$$

and consequently $\mathbb{H} := \mathrm{E}^{H^-}|_{H^+}$ corresponds to H_Θ under the isomorphism defined by $f \leftrightarrow \int f d\hat{w}_-$.

Proof of Theorem 5.3. It follows from Theorem 5.2 that $\Big||\hat{W}_\nu(e^{i\theta})| - |W_\nu(e^{i\theta})|\Big| \to 0$ uniformly in θ as $\nu \to \infty$, and hence, by Lemma A.1, there are positive real numbers μ_1 and μ_2 such that

$$\mu_1 \leq |\hat{W}_\nu(e^{i\theta})| \leq \mu_2 \tag{B.10}$$

for all θ and sufficiently large ν. Therefore, since

$$\|\hat{W}_\nu - \hat{W}\|_\infty \leq \|\hat{W}_\nu\|_\infty \|\hat{W}_\nu^{-1}(\hat{W}_\nu - \hat{W})\|_\infty,$$

(A.13) and (4.26) imply that

$$\|\hat{W}_\nu - \hat{W}\|_\infty \leq 2\mu_2 \sum_{k=\hat{n}+1}^{\nu} \frac{\hat{\sigma}_k}{1-\hat{\sigma}_k}, \tag{B.11}$$

for sufficiently large ν, where $\hat{\sigma}_1, \hat{\sigma}_2, \dots, \hat{\sigma}_\nu$ are the singular values (5.1) determined from the covariance estimates (2.6).

It is well-known (see, e.g., [56, p. 204]) that the singular value σ_k of the Hankel operator H_Θ, defined by (B.8) equals the infimum of $\|H_\Theta - K\|$ over all operators $K : \mathcal{H}^2_+ \to \mathcal{L}^2 \ominus \mathcal{H}^2_+$ of finite rank at most k. Recall that $\Theta(z) = W_\nu(z)/W_\nu(z^{-1})$. The singular value $\hat{\sigma}_k$ of $H_{\hat{\Theta}}$, where $\hat{\Theta}(z) = \hat{W}_\nu(z)/\hat{W}_\nu(z^{-1})$, is described analogously. Therefore, since

$$\|H_{\hat{\Theta}} - K\| \leq \|H_{\hat{\Theta}} - H_\Theta\| + \|H_\Theta - K\| \leq \|\hat{\Theta} - \Theta\|_\infty + \|H_\Theta - K\|,$$

we have $\hat{\sigma}_k \leq \|\hat{\Theta} - \Theta\|_\infty + \sigma_k$. But, for $k > n$, $\sigma_k = 0$, and hence $\hat{\sigma}_k \leq \|\hat{\Theta} - \Theta\|_\infty$. Consequently, (B.11) yields

$$\|\hat{W}_\nu - \hat{W}\|_\infty \leq M_1 \nu \|\hat{\Theta} - \Theta\|_\infty, \tag{B.12}$$

where $M_1 := 2\mu_2(1-\hat{\sigma}_{\hat{n}+1})^{-1}$. However,

$$\hat{\Theta}(z) - \Theta(z) = \hat{W}_\nu(z^{-1})^{-1} \left\{ \hat{W}_\nu(z) - W(z) - \Theta(z)[\hat{W}_\nu(z^{-1}) - W(z^{-1})] \right\},$$

so, since $\|\hat{W}_\nu(z^{-1})\|_\infty$ is uniformly bounded by (B.10), and $\|\Theta\|_\infty$ is constant,

$$\|\hat{\Theta} - \Theta\|_\infty \leq M_2 \|W - \hat{W}_\nu\|_\infty,$$

which together with (B.12) yields

$$\|\hat{W}_\nu - \hat{W}\|_\infty \leq M_1 M_2 \nu \|W - W_\nu\|_\infty + M_1 M_2 \nu \|W_\nu - \hat{W}_\nu\|_\infty$$

for sufficiently large ν. Consequently the theorem follows from Theorem 3.4 and Corollary B.1. □

Acknowledgment. We would like to thank Gy. Michaletzky, L. Baratchart, A. Gombani, W. B. Gragg, G. Picci and T. Söderström for stimulating discussions and for providing us with appropriate references. We are also indebted to the anonymous referees for several useful suggestions.

