Chapter 6

# SINGULAR-VALUE-DECOMPOSITION ALGORITHMS FOR LINEAR SYSTEM APPROXIMATION AND SPECTRUM ESTIMATION[1]

S. Y. Kung and K. S. Arun

## 1. INTRODUCTION

Linear system approximation has found important applications in model reduction, system identification, spectrum estimation, among others. This chapter covers the deterministic and stochastic aspects of the linear system approximation theory with special emphasis on spectrum estimation applications.

There is a fundamental difference between the identification and approximation problem formulations: The **approximation problem** formulation realistically takes into account the fact that a randomly pertur-

**Advances in Statistical Signal Processing, vol. 1, pages 203–250**

**ISBN: 0-89232-570-4**

bed signal is never completely recoverable. Thus, it offers a more robust solution domain. The real world is hardly ever linear and rational, and so any modelling procedure should attempt to find only an *approximate* and not an exact match to given data. In situations where the underlying system *is* linear and rational, we might still wish to approximate it by a reduced-order model in order to reduce computational complexity. And even if the underlying system is low-order, the given data are usually perturbed by estimation or measurement errors, making the apparent order much larger than the true system order. In such cases, instead of finding an exact high-order fit to the perturbed data, a low-order approximation is more desirable, in order to smooth out the perturbations. In fine, it appears that the modelling problem is always an approximation problem, and in Akaike's words [2a], the subject of statistical identification is essentially concerned with the art of approximation.

The system-approximation problems addressed in this review chapter may be categorized into three classes:

1. deterministic linear system approximation;
2. stochastic linear system approximation; and
3. sinusoidal approximation (or harmonic retrieval).

## 1.1 Deterministic Linear System Approximation

For deterministic systems, noniterative approximation methods can be roughly categorized into the dominant component methods, the moments matching methods, and the Hankel approximation methods. The Hankel approach, based on a Hankel matrix formed from the impulse-response sequence associated with a linear system, has recently become very popular for model reduction, system identification, etc. The mathematical foundation of the Hankel approach lies in the well-known Kronecker's Theorem [29], which says that the rank of the Hankel matrix is equal to the order of the system $H(z)$ (cf. Section 3.1). The situation is, however, very different when the measurements are corrupted by noise. The Hankel matrix formed from the impulse-response (noisy) impulse response coefficients will, in general, have a rank much exceeding the true system order. In this case, one should resort to an approximate system realization which produces some smoothened version of this sequence, rather than exactly reproducing the noisy impulse-response sequence.

### *1.1.1 Optimal Hankel-Norm Approximation*

It is natural to ask if there is an optimal approximation in any sense, within this Hankel approximation framework. The answer is in the affirmative.

For the single-input-output systems case, Adamjan et al. [1] developed a closed-form optimal solution for model-reduction problems with a Hankel-norm error criterion, where the Hankel-norm error is defined to be the spectral norm of the difference between the Hankel matrices associated with the original and the approximate systems. It has been shown that the spectral norm of the Hankel matrix associated with a stable lies between the more conventional $L_2$ and $L_\infty$ norms. Hence the Hankel-norm criterion can be viewed as a compromise between these two more popular error measures.

More significantly, the theory of bounded Hankel operators has had a far-reaching impact on mathematical system theory. For example, the eigenvalues of the Hankel operator can be shown [17] to be related to the eigenvalues of the so-called Nevanlinna matrix, classically associated with the Nevanlinna algorithm. In fact, the Adamjan et al. AAK theory [1] has provided a unifying approach to many seemingly unrelated fundamental mathematical questions, for example, approximations of Hankel operators in spectral norm, approximations in different norms, the classical Schur and Nevanlinna–Pick interpolation problems, bounded extensions of one-sided Fourier series, decompositions of periodic functions into two analytic functions bounded in the open disk, among others.

Many investigators have worked on the engineering meaning of the mathematical results of [1]. Relationships between the minimal Hankel-norm approximation and rational function approximation, $L_1$ sensitivity optimization in compensator design, etc., have resulted from these investigations. A polynomial approach that elucidates the singular value/vector properties of Hankel matrices, leads to a general theory and algorithm for optimal Hankel-norm approximation of multivariable systems, first proposed by [35]. In the same process, a fast algorithm based on a formulation employing polynomial-Hamiltonian-type systems can be derived. Recently, Glover presented an excellent review paper discussing all optimal Hankel-norm approximations of linear multivariable systems and their $L_\infty$ error bounds, in which a complete characterization of all approximations that minimize the Hankel-norm is derived [19]. The interested reader is also referred to the other related reference papers cited in [19].

### *1.1.2 Principal Components Approximation*

While the computational complexity of the optimal Hankel-norm algorithms can be reduced further, the main concern in practical implementations is the algorithm's numerical behavior and stability. Much research needs to be done to produce a numerically-robust version of the AAK algorithm.

On the other hand, there are several related approximation methods [33], [39], and [58], which are based on the singular-value decomposition (SVD) of the Hankel matrix. They all use the singular vectors corresponding to the dominant singular values only; and therefore are termed principal components (PC) methods. More importantly, this class of PC algorithms has consistently demonstrated very satisfactory numerical results for system approximation and identification. The main research left to be done regarding these algorithms is a formal approximation error analysis. This topic is also addressed in this chapter.

Since numerical performance is the major factor dictating the real-world practicality of the algorithms, we focus on the numerically more attractive PC-type approach instead of the mathematically more elegant AAK-type approach.

## 1.2 Stochastic Linear System Approximation

The stochastic realization problem has recently received a great deal of attention, since most applicational problems encountered in system identification, signal estimation, control system design, time series modeling, etc., are inherently stochastic.

### *1.2.1 Canonical Components Approximation*

The canonical correlations (c.c.) criterion for stochastic system approximation was first proposed by Akaike [26]. The c.c. coefficients between the past and the future of a time-series (cf. [21]), provide a measure of the mutual information between the past and the future in the Kullback-Leibler sense [30] and [16]. If the time-series has a finite-order Markovian representation (i.e., it can be represented as the output of a linear, rational model driven by white noise), then this information interface will have finite dimension. In fact, the number of nonzero c.c. coefficients will be exactly equal to the order of the model, and Akaike demonstrated how the minimal-order state-vector could be constructed from the canonical decomposition of the past with respect to the future [26]. Since the components of this state vector that have small c.c. coefficients (with respect to the future of the output stochastic process) also have little mutual information with the future, Akaike suggested that the low-order approximant be constructed from only these components with the dominant c.c. coefficients [3]. Subsequently, many approximation algorithms have been proposed, based on the c.c. criterion, [7], [11], and [53].

The c.c. criterion itself is the same criterion that was used in [56] for the extrapolation of a stationary random process, which is basically the

problem of predicting the future of a time-series from its past. Since the state-vector (of a linear rational system driven by white noise) summarizes all the information in the past of the output process that is necessary for the future, the c.c. criterion was suggested by Akaike for picking the state-vector of the low-order approximate model.

### *1.2.2 Principal Components Approximation*

It can be shown that the canonical components of the past with the longest c.c. coefficients (with respect to the future) do not necessarily make the minimum-variance prediction of the future. That is the case because the mutual information between a canonical component (of the past with respect to the future) and the future does not measure the (variance of the) canonical component's contribution towards the prediction of the future [27]. Instead the contribution of a state-component to the prediction of the future is better guaged by a least-squares predictive-efficiency measure of the kind used in multivariate statistical analysis by Rao [48].

The predictive efficiency criterion measures the error-variance in the prediction of the future. Using the predictive efficiency criterion for stochastic system approximation leads to the derivation of a principal-components algorithm [ ] and [27], which is a stochastic version of the principal-components method of Kung [33]. The PC approach has demonstrated better numerical performance than the c.c. methods of Desai and Pal [11] and White [53]. The reasons for the same, along with a numerical sensitivity analysis, is available in a recent dissertation [27]. We will focus primarily on the PC approach in this exposition.

## 1.3 Sinusoidal Approximation

In a majority of modern signal-processing applications, such as radar and sonar array processing, a key problem is estimating the locations of spectral lines or spectral peaks, which often represent important physical quantities, such as speed and angular direction. In this context, an important measure of performance is frequency resolution, that is, the ability to distinguish and identify spectral lines that are closely spaced in frequency. Ironically, for these signal-processing applications, the spectral estimate has to be based on short data records. As is well known, however, the frequency resolution achievable in the discrete Fourier transform is equal to the reciprocal of the data length used. Hence, the only means for achieving high resolution from a short data segment is by extrapolating the data based on certain prior knowledge.

The usual approach is to approximate the given data-record by a linear

combination of sinusoids, and to determine the sinusoid frequencies that provide the best match. The prior information used here is that signals with line-spectra are periodic, and can be uniquely extrapolated from a short record. If the given data record were truly the sum of $p$ sinusoidal signals, then only $2p$ data points would be sufficient to determine with (theoretically) infinite resolution the $p$ frequencies and $p$ amplitudes. When the problem is one of retrieving hidden periodicities from *noisy* data, however, one can at best *approximately* fit sinusoids to the data, and then, the larger the record length, the better the approximation.

### 1.3.1 *Linear Prediction-Approximation*

The first such sinusoidal approximation method was proposed by Prony, who used linear prediction (l.p) parameters to parameterize the sinusoidal model [46]. The parameters were obtained by finding a least-squares solution to the linear-prediction equations that is exactly satisfied by true sinusoidal signals. It turns out that the covariance sequence (of sinusoidal signals) also satisfies similar linear-prediction equations exactly. Obtaining a least-squares solution to a large number of l.p. equations in the raw data, is asymptotically equivalent (as record-length and number of equations increase to infinity) to *exactly* solving the first $p$ l.p. equations in the covariances [9]. The latter approach goes under the fancy name of maximum-entropy method (MEM) for high-resolution spectral estimation.

For the more practical situation, when the covariances are perturbed, Beex and Scharf [8] suggested a least-squares solution to a large number of l.p. equations in the covariance lags, and called their method, the "covariances-of-covariances" method. Improvements in numerical performance were achieved by Tufts and Kumaresan [50], who used an SVD-based least-squares solution of the l.p. equations in the covariances, after retaining only the principal-components of the covariance matrix.

### 1.3.2 *State-Space Based Principal-Components Approximation*

It turns out that the sinusoidal model is a very special case of the general linear rational model, and that linear-system approximation methods can be directly employed. In fact, deterministic system approximation methods can be used for sinusoidal approximation of raw time-series data, and stochastic system approximation methods can be used for the retrieval of harmonics from covariance lags.

Instead of using the l.p. parameters to characterize the sinusoidal model, state-space parameters can be used in order to reduce the sensitivity of the sinusoidal frequencies to the parameters, and to improve numerical performance and resolution capability.

Using a state-space based sinusoidal model, a key finite rank property can be shown for the data Hankel and Toeplitz covariance matrices. Therefore, the SVD factorization and linear system approximation schemes discussed above can be employed to produce high-resolution estimates. Under this theoretical foundation, two approximate realization methods have been proposed. First, the Toeplitz (covariance) approximation method [34] offers a robust principal-components-based spectral estimate from noisy covariance sequence. In the second case [31], the principal components approach is used to develop a direct-data approximation method for estimation directly from time-series data [32]

## 1.4 Application to Power-Spectrum Estimation

Spectral analysis forms the basis of a major part of signal processing, typically for distinguishing and tracking signals of interest, and for extracting relevant information from the data. The power spectrum of a stochastic process represents the distribution of power over frequencies and is usually defined in terms of its autocovariance sequence. Suppose that $y(t)$ is a zero-mean, wide-sense-stationary, discrete-time stochastic process, then its autocovariance sequence is defined as

$$r(m) = \mathbf{E}\{y(t) \cdot y(t+m)^*\},$$

where $\mathbf{E}\{\cdot\}$ denotes the expectation operator, and $^*$ denotes complex conjugation.

The power spectrum $P(\omega)$ is related to the infinite autocovariance sequence $\{r(m)\}$ of the process by the discrete Fourier transform

$$P(\omega) = \sum_{m=-\infty}^{+\infty} r(m) \exp\{-j\omega m\}, \qquad \omega \in (-\pi, \pi).$$

It can be shown to be equivalent to

$$P(\omega) = \lim_{N \to \infty} \mathbf{E}\left\{ \frac{1}{N} \left| \sum_{n=0}^{N-1} y(n) \exp(-jn\omega) \right|^2 \right\}$$

The spectrum estimation problem is one of estimating the power spectrum of a discrete-time stochastic process $y(t)$, given either a finite record (of one sample sequence) of the process, or estimates of its first few covariance lags. A key performance criterion is frequency resolution, the ability to reproduce sharp details in the spectrum. This is a qualitative measure of the fineness with which the spectrum is observed. For a majority of applications it is crucial to be able to resolve spectral peaks that are closely spaced in frequency. Two peaks, very close to each other, should be reproduced in the spectral estimate as two separate peaks instead of being smoothed into a single peak

### 1.4.1 *Fourier Transform Methods and Their Limitations*

Since the power spectrum of a process is equal to the Fourier transform of its infinite covariance sequence, the simplest spectrum-estimate based on a finite segment of the covariance is the Blackman–Tukey estimate

$$\hat{P}_{\mathrm{BT}}(\omega) = \sum_{m=-N}^{N} r(m) \exp\{-j\omega m\}.$$

$\hat{P}_{\mathrm{BT}}(\omega)$ is the Fourier transform of the available covariance segment assuming that the lags outside the observation interval are zero. This is equivalent to multiplying the true covariance sequence by a rectangular window; and it is well known that this causes smoothing in the spectral estimate and limits the frequency resolution to roughly the reciprocal of the observation length. The details in the spectrum are lost, and closely spaced peaks are smoothed into a single peak.

The basic problem with Fourier transform methods is that when the spectrum has sharp variations and closely spaced peaks, the large-lag covariances (that are neglected) are also significant. Moreover, when the signal is periodic, the covariance sequence is also periodic, and a zero extension outside the observation segment is rather unnatural. Hence, modern methods use additional prior information to provide a "smooth" extrapolation of the covariance outside the finite interval, and improve resolution over the fundamental limit. The use of stochastic models is one such approach.

### 1.4.2 *Model-Based Methods*

In certain applications, the physical system generating the signal can be modeled well by a linear rational system of low order. In speech, for instance, it is well known that a good model for the vocal tract is an all-pole system [37]. In general, linear rational models can be found to approximately fit the given covariances and extrapolate them outside the observation interval to infinity. These models will not only provide an infinite covariance extension (without an abrupt transition to zero), but will also smooth the perturbations in the given covariance.

Hence, current methods model the process $y(t)$ as the output of a linear, rational system driven by white noise.

If the transfer function of the model is $H(z)$, then the power spectrum of the output process is simply

$$P(\omega) = S(z)\big|_{z=\exp\{j\omega\}},$$

where $S(z) = \rho H(z)H(z^{-1})$, and $\rho$ is the variance of the input white-noise process. Thus, the power spectrum is specified completely by the

parameters of the model, and the spectrum estimation problem is reduced to that of simply estimating the model parameters. It is shown in Section 4 that when the model is exact, and the covariances are known perfectly, then the model parameters can be estimated from only $2p$ covariances, where $p$ is the model order. Thus, the power spectrum is exactly reproduceable from only $2p$ covariances. At the same time, the covariance sequence of the model output extends[2] to infinity, so that theoretically, infinite resolution is achievable.

