March 1988 LIDS-P-1761

# Improved Frequency Estimation by Identification of Signal and Noise Subspaces

Toshio M. Chin

February, 1988

**Abstract:**
Three sinusoidal decomposition methods are described. They are the total least squares principal eigenvector method, the Pisarenko harmonic decomposition method, and the MUSIC method. Efforts are made to identify the common features among the three.

---

This report was prepared for the author's Area Examination. Work supported by the National Science Foundation under Grant ECS-8700903.

# Introduction

We consider the problem of identifying the sinusoidal components in a short time series. We assume that the samples are equally spaced in time. The necessity to deal with a short sequence of samples arises in applications such that the process generating the signal is short-lived (e.g., NMR spectroscopy) or that the process is time-varying (e.g., sonar tracking of a moving object). In this report we compare three spectral estimation methods that are designed to work well with short data sequences.

## Traditional methods and their short comings

Common spectral estimation methods are based on the discrete Fourier transform (DFT) and are computationally efficient because of the availability of the fast Fourier transform (FFT) algorithm. Let $\{x_n\}$ be the sequence of signal values sampled at the interval $\tau$, and $N$ samples are observed. Then the "Periodogram" provides us with an estimate of the power spectral density (PSD) function for $-\pi/\tau \leq \omega \leq \pi/\tau$:

$$P_P(\omega) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x_n e^{-j\omega n}\right|^2$$

We can also obtain PSD from the estimate of the correlation process, $r_k \equiv E(x_{n+k}x_n^*)$:

$$P_{BT}(\omega) = \sum_{k=-K}^{K} r_k e^{-j\omega k}$$

which is often called the Blackman-Tukey estimator. The sequence $\{r_k\}$ is usually estimated by some weighted averages of the correlations among the data $\{x_n\}_{n=0}^{N-1}$. Typically, these spectral estimators are computed digitally, and samples of $P_P(\omega)$ and $P_{BT}(\omega)$ are computed by DFT taking the places of the Fourier transform.

Methods utilizing DFT have great disadvantages when dealing with a short data sequence [6]. First, they have a limited frequency resolution. Second, they produce a distorted spectrum (the plot of the magnitude of PSD against frequency) such that "sidelobes" of larger peaks tend to mask smaller peaks nearby. The main sources for these problems are:

- DFT models the signals as a linear combination of sinusoids with *harmonically related* (equally spaced) frequencies.
- DFT implicitly models the entire sample sequence $\{x_n\}_{n=-\infty}^{\infty}$ as a periodic extension of the observed sequence $\{x_n\}_{n=0}^{N-1}$. This is not a realistic model of the signal in general.

To see these problems in more detail, consider the DFT pair $\{y_n\}$ and $\{Y_m\}$:

$$Y_m = \sum_{n=0}^{N-1} y_n e^{-j2\pi mn/N} \tag{1}$$

$$y_n = \frac{1}{N} \sum_{m=0}^{N-1} Y_m e^{j2\pi mn/N} \tag{2}$$

Note in (2) that $\{y_n\}$ is decomposed as a linear combination of harmonically related sinusoids. Also, the implicit periodic extension, $y_{n+N} = y_n$, can be observed. The spectral resolution is given by the interval in the frequency domain:

$$\triangle\omega = \frac{2\pi}{N\tau}.$$

Thus, for small $N$ the resolution is not good.

Zero-padding the data, thereby increasing the appearent $N$, before the application of DFT does not improve the real resolution. To see this, note that $\{Y_m\}$ in (1) are not the samples of the Fourier transform of the true signal sequence, unless $y_n$'s are zero outside of the observed time frame $n = 0, \ldots, N-1$, i.e.,

$$\begin{aligned} Y(\omega) &= \sum_{n=-\infty}^{\infty} y_n e^{-j\omega n} \\ Y_m &\neq Y(m\triangle\omega). \end{aligned}$$

Rather, $\{Y_m\}$ are the samples of the Fourier transform of the truncated sequence:

$$\begin{aligned} Y'(\omega) &= \sum_{n=0}^{N-1} y_n e^{-j\omega n} \\ Y_m &= Y'(m\triangle\omega) \end{aligned}$$

By zero-padding the data sequence, one can achieve only a better approximation of $Y'(\omega)$ but not of $Y(\omega)$. It can be shown that convolving $Y(\omega)$ with a "*sinc*"-like,

periodic function yields $Y'(\omega)$ [8]. Such a convolution transforms a peak in $Y(\omega)$ into several lobes in $Y'(\omega)$. The heights and locations of the peaks in the true spectrum are obscured by overlapping of these lobes.