References

1. M. Aoki. *State Space Modeling of Time Series.* Springer Verlag, 1987.
2. K. N. Berk. Consistent autoregressive spectral estimates. *Annals Statistics*, 2:489–502, 1974.
3. J. P. Burg. *Maximum entropy spectral analysis.* PhD thesis, Stanford University, Dept. Geophysics, Stanford CA., 1975.
4. C. I. Byrnes, S. V. Gusev, and A. Lindquist. A convex optimization approach to the rational covariance extension problem. *SIAM Journal on Control and Optimization*, 37:211–229, 1999.
5. C. I. Byrnes, A. Lindquist, S. V. Gusev, and A. V. Matveev. A complete parametrization of all positive rational extensions of a covariance sequence. *IEEE Trans. Automatic Control*, AC-40:1841–1857, 1995.
6. C. I. Byrnes, A. Lindquist, and Y. Zhou. On the nonlinear dynamics of fast filtering algorithms. *SIAM Journal on Control and Optimization*, 32:744–789, 1994.

7. P.E. Caines and M. Baykal-Gürsoy. On the L^∞ consistency of L^2 estimators. *Systems & Control Letters*, 12:71–76, 1989.
8. G. Cybenko. *Error Analysis of Some Signal Processing Algorithms.* PhD thesis, Princeton University, 1978.
9. A. Dahlén, A. Lindquist, and J. Mari. Experimental evidence showing that stochastic subspace identification methods may fail. *Systems and Control Letters*, 34:303–312, 1998.
10. U. B. Desai and D. Pal. A realization approach to stochastic model reduction and balanced stochatic realizations. In *Proc. 21st IEEE CDC*, pages 1105–1112, 1983.
11. U. B. Desai and D. Pal. A transformation approach to stochastic model reduction. *IEEE Trans. Automatic Control*, AC-29:1097–1100, 1984.
12. J. Durbin. Efficient estimation of parameters in moving average models. *Biometrika*, 46:306–316, 1959.
13. J. Durbin. The fitting of time-series models. *Rev. Inst. Int. Stat.*, pages 223–243, 1960.
14. P. Duren. *Theory of $\mathcal{H}^p$ spaces.* Academic Press, 1970.
15. P. Enqvist. PhD thesis, Royal Institute of Technology, to appear.
16. G. Freud. *Orthogonale Polynome.* Birkhäuser Verlag, 1969.
17. J. J. Fuchs. ARMA order estimation via matrix perturbation theory. *IEEE Trans. Automatic Control*, AC-32:358–361, 1987.
18. T. Georgiou. Realization of power spectra from partial covariance sequences. *IEEE Trans. Ac., Speech and Signal Processing*, ASSP-35:438–449, 1987.
19. L. Geronimus. *Orthogonal Polynomials.* Consultant Bureau, 1961.
20. M. Gevers. Towards a joint design of identification and control. In J. Willems and H. Trentelman, editors, *Essays on Control: Perspectives in the Theory and its Applications*, 1993.
21. K. Glover. All optimal Hankel-norm approximations of linear multivariable systems and their l_∞ -error bounds. *Int. J. Contr.*, 39:1115–1193, 1984.
22. G. H. Golub and C. F. van Loan. *Matrix Computations.* Johns Hopkins, 1989.
23. W. B. Gragg. Positive definite Toeplitz matrices, the Arnoldi process for isometric operators, and Gaussian quadrature on the unit circle. In E. S. Nikolaev, editor, *Numerical Methods in Linear Algebra*, pages 16–32. Moscow U. P., 1982.
24. U. Grenander and G. Szegö. *Toeplitz forms and their applications.* Univ. California Press, 1958.
25. E. J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems.* John Wiley & Sons, 1988.
26. C. Hurvich and C.L. Tsai. Regression and time series model selection in small samples. *Biometrika*, pages 297–307, 1989.
27. W. Jones and E. Saff. Szegö polynomials and frequency analysis. In *Approximation Theory*, pages 341–352. Dekker Inc., 1992.
28. S. Y. Kung. A new identification method and model reduction algorithm via singular value decomposition. In *12th Asilomar Conf. on Circuits, Systems and Comp.*, pages 705–714, 1978.
29. W. E. Larimore. System identification, reduced ordered filtering and modeling via canonical variate analysis. In *Proc. of the American Control Conference*, 1983.
30. A. Lindquist and G Picci. Realization theory for multivariate stationary gaussian processes. *SIAM J. Control and Optimization*, 23:809–857, 1985.
31. A. Lindquist and G. Picci. A geometric approach to modelling and estimation of linear stochastic systems. *J. of Math. Systems, Estimation and Control*, 1:241–333, 1991.
32. A. Lindquist and G. Picci. Canonical correlation analysis, approximate covariance extension, and identification of stationary time series. *Automatica*, 32(5):709–733, 1996.
33. L. Ljung and B. Wahlberg. Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Adv. Appl. Prob.*, 24:412–440, 1992.
34. L. Ljung and Z. Yuan. Asymptotic properties of black box identification of transfer functions. *IEEE Trans. Automatic Control*, AC-26:514–530, 1985.
35. J. Mari. *Rational Modeling of Time Series and Applications to Geometric Control.* PhD thesis, Royal Instiute of Technology, 1998.
36. D. Q. Mayne and F. Firoozan. Linear identification of ARMA processes. *Automatica*, 18:461–466, 1982.