In practice, the models are never perfect, and the available covariances are perturbed by estimation or measurement errors. Then the problem is one of approximately fitting a model to the given covariances. The spectral estimation problem is thus reduced to the problem of stochastic system identification or approximate stochastic modeling.

### *1.4.3 Applicational Problems*

Some typical applications where the problem of spectrum estimation is encountered are: interference spectrometry; the design of Wiener filters for signal recovery and image restoration; the design of channel equalizers in communication systems; the determination of formant frequencies (location of spectral peaks) in speech analysis; the retrieval of hidden periodicities from noisy data (locating spectral lines); the estimation of source-energy distribution as a function of angular direction in passive underwater sonar; the estimation of the brightness distribution (of the sky) using aperture synthesis telescopy in radio astronomy; and many others. In the last two applications, the quantity of interest is the spatial power spectrum (as a function of angular direction), and the available data are measurements from an array of sensors/telescopes. At each instant of time, the measurements form a spatial series, so that the problem is one of estimating the spatial power spectrum from an ensemble of spatial series. High resolution is an important requirement here, to be able to locate multiple sources (submarines or stars) that are close to one another [20], [14], and [10]. Resolution is also crucial in the related problem of retrieving hidden periodicities in a time series. This problem arises in Doppler radar, geophysics [47] and [51], meteorology, tidal analysis, neurophysics, and astronomy [57] and [52]. If the frequencies of the embedded sinusoids are close, good resolution capability is needed to detect two sinusoids instead of one.

### *1.4.4 Other Applications of System Approximation*

Apart from spectrum estimation applications, approximate models are used in a variety of other problems. A few representative examples are: speech recognition and coding, the design of plant control, econometrics

and meteorology for forecasting, image coding and classification, and source-wavelet modeling in seismic data processing.

## 2. NOTATION AND PROBLEM FORMULATION

### 2.1 The ARMA Difference Equation Representation

The most general, causal, linear, rational, discrete-time model is the so-called autoregressive, moving-average (ARMA) model. A $p^{\text{th}}$-order ARMA model relating the input $v(t)$ to the output $y(t)$ is described by the difference equation

$$y(t) = \sum_{k=1}^{p} a_k y(t-k) + \sum_{k=1}^{p} b_k v(t-k) + v(t). \tag{1}$$

The transfer function of this model is

$$H(z) = \frac{B(z)}{A(z)}, \tag{2}$$

where

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}, \qquad \text{and} \qquad B(z) = 1 + \sum_{k=1}^{p} b_k z^{-k}.$$

The poles of the model are the roots of the denominator polynomial $A(z)$, and the zeros are the roots of the numerator polynomial $B(z)$. The model is said to be stable if all the poles are within the unit circle on the $z$ plane. Then, the model's impulse response $i(k)$ will eventually decay to zero, as $k$ goes to infinity, and the infinite series $\sum_{k=1}^{\infty} i(k) z^{-k}$ will converge to the transfer function $H(z)$ at all points $z$ outside the unit circle.

Similarly, if we denote the covariance of the output process $y(t)$ by $r(m) = \mathbf{E}\{y(t)y(t+m)\}$, (where $\mathbf{E}$ stands for the expectation operator), then the output power spectrum is given by the doubly infinite summation

$$S(z) = \sum_{m=-\infty}^{\infty} r(m) z^{-m},$$

evaluated at $z = \exp(j\omega)$. When the input $v(t)$ is a white noise process, of variance $\rho$, then using the definition of the impulse response, $y(t) = \sum_{k=0}^{\infty} i(k)v(t-k)$, it can be easily seen that the covariance $r(m)$ is obtained by the convolution of the impulse response with its mirror

image

$$r(m) = \rho \sum_{k=0}^{\infty} i(k)i(k+m) \tag{3}$$

In $z$-transform language, this translates to $S(z) = \rho H(z)H(z^{-1})$.

Numerically, difference-equation parameterization of the ARMA model is a bad choice for pole-zero estimation or power spectrum estimation (especially when the problem is to locate peaks in the spectrum). It is well known in digital filtering theory [42] that the poles and zeros of a system are highly sensitive to perturbations in the difference-equation parameters. It was first shown by [26] that if the poles (or zeros) are tightly clustered, small errors in the difference-equation parameters can cause large variations in the poles (or zeros). Hence, parameter-estimation errors and finite-precision errors are greatly amplified in the pole estimates (spectral peak locations). In high-resolution spectrum-estimation, the poles are generally close together, and in such situations, the numerical sensitivity of the difference-equation approach hampers its resolution capability, as demonstrated in the simulations of [31] and [32].

## 2.2 State-Space Representation

In state-space notation, the $p$th order ARMA model for $y(t)$ is

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{F}\mathbf{x}(t) + \mathbf{T}v(t) \\ y(t) &= \mathbf{h}\mathbf{x}(t) + v(t) \end{aligned} \tag{4}$$

where $\mathbf{x}(t)$ is a $p \times 1$ state vector process and $\mathbf{F}$, $\mathbf{T}$, and $\mathbf{h}$ are constant matrices of sizes $p \times p$, $p \times 1$ and $1 \times p$, respectively. Henceforth, bold letters, upper case Greek letters, and script letters are used to denote matrices and vectors, and the transpose operator are denoted by a prime.

In the state-space model, the poles can be made relatively insensitive to state-space parameters by choosing an appropriate coordinate framework for the state. In other words, there is a flexibility in the choice of state coordinates, which proves beneficial to us. In general, the poles of the model are the eigenvalues of the state-feedback matrix $\mathbf{F}$ (c.f. Eq. (4)). In numerical analysis, Osborne [43] showed that the eigenvalues of a matrix are relatively well conditioned when (for each $i$) the $i^{\text{th}}$ row and the $i^{\text{th}}$ column of the matrix have the same vector norm. Fortunately, every system can be put in a coordinate framework wherein the $\mathbf{F}$-matrix satisfies the above condition. In the so-called balanced coordinates [41] and [39], the $\mathbf{F}$-matrix is sign-symmetric, that is, there exists a sign matrix $\mathbf{S}$ (whose diagonal entries are $+1$ or $-1$, and whose non diagonal entries are 0) that relates $\mathbf{F}$ to its transpose $\mathbf{F}'$ as $\mathbf{F}' = \mathbf{SFS}$. In this coordinate

framework, the **F**-matrix satisfies Osborne's condition, and the poles are relatively insensitive to perturbations in the state-space parameters. This means the problem of pole estimation via balanced state-space parameters is well conditioned, and high resolution can be achieved. A detailed sensitivity analysis is discussed in [27].

### 2.2.1 *The Impulse-Response Function and State-Space Parameters*

In terms of the state-space parameters, the transfer function of the system is given by

$$H(z) = \mathbf{h}(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{T} + 1,$$

and the poles of the model are the eigenvalues of **F**, while the zeros are the eigenvalues of the matrix (**F**-**Th**). It can be shown that the relationship between the impulse response of the model and the state-space parameters is

$$i(k) = \mathbf{h}\mathbf{F}^{k-1}\mathbf{T}, \qquad k > 0 \tag{5}$$

### 2.2.2 *The Covariance Function and State-Space Parameters*

Similarly, when the input $v(t)$ is a white-noise process of variance $\rho$, the covariance of the output process $y(t)$ is given by

$$r(m) = \begin{cases} \mathbf{h}\mathbf{P}\mathbf{h}' + \rho, & m = 0 \\ \mathbf{h}\mathbf{F}^{m-1}\mathbf{g}, & m > 0 \end{cases} \tag{6}$$

where

$$\mathbf{g} = \mathbf{F}\mathbf{P}\mathbf{h}' + \rho\,\mathbf{T} \tag{7}$$

and where $\mathbf{P} = \mathbf{E}\{\mathbf{x}\mathbf{x}'\}$ is the $p \times p$ state covariance matrix that satisfies the Lyapunov equation

$$\mathbf{P} = \mathbf{F}\mathbf{P}\mathbf{F}' + \rho\,\mathbf{T}\mathbf{T}' \tag{8}$$

## 2.3 Canonical and Balanced Forms

For a given transfer function, the parameter-triple $(\mathbf{F}, \mathbf{T}, \mathbf{h})$ is unique (if the order $p$ is minimal) modulo a similarity (coordinate) transformation. For any invertible $p \times p$ matrix $\mathbf{Q}$, the transformed triple $(\mathbf{Q}^{-1}\mathbf{F}\mathbf{Q}, \mathbf{Q}^{-1}\mathbf{T}, \mathbf{h}\mathbf{Q})$ is also a valid choice for the state-space parameters, and it corresponds to a transformed coordinate system for the state. The new state is $\mathbf{Q}^{-1}\mathbf{x}$ instead of $\mathbf{x}$, but they both span the same space, and the input–output relationship and system transfer function is left unchanged.

### 2.3.1 Canonical Form

An interesting choice of coordinates leads to a canonical form realization [25] with the state as

$$\mathbf{x}(t) = (w(t-1)w(t-2)\cdots w(t-p))'$$

where

$$W(z) = \frac{1}{A(z)}\,V(z) \qquad \text{and} \qquad Y(z) = B(z)\,W(z),$$

that is,

$$w(t) = \sum_{k=1}^{p} a_k w(t-k) + v(t)$$

and

$$y(t) = \sum_{k=1}^{p} b_k w(t-k) + w(t).$$

If we substitute the first equation into the second, we get:

$$y(t) = \sum_{k=1}^{p} (a_k + b_k) w(t-k) + v(t).$$

In state-space notation, this translates to

$$\mathbf{x}(t+1) = \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_p \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \cdot \mathbf{x}(t) + \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \cdot v(t)$$

$$y(t) = (a_1 + b_1 \quad a_2 + b_2 \quad a_3 + b_3 \cdots a_p + b_p)\,\mathbf{x}(t) + v(t)$$

This is a form that directly relates the state-space model to the difference-equation parameters.

### 2.3.2 Balanced Form

A different set of coordinates that we will be using in this paper are the so-called internally balanced coordinates [39]. The balanced realization for the ARMA model is a special case of the principal-axis realization introduced in [41]. The principal-axis realization is characterized by both the observability grammian **W** and the controllability grammian **K** being diagonal. In general, these grammians are the solutions of the following

two Lyapunov equations [25]:

$$\begin{aligned} \mathbf{K} &= \mathbf{F}\mathbf{K}\mathbf{F}' + \mathbf{T}\mathbf{T}' \\ \mathbf{W} &= \mathbf{F}'\mathbf{W}\mathbf{F} + \mathbf{h}'\mathbf{h} \end{aligned} \tag{9}$$

and are also explicitly given by

$$\mathbf{K} = (\mathbf{T} \quad \mathbf{F}\mathbf{T} \quad \mathbf{F}^2\mathbf{T} \quad \mathbf{F}^3\mathbf{T} \cdots) \begin{pmatrix} \mathbf{T}' \\ \mathbf{T}'\mathbf{F}' \\ \mathbf{T}'\mathbf{F}'^2 \\ \mathbf{T}'\mathbf{F}'^3 \\ \vdots \end{pmatrix}$$

$$= \mathscr{C} \mathscr{C}'$$

and

$$\mathbf{W} = (\mathbf{h}' \quad \mathbf{F}'\mathbf{h}' \quad \mathbf{F}'^2\mathbf{h}' \quad \mathbf{F}'^3\mathbf{h}' \cdots) \begin{pmatrix} \mathbf{h} \\ \mathbf{h}\mathbf{F} \\ \mathbf{h}\mathbf{F}^2 \\ \mathbf{h}\mathbf{F}^3 \\ \vdots \end{pmatrix}$$

$$= \Theta' \Theta$$

In linear systems terminology, the matrices $\Theta$ and $\mathscr{C}$ are known as the observability matrix and controllability matrix, respectively. Note that these matrices and the two grammians are not unique for a given transfer function, and change with the state coordinates. A transformation of the state from $\mathbf{x}$ to $\mathbf{Q}^{-1}\mathbf{x}$ changes the observability matrix to $\Theta\mathbf{Q}$ and the controllability matrix to $\mathbf{Q}^{-1}\mathscr{C}$, while changing the grammians to $\mathbf{Q}^{-1}\mathbf{K}\mathbf{Q}'^{-1}$ and $\mathbf{Q}'\mathbf{W}\mathbf{Q}$. A transformation $\mathbf{Q}$ that simultaneously diagonalizes both the grammians can always be found, and a principal-axis realization always exists [41]. In fact, for any given transfer function, many such principal-axis realizations exist, and the balanced realization is one of them.

A realization is said to be internally balanced [39] if the grammians $\mathbf{K}$ and $\mathbf{W}$ are not only diagonal, but are also *equal* to each other:

$$\mathbf{K} = \mathbf{W} = \Sigma, \qquad \text{where } \Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p)$$

A balanced realization has the following well-known properties [33]

and [44]:

1. Its parameters have a sign-symmetry, that is, there exists a sign matrix **S** (a diagonal matrix composed of only +1s and −1s) that relates the model parameters as

$$\mathbf{F}' = \mathbf{SFS}$$
$$\mathbf{T} = \mathbf{Sh}'.$$

2. The spectral norm[3] of **F** is bounded by one, that is,

$$\|\mathbf{F}\|_s \leq 1$$

These properties make the balanced realization particularly insensitive to parameter perturbations, so that parameter estimation errors do not get excessively magnified in the power spectrum estimate (cf. [27]).

## 2.4 Problem Formulation: Partial-State Selection

Intuitively, the state of a linear, rational (minimal-order) system is a summary of the information in the past input history that is both necessary and sufficient to predict the future output. In the stochastic case, where the input is white noise, and if in addition, the model is minimum-phase, then the state can also be interpreted as a summary of the past *output* history (instead of past input history) with regards to the prediction of the future output. For the sinusoidal models as well (which are zero-input models), the state summarizes the past of the output signal for *exactly* generating the future of the output signal.

In all cases, the system's state is an information-interface and its dimension is equal to the order of the system. When the system has to be approximated by a lower order model, it is thus a question of compressing this information-interface into a lower dimensional one, that is, a question of approximating the full-order state vector by a smaller sized "partial state" that contains most of the information in the full-order state.

Our general approach to the three system-approximation problems addressed in this review is as follows:

1. formulating the problem as a partial-state selection problem,
2. constructing a partial-state from the significant components of the full-order state, that is, the components that have "maximum" information about the future output, and
3. realizing the reduced-order model from the partial state.

For the second step of picking a partial state, we develop a predictive-

efficiency criterion that measures the efficiency of the partial state in predicting the future output. Fundamentally, the criterion of predictive efficiency is the same in all three related approximation problems. It hinges upon a measure of the predicting capability of the partial state for the future output. In the deterministic system approximation problem, it is a **least-squares** measure; and in the stochastic case, it is a **minimum-variance** measure.

## 2.5 Our Approach: Principal Components Approximation

It turns out that the predictive efficiency measure of the full-order state as well as for candidate partial-state vectors can be computed from the singular values of certain matrices. The "best" partial state is that constructed from the principal components in the singular-value decomposition, that is, from the singular vectors corresponding to the largest singular values.