## High resolution methods

The so-called "high resolution" spectral estimation methods are developed mainly for the purpose of distinguishing closely-placed frequencies in the spectrum. A variety of such methods exist, but this report focuses on three methods that employ similar mathematical technique to deal with the noise. These are: the Total Least Squares—Principal Eigenvector (TLS) method, Pisarenko Harmonic Decomposition (PHD), and MUSIC method. In all three methods the *ideal* signal is modeled as a linear combination of exponential sinusoids:

$$x_n = \sum_{i=1}^{L} p_i e^{j\omega_i n} \tag{3}$$

We are interested in estimating the number of the sinusoidal components ($L$), their frequencies ($\omega_i$), and their complex amplitudes (magnitudes and phases, $p_i$). Unlike the situation in the DFT-based methods,

- $\omega_i$'s are not harmonically related.
- The signal model is valid for all $n$, and it does not introduce any artifical periodicity, i.e., there is no periodic extension.

The three methods differ in the manners they incorporate the noise into the signal model:

TLS

$$x_n \approx \sum_{i=1}^{L} p_i e^{j\omega_i n} \tag{4}$$

PHD

$$r_k \approx \sum_{i=1}^{L} \overline{p_i}^2 e^{j\omega_i k} \tag{5}$$

MUSIC

$$x_n = \sum_{i=1}^{L} p_i e^{j\omega_i n} + n_n \tag{6}$$

While TLS and PHD treat the noise as some unspecific perturbations in the data samples, MUSIC requires an explicit noise model. $\{n_n\}$ is a zero-mean complex random process uncorrelated with $\{p_i\}$. The complex amplitudes $p_i$'s are considered to be a zero-mean complex random variables in PHD and MUSIC, but TLS assumes no prior statistical information on them.

PHD models the correlation sequence $\{r_k\}$ instead of the raw data sequence, implying that $\{x_n\}$ *must* be wide-sense stationary. This stationarity requirement restricts $\{p_i\}$ to be an uncorrelated sequence:

$$\begin{aligned} r_k &= E(x_{n+k}x_n^*) \\ &= \sum_{i=1}^{L}\sum_{m=1}^{L} E(p_i p_m^*) e^{j(\omega_i-\omega_m)n} e^{j\omega_i k} \end{aligned} \tag{7}$$

Since $r_k$ cannot be dependent on the value of $n$, we must have

$$E(p_i p_m^*) = 0$$

for $i \neq m$. Thus, (7) yields (5) where $\overline{p_i}^2 \equiv E(|p_i|^2)$. Since we have taken the norm of $p_i$, the phase information is lost in PHD.

TLS and PHD are based on a computational technique called Prony's method which implies that the behavior of the signal ($\{x_n\}$ or $\{r_k\}$) can be described by a linear predictive model. Interestingly, this hidden model specifies the noise process in PHD:

PHD (implied models):

$$\begin{aligned} r_k &= \sum_{i=1}^{L} \overline{p_i}^2 e^{j\omega_i k} + \eta\delta_k \\ x_n &= \sum_{i=1}^{L} a_i x_{n-i} + n_n \end{aligned}$$

where $\delta_k$ is the impluse function, $\{a_i\}$ are the linear predictive coefficients, and $n_n$ is a white noise with variance $\eta$. Prony's method and its implications are discussed

later in more details. Let us note that the MUSIC procedure is not involved with such an underlying model.

The goal of this report is to compare the three high resolution spectral estimators from a common perspective.

# An Unified Viewpoint

A common theme exists in the noise reduction techniques used by TLS, PHD, and MUSIC. This section previews the techniques and defines some terminologies in order to establish a basis for comparison of the three methods in later sections.

## The rank $L$ data matrices

Throughout this report, let us denote the number of the sinusoidal components in the signal and the number of samples in the time series as $L$ and $N$, respectively. We define the *data vector* $\mathbf{x}$ as a subsequence of $K$ data samples where $L < K \leq N$:

$$\mathbf{x}_m = [x_{m+1}, \ldots, x_{m+K}]^T, \qquad m = 0, \ldots, N-K.$$

The value of $K$ differs among the three methods. The *data matrix* $\mathbf{X}$ is a matrix derived from the data vectors. Let us describe $\mathbf{X}$ in each method.