37. H. Mhaskar and E. Saff. The distribution of zeros of asymptotically extremal polynomials. *J. Approx. Theory*, 3:279–300, 1991.
38. B. C. Moore. Singular value analysis of linear systems. In *Proc. IEEE CDC*, pages 66–73, 1978.
39. P. Nevai. Research problems in orthogonal polynomials. In C. Chui, L. Schumaker, and J. Ward, editors, *Approximation Theory VI*, 1989.
40. E. Nikishin and V. Sorokin. Rational approximations and orthogonality. In *Translations of Mathematical Monographs*, volume 92, 1991.
41. P. Van Overschee. *Subspace Identification, Theory - Implementation - Application.* PhD thesis, Katholieke Universiteit Leuven, 1995. Kluwer book with same title by Van Overschee and De Moor.
42. P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. In *Proc. 30th Conference on Decision and Control, Brighton*, 1991.
43. P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems - Theory Implementation Applications.* Kluwer Academic Publishers, 1996.
44. B. Porat. *Digital Processing of Random Signals, Theory & Methods.* Prentice Hall, 1994.
45. I. Schur. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. *J. für die Reine und Angewandte Mathematik*, 147:205–232, 1917.
46. J. Sorelius, T. Söderström, P. Stoica, and M. Cedervall. Order estimation method for subspace-based system identification. In *Proc. SYSID '97*, 1997.
47. G. Szegö. Beiträge zur Theorie der Toeplitzschen Formen, I, II. *Mathematische Zeitschrift*, 6:167–202, 1920.
48. G. Szegö. Über die Randwerte analytischer Funktionen. *Mat. Annalen*, 84:232–244, 1921.
49. G. Szegö. *Orthogonal Polynomials.* American Mathematical Society, Colloqium Publications, 1939 (4th edition 1975).
50. B. Wahlberg. *On the Identification and Approximation of Linear Systems.* PhD thesis, Linköping University, 1987. Linköping Studies in Science and technology. Dissertations No. 163.
51. B. Wahlberg. Estimation of autoregressive moving-average models via high-order autoregressive approximations. *Journal of Time Series Analysis*, 10:283–299, 1989.
52. B. Wahlberg and L. Ljung. Hard frequency-domain model error bounds from least-squares like identification techniques. *IEEE Trans. Automatic Control*, 37:900–912, 1992.
53. W. Wang and M. G. Safonov. A tighter relative error bound for balanced stochastic truncation. *Systems and Control Letters*, 14:307–317, 1990.
54. P. Whittle. Estimation and information in stationary time series. *Ark. Mat. Astr. Fys.*, 2:423–434, 1953.
55. H. Wold. *A study in the analysis of Stationary Time Series.* Almqvist and Wiksell, 1938.
56. N. Young. *A Introduction to Hilbert Space.* Cambridge University Press, 1988.