The use of singular-value decomposition (SVD) is a plus-point for the estimation scheme, because of the good numerical behavior of SVD algorithms [28]. Moreover the SVD of a matrix can be computed efficiently on a multiprocessor matrix array with only localized communication between processors, using Given's rotations and repeated triangularization [15]. This makes an SVD-based scheme suitable for VLSI implementation and gives it a computational speed advantage over competing schemes.

When the given information is exact, (impulse-response or covariance-lag or sinusoidal data/covariance), then certain matrices constructed from the given data have rank equal to the system order, say $p$. But, perturbations in the given information destroy the rank structure of the matrix, and instead of only $p$ singular values being nonzero, all the singular values will be nonzero. It is well known in numerical analysis [49], however that the singular values of a matrix are stable and relatively insensitive to perturbations in the matrix. This means that the singular values (apart from the $p$ largest ones) will not have deviated much from zero, and the singular-value distribution can display the underlying model order. Moreover, as we shall see later, the singular values also display the magnitude of the approximation error in reduced-order models of different orders. For instance, the $(k+1)$st singular value is a lower bound on the error in approximating the full-order system by a $k$th-order model.

In addition, the singular values and vectors are relatively insensitive to perturbations in the matrix entries as well as to finite-precision errors in the computation. This makes model estimates based on SVD-principal components, numerically robust and reliable.

# 3. DETERMINISTIC SYSTEM APPROXIMATION

## 3.1 Background in Exact Realization

We saw in Eq. (5), that the impulse-response is related to the state-space parameters as

$$i(t) = \mathbf{h}\mathbf{F}^{t-1}\mathbf{T}, \quad t > 0.$$

This indicates that the infinite Hankel matrix $H$ formed from the impulse-response sequence can be factorized as:

$$\begin{pmatrix} i(1) & i(2) & i(3) & \cdots \\ i(2) & i(3) & i(4) & \cdots \\ i(3) & i(4) & i(5) & \cdots \\ i(4) & i(5) & i(6) & \cdots \\ \vdots & & & \end{pmatrix} = \begin{pmatrix} \mathbf{h} \\ \mathbf{hF} \\ \mathbf{hF}^2 \\ \mathbf{hF}^3 \\ \vdots \end{pmatrix} (\mathbf{T} \quad \mathbf{FT} \quad \mathbf{F}^2\mathbf{T} \quad \mathbf{F}^3\mathbf{T} \quad \cdots) \qquad (10)$$

$$\mathscr{H} = \Theta \cdot \mathscr{C}$$

Observe that the factors (the observability matrix $\Theta$ and the controllability matrix $\mathscr{C}$) have only $p$ rows (columns) and so the impulse-response Hankel $\mathscr{H}$ must have rank $\leq p$. In fact, if the system is of minimal order, then the $p$ columns of $\Theta$ ($p$ rows of $\mathscr{C}$) are linearly independent, and consequently, the Hankel matrix has rank equal to the order $p$. This rank property can be traced back to [29], who noted that an impulse-response sequence admits a finite dimensional realization of (minimal) order $p$, if and only if the infinite Hankel matrix formed from the sequence has rank equal to $p$. In fact, as long as the Hankel matrix has both dimensions larger than $p$, its rank must equal $p$.

First, note that since the state-space parameters are unique only up to a similarity transformation, the matrices $\Theta$ and $\mathscr{C}$ are also not unique. In fact, for any invertible $p \times p$ matrix $\mathbf{Q}$, the transformed matrices $\Theta\mathbf{Q}$ and $\mathbf{Q}^{-1}\mathscr{C}$ are valid observability and controllability matrices in a new coordinate system. Thus, any factorization of the Hankel into $\mathscr{C}$ and $\Theta$ will do, just as long as the column and row dimension of $\Theta$ and $\mathscr{C}$, respectively, are $p$. To obtain the state-space parameters, we need to start with a Hankel matrix of size $(p+1) \times p$ constructed from $i(1), i(2), \ldots, i(2p)$. The actual factorization into $\Theta$ and $\mathscr{C}$ may be done using QR factorization, LU decomposition, or singlar-value decomposition. We obtain a $(p+1) \times p$-sized observability matrix $\Theta$ whose first column is the output matrix $\mathbf{h}$. Similarly, the first column of the controllability matrix $\mathscr{C}$ will be $\mathbf{T}$. To obtain the corresponding $\mathbf{F}$ matrix, we need the following property

of $\Theta$:

$$\begin{pmatrix} \mathbf{h} \\ \mathbf{hF} \\ \mathbf{hF}^2 \\ \mathbf{hF}^3 \\ \vdots \end{pmatrix} \mathbf{F} = \begin{pmatrix} \mathbf{hF} \\ \mathbf{hF}^2 \\ \mathbf{hF}^3 \\ \mathbf{hF}^4 \\ \vdots \end{pmatrix} \tag{11}$$

Hence, if we denote the first $p$ rows of $\Theta$ by $\Theta_1$, and the last $p$ rows (second row to last row) of $\Theta$ by $\Theta_2$, then Eq. (11) can be rewritten in the new notation as

$$\Theta_1 \cdot \mathbf{F} = \Theta_2,$$

and so, the complete algorithm is the following.

**Algorithm 1: Ho-Kalman Algorithm.** *Factorize a* $(p+1) \times p$ *sized Hankel matrix* $\mathcal{H}$ *into* $\Theta \cdot \mathcal{C}$.

$$\mathbf{h} = \textit{first row of } \Theta; \ \mathbf{T} = \textit{first column of } \mathcal{C}; \ \mathbf{F} = \Theta_1^{-1} \cdot \Theta_2.$$

An algorithm on these lines was first suggested by Ho and Kalman [22].

## 3.2 The Notion of State

Intuitively, the state of a (minimal-order) system is a summary of the information in the past input history that is both necessary and sufficient to predict the future output. In fact, from the state-transition equation (4)

$$\mathbf{x}(t+1) = \mathbf{F}\mathbf{x}(t) + \mathbf{T}v(t)$$

we can see that the state is a linear combination of the past inputs:

$$\mathbf{x}(t) = (\mathbf{T} \quad \mathbf{FT} \quad \mathbf{F}^2\mathbf{T} \quad \cdots) \begin{pmatrix} v(t-1) \\ v(t-2) \\ v(t-3) \\ \vdots \end{pmatrix}$$

$$= \mathcal{C}\mathbf{V}^-(t).$$

Here, $\mathcal{C}$ is the controllability matrix defined in the previous section and $\mathbf{V}^-(t)$ is the vector of past inputs. Moreover, from the output Eq. (4),

$$y(t) = \mathbf{h}\mathbf{x}(t) + v(t).$$

we can see that if the future input is zero, that is, if $v(k) = 0 \; \forall \; k \geq t$, then

$$\begin{pmatrix} y(t) \\ y(t+1) \\ y(t+2) \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{h} \\ \mathbf{hF} \\ \mathbf{hF}^2 \\ \vdots \end{pmatrix} \cdot \mathbf{x}(t)$$

or

$$\mathbf{Y}^+(t) = \Theta \mathbf{x}(t)$$

where $\mathbf{Y}^+(t)$ is the vector of present and future outputs, and $\Theta$ is the observability matrix defined in the previous section.

Hence, the controllability matrix $\mathscr{C}$ maps the past input $\mathbf{V}^-$ into the state $\mathbf{x}$, and the observability matrix $\Theta$ maps the state vector into the future output $\mathbf{Y}^+$. Therefore, the Hankel matrix $\mathscr{H}$ constructed from the impulse response (c.f. eq. (10)), which is a composite of $\Theta$ and $\mathscr{C}$, is an operator from the past input to the future output. In fact, from the definition of impulse resonse, we have

$$\begin{pmatrix} y(t) \\ y(t+1) \\ y(t+2) \\ y(t+3) \\ \vdots \end{pmatrix} = \begin{pmatrix} i(1) & i(2) & i(3) & \cdots \\ i(2) & i(3) & i(4) & \cdots \\ i(3) & i(4) & i(5) & \cdots \\ i(4) & i(5) & i(6) & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \cdot \begin{pmatrix} v(t-1) \\ v(t-2) \\ v(t-3) \\ v(t-4) \\ \vdots \end{pmatrix}$$

$$+ \begin{pmatrix} i(0) & & & \\ i(1) & i(0) & & \\ i(2) & i(1) & i(0) & \\ i(3) & i(2) & i(1) & i(0) \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \cdot \begin{pmatrix} v(t) \\ v(t-1) \\ v(t-2) \\ v(t-3) \\ \vdots \end{pmatrix}$$

$$\mathbf{Y}^+(t) = \mathscr{H} \mathbf{V}^-(t) + \mathbf{L} \mathbf{V}^+(t), \tag{12a}$$

which can be rewritten (using $\mathscr{H} = \Theta \mathscr{C}$) as

$$\mathbf{Y}^+ = \Theta \mathbf{x} + \mathbf{L} \mathbf{V}^+, \qquad \text{where } \mathbf{x} = \mathscr{C} \mathbf{V}^- \tag{12b}$$

Hence, $\mathscr{H}$ is a two-stage operator that maps the past input $\mathbf{V}^-$ into the state $\mathbf{x}$, and the state $\mathbf{x}$ into the future $\mathbf{Y}^+$. Consequently, the rank of $\mathscr{H}$ is equal to the size of the state vector, which, in turn, is equal to the model order $p$.

### 3.3 Partial State Selection and Criterion

The situation is very different, however, when the measurements of $\{i(k)\}$ are corrupted by noise. The Hankel matrix formed from perturbed impulse-response measurements will have full rank and will no longer be factorizable into $\Theta$ and $\mathscr{C}$. Moreover, in the real world, the system under study will almost never be an exact ARMA system. The almost unavoidable presence of such imperfections in the assumed model, additive noise in the data, finite precision errors, etc., will make the Hankel matrix a full-rank matrix, and the apparent dimension of the state much larger than the model order $p$.

But it might be possible to approximate the system (or the given impulse-response measurements) by a $p$th-order model with relatively low modeling error. It is desirable to use a low-order approximate model, not only to reduce the computational complexity,[4] but also to "smooth" the perturbations in the data. Assume that we know a priori that the underlying system is actually a $p$th order ARMA model (or is sufficiently close to one), then the problem is one of constructing a $p \times 1$-sized partial state vector from the most "important" components of the full-order state. Since the purpose of the state is to summarize the past information for the future, it is therefore desirable that the "partial state" contain as much useful information as possible regarding the future output.

Mathematically, we need to determine the $p$ rows and $p$ columns of mappings $\mathscr{C}$ and $\Theta$, respectively, such that

$$\mathbf{x}_{\text{partial}} = \mathscr{C}\mathbf{V}^- \qquad \text{and} \qquad \Theta\mathbf{x}_{\text{partial}} \overset{\text{approx}}{=} \mathscr{H}\mathbf{V}^-$$

for every $\mathbf{V}^-$, where the rank of $\mathscr{H}$ is much larger than $p$.

A criterion for the quality of the approximation is the maximum least-squares error,

$$\sup_{\|\mathbf{V}^-\|_2 = 1} \|\mathscr{H}\mathbf{V}^- - \Theta\mathscr{C}\mathbf{V}^-\|_2$$

which by definition is the spectral norm of the error matrix

$$\|\mathscr{H} - \Theta\mathscr{C}\|_s$$

Since the only restriction on the approximant $\Theta\mathscr{C}$ is that it have rank $p$, the solution to the above minimization problem is obtained from the principal components in the SVD of $\mathscr{H}$. Let the SVD of $\mathscr{H}$ be

$$\mathscr{H} = \sum_{k=1}^{n} \sigma_k \mathbf{u}_k \mathbf{v}_k' = \mathbf{U}\Sigma\mathbf{V}', \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n.$$

Then, the optimal approximant is

$$\Theta\mathscr{C} = \sum_{k=1}^{p} \sigma_k \mathbf{u}_k \mathbf{v}_k' \underset{\text{def}}{\Delta} \mathbf{U}_1 \Sigma_1 \mathbf{V}_1'$$

and the error in approximating the full-order system (corresponding to the given impulse response data) by the $p$th-order model $(\Theta, \mathscr{C})$ is

$$\sup_{\|\mathbf{V}\|_2 = 1} \|\mathscr{H}\mathbf{V}^- - \Theta\mathscr{C}\mathbf{V}^-\|_2 = \sigma_{p+1},$$

the $(p+1)^{\text{th}}$ singular value (in descending order) of $\mathscr{H}$. Thus, though the rank of $\mathscr{H}$ may be very large indicating that a large order is needed for exact modeling, the distribution of the singular values of $\mathscr{H}$ provide information regarding the error in approximating the data by models of smaller order. Simultaneously, the principal singular vectors of $\mathscr{H}$ provide the optimal low-order approximant.

The next step involves determining the model parameters from the components selected for the partial state. There is an inherent problem in this estimation step, because the partial state is not a "true" state of a linear time-invariant system. This discrepancy manifests itself in many fashions. The simplest way of perceiving the problem is to note that $\Theta\mathscr{C}$ is no longer a Hankel matrix, or equivalently that the factors $\Theta$ and $\mathscr{C}$ do not have the desired structure of the observability and controllability matrices, respectively, of a time-invariant system. If $\Theta$ had the desired structure, there would have existed an exact solution $\mathbf{F}$ to the matrix equation (11)

$$\Theta_1 \mathbf{F} = \Theta_2,$$

where $\Theta_1$ and $\Theta_2$ are formed from $\Theta$ by deleting the last and first row, respectively. Because of the lack of structure, however, no exact solution exists and we have to resort to a least-squares solution that minimizes $\|\Theta_1\mathbf{F} - \Theta_2\|_F$. This leads to the least-squares estimate of the $\mathbf{F}$ matrix:

$$\mathbf{F} = \Theta_1^{\dagger}\Theta_2, \tag{13}$$

where $\Theta_1^{\dagger}$ is the pseudoinverse of $\Theta_1$ and is given by

$$\Theta_1^{\dagger} = (\Theta_1'\Theta_1)^{-1}\Theta_1'.$$

In summary, the two steps of the principal Hankel components (PHC) method are as follows:

**Algorithm 2. Principal Hankel Components (PHC) Algorithm [33]**

STEP 1. *Perform an SVD of* $\mathscr{H}$. *arrange the singular values* $\{\sigma_k\}$ *in*

*decreasing order; and partition as*:

$$\mathcal{H} = (\mathbf{U}_1 \quad \mathbf{U}_2) \cdot \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{V}_1' \\ \mathbf{V}_2' \end{pmatrix},$$

*where the* $p \times p$ *matrix* $\Sigma_1$ *contains the dominant singular values and* $\Sigma_2$ *contains the smaller singular values. Our first estimate of* $\Theta\mathcal{C}$ *is*

$$\Theta\mathcal{C} = \mathbf{U}_1\Sigma_1\mathbf{V}_1'$$

*and the (minimal) approximation error in the spectral norm is*

$$\|\mathcal{H} - \Theta\mathcal{C}\|_s = \sigma_{p+1}.$$

STEP 2. *Though the component selection step provides estimates of the product* $\Theta\mathcal{C}$, *the actual choice of the factors is not unique. Different choices will only lead to coordinate transformations of the partial state, however and not effect the input–output behavior of the final model in any way. Yet, the choice of coordinates can be crucial for numerical reasons. A good choice of coordinates for numerical stability is the balanced coordinate system. A realization in balanced coordinates is obtained by choosing the controllability-type and observability-type maps as*

$$\Theta = \mathbf{U}_1 \cdot \Sigma_1^{1/2}$$

*and*

$$\mathcal{C} = \Sigma_1^{1/2} \cdot \mathbf{V}_1'.$$

*The state-space parameters can be derived from the matrices* $\Theta$ *and* $\mathcal{C}$, *as*:

$$\mathbf{h} = \text{first row of } \Theta,$$
$$\mathbf{T} = \text{first column of } \mathcal{C},$$
$$\mathbf{F} = \Theta_1^{\dagger}\Theta_2$$

*The minimal error* $\|\Theta_1\mathbf{F} - \Theta_2\|_s$ *can be shown to be* $0(\sigma_{p+1})$ [33]

## 3.4 Features of the PHC Algorithm

### 3.4.1 *Computational Aspects*

Since the columns of $\mathbf{U}_1$ are orthonormal, the pseudoinverse of $\Theta = \mathbf{U}_1\Sigma_1$ is, in fact, simply $\Sigma_1^{-1/2}\mathbf{U}_1'$, that is,

$$\Theta^{\dagger} = \Sigma_1^{-1}\Theta'.$$

Using the matrix-inversion lemma, it can be shown that the pseudo-inverse of $\Theta_1$ is only a rank-one update of $\Sigma_1^{-1}\Theta_1'$ [33]. This reduces the computation of $\mathbf{F}$ to ordinary matrix multiplications without any matrix inversion.