1. TLS: $\mathbf{B}$ is a matrix where rows are $N-K+1$ distinct observation vectors (transposed).

$$\mathbf{X} = \mathbf{B} \equiv \begin{bmatrix} \mathbf{x}_0^T \\ \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_{N-K}^T \end{bmatrix} \tag{8}$$

The number of rows must be at least $L$, i.e., $K \leq N-L+1$. Inspecting the matrix carefully, we see that the columns are also data vectors (of dimension $N-K+1$).

2. PHD: $\mathbf{R}$ is the transpose of the correlation matrix. It is similar to $\mathbf{B}$ except that its elements are $\{r_k\}$ instead of $\{x_n\}$, i.e., its rows are subsequences of correlations.

$$\mathbf{X} = \mathbf{R} \equiv [E(\mathbf{x}\mathbf{x}^H)]^T = E(\mathbf{x}^*\mathbf{x}^T) \tag{9}$$

The rows of $\mathbf{R}$ are some linear combinations of transposed data vectors.

3. MUSIC: $\mathbf{Q}$ is the $K \times K$ correlation matrix.

$$\mathbf{X} = \mathbf{Q} \equiv E(\mathbf{x}\mathbf{x}^H) \tag{10}$$

Note $\mathbf{R} = \mathbf{Q}^T$. The columns of $\mathbf{Q}$ are linear combinations of data vectors.

In the ideal situation where there is no observation noise or modeling error, each data matrix described above has rank $L$. To see this, we first note that an ideal observation vector $\tilde{\mathbf{x}}$ is in a $L$ dimensional subspace. Let us define the *sinusoidal vector* as

$$\mathbf{s}_i = [e^{j\omega_i}, \ldots, e^{jK\omega_i}]^T, \qquad i = 1, \ldots, L,$$

which are linearly independent when $\{\omega_i\}$ are distinct. From (3), $\tilde{\mathbf{x}}$ is in the subspace, $S_s$, spanned by $\{\mathbf{s}_i\}_{i=1}^L$:

$$\tilde{\mathbf{x}} \in S_s \equiv span(\mathbf{s}_1, \ldots, \mathbf{s}_L) \tag{11}$$

We call this $L$ dimensional subspace the *signal subspace.*

For TLS, it is easy to see that $rank(\mathbf{B}) = L$ because each of its row and column spaces is spanned by the corresponding sinusoidal vectors. The same is true for PHD, i.e., $rank(\mathbf{R}) = L$ from (5).

For MUSIC, we must make some assumption about the second order statistics of $\{p_i\}$, the amplitudes of the sinusoids, to ensure that $rank(\mathbf{Q}) = L$. Let us define the *sinusoidal matrix* as

$$\mathbf{S} = [\mathbf{s}_1 | \cdots | \mathbf{s}_L]$$

Then,

$$\mathbf{x} = \mathbf{S}\mathbf{p}, \tag{12}$$

where $\mathbf{p} = [p_1, \ldots, p_L]^T$. Since $\mathbf{S}$ has rank $L$, the rank of $\mathbf{Q}$ depends on the state of the matrix $\mathbf{P} \equiv E(\mathbf{p}\mathbf{p}^H)$:

$$\mathbf{Q} = \mathbf{S}\, E(\mathbf{p}\mathbf{p}^H)\, \mathbf{S}^H = \mathbf{S}\mathbf{P}\mathbf{S}^H$$

A sufficient condition for $rank(\mathbf{Q}) = L$ is that $\mathbf{P}$ is invertible. We assume that this condition holds; it implies that $\{p_i\}$ are not strongly correlated from each other.

## The theme in noise reduction

A data matrix $\mathbf{X}$ is ideally of rank $L$, but because of the noise it tends to have a full rank. Recovering the true rank of $\mathbf{X}$ is important because it yields an estimate for $L$, the number of the sinusoids in the signal. Furthermore, by analyzing the dimensionality of the data matrix we can optimally estimate the $L$ dimensional signal subspace. Specifically, when the signal to noise ratio is reasonably high, the noise perturbs an ideal data vector $\tilde{\mathbf{x}}$ out of the signal subspace only by a relatively small amount, so that the perturbed vectors cluster around the signal subspace. Thus, $\mathbf{X}$ is almost rank $L$ and thus is rank deficient. The singular value decomposition is useful in the analysis of "near rank deficiency" of a matrix [1]. Let

$$\chi \equiv rank(\mathbf{X}) > L.$$

$\mathbf{X}$ can be decomposed as

$$\mathbf{X} = \sum_{i=1}^{\chi} \sigma_i \mathbf{u}_i \mathbf{v}_i^H, \tag{13}$$

where $\sigma_i, \mathbf{u}_i, \mathbf{v}_i$ are the singular values, left singular vectors, and right singular vectors, respectively. "$H$" denotes the conjugate transpose. An important observation here is that since $\mathbf{X}$ is almost $L$ dimensional,

$\sigma_{L+1}, \ldots, \sigma_\chi$ are much smaller than the other singular values.