### *3.4.2 Error Analysis*

We saw that $\|\mathscr{H} - \Theta\mathscr{C}\|_2 = \sigma_{p+1}$ and that $\|\Theta_1\mathbf{F} - \Theta_2\|_F = o(\sigma_{p+1})$. Thus, the singular values of $\mathscr{H}$ display the order of magnitude of the error in approximating the given data by low-order models (of all possible orders). Thus, when the underlying system is truly ARMA, the singular-value distribution will indicate its order, if the perturbations are small. Moreover, when $\sigma_{p+1}$ is small, the modeling error is guaranteed to be small, because in a statistical sense [33], we can show that

$$\sum_{k=1}^{2p+1} |i(k) - \mathbf{h}\mathbf{F}^{k-1}\mathbf{\Gamma}|^2 \leq (p+2)^{1/2}\,\sigma_{p+1}$$

### *3.4.3 Obtaining a Balanced Realization*

When the given impulse-response coefficients correspond exactly to a $p$th-order ARMA model, then $\sigma_{p+1} = 0$, and there is no modeling error. Then, $\mathbf{U}_1\Sigma_1^{1/2}$ (and $\Sigma_1^{-1/2}\mathbf{V}_1'$) will be the true observability matrix (and controllability matrix) of the identified model $(\mathbf{F}, \mathbf{\Gamma}, \mathbf{h})$. Therefore, in the exact realization problem, the identified model satisfies

$$\Theta'\Theta = \mathscr{C}\mathscr{C}' = \Sigma_1.$$

This indicates that the two grammians are equal and diagonal, and so the model is in balanced coordinates. Consequently, the transfer function, the poles, and the power spectrum are less effected by parameter errors and numerical errors (c.f. Appendix A). This is an inherent advantage of identification methods that use balanced state-space parameters instead of difference-equation parameters.

## 3.5 Simulation Examples: Seismic Applications of the PHC Algorithm

We have very successfully applied the PHC algorithm to the modeling of real seismic data. In Figures 1 and 2 we depict third- through tenth-order ARMA realizations (shown dashed) for two seismic source signatures (shown solid). To help illustrate the relationship between **singular values** and the **approximation errors,** plots of singular values are given in Figure 1(b) and Figure 2(b).

Figure 1 is for a Bolt Standard Test Shot of air gun model 600B – 20 in$^3$. The sharp peaks, which are probably not very realistic, make this a very difficult wavelet to model. By the PHC algorithm, the wavelet's shape is already satisfactorily obtained with a fourth-order model. (Note that a seventh-order model gives a nearly perfect fit.)

Data from another air gun is modeled as shown in Figure 2(a). The air

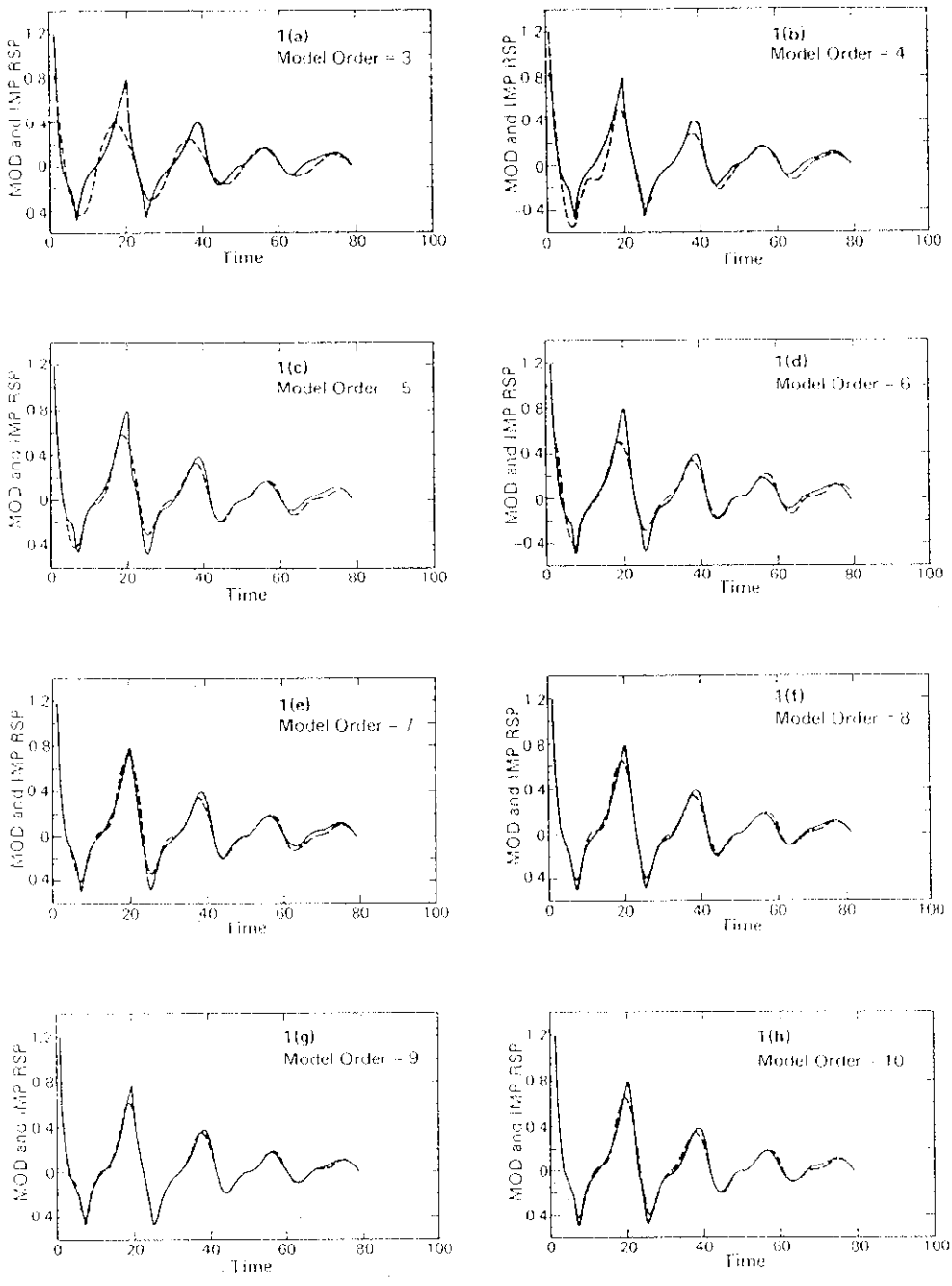


*Figure 1* Third- through tenth-order models for Bolt air gun wavelet

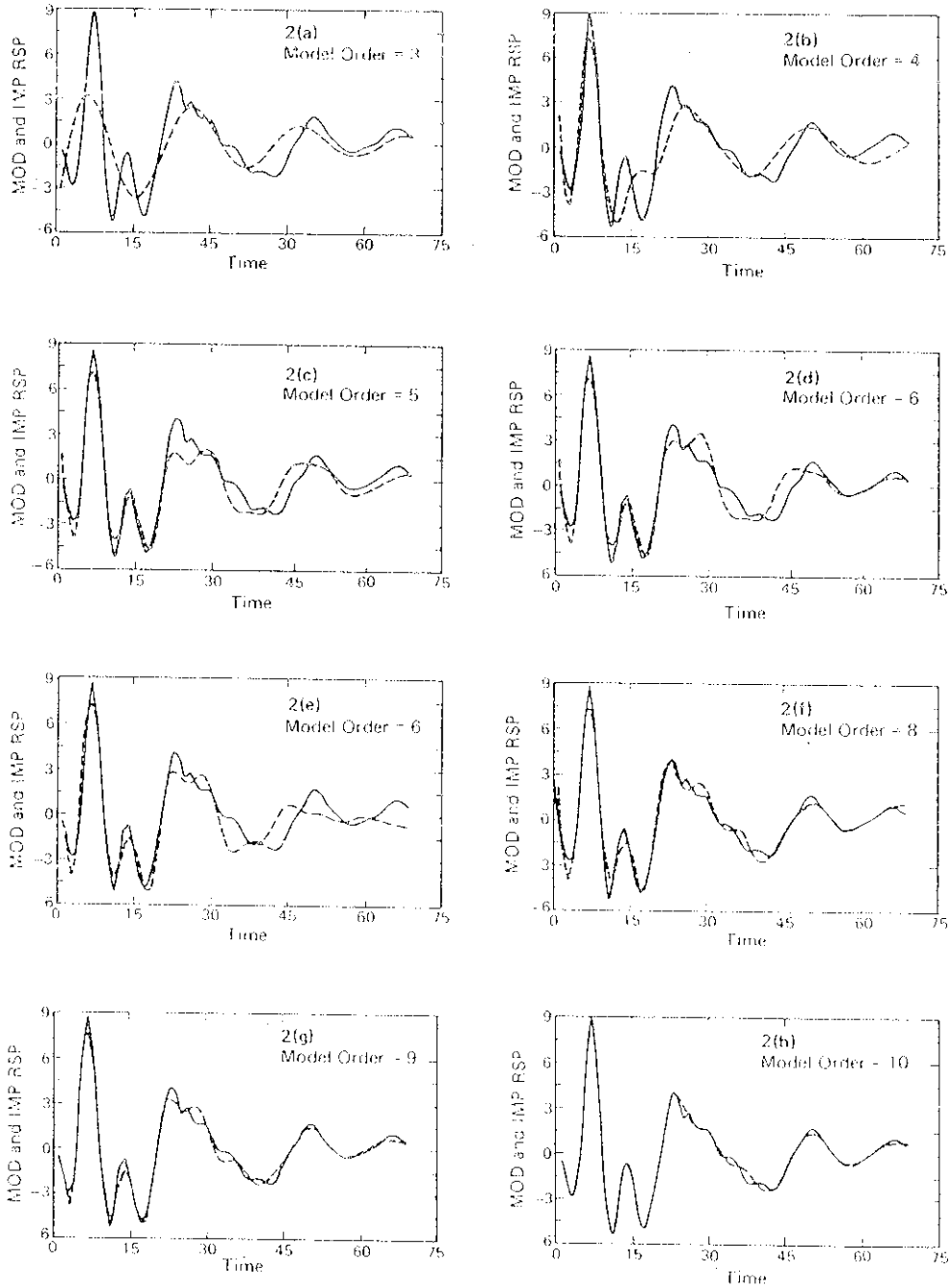


*Figure 2* Third- through tenth-order models for a second air gun wavelet.

gun's impulse response is quite difficult, but a tenth-order ARMA model captures the correct waveshape quite well.

It appears from these examples, as well as many others [38], including some water gun wavelet modelings, that the PHC algorithm can model seismic sources quite well by at most tenth-order ARMA systems. The $\{\mathbf{F}, \mathbf{h}, \mathbf{T}\}$ matrices for the tenth-order models are "internally balanced" (with approximation of less than 1.0 E-03), hence *the triples* $\{\mathbf{F}, \mathbf{h}, \mathbf{T}\}$ *for all lower order models are embedded in these matrices as well*. For an example, the $4 \times 4$ principal minor of $\mathbf{F}$ is the state-feedback matrix for the fourth-order model, and so on.

REMARKS.

1. *In practice, only finite* impulse-response data are available. The PHC algorithm can adapt to the finite data situation easily.
2. The algorithm, in general, achieves very good accuracy of approximation at a reasonably low order.
3. The poles of the approximant systems are all stable.
4. Both PHC Steps 1 and 2 are reported to be numerically stable and well conditioned. This is because the use of SVD for the system approximation problem gives good numerical properties to the PHC algorithm.
5. In a straightforward manner, the algorithm can be extended to the *multivariable case*. See [36] for an applicational example of multivariable modeling of macroeconomic systems (based on the PHC algorithm) as well as a closed-loop simulation result of the (reduced-order) controller.

## 4. STOCHASTIC SYSTEM APPROXIMATION

### 4.1 Background in Exact Stochastic Realization [12]

Recollect from Eq. (6) that when the input to the system is a white-noise process, the covariance of the output process satisfies

$$r(m) = \mathbf{h}\mathbf{F}^{m-1}\mathbf{g},$$

which is very similar to Eq. (5) for the impulse response. As a result, the Hankel matrix $\mathbf{H}$ built from the covariance sequence is also factorizable into an observability matrix $\Theta$ and a controllability like matrix $\mathcal{G}$, as shown below.

$$\begin{pmatrix} r(1) & r(2) & r(3) & \cdots \\ r(2) & r(3) & r(4) & \cdots \\ r(3) & r(4) & r(5) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \mathbf{h}' \\ \mathbf{h}'\mathbf{F}' \\ \mathbf{h}'\mathbf{F}^{2\prime} \\ \vdots \end{pmatrix} (\mathbf{g}, \mathbf{F}\mathbf{g}, \mathbf{F}^2\mathbf{g}, \cdots) \tag{14}$$

$$\mathbf{H} = \mathbf{\Theta}\mathbf{\mathcal{G}}.$$

Thus a covariance Hankel matrix with both row and column dimension not less than $p$ will have rank equal to $p$.

On the lines of Algorithm 1, the parameters $\mathbf{F}$, $\mathbf{g}$, and $\mathbf{h}$ (corresponding to a particular choice of coordinates) can be obtained by factoring a $(p+1) \times p$-sized Hankel matrix built from the $2p$ covariance lags $r(1)$, $r(2)$, $\cdots$, $r(2p)$. Thus, the system poles are readily obtained as the eigenvalues of matrix $\mathbf{F}$.

If one wants the input matrix $\mathbf{T}$ as well, an extra step is necessary, akin to spectral factorization. At this stage, $\mathbf{F}$ and $\mathbf{h}$ (equivalently, the system poles) are available, and only $\mathbf{T}$ (the system zeros) need be evaluated. The covariance sequence $r(m)$ is unaffected, however, by a reflection of some (or all) zeros with respect to the unit circle on the $z$ plane. Consequently, the triple $(\mathbf{F}, \mathbf{g}, \mathbf{h})$, which is determined entirely by the covariance lags, is common to a number of stable models with the same set of poles, and with some (or all) zeros reflected across the unit circle. Thus each of these models (with the same covariance sequence) can be put in a coordinate system where they share the same $(\mathbf{F}, \mathbf{g}, \mathbf{h})$ and can differ only in the values of the input matrix $\mathbf{T}$ and the input variance $\rho$. We thus have a number of competing models (with the same $\mathbf{F}$, $\mathbf{g}$, and $\mathbf{h}$) driven by different white-noise sequences that generate the same process $y(t)$ and the same output covariance sequence $r(m)$. Of all these numerous choices, we concentrate on identifying the *minimum-phase* model, which has all zeros within the unit circle.

In Faurre's pioneering work [12, 13], the state-variance $\mathbf{P}$ of the minimum-phase model is obtained as the *smallest*,[5] symmetric positive-definite solution to the algebraic Riccati equation:

$$\mathbf{P} = \mathbf{F}\mathbf{P}\mathbf{F}' + (\mathbf{g} - \mathbf{F}\mathbf{P}\mathbf{h}')(r(0) - \mathbf{h}\mathbf{P}\mathbf{h}')^{-1}(\mathbf{g} - \mathbf{F}\mathbf{P}\mathbf{h}')' \tag{15}$$

We can quite easily verify that the state variance of the minimum-phase model as well as of all the non-minimum-phase models that share the triple $(\mathbf{F}, \mathbf{g}, \mathbf{h})$ must satisfy this Riccati equation. First, recall that the state-variance $\mathbf{P}$ satisfies the Lyapunov equation (8): $\mathbf{P} = \mathbf{F}\mathbf{P}\mathbf{F}' + \rho\mathbf{T}\mathbf{T}'$, and that by definition, $\mathbf{g} = \mathbf{F}\mathbf{P}\mathbf{h}' + \rho\mathbf{T}$ (see Eq. (7)). Moreover, the output variance $\rho$ is related to the state-space parameters via the first part of Eq.

(6): $r(0) = \mathbf{hPh}' + \rho$. Combining these three equations, we get the Riccati equation.

Since the constants $r(0)$, $\mathbf{F}$, $\mathbf{g}$, and $\mathbf{h}$ in the Riccati equation are common to all the models (minimum-phase, as well as non-minimum-phase) they will all have state variances that satisfy the Riccati equation, though their input matrices $\mathbf{T}$ will be different. Among these many choices, the minimum-phase model will have the smallest state variance $\mathbf{P}_*$, and the maximum-phase model (with all zeros outside the unit circle) will have the largest state-variance $\mathbf{P}^*$ [13], [4], [55].[6]

From $\mathbf{P}_*$ we may obtain the input parameters for the minimum phase model easily. In summary, Faurre's algorithm for exact stochastic realization is as follows.

**Algorithm 3: Faurre's Algorithm.** *Factorize a* $(p+1) \times p$*-sized covariance Hankel matrix* $\mathbf{H}$ *as*:

$$\mathbf{H} = \Theta \cdot \mathcal{G}$$

$$\mathbf{F} = \Theta_1^{-1} \cdot \Theta_2$$

$$\mathbf{g} = 1^{st} \text{ row of } \mathcal{G}$$

$$\mathbf{h} = 1^{st} \text{ col of } \Theta$$

*Obtain the state-variance* $\mathbf{P}_*$ *of the minimum-phase model as the smallest solution of the Riccati equation*

$$\mathbf{P} = \mathbf{FPF}' + (\mathbf{g} - \mathbf{FPh}')(r(0) - \mathbf{hPh}')^{-1}(\mathbf{g} - \mathbf{FPh}')'$$

Then,

$$\rho_* = r(0) - \mathbf{hP}_*\mathbf{h}'$$

$$\mathbf{T}_* = (\mathbf{g} - \mathbf{FP}_*\mathbf{h}')/\rho_*.$$

## 4.2 The Notion of State as Derived from Predictor Space Concepts

We first need to define some more notation. For a zero-mean random vector $\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$, Span($\mathbf{Y}$) will denote the Hilbert space of all random variables that are linear combinations of $\{y_1, y_2, \ldots, y_n\}$. The inner product on this space of zero-mean random variables is the cross covariance, and its dimension (upper bounded by $n$) is the largest number of mutually uncorrelated random variables in the space.

We will need the notation $\mathbf{x}\backslash\mathbf{Y}$ to denote the linear, minimum-variance estimator of zero-mean random vector $\mathbf{x}$ from the zero-mean random vector $\mathbf{Y}$. It is also the orthogonal projection $\mathbf{x}$ onto the subspace Span($\mathbf{Y}$). From elementary estimation theory [ ], we know that

$$\mathbf{x}\backslash\mathbf{Y} = \mathbf{E}\{\mathbf{xY}'\} \cdot (\mathbf{E}\{\mathbf{YY}'\})^{*-1}\mathbf{x}$$

When the input to our linear system of Eq. (4) is a white-noise process, the past and future inputs ($\mathbf{V}^-$ and $\mathbf{V}^+$) are uncorrelated, and the two components of the future output vector $\mathbf{Y}^+$ in Eq. (12a):

$$\mathbf{Y}^+ = \mathscr{H}\mathbf{V}^- + \mathbf{L}\mathbf{V}^+$$

are orthogonal. As a result, the orthogonal projection of $\mathbf{Y}^+$ on Span($\mathbf{V}^-$) must be $\mathscr{H}\mathbf{V}^-$ itself, that is,

$$\mathbf{Y}^+ \backslash \mathbf{V}^- = \mathscr{H}\mathbf{V}^-.$$

We saw in the last section, however, that this information is completely summarized in the state, since

$$\mathscr{H}\mathbf{V}^- = \Theta\mathbf{x}, \qquad \text{where } \mathbf{x} = \mathscr{C}\mathbf{V}^-.$$

Therefore, $\mathbf{Y}^+ \backslash \mathbf{V}^- = \Theta\mathbf{x}$, which is a mathematical statement of the fact that the state contains all the information in the past input for predicting the future output.

### 4.2.1 *Minimum-Phase Model (Innovations Representation)*

When the model is minimum phase, we can show that the state can be also interpreted as a summary of the past *output* history (instead of past input history) with regard to the prediction of the future output.

Since the minimum-phase model has all zeros within the unit circle, it has a stable inverse. The inverse filter is obtained by simply rearranging the model equations (4) as

$$\begin{aligned} \mathbf{x}(t+1) &= (\mathbf{F} - \mathbf{T}\mathbf{h})\mathbf{x}(t) + \mathbf{T}y(t) \\ v(t) &= -\mathbf{h}\mathbf{x}(t) + y(t). \end{aligned} \tag{16}$$

The minimum-phase property ensures that the zeros of the model, which are the eigenvalues of $(\mathbf{F} - \mathbf{T}\mathbf{h})$, lie within the unit circle. But these eigenvalues are precisely the poles of the inverse filter, so this implies that the inverse filter is stable. Thus the state process $\mathbf{x}(t)$ as well as the input process $v(t)$ can be obtained causally from the output $y(t)$ using the above filter.

The state transition equation of the inverse filter indicates that

$$\mathbf{x}(t) = (\mathbf{T} \quad (\mathbf{F} - \mathbf{T}\mathbf{h})\mathbf{T} \quad (\mathbf{F} - \mathbf{T}\mathbf{h})^2\mathbf{T} \quad (\mathbf{F} - \mathbf{T}\mathbf{h})^3\mathbf{T} \quad \cdots) \begin{pmatrix} y(t) \\ y(t+1) \\ y(t-2) \\ y(t-3) \\ \vdots \end{pmatrix} \tag{17}$$

$$= \Psi\mathbf{Y}^-(t)$$

Hence, Eq. (12) can be rewritten for the minimum phase model as

$$\mathbf{Y}^+ = \Theta\Psi\mathbf{Y}^- + \mathbf{L}\mathbf{V}^+ \tag{12c}$$

When the input is white, the two terms are orthogonal, since the future input $\mathbf{V}^+$ is uncorrelated with the past output $\mathbf{Y}^-$. Hence,

$$\mathbf{Y}^+ \backslash \mathbf{Y}^- = \Theta\Psi\mathbf{Y}^- = \Theta\mathbf{x}.$$

Thus, for the minimum phase model, we have

$$\mathbf{x} = \Psi\mathbf{Y}^- \qquad \text{and} \qquad \mathbf{Y}^+ \backslash \mathbf{Y}^- = \Theta\mathbf{x}, \tag{18}$$

which means that the state of the minimum-phase model summarizes the past *output* history for predicting the future output.

As a footnote, Eq. (18) indicates that the projection of $y(t)$ on the past space Span($\mathbf{Y}^-$) is nothing but

$$y(t) \backslash \mathbf{Y}^- = \mathbf{h}\mathbf{x}(t).$$

Therefore, the part of $y(t)$ that cannot be predicted from the past $\mathbf{Y}^-$ is simply

$$y(t) - \mathbf{h}\mathbf{x}(t) = v(t).$$

Thus, the input white noise to a minimum-phase model is the *innovations* process [23] for the output, and consequently, the minimum-phase ARMA model is also called the **innovations representation** (IR) of the output process [5], [18].

As a second footnote, the state of the minimum-phase model is completely reproduceable from the infinite past $\mathbf{Y}^-$ as in Eq. (17), and the inverse filter of Eqs. (16) is the *steady-state Kalman filter* for the model. The state $\mathbf{x}(t)$ of the minimum-phase model is, consequently, also the state of the Kalman filter, which is an estimate of the state of every other non-minimum-phase model (with the same $\mathbf{F}$ and $\mathbf{h}$ matrices) estimated from the infinite past $\mathbf{Y}^-$. The state variance of the minimum-phase model has to be smaller than the state variance of every non-minimum-phase model with the same $\mathbf{F}$ and $\mathbf{h}$ matrices. This explains why $\mathbf{P}_*$ is the smallest solution of the algebraic Riccati equation (15), and provides a justification for Faurre's algorithm (c.f. Algorithm 2).

Since Faurre's algorithm generates the minimum-phase model from the covariance information, it also gives us a Kalman filter directly from the covariance, and supports the claim in [5], [18], and [6] that the Kalman filter can be constructed directly from the covariance data without an underlying model assumption.

## 4.3 Partial-State Selection

Using elementary estimation theory, one can verify that

$$\mathbf{Y}^+ \backslash \mathbf{Y}^- = \mathbf{H}\mathbf{R}^{-1}\mathbf{Y}^-$$

where $\mathbf{R} = \mathbf{E}(\mathbf{Y}^-\mathbf{Y}^{-\prime})$ and $\mathbf{H} = \mathbf{E}(\mathbf{Y}^+\mathbf{Y}^{-\prime})$ are the Toeplitz and Hankel matrices formed from the covariance lags of the output process. $\mathbf{H}$ is given from Eq. (14) and $\mathbf{R}$ is given by

$$\mathbf{R} = \begin{pmatrix} r(0) & r(-1) & r(-2) & \cdots \\ r(1) & r(0) & r(-1) & \cdots \\ r(2) & r(1) & r(0) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Thus $\mathbf{H}\mathbf{R}^{-1}$ must be factorizable into $\Theta$ and $\Psi$, and consequently it must have rank equal to the size of the state vector (i.e. equal to the model order p).

When the covariance lags are estimated from a finite record of the stochastic process or are directly measured, however, then the perturbations in the lags will distort the rank structure of $\mathbf{H}\mathbf{R}^{-1}$. It will have full rank, making the apparent state size much larger than the true model order. The problem is once again that of constructing a partial state from those components in Span($\mathbf{Y}^-$) that contain the most information regarding $\mathbf{Y}^+$. The partial state must effectively summarize the information interface between $\mathbf{Y}^+$ and $\mathbf{Y}^-$.

Note that the problem is one of compressing $\mathbf{Y}^-$ while retaining maximal information not about $\mathbf{Y}^-$ but about $\mathbf{Y}^+$. Hence, the direct principal-components analysis (i.e., Karhunen–Loeve decomposition of $\mathbf{Y}^-$) of $\mathbf{Y}^-$ will not suffice [56], [24][7] for the partial-state selection problem. The compression of $\mathbf{Y}^-$ into its principal components is not appropriate because it is based on the selection of components containing the maximum information about $\mathbf{Y}^-$ itself, whereas only specific information about $\mathbf{Y}^+$ is of interest in the partial-state selection problem.

There exist in the statistical literature, however, generalizations of the concept of principal components (of a random vector) to the problem of compressing the information interface between two random vectors (which will henceforth be referred to as the 2-vector problem for the sake of brevity).

## 4.4 The Predictive Efficiency Criterion

A criterion used for the 2-vector problem in statistical literature, is the predictive-efficiency criterion that minimizes the total variance of the error in predicting $\mathbf{Y}^+$ from a candidate summary $\mathbf{x}_{\text{partial}} = \Psi\mathbf{Y}^-$. The

problem is then, of picking the $p \times 1$ partial state $\mathbf{x}_{\text{partial}} = \Psi \mathbf{Y}$ to minimize $\mathbf{E}\|\mathbf{Y}^+ - \mathbf{Y}^+\backslash\mathbf{x}\|^2$.

The inherent constraint here is that $\Psi$ should have only $p$ rows [8]. Such a criterion was first used by Rao in multivariate statistics for the 2-vector problem [48]. Since $\mathbf{x} = \Psi\mathbf{Y}^+$, it can be shown using elementary estimation theory that

$$\begin{aligned} \mathbf{Y}^+\backslash\mathbf{x} &= \mathbf{E}(\mathbf{Y}^+\mathbf{x}')(\mathbf{E}(\mathbf{x}\mathbf{x}'))^{-1}\mathbf{x}, \\ \mathbf{Y}^+\backslash\mathbf{x} &= \mathbf{H}\Psi'(\Psi\mathbf{R}\Psi')^{-1}\mathbf{x}, \end{aligned} \tag{19}$$

and the prediction error to be minimized is

$$\text{trace}(\mathbf{R} - \mathbf{H}\Psi'(\Psi\mathbf{R}\Psi')^{-1}\Psi\mathbf{H}')$$

Equivalently, we have to choose a $p \times n$ matrix $\Psi$ that maximizes

$$\text{trace}((\Psi\mathbf{H}'\mathbf{H}\Psi')(\Psi\mathbf{R}\Psi')^{-1}).$$

The solution to this optimization problem is: The $p$ rows of $\Psi$ must be a basis for the space spanned by the $p$ generalized eigenvectors of the matrix pencil $(\mathbf{H}'\mathbf{H}, \mathbf{R})$, corresponding to the $p$ largest generalized eigenvalues.