Let us distinguish the two groups of smaller and larger singular values by calling them "*near-zero*" and "*principal*" singular values. The optimal rank $L$ approximation, $\mathbf{X}'$, of $\mathbf{X}$ is obtained by reconstructing a matrix from the $L$ "principal components" in $\mathbf{X}$:

$$\mathbf{X}' = \sum_{i=1}^{L} \sigma_i \mathbf{u}_i \mathbf{v}_i^H. \tag{14}$$

$\mathbf{X}'$ is optimal in the sense that it minimizes the Frobenius norm [1]:

$$\|\mathbf{X} - \mathbf{X}'\|_F \equiv \left[ \sum_i \sum_j (x_{ij} - x'_{ij})^2 \right]^{1/2}.$$

Thus, the singular vectors associated with the "principal" singular values can be used to estimate the signal subspace. For example, for $\mathbf{X} = \mathbf{Q}$ (MUSIC), the signal subspace can be estimated as

$$\mathcal{S}_s = span(\mathbf{u}_1, \ldots, \mathbf{u}_L)$$

The corresponding *noise subspace* is given by

$$\mathcal{N}_s = span(\mathbf{u}_{L+1}, \ldots, \mathbf{u}_\chi)$$

$$\mathcal{N}_s \perp \mathcal{S}_s$$

Note that $\mathcal{N}_s$ is not the subspace spanned by the noise process. A random noise process spans the entire space. Only the noise components in $\mathcal{N}_s$ are identified, and the components in $\mathcal{S}_s$ are treated as parts of the signal. Nevertheless, the noise perturbs the ideal data vectors out of the signal subspace in random directions so that the average effect over many data vectors should be small. We can therefore expect a fairy accurate estimate of the signal subspace.

Once the signal subspace is determined, the set of basis vectors $\{\mathrm{s}_i\}_{i=1}^{L}$ must be identified. The identification of the spectral frequencies $\{\omega_i\}_{i=1}^{L}$ directly follows this. The rest of this report attempts to contrast the ways TLS, PHD, and MUSIC identify those frequencies.

## A variation of the theme

TLS and PHD do not explicitly use the signal and noise subspaces as described above. To them, the null space $Null(\mathbf{X})$ is important. In particular, the identification of the vector $\mathbf{c}$ such that $\mathbf{Xc} = \mathbf{0}$ is a key step in Prony's method on which TLS and PHD are based. Because an over-ranked $\mathbf{X}$ tends to have no non-trivial null vector, we make the approximation that $\mathbf{X}'\mathbf{c} = 0$, which leads to

$$\mathbf{c} \in Null(\mathbf{X}') = span(\mathbf{v}_{L+1}, \ldots, \mathbf{v}_\chi) \tag{15}$$

where $\{\mathbf{v}_i\}$ are the right singular vectors of $\mathbf{X}$, (13). $\mathbf{c}$ is related to the noise subspace as

$$\mathbf{c}^* \in \mathcal{N}_s = span(\mathbf{v}_{L+1}^*, \ldots, \mathbf{v}_\chi^*)$$

This relation implies the connection between Prony's method and MUSIC which is constructed on the concept of the signal and noise subspaces. We discuss this connection further in the last section.

Another variation of the theme

In PHD and MUSIC, the rank $L$ approximation is not the one given in (14); it is

$$\mathbf{X}' = \mathbf{X} - \eta\mathbf{I} = \sum_{i=1}^{L}(\sigma_i - \eta)\mathbf{u}_i\mathbf{v}_i^H \tag{16}$$

where

$$\eta \equiv \sigma_{L+1} \approx \cdots \approx \sigma_\chi$$

(Note that $\mathbf{X}$ in PHD or MUSIC is Hermitian.) The assumption that the "near-zero" singular values are almost equal is acceptable for a noise uncorrelated with the signal. $\mathbf{X}'$ is not an optimal approximation in the sense of the Frobenius distance, but it yields a physically interpretable noise process as described in later sections.