If $\mathbf{R}$ is invertible, as is the case when the model is strictly stable, and if $\mathbf{u}_k$ is an eigenvector of $\mathbf{H}\mathbf{R}^{-1}\mathbf{H}'$, that is,

$$(\mathbf{H}\mathbf{R}^{-1}\mathbf{H}')\mathbf{u}_k = \lambda_k \mathbf{u}_k$$

then $\mathbf{R}^{-1}\mathbf{H}'\mathbf{u}_k$ is a generalized eigenvector of the pair $(\mathbf{H}'\mathbf{H}, \mathbf{R})$ corresponding to the same eigenvalue $\lambda_k$, that is,

$$\mathbf{H}'\mathbf{H}(\mathbf{R}^{-1}\mathbf{H}'\mathbf{u}_k) = \lambda_k \mathbf{R}(\mathbf{R}^{-1}\mathbf{H}'\mathbf{u}_k)$$

Thus, $\Psi$ can be obtained from the principal components of the matrix $\mathbf{H}\mathbf{R}^{-1}\mathbf{H}'$. Let the eigendecomposition (or SVD) of $\mathbf{H}\mathbf{R}^{-1}\mathbf{H}'$ be

$$\mathbf{H}\mathbf{R}^{-1}\mathbf{H}' = \mathbf{U}\Psi^2\mathbf{U}' = \mathbf{U}_1\Psi_1^2\mathbf{U}_1' + \mathbf{U}_2\Psi_2^2\mathbf{U}_2'$$

and let subscript 1 denote the principal components, as before. Then the predictive-efficiency criterion is optimized when

$$\Psi = \mathbf{A}\mathbf{U}_1'\mathbf{H}\mathbf{R}^{-1}$$

where $\mathbf{A}$ is *any* $p \times p$ invertible matrix. For reasons that will become clear later, we call this solution the unweighted principal components (UPC) solution.

## 4.5 The UPC Algorithm

After picking the partial-state components using the predictive efficiency criterion, we still have to obtain the corresponding model-parameter estimates. The parameter-estimation step is taken from the

deterministic-modeling algorithm of Section 3.3. It is assumed here that the model order $p$ is estimated (or given) prior to the model parameter estimation. From that point on, the rest of the unweighted principal components (UPC) algorithm is:

**Algorithm 4: The UPC Algorithm**

STEP 1. *Perform an eigendecomposition of*

$$\mathbf{HR}^{-1}\mathbf{H}' = \mathbf{U}\Sigma^2\mathbf{U}' = \mathbf{U}\Sigma_1^2\mathbf{U}_1' + \mathbf{U}_2\Sigma_2^2\mathbf{U}_2'$$

*and retain only the principal components (denoted by subscript 1). Now* $\Psi$ *can be any basis from the row span of* $\mathbf{U}_1'\mathbf{HR}^{-1}$, *that is,*

$$\Psi = \mathbf{A}\mathbf{U}_1'\mathbf{HR}^{-1} \qquad \text{for an invertible } p \times p \text{ matrix } \mathbf{A}.$$

*But, different choices of* $\mathbf{A}$ *will only correspond to coordinate transformations of the partial state. The partial state will be in balanced coordinates if we choose*

$$\Psi = \Sigma_1^{-1/2}\mathbf{U}_1'\mathbf{HR}^{-1}.$$

*Then, Eq. (19) indicates that*

$$\mathbf{Y}^+\mathbf{x}_{\text{partial}} = \mathbf{H}\Psi'(\Psi\mathbf{R}\Psi')^{-1}\mathbf{x}_{\text{partial}}$$

*which implies that the observability matrix is*

$$\begin{aligned}\Theta &= \mathbf{HR}^{-1}\mathbf{H}'\mathbf{U}_1\Sigma_1^{1/2}(\Sigma_1^{1/2}\mathbf{U}_1'\mathbf{HR}^{-1}\mathbf{H}'\mathbf{U}_1\Sigma_1^{1/2})^{-1} \\ &= \mathbf{U}_1\Sigma_1^{-1/2}.\end{aligned}$$

*The state variance is* $\Psi\mathbf{R}\Psi' = \Sigma_1$ *and the observability grammian, which is* $\Theta'\Theta$ *is also equal to* $\Sigma_1$. *Hence the partial-state is in balanced coordinates.*

STEP 2: *The partial state is not a "true" state of a linear time-invariant system, and the* $\Theta$ *and* $\Psi$ *matrices do not have the required structure. Hence, as in the deterministic algorithm of Section 3.3, we resort to a second approximation and* $\mathbf{F}$ *is obtained as the least-squares solution of*

$$\Theta_1\mathbf{F} = \Theta_2,$$

*where* $\Theta_1$ *and* $\Theta_2$ *are as defined previously. Moreover, from Eqs. (10) and (17), we can see that* $\mathbf{h}$ *and* $\mathbf{T}$ *are the first row and column of* $\Theta$ *and* $\Psi$, *respectively. Therefore, the parameter estimates are:*

$$\begin{aligned}\mathbf{h} &= \textit{first row of } \Theta \\ \mathbf{T} &= \textit{first column of } \Psi \\ \mathbf{F} &= \Theta_1^{\dagger}\Theta_2,\end{aligned}$$

*where the superscript* $\dagger$ *stands for the pseudoinverse.*

## 4.6 Comparison to the Canonical Correlations Criterion and Akaike's Method

The predictive-efficiency criterion sheds new light on Akaike's canonical (c.c.) correlations criterion [3] and clearly illustrates the disadvantage of the inherent normalization in c.c. analysis.

It can be shown that the partial-state components selected by the c.c. approach maximize the predictive efficiency of the partial state for a *normalized* future vector. The normalized problem is one of picking $\mathbf{x}_{\text{partial}} = \Psi\mathbf{Y}^-$ to minimize

$$\mathbf{E}\|\mathbf{R}^{-1/2}\mathbf{Y}^+ - \mathbf{R}^{-1/2}\mathbf{Y}^+\backslash\mathbf{x}\|^2 = \text{trace}(\mathbf{I} - \mathbf{R}^{-1/2}\mathbf{H}\Psi'(\Psi\mathbf{R}\Psi')^{-1}\Psi\mathbf{H}'\mathbf{R}^{-1/2}),$$

and the solution is obtained from the principal generalized eigenvectors of the matrix pair $(\mathbf{H}\mathbf{R}^{-1}\mathbf{H}', \mathbf{R})$ or, equivalently, from the principal components of the matrix $\mathbf{R}^{-1/2}\mathbf{H}\mathbf{R}^{-1/2}$. Thus, Akaike's method maximizes the predictive efficiency of the partial state for the normalized future vector $\mathbf{Z}^+ = \mathbf{R}^{-1/2}\mathbf{Y}^+$. This indicates that the partial state components picked by Akaike's method do not do the best job of predicting the future. It is this inherent normalization in the c.c. criterion that causes it to ignore the strength of the modes in the following counterexample.

### 4.6.1 *A Counter Example*

Assume we are given the exact covariances of a fourth-order model whose impulse response is

$$i(t) = (1-\epsilon)\gamma^t \cos\omega_1 t + \epsilon\gamma^t \cos\omega_2 t, \qquad \epsilon \ll 1, \quad \gamma < 1,$$

so that one mode is much stronger than the other. We wish to reduce the model order from 4 to 2. Intuitively, it seems we should pick the stronger mode, but it turns out that when the poles are sufficiently close to the unit circle ($\gamma \to 1$), the 4 nonzero c.c. coefficients will all be equal, and will not display the difference in the amplitudes $(1-\epsilon)$ and $\epsilon$, so that the c.c. approximation may not pick the stronger mode.

In the following analysis, we have to reduce the input white-noise variance $\rho$ to zero (as the pole-radius $\gamma$ is increased to unity), in order to keep the variance $r(0)$ of the output process constant. When $\gamma$ is sufficiently close to one, $\rho = 2(1-\gamma^2)$ will ensure that $r(0) = 1$.

CLAIM. As $\gamma \to 1$, and $\rho = 2(1-\gamma^2)$, the c.c. coefficients converge to

$$(1, 1, 1, 1, 0, 0, \ldots)$$

For the proof of this claim, the reader is referred to [27].

The above counterexample illustrates the effect of the inherent nor-

malization in c.c. analysis: The c.c. coefficients are normalized correlations that do not contain any information pertaining to the strength of the component modes, and hence do not effectively measure the component's significance to the model. In the limiting case, ($\gamma = 1$), c.c. analysis could very well pick the weaker mode. This lack of strength information in the c.c. coefficients makes them very sensitive to small perturbations. When the impulse response consists of only one damped sinusoid, only two c.c. coefficients are nonzero, and the rest are zero. When we add a small perturbation of the form $\epsilon\gamma^t \cos \omega_2 t$ (possibly 60-Hz leakage from the power lines) to the impulse response, the third and fourth c.c. coefficients are dramatically changed from 0 to nearly 1, irrespective of how small $\epsilon$ is. This makes the approximate model obtained by the c.c. analysis very sensitive to perturbations.

## 4.7 Relation to the Deterministic PHC Method

The connections between the UPC algorithm and the PHC algorithm (for deterministic modeling) run much deeper than apparent at first sight. Not only do the methods share a similar second step, but also the first step approximations (the partial-state selection step) are closely related. Recall that for the minimum phase model, $\mathbf{Y}^+ \backslash \mathbf{Y}^- = \Theta \mathbf{x}$, which in turn is equal to $\mathscr{H}\mathbf{V}^-$ because $\mathbf{x} = \mathscr{C}\mathbf{V}^-$. Moreover, we saw that $\mathbf{Y}^+ \backslash \mathbf{Y}^- = \mathbf{H}\mathbf{R}^{-1}\mathbf{Y}^-$. Combining the two, we get

$$\mathbf{H}\mathbf{R}^{-1}\mathbf{Y}^- = \mathscr{H}\mathbf{V}^-.$$

Therefore, the covariance matrices of the two vectors must also be the same. And thus, we come to the rather surprising result:

$$\mathbf{H}\mathbf{R}^{-1}\mathbf{H}' = \rho\mathscr{H}\mathscr{H}', \tag{20}$$

where $\rho$ is the variance of the input white-noise process.

Thus, an eigendecomposition of $\mathbf{H}\mathbf{R}^{-1}\mathbf{H}$ is equivalent to the SVD of the impulse-response Hankel $\mathscr{H}$. A perturbed covariance sequence corresponds to a model whose order is much larger than the true system order. Then, the matrix $\mathbf{H}\mathbf{R}^{-1}\mathbf{H}'$ constructed from the perturbed covariances will factorize into a product of full-rank Hankel matrices $\mathscr{H}$, and a principal components approximation of $\mathbf{H}\mathbf{R}^{-1}\mathbf{H}$ will correspond to a principal components approximation of the perturbed $\mathscr{H}$. Hence, the UPC algorithm is a stochastic generalization of the PHC algorithm, and allows us to obtain the PHC model estimates directly from covariance data instead of impulse-response measurements.

## 4.8 Simulation Example

In this section, we present a simulation example to demonstrate the approximation performance of the UPC method. In this example, we consider the problem of estimating a second-order AR spectrum from a short record of the AR process in additive white noise. The stochastic

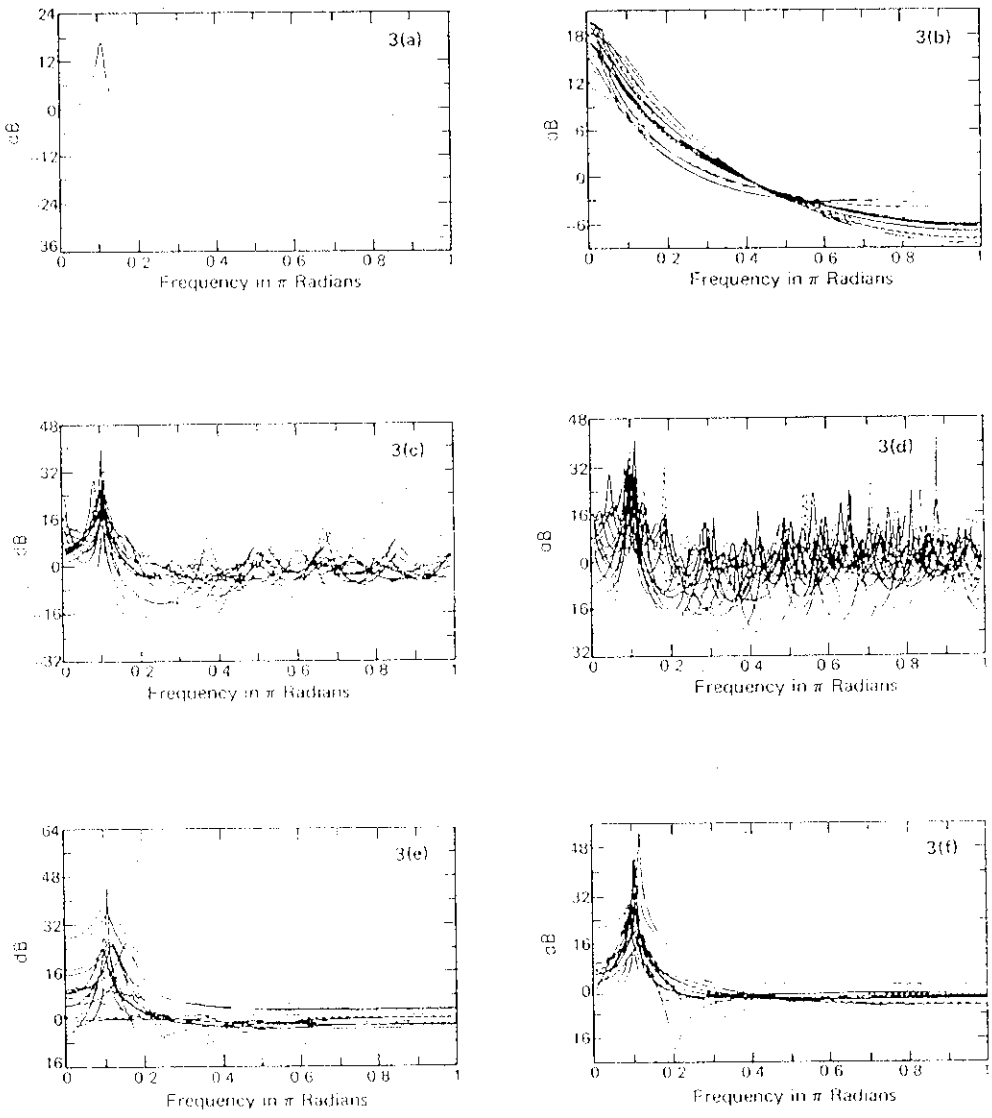


*Figure 3*

process was generated using the following model:

$$y(n) = 1.864\,y(n-1) - 0.96\,y(n-2) + v(n),$$

which has two symmetric poles approximately at $0.98\exp\{\pm j0.1\pi\}$, driven by a pseudorandom white-noise sequence of unit variance. Another statistically independent pseudorandom white sequence of variance 10.0 was added to the output. Twenty such statistically independent data records of the AR + noise time-series, each of length 64, were generated.

The first covariance lags $\{r(0), r(1), \ldots, r(25)\}$ were estimated using the unbiased estimator from each 64-point record seperately, and the model parameters were estimated from these covariance estimates using a variety of different methods. The maximum entropy method (MEM) [9] was first used to obtain AR models of orders 2, 12, and 25, that exactly match the first 2, 12, and 25 lags, respectively. The true power spectrum is plotted in Figure 3(a), and the 20 MEM spectral estimates are plotted one over the other in Figure 3(b–d).

MEM does not perform very well, because an AR(2) + white-noise sequence needs an ARMA(2) model or a high-order AR model. Hence, the AR(2) model that exactly matches the first two lags, fails to resolve the two peaks (that are $0.2\pi$ radians apart) in every trial. Instead, it puts both the poles on the real line, and detects only one peak at zero. On the other hand, high-order AR fits that use more lags, give rise to spurious peaks, though the true peaks at $\pm 0.1\pi$ are resolved well. Sometimes, the strength of the spurious peak can be larger than the true peak, and the spectral shape is not reproduced with any degree of fidelity.

The next two plots in Figure 3 are the spectral estimates obtained by the algorithm of Desai and Pal [11] for the c.c. approach, and by the UPC method. Here, $13 \times 13$ matrices were employed that used all the 25 covariance lags. The model order $p$ was assumed to be predetermined to be 2, and a rank-2 approximation was used in all trials.

The spectral shape is well reproduced by both methods. But, the c.c. approach fails to resolve the two peaks in two of the 20 trials. This loss in resolution capability can be attributed to the numerical sensitivity problems inherent in the c.c. approach (c.f. [27]) when the poles are close together.

## 5. SINUSOIDAL APPROXIMATION: THE HARMONIC RETRIEVAL PROBLEM

The problem of retrieving sinusoids (with frequencies close to each other) from perturbed time-series or covariance information is of special interest in a vast range of signal-processing applications. Very often the

covariance sequence may have to be estimated from time-series data, as in Doppler processing in radar. It is not uncommon, however, to encounter applications in which the (time-series) data are not measurable while the covariance information is directly available. Such situations arise in astronomical star bearing estimation, interferometry, and passive sonar applications.

We first provide the background in exact sinusoidal modeling, that is, the determination of the sinusoid frequencies and amplitudes from exact data, both time-series and covariance lags. We then use the principal components approach developed in the preceding sections, to derive two sinusoidal approximation methods. The first called Toeplitz approximation method (TAM) is a special case of the stochastic system approximation (UPC) method and performs harmonic retrieval from noisy covariances.

The second called direct data approximation (DDA) is an application of the deterministic system approximation PHC method and performs harmonic retrieval directly from time-series data.

## 5.1 Background in Exact Sinusoidal Modeling

A signal composed of $p/2$ sinusoids admits a special ARMA representation

$$y(t) = \sum_{i=1}^{p} a_i y(t-k), \tag{21}$$

where the roots of the $A(z)$ polynomial are on the unit circle in the $z$ plane, at $\exp(j\omega_1)$, $\exp(j\omega_2)$, ... $\exp(j\omega_p)$, where $\omega_i$s are the desired sinusoid frequencies. It can be shown that the covariances also satisfy a similar recurrence relation

$$r(m) = \sum_{i=1}^{p} a_i r(m-i) \tag{22}$$

Hence both the data and the covariances are *exactly* (linearly) predictable from $p$ past values, and the $1,p$ parameters can be easily obtained from exact time-series or covariance data [46].

When the signal is composed of $p/2$ sinusoids corrupted by additive white noise, and if *exact* covariances of the noise-corrupted signal are available, Pisarenko's method is applicable. The observed Toeplitz covariance matrix $\mathbf{R}$ is the sum of the signal-covariance matrix $\mathbf{R}_s$ and the noise-covariance matrix $\mathbf{R}_n$. The former has rank $p$ (if its row and column dimensions are at least $p+1$), however, because of the covariance-recurrence relation (22); while the latter matrix is $\sigma^2$ times the

identity matrix (where $\sigma^2$ is the variance of the additive white noise). Hence, we can make the following observations [45]:

1. The Toeplitz matrix $\mathbf{R}$ constructed from the *exact* covariance lags, can be written as
$$\mathbf{R} = \mathbf{R}_s + \sigma^2\mathbf{I},$$
where $\mathbf{R}_s$ has rank $p$, and is semi-positive-definite. Hence, if the row and column dimension of $\mathbf{R}$ is $N > p$, then $\sigma^2$ is the minimum eigenvalue of $\mathbf{R}$, and it has multiplicity $(N - p)$
2. Moreover, if $N = p + 1$, then
$$[1, \quad -a_1, \quad -a_2, \ldots, -a_p]\mathbf{R}_s = 0$$
$$\Leftrightarrow \mathbf{a}\mathbf{R} = \mathbf{a}(\mathbf{R}_s + \sigma^2\mathbf{I}) = \sigma^2\mathbf{a}.$$
Therefore, the l.p. parameter vector $\mathbf{a}$ is the eigenvector of $\mathbf{R}$, associated with the minimum eigenvalue $\sigma^2$, when the size of $\mathbf{R}$ is $N = p + 1$.

All these results combined lead to Pisarenko's spectral estimation method, summarized below:

1. Diagonalize a sufficiently large matrix $\mathbf{R}$, and determine the number of sinusoids by examining the multiplicity of the minimum eigenvalue.
2. Compute the eigenvector $\mathbf{a}$ of the $(p + 1) \times (p + 1)$ leading principal minor of $\mathbf{R}$ corresponding to the minimum eigenvalue.
3. Solve for the roots of $A(z) = 0$, to obtain the location of the spectral lines.
4. The amplitude (power) corresponding to each of the sinusoidal components can be derived by solving a linear system of equations [45]

## 5.2 The Notion of State for the Sinusoidal Model

The state-space representation of the special model (21) for sinusoidal signals is a special case of (4) (with $\Gamma = 0$ and no input):

$$\begin{aligned} \mathbf{x}(k + 1) &= \mathbf{F}\mathbf{x}(k) \\ y(k) &= \mathbf{h}\mathbf{x}(k), \end{aligned} \tag{17}$$

where the eigenvalues of $\mathbf{F}$ are of unit magnitude and equal the roots of $A(z)$. The sinusoidal signal $y(t)$ is the model's zero-input response to some nonzero initial condition $\mathbf{x}(0)$. In fact, we have

$$y(t) = \mathbf{h}\mathbf{F}^t\mathbf{x}(0), \qquad t \geq 0$$

and

$$r(m) = \mathbf{h}\mathbf{P}\mathbf{F}'^{m}\mathbf{h}' = \mathbf{h}\mathbf{F}^{m}\mathbf{P}\mathbf{h}', \qquad m \geq 0, \tag{23}$$

where the state-variance satisfies $\mathbf{P} = \mathbf{F}\mathbf{P}\mathbf{F}'$. Comparing these equations with Eqs. (5) and (6) for the general ARMA model (4), we can conclude that the state of the sinusoidal model contains *all* the information required to generate the future output, and that now, the future output $\mathbf{Y}^+$ is *exactly* predictable from the state. In fact, we have

$$\mathbf{Y}^+ = \begin{pmatrix} y(t) \\ y(t+1) \\ y(t+2) \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{h} \\ \mathbf{h}\mathbf{F} \\ \mathbf{h}\mathbf{F}^2 \\ \vdots \end{pmatrix} \cdot \mathbf{X} = \Theta\mathbf{x},$$

while for the general ARMA system we had Eq. (12):

$$\mathbf{Y}^+ = \Theta\mathbf{x} + \mathbf{L}\mathbf{V}^+$$

Hence, for the sinusoidal model, we have some special rank properties that are not enjoyed by the general ARMA model. For instance, the Toeplitz covariance matrix $\mathbf{R} = \mathbf{E}(\mathbf{Y}^+\mathbf{Y}^{+\prime})$ now equals $\Theta\mathbf{P}\Theta'$ and hence has rank equal to the state dimension (model order) $p$, as long as the row and column dimension of $\mathbf{R}$ are both at least $p$. This factorization of $\mathbf{R}$ is easily deducible from the Eqs. (23) for the covariance lags.

Similarly, the Hankel matrix built from the sinusoidal data themselves, is also factorizable, as can be easily verified from the equation

$$(\mathbf{Y}^+(t)|\mathbf{Y}^+(t+1)|\mathbf{Y}^+(t+2)|\cdots) = \Theta \cdot (\mathbf{x}(t)\,|\,\mathbf{x}(t+1)|\mathbf{x}(t+2)|\cdots)$$

This factorization of the data-Hankel is also deducible from Eqs. (23) since the sinusoidal data $y(t)$ behaves analogously to the impulse response $i(t)$ of the general ARMA model (4). In fact, it turns out that the double Hankel matrix $\mathbf{D}$ constructed from a data record of length $L$

$$\mathbf{D} = \begin{pmatrix} y(1) & y(2) & \cdots & y(L-N+1) & y(L) \\ y(2) & y(3) & \cdots & y(L-N+2) & y(L-1) \\ y(3) & y(4) & \cdots & y(L-N+3) & y(L-2) \\ \vdots & & & & \vdots \\ y(N) & y(N+1) & \cdots & y(L) & y(L-N+1) \end{pmatrix}$$

$$\begin{pmatrix} y(L-1) & \cdots & y(N) \\ y(L-2) & \cdots & y(N-1) \\ y(L-3) & & y(N-2) \\ \vdots & & \vdots \\ y(L-N) & & y(1) \end{pmatrix}$$

also has rank $p$ and factorizes into an observability matrix and a right factor $\mathscr{X}$. The proof uses the fact that the eigenvalues of $\mathbf{F}$ are on the unit circle.

## 5.3 Partial State Selection and Criterion

For the problem of robust harmonic retrieval from noise-corrupted information wherein the matrices $\mathbf{R}$ and $\mathbf{D}$ no longer have the low-rank structure, we have to resort to approximate modeling. It is then a question of selecting a partial state that best predicts the future output $\mathbf{Y}^+$. As before, the predictive-efficiency criterion takes two similar but different forms for the deterministic and stochastic cases. In the former case, the problem is of minimizing a sum-of-squares error

$$\sum_t \|\mathbf{Y}^+(t) - \Theta\mathbf{x}_{\text{partial}}(t)\|_2^2 = \|\mathbf{D} - \Theta\mathscr{X}\|_2$$

and in the stochastic case, the problem is of minimizing a sum of error variances

$$\mathbf{E}\|\mathbf{Y}^+ - \Theta\mathbf{x}_{\text{partial}}\|_2 = \|\mathbf{R}^+ - \Theta\mathbf{P}\Theta'\|_2$$

The inherent constraint in these minimization problems is that $\Theta$ have only $p$ columns, so that the partial state vector has only $p$ components. Just as in the system-approximation problem, *the solution is built from the principle components in the SVD of* $\mathbf{R}$ *or* $\mathbf{D}$. Accordingly we have two sinusoidal approximation methods, one called the Toeplitz approximation method (TAM) for robust harmonic retrieval from perturbed covariance, and the other called direct data approximation (DDA) for robust harmonic retrieval from noisy time-series data [31].

## 5.4 Toeplitz Approximation Method (TAM) [34]

Let the SVD of $\mathbf{R}$ be

$$\mathbf{R}^+ = U\Sigma^2 V = (\mathbf{U}_1 \quad \mathbf{U}_2)\begin{pmatrix}\Sigma_1^2 & 0\\ 0 & \Sigma_2^2\end{pmatrix}\begin{pmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{pmatrix}$$

where $\Sigma_1$ is $p \times p$ and $\Sigma_2$ is $(N-p)\times(N-p)$. The observability matrix is obtained from the principal singular vectors $\mathbf{U}_1$ and the principal singular values $\Sigma_1^2$. In the presence of white noise, though the singular vectors are unchanged, the singular values are effected. In fact, all the singular values are increased by the noise variance, and so the smallest singular value has to be subtracted to compensate for this effect, that is

$$\hat{\Sigma}_1^2 = \Sigma_1^2 - \sigma_N^2\mathbf{I}.$$

where $\Sigma_N^2$ is the smallest singular value of $R$. Hence, the observability-type matrix is given by

$$\Theta = \mathbf{U}_1 \cdot \hat{\Sigma}_1$$

while

$$\mathbf{P} = \mathbf{I}.$$

Then, the state-space parameters can be estimated as follows:

$$\mathbf{F} = \Theta_1^{\dagger} \cdot \Theta_2,$$

$$\mathbf{h} = \text{the first row of } \Theta.$$

The eigenvalues of $\mathbf{F}$ give the frequencies of the sinusoids.

## 5.5 Direct Data Approximation (DDA)

Here, SVD on $\mathbf{D}$ is performed, instead of on $R$, resulting in

$$\mathbf{D} = \mathbf{U}\Sigma\mathbf{V}'.$$

The effect of noise is suppressed by subtracting the smallest singular value squared:

$$\hat{\Sigma}^2 = \Sigma^2 - \sigma_N^2\mathbf{I},$$

where $\sigma_N$ is the smallest singular value. Then the observability matrix $\Theta$ is formed from the principal singular vectors and singular values:

$$\Theta = \mathbf{U}_1\hat{\Sigma}_1^{1/2}$$

and

$$\Theta = \hat{\Sigma}_1^{1/2}\mathbf{V}_1'$$

The state-space parameters are computed from $\Theta$ and $?$ as

$$\mathbf{F} = \Theta_1^{\dagger} \cdot \Theta_2,$$

$$\mathbf{x}(1) = \text{first column of } ?$$

$$\mathbf{h} = \text{first row of } \Theta$$

From the state space parameters, one can retrieve the sinusoidal information (frequencies and amplitudes) after diagonalizing $F$, just as in TAM. In addition, we can also obtain the phase information from the transformed coordinates.

Recall that the left singular vectors of $\mathbf{D}$ are the same as the eigenvectors of $\mathbf{DD}'$. Hence, the approximation of $\mathbf{D}$ is theoretically equivalent to the Toeplitz approximation of $\mathbf{DD}'$. Working on $\mathbf{D}$, however, avoids the numerical problems associated with the increased condition number of $\mathbf{DD}'$. Therefore, DDA is preferable for numerical reasons.

*Table 1*

| *Zero-Phase Sinusoid of Frequency →* | *0.005* | *0.001* | *0.0005* | *0.0001* |
|---|---|---|---|---|
| Direct Data Tufts and Kumaresan method | R | R | × | × |
| DDA | R | R | R | R |

## 5.6 Simulation Example

This section provides a simulation example that clearly indicates the better numerical behaviour of the PC approach. The simulations were performed using double-precision arithmetic on a 36-bit PDP-10 computer. The IMSL routine ZRPOLY was used for polynomial rooting, and the EISPACK routine RG was used for eigendecomposition of **F**. The problem considered here is the estimation of the frequency of a single sinusoid (effectively, the problem of resolving two closely spaced complex exponentials) from eight *uncorrupted* data samples. Assuming the sampling frequency is 2 Hz, if the frequency of the sinusoid is chosen to be $f$ Hz, then the problem is of resolving complex exponentials $2f$ Hz apart. In theory, all methods should be able to resolve the exponentials independent of how small $f$ is. In practice, however, the finite word length of the processor will limit the resolution capability. The results in terms of success (R) and failure[9] (×) to resolve the spectral lines (for different frequency separation) are tabulated in Table 1. The direct-data version of the Tufts and Kumaresan method is used with the predictor polynomial of size 4 and for DDA, the matrix **D** is of dimension $4 \times 10$. In this example, the resolution of both methods is limited only by finite precision. The results indicate that DDA is less sensitive to roundoff errors. The balanced realization is a numerically sound choice for identification and spectral estimation, compared to the canonical realization (difference-equation representation), where it is known that the poles are highly sensitive to parameter perturbations, and that polynomial rooting is numerically ill-conditioned [54]. A detailed sensitivity analysis is available in [27].