# Background — Prony's Method

Prony's method is a computational technique for modeling equally spaced samples with a linear combination of complex exponentials. It is a center-piece in both TLS and PHD spectral estimators. Some specializations are required when applying Prony's methods to TLS and PHD because the samples are to be modeled with sinusoids (i.e., an additional constraint that all exponentials must have unit norm.). Special issues involving the application of Prony's method to spectral estimation are discussed later in this section. We first review the basics of Prony's method.

## Prony's Method [5]

Let us consider a sequence of data samples $\{y_n\}_{n=1}^N$, and suppose that we want to approximate the sequence as a linear combination of exponents of $L$ complex numbers $\{z_i\}_{i=1}^L$:

$$y_n \approx \sum_{i=1}^{L} a_i z_i^n, \qquad n = 1, \ldots, N, \tag{17}$$

where the coefficients $\{a_i\}$ are also complex numbers (the amplitudes and phases of the exponentials). Let us assume for the moment that $N = 2L$. Solving for the $2L$ unknowns, $z_i$'s and $p_i$'s, from the same number of non-linear equations is computationally unattractive. With Prony's method, however, all computational stages except one are typical matrix algebra; the only other stage is the identification of the zeros of a polynomial.

Let us define an $L^{th}$ order polynomial $\psi(z)$ (with coefficients $c_i$, $i = 1, \ldots, L$) whose roots are the unknown exponents:

$$\begin{aligned} \psi(z) &\equiv \prod_{i=1}^{L} (z - z_i) \\ &\equiv c_0 z^L + c_1 z^{L-1} + \cdots + c_L, \qquad c_0 = 1. \end{aligned} \tag{18}$$

Then,

$$\sum_{i=0}^{L} c_i y_{n-i} \approx \sum_{i=0}^{L} \sum_{k=1}^{L} c_i a_k z_k^{n-i}$$

$$
\begin{aligned}
&= \sum_{k=1}^{L} a_k z_k^{n-L} \sum_{i=0}^{L} c_i z_k^{L-i} \\
&= \sum_{k=1}^{L} a_k z_k^{n-L} \psi(z_k) \\
&= 0. \qquad (19)
\end{aligned}
$$

This shows that the coefficients of the polynomial $\psi(z)$ are the coefficients of a linear predictive (LP) model:

$$y_n \approx \sum_{i=1}^{L} (-c_i) y_{n-i}. \qquad (20)$$

Thus, the N data samples $\{y_n\}$ provide us with the following matrix equation:

$$\begin{bmatrix} y_1 & y_2 & \cdots & y_L \\ y_2 & y_3 & \cdots & y_{L+1} \\ \vdots & \vdots & & \vdots \\ y_{N-L} & y_{N-L+1} & \cdots & y_{N-1} \end{bmatrix} \begin{bmatrix} c_L \\ c_{L-1} \\ \vdots \\ c_1 \end{bmatrix} \approx \begin{bmatrix} -y_{L+1} \\ -y_{L+2} \\ \vdots \\ -y_N \end{bmatrix}. \qquad (21)$$

We write this equation as

$$\mathbf{A}\bar{\mathbf{c}} \approx \mathbf{b}. \qquad (22)$$

In principle, therefore, the data samples can be used to evaluate $\bar{\mathbf{c}}$ from which the exponents $\{z_i\}$ can be found. In particular, if we assume that $N = 2L$ (making $\mathbf{A}$ a square matrix) and that $\mathbf{A}$ is invertible (dependent on the particluar data sequence, of course, but practically acceptable), then the coefficients of $\psi(z)$ can be estimated as $\bar{\mathbf{c}} = \mathbf{A}^{-1}\mathbf{b}$, and $\{z_i\}_{i=1}^{L}$ can be obtained by finding the zeros of the polynomial. Once $z_i$'s are known, (17) provides us with $2L$ linear equations to solve for $a_i$'s. This overdetermined system of equations is usually solved by a least squares method. This completes the exponential fit procedure. Note again that the only nonlinear computation involved is the determination of the zeros of $\psi(z)$.

## Extended Prony's method

To solve for $\bar{\mathbf{c}}$ in (22) we have just assumed $N = 2L$. A more realistic situation is $N > 2L$, i.e., more data samples are used to "average out" the noise. This makes

(21) an overdetermined system of equations. In the so-called "extended Prony's method," the equations are solved by a least square fit. A key contribution of the TLS method discussed in the next section is that it introduces the *total least squares* criterion to solve for the overdetermined system of equations.

## Sinusoidal fit with Prony's method

In TLS and PHD, the data samples are modeled with sinusoids, i.e., $|z_i| = 1$, but Prony's method does not constrain the exponents to be of unit norm. The following examples illustrate the danger in applying Prony's method to sinusoidal fit carelessly:

example 1

Given $L = 1, y_1 = 1$, and $y_2 = 2$ $(N = 2L)$,

$$\psi(z) = z - 2 = 0$$
$$z_1 = 2$$
$$y_n = \tfrac{1}{2}2^n$$

Obviously, sinusoidal fit is not successful. If more data points are available, however, Prony's method works:

example 2

Given additional data samples $y_3 = 1$ and $y_4 = 2$ and $L = 2$,

$$\psi(z) = z^2 - 1 = 0$$
$$z_1 = 1 = e^{j0},\ z_2 = -1 = e^{j\pi}$$
$$y_n = \tfrac{3}{2} + \tfrac{1}{2}\cos \pi n$$

The two examples have the same ratio of the number of samples to the number of sinusoids. The difference between the two sample sequences in these examples seems best described qualitatively: the second sequence shows periodicity while the first one does not. Here is another example illustrating the lack of periodicity in the data sequence leads to a failure. The same number of samples are generated from the same waveform as in *example 2*, but this time the sampling rate is doubled:

example 3

Given $y_1 = 1.5$, $y_2 = 1$, $y_3 = 1.5$, $y_4 = 2$,

$$y_n = 0.606\,(1.348)^n - 4.617\,(-0.148)^n$$

These examples seem to indicate that the applicability of Prony's method to sinusoidal decomposition must be judged based on some qualitative observation on the periodicity in the data sequence.

In TLS sinusoids are fit to the raw data samples, (4), while in PHD they are fit to the estimates of the correlation sequence, (5). The later is theoretically safer than the former, because a sequence with the Hermitian property ($r_{-k} = r_k^*$) can be expressed as a linear combination of a small number of sinusoids. See the section on PHD for more details. The rest of the report assumes that the data set $\{x_n\}$ is determined "fittable" with sinusoids.

## Resolution and LP model order

We have shown that Prony's method implicitly assumes an $L^{th}$ order LP model of the samples, (20). This roughly means that the method performs a decomposition based on the behavior of the signal within the time frame of $L+1$ samples. Noting that sinusoids with similar frequencies must be observed for a relatively long time to distinguish themselves, we argue that the higher the model order the better the ability of the estimator to resolve the frequencies. The order of the LP model in (20) is hence raised to $M$ ($L \le M \le N - L$) to improve the frequency resolution. We now have a rearranged equation $\mathbf{A}\bar{\mathbf{c}} = \mathbf{b}$

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_M \\ x_2 & x_3 & \cdots & x_{M+1} \\ \vdots & \vdots & & \vdots \\ x_{N-M} & x_{N-M+1} & \cdots & x_{N-1} \end{bmatrix} \begin{bmatrix} c_M \\ c_{M-1} \\ \vdots \\ c_1 \end{bmatrix} \approx \begin{bmatrix} -x_{M+1} \\ -x_{M+2} \\ \vdots \\ -x_N \end{bmatrix} \tag{23}$$

and a $M^{th}$ order polynomial

$$\psi(z) \;=\; c_0 z^M + c_1 z^{M-1} + \cdots + c_M, \qquad c_0 = 1. \tag{24}$$

$L$ of the $M$ zeros of the polynomial is the exponents being sought; the rest of the zeros are not related with the signal and are called "extranous zeros." A popular

method for segregating the signal and extraneous zeros is described by Kumaresan [7]:

> Suppose $\{x_n\}$ are in fact samples of a linear combination of sinusoids (no noise). If the system of equations (22) is solved so that $\bar{\mathbf{c}}$ is of the *minimum norm* among the possible solutions, then
>
> - $|z_i| = 1$ for the signal zeros
> - $|z_i| < 1$ for the extraneous zeros.

Still, identification of the signal zeros from a noisy data set can be difficult especially when a wrong value of $M$ is chosen [4]. Kay and Marple [6] suggests to keep $M \leq N/2$ to avoid mistaking extraneous zeros as signal zeros. Contradictingly, the case where $M = N - L$ (note $M > L$), called the "Kumaresan-Prony" case, yields a good result and is computationally attractive [4][14]. Typically, during the theoretical development of a spectral estimator, the LP model order $M$ is kept as an implementational variable, so that the issue of choosing an optimal $M$ is left unanswered. That is the case for TLS.

A possible disadvantage of using a higher order LP model is that zeros must be found from a higher order $\psi(z)$, which may be computationally unattractive. PHD, older among the high resolution methods, does not utilize a higher order model, i.e., $M = L$.

# TLS: A Principal Eigenvector Method

The Principal Eigenvector spectral estimator is a direct application of the extended Prony's method. The essence of the method is its optimal solution to the system of equations (23). In what is commonly referred to as *the* Principal Eigenvector method, Tufts and Kumaresan [13] have imposed a least squares criterion upon (23) to solve for $\bar{\mathbf{c}}$. Later, Rahman and Yu [11] have rearranged (23) into a homogeneous set of equations so that a total least squares criterion can be applied. We refer to this as TLS. In this section we discuss both approaches and compare them.

## The least squares formulation [13]

If (23) is overdetermined, $\bar{\mathbf{c}}$ cannot be solved exactly in general. If, on the other hand, (23) is underdetermined, there are infinitely many solutions. Hence, the following least squares constraints are applied:

> Find $\hat{\bar{\mathbf{c}}}$ that minimizes $\|\mathbf{b} - \mathbf{A}\hat{\bar{\mathbf{c}}}\|$. If there are more than one such $\hat{\bar{\mathbf{c}}}$, choose the one that has the minimum norm.

We have discussed the advantage of the minimum norm solution in segregating the signal zeros of $\psi(z)$ from the extraneous zeros. The least squares constraints can be restated as follows:

> Find the matrix $\mathbf{A}'$ such that $\|\mathbf{A}' - \mathbf{A}\|_F$ is minimum and $\mathbf{b} \in range(\mathbf{A}')$.
> Find the minimum norm solution to $\mathbf{A}'\hat{\bar{\mathbf{c}}} = \mathbf{b}$.

This alternative statement indicates that the least squares criterion correct for the perturbations in $\mathbf{A}$ but not those in $\mathbf{b}$.

## Psudo-inverse and minimum norm solution

In the ideal case (no noise), the minimum norm solution is obtained using the "psudo-inverse" of the matrix $\mathbf{A}$ [4]:

$$\hat{\bar{\mathbf{c}}} = \mathbf{A}^{-P}\mathbf{b}. \tag{25}$$

In general the psudo-inverse $\mathbf{A}^{-P}$ is defined by

$$\mathbf{A}^{-P} = \sum_{i=1}^{rank(A)} \sigma_i^{-1} \mathbf{v}_i \mathbf{u}_i^H \tag{26}$$

where $\sigma, \mathbf{u}_i, \mathbf{v}_i$ are the singular values, left singular vectors, and right singular vectors of $\mathbf{A}$, respectively. When $\mathbf{A}$ has a full column rank, the expression in (26) can be shown to be equal to $(\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$, which is a common but not a general definition of the psudo-inverse.

## The principle behind Principal Eigenvector

A small perturbation in $\mathbf{A}$ tends to increase the rank of the matrix and create small singular values which should have been zero. This leads to a large perturbation in the psudo-inverse because the singular values are inverted in (26). One way to avoid this problem is to make a lower rank approximation of $\mathbf{A}$ before taking the psudo-inverse. The lower rank approximation $\mathbf{A}'$ of $\mathbf{A}$ is computed from the singular values and vectors of $\mathbf{A}$:

$$\mathbf{A}' = \sum_{i=1}^{L} \sigma_i \mathbf{u}_i \mathbf{v}_i^H \tag{27}$$

The value of $L$ is estimated by the number of "principal" singular values.

$\mathbf{A}'$ in (27) is the closest rank $L$ approximation to $\mathbf{A}$ in the sense of Frobenius norm. For any rank $L$ matrix $\mathbf{A}''$ (which has the same physical dimensions as the matrix $\mathbf{A}$), we have [1]

$$\|\mathbf{A}'' - \mathbf{A}\|_F \geq \sum_{i=L+1}^{rank(\mathbf{A})} \sigma_i^2.$$

$\mathbf{A}'' = \mathbf{A}'$ achieves the lower bound; hence, $\mathbf{A}'$ is the closest rank $L$ matrix to $\mathbf{A}$.

The principle underlying the lower rank approximation is that the perturbation in the data affect the singular values and vectors very little. Specifically, the singular values and vectors of the ideal data matrix $\tilde{\mathbf{A}}$ are assumed to be related to those of the perturbed matrix $\mathbf{A}$ by:

$$\begin{cases} \sigma_i \approx \tilde{\sigma}_i & i = 1, \ldots, L \\ \sigma_i \approx 0 & i = L+1, \ldots, rank(\mathbf{A}) \end{cases} \tag{28}$$

$$\left.\begin{array}{l} \mathbf{u}_i \approx \tilde{\mathbf{u}}_i \\ \mathbf{v}_i \approx \tilde{\mathbf{v}}_i \end{array}\right\} \qquad i = 1, \ldots, rank(\mathbf{A}) \tag{29}$$

where $\sigma_i, \mathbf{u}_i, \mathbf{v}_i$ are the singular values/vectors of $\mathbf{A}$ and $\tilde{\sigma}_i, \tilde{\mathbf{u}}_i, \tilde{\mathbf{v}}_i$ are those of $\tilde{\mathbf{A}}$. There are some experimental evidence supporting this assumption [14]. Also, the assumption can be considered to be an extension of the observation that the eigenvectors corresponding with large, well-separated eigenvalues are relatively insensitive to matrix element perturbations [15]. (Note that for a conjugate symmetric matrix, eigenvalue/vectors and identical to singular value/vectors.)

## Highlights of the least squares approach

- Estimate $L$ from the number of principal singular values of $\mathbf{A}$.
- Calculate $\mathbf{A}'^{-P}$, the psudo-inverse of the lower rank approximation of $\mathbf{A}$, by replacing "$rank(\mathbf{A})$" with "$L$" in (26).
- Find $\hat{\bar{\mathbf{c}}}$ using (25).

## TLS—total least square approach [11]

One major drawback of the least squares approach described above is that it only handles the perturbations in the matrix $\mathbf{A}$ but not in $\mathbf{b}$. As you can see from (23), both matrices are composed of the data samples, and correcting perturbations in just one of them is biased. The total least squares approach rearranges (23) into a homegeneous matrix equation:

$$[\mathbf{A} \; -\mathbf{b}] \begin{bmatrix} \bar{\mathbf{c}} \\ 1 \end{bmatrix} = \mathbf{0},$$

which we write as:

$$\mathbf{Bc} = \mathbf{0} \tag{30}$$

The dimension of $\mathbf{B}$ is $(N - M) \times (M + 1)$, and the last element of the $M + 1$ dimensional $\mathbf{c}$ must be 1. Let us consider the case in which the system of equations is overdetermined. With noisy data set, $\mathbf{B}$ tends to have a full column rank

($rank(\mathbf{B}) = M+1$). Then, no non-trivial solution to the equations exists in general. The total least square criterion is thus imposed upon the equations so that a unique $\mathbf{c}$ can be obtained:

> Find the rank $L$ matrix $\mathbf{B}'$ which minimizes $\|\mathbf{B}' - \mathbf{B}\|_F$. In the null space of $\mathbf{B}'$, find the minimum norm vector $\hat{\mathbf{c}}$, whose last element is 1.

Again, the advantage of the minimum norm solution is in the segregation of the signal zeros from extraneous zeros. $\mathbf{B}'$ is approximated based on the "Principal Eigenvector principle" (28)(29) and is given as:

$$\mathbf{B}' = \sum_{i=1}^{L} \sigma_i \mathbf{u}_i \mathbf{v}_i^H \tag{31}$$

Its null space is given by:

$$Null(\mathbf{B}') = span(\mathbf{v}_{L+1}, \ldots, \mathbf{v}_{rank(\mathbf{B})}). \tag{32}$$

The minimum norm null vector (with the last element = 1) can be obtained from the algorithm given by Golub and van Loan [2][1]: Let

$$\beta \equiv rank(\mathbf{B}), \qquad \alpha \equiv \sum_{i=L+1}^{\beta} |v_{i,M+1}|^2.$$

If $\alpha = 0$ there is *no solution*; otherwise,

$$\hat{\mathbf{c}} = \frac{1}{\alpha} \sum_{i=L+1}^{\beta} v_{i,M+1}^* \mathbf{v}_i \tag{33}$$

which ensures that the last element of $\hat{\mathbf{c}}$ is 1. $v_{ij}$ is the $j^{th}$ element of $\mathbf{v}_i$. This algorithm is also used to choose a unique $\mathbf{c}$ when (30) is underdetermined.

We can summarize TLS spectral estimation as follows:

1. Choose a LP model order $M$, $L \leq M \leq N - L$.
2. Form