# 6. CONCLUSIONS

This paper addressed three types of system-approximation problems: **deterministic, stochastic,** and **sinusoidal** model approximations. We

formulated each of the three approximation problems as a partial-state selection problem and derived a **principal components approach** applicable to all of them. The crux of the three approximation problems addressed here, is the selection of a **partial state** that contains the most information regarding the future output. The amount of information was quantified by a **predictive efficiency measure** that took the form of a least-squares prediction error in the deterministic case, and that of total prediction error variance in the stochastic case.

By maximizing the predictive efficiency of the partial state, we derived the principal-components approach, and more specifically, the principal Hankel components (PHC) algorithm for deterministic system approximation, and the unweighted principal components (UPC) algorithm for stochastic system approximation. For the related problem of harmonic retrieval or sinusoidal approximation, the UPC algorithm specializes to the Toeplitz approximation method (TAM), and the PHC algorithm specializes to the direct data approximation (DDA) method.

Since the criterion used by all these methods is the same, it is not surprising that the four methods are closely related. For instance, the target matrices used for SVD-based approximation by PHC, UPC, TAM, and DDA are $\mathcal{H}$, $\mathbf{H}\mathbf{R}^{-1}\mathbf{H}'$, $\mathbf{R}$, and $\mathbf{D}$, respectively. We have shown, however, that

$$\mathbf{H}\mathbf{R}^{-1}\mathbf{H}' = \rho\mathcal{H}\mathcal{H}',$$

and that $\mathbf{D}\mathbf{D}'$ is an estimate of $\mathbf{R}$ in the sinusoidal problem.

The underlying state-space representation used in the derivation of these methods lends itself well to parameter estimation from the SVD factorization. Moreover, SVD offers good numerical stability. Therefore the algebraic and numerical significance of using SVD and the state-space approach for these approximation problems appears to be promising. From both the research and application perspectives, it promises to become an important area for future in-depth exploration.

## NOTES

1. This research was supported in part by the National Science Foundation under Grant ECS-85-12479 and by the Office of Naval Research under Grant N00014-81-K-0191.

2. The output covariance of a linear rational model satisfies certain recurrence relations (called higher order Yule-Walker equations), so that only $2p$ covariances specify the infinite covariance extension.

3. The spectral norm of a matrix is defined as

$$\|A\|_2 = \sup_{\|x\|_2 = 1} \|Ax\|_2$$

where $\|x\|_2$ denotes the Euclidean norm of a vector $x$ that is, $\|x\|_2^2 = x'x$.

4 For instance, in seismic signal processing, it is desirable to model the source wavelet by as small a model order as possible (as long as the modeling error is small) to reduce the computational burden in the subsequent deconvolution.

5 A symmetric matrix $A$ is said to be bigger than symmetric matrix $B$ if $A - B$ is semipositive definite.

6 An intuitive proof of this statement is available in Section 4.2.1.

7 Note that the covariance matrix $\mathbf{R}$ is not expected to have rank equal to the model order, even when the lags are exact. Hence, in the perturbed situation, a principal-components approximation of $\mathbf{R}$ is not justified.

8 Without such a constraint, no size compression is required, and the entire past $Y$ can be used as the state.

9 A trial is considered a failure if the frequencies of the exponentials are both identified to be 0 Hz, that is, both the $z$-plane roots are on the positive real axis.

# REFERENCES


[1] Adamjan, V. M., Arov, D. Z., and Krein, M. G. "Analytic Properties of Schmidt Pairs for a Hankel Operator and the Generalized Schur-Takagi Problem," *Math USSR Sbornik*, **15**, 1 (1971), pp. 31-73.

[2a] Akaike, H. "A New Look at Statistical Model Identification," *IEEE Trans. Automatic Control*, **AC-19**, 6 (Dec. 1974), pp. 716-723.

[2b] Akaike, H. "Markovian Representation of Stochastic Processes and Its Application to the Analysis of Autoregressive Moving Average Processes," *Annals of the Institute of Statistical Mathematics*, **26** (1974), pp. 363-387.

[3] Akaike, H. "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion," in *System Identification: Advances and Case Studies*, Chap. 2, R. K. Mehra and D. G. Lainiotis (Eds.), Academic, New York, 1976.

[4] Anderson, B. D. O. "Algebraic Properties of Minimum Degree Spectral Factors," *Automatica*, **9**, (4 July 1973), pp. 491-500.

[5] Anderson, B. D. O., and Kailath, T. "The Choice of Signal Process Models," *J. Math. Analysis and Applications*, **35**, 3 (September 1971), pp. 659-668.

[6] Anderson, B. D. O., and Moore, J. B. *Optimal Filtering*, Information and Systems Sciences Series, Prentice-Hall, Englewood Cliffs, N.J., 1979.

[7] Baram, Y. "Realization and Reduction of Markovian Models from Nonstationary Data," *IEEE Trans. Automatic Control*, **AC-26**, 6 (December 1981), pp. 1225-1231.

[8] Beex, A. A., and Scharf, L. L. "Covariance Sequence Approximation for Parametric Spectrum Modeling," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **ASSP-29**, 5 (October 1981), pp. 1042-1051.

[9] Burg, J. P. "*Maximum Entropy Spectral Analysis*," Ph.D. dissertation, Stanford Univ., Stanford, Calif., 1975.

[10] Cohen, M. H., Jauncey, D. L., Kellermann, K. I., and Clark, B. G. "Radio Interferometry at One-Thousandth Second of Arc," *Science*, **162** (Oct. 4, 1968), pp. 88-94.

[11] Desai, U. B., and Pal, D. "A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realization," *Proc. 16th Annual Conf. Information Sciences and Systems*, Princeton University, N.J., March 1982, pp. 613-620.

[12] Faurre, P. *Symposium on Optimization, Nice*, Lecture Notes in Mathematics, Vol. 132, Springer, New York, 1969.

[13] Faurre P. "Stochastic Realization Algorithms," in *System Identification: Advances and Case Studies* Chap. 1, R. K. Mehra, and D. G. Lainiotis (Eds). Academic, New York, 1976.

[14] Fomalont, E. B. "Fundamentals and Deficiencies of Aperture Synthesis," in *Image Formation from Coherence Functions in Astronomy*, D. Reidel (Ed.). Schooneveld C. V., 1979.

[15] GalEzer, R. J. *Wavefront Array Processor and its Applications*, Ph.D. dissertation, Univ. of Southern California, Los Angeles, Dec. 1982.

[16] Gelfand, I. M., and Yaglom, A. M. "Calculation of the Amount of Information about a Random Function Contained in Other Such Function," *American Mathematical Society Translations, Series 2*, **12** (1959), pp. 199–246.

[17] Genin, Y., and Kung, S. Y. "A Two-Variable Approach to the Model Reduction Problem with Hankel Norm Criterion," *IEEE Trans. Circuits Systems*, **CAS-28**, 9 (September 1981), pp. 912–924.

[18] Gevers, M., and Kailath, T. "An Innovations Approach to Least Squares Estimation Part VI: Discrete-time Innovations-Representations and Recursive Estimation," *IEEE Trans. Automatic Control*, **AC-18** (December 1973), pp. 588–600.

[19] Glover, K. "All Optimal Hankel-Norm Approximation of Linear Multivariable Systems and their $L^\infty$-error Bounds," *Int. J. Control*, **39**, 6 (1984), pp. 1115–1193.

[20] Halpeny, O. S., and Childers, D. G. "Composite Wavefront Decomposition via Multidimensional Digital Filtering of Array Data," *IEEE Trans. Circuits Systems*, **CAS-22**, 6 (June 1975), pp. 552–562.

[21] Hotelling, H. "Relations Between Two Sets of Variates," *Biometrika*, **28**, (1936), pp. 321–372.

[22] Ho, B. L., and Kalman, R. E. "Effective Construction of Linear, State-Variable Models from Input/Output Functions," *Regelungstechnik*, **14** (1966), pp. 545–548.

[23] Kailath, T. "The Innovations Approach to Detection and Estimation Theory," *Proc. IEEE*, **58** (May 1970), pp. 680–695.

[24] Kailath, T. "A View of Three Decades of Linear Filtering Theory," *IEEE Trans. Inform. Theory*, **IT-20**, 2 (Mar. 1974), pp. 145–181.

[24] Kailath, T. *Linear Systems*, Prentice-Hall, Englewood Cliffs, N.J., 1980.

[26] Kaiser, J. F. "Some Practical Considerations in the Realization of Linear Digital Filters," *Proc. 3rd Allerton Conf. Circuits and Systems Theory*, (Univ. of Illinois, Oct. 20–22, 1965), pp. 621–633.

[27] Karalamangala, A. S. "A Principal Components Approach to Approximate Modeling and ARMA Spectral Estimation," Ph.D. dissertation, Univ. Southern California, Dec. 1984.

[28] Klemma, V. C., and Laub, A. J. "The Singular Value Decomposition: Its Computation and Some Applications," *IEEE Trans. Automatic Control*, **AC-25** (1980), pp. 164–176.

[29] Kronecker, L. "Zur theorie der elimination einer variabeln aus zwei algebraischen gleichungen," in *Trans. Royal Prussian Academy of Sciences, collected works*, 1881.

[30] Kullback, S., and Leibler, R. A. "On information and Sufficiency," *Annals of Mathematical Statistics*, **22** (1951), pp. 79–86.

[31] Kung, S. Y., Arun, K. S., and BhaskarRao, D. V. "State Space and Singular Value Decomposition Based Approximation Methods for the Harmonic Retrieval Problem," *J. Optical Society of America*, Dec. 1983.

[32] Kung, S. Y., BhaskarRao, D. V., and Arun, K. S. "New State Space and Singular Value Decomposition Based Approximate Modeling Methods for Harmonic Retrieval," *ASSP Spectral Estimation Workshop II* (IEEE, Tampa, FL, Nov. 1983), pp. 266–271.

[33] Kung, S. Y. "A New Identification and Model Reduction Algorithm via Singular Value Decomposition." *Proc. 12th Asilomar Conf. Circuits, Systems and Computers* (IEEE, Pacific Grove, Calif., Nov. 1978), pp. 705–714.

[34] Kung, S. Y. "A Toeplitz Approximation Method and some Applications." *Proc. Int. Symp. the Mathematical Theory of Networks and Systems* (Santa Monica, Calif., August 5–7, 1981), pp. 262–266.

[35] Kung, S. Y., and Lin, D. "Optimal Hankel-Norm Model Reduction: Multivariable Systems." *IEEE Trans. Automatic Control*, **AC-26** (August 1981).

[36] Maciejowski, J. M., and Vines, D. A. "The Use of Multivariable Frequency-Domain Techniques to Investigate Macroeconomic Policy Options." *Proc.*, JACC, San Francisco, Calif., Aug. 1980.

[37] Markel, J. D., and Gray, A. H., Jr. *Linear Prediction of Speech*. Springer, New York, 1976.

[38] Mendel, J. M., and Kung, S. Y. "Computer Programs for Wavelet Modeling." USC Geo-Signal Processing Program Report, Univ. of Southern California, May 1981.

[39] Moore, B. C. "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction." *IEEE Trans. Automatic Control*, **AC-26**, 1 (February 1981), pp. 17–31.

[40] Mullis, C. T., and Roberts, R. A. "The Use of Second-Order Information in the Approximation of Discrete-time Linear Systems." *IEEE Trans. Acoustics, Speech, Signal Processing*, **ASSP-24**, 3 (June 1976), pp. 226–238.

[41] Mullis, C. T., and Roberts, R. A. "Synthesis of Minimum Round-off Noise Fixed Point Digital Filters." *IEEE Trans. Circuits Systems*, **CAS-23** (1976), pp. 551–562.

[42] Oppenheim, A. V., and Schafer, R. W. *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, N.J., 1975.

[43] Osborne, E. E. "On Pre-conditioning of Matrices." *J. Association of Computing Mathematics*, **7** (1960), pp. 338–345.

[44] Pernebo, L., and Silverman, L. M. "Model Reduction via Balanced State Space Representations." *IEEE Trans. Automatic Control*, **AC-27**, 2 (April 1982), pp. 382–387.

[45] Pisarenko, V. F. "The Retrieval of Harmonics from a Covariance Function." *Geophys. J. R. Astron. Soc. Can.*, Vol. **33** (1973), pp. 347–366.

[46] Prony, G. R. B. "Essai Experimentale et Analytique ..." *Paris Journal de L'Ecole Polytechnique*, **1**, 2 (1795), pp. 24–76.

[47] Radoski, H. R., Zawalick, E. J., and Fougere, P. F. "The Superiority of Maximum Entropy Power Spectrum Techniques Applied to Geomagnetic Micropulsations." *Physics of the Earth and Planetary Interiors*, **12** (August 1976), pp. 208–216.

[48] Rao, C. R. "The Use and Interpretation of Principal Component Analysis in Applied Research." *Sankhya Series A*, **26** (1964), pp. 329–358.

[49] Stewart, G. W. *Introduction to Matrix Computations*. Academic, New York, 1973.

[50] Tufts, D. W., and Kumaresan, R. "Estimation of Frequencies of Multiple Sinusoids: Making Linear Prediction Perform Like Max Likelihood." *Proc. IEEE*, **70** (1982), pp. 975–989.

[51] Ulrych, T. J. "Maximum Entropy Power Spectrum of Long Period Geomagnetic Reversals." *Nature*, **235** (1972), pp. 218–219.

[52] Walker, G. "On Periodicity in Series of Related Terms." *Proc. Royal Society of London, Series A*, **131A** (1931), 518.

[53] White, J. "Stochastic State Space Models from Empirical Data." *Int. Conf. Acoustics, Speech, and Signal Processing*, IEEE, Boston, Mass., April 1983, pp. 243–246.

[54] Wilkinson, J. H. *The Algebraic Eigenvalue Problem*. Monographs on Numerical Analysis, Oxford University Press, New York, 1965.

[55] Willems, J. C., "Least Squares Stationary Optimal Control and the Algebraic Riccati Equation," *IEEE Trans. Automatic Control*, **AC-16**, (Dec. 1971), pp. 621–634.

[56] Yaglom, A. M., "Outline of Some Topics in Linear Extrapolation of Stationary Random Processes," ***Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability***, University of California Press, Berkeley, 1965, pp. 259–278.

[57] Yule, G. U., "On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers," ***Philosoph. Trans. of the Royal Society of London, Series A*, 226A** (1927), pp. 267–298.

[58] Zeiger, H. P., and McEwen, A. J., "Approximate Linear Realizations of Given Dimension Via Ho's Algorithm," *IEEE Trans. Automatic Contr.* **AC-19** (April 1974), p. 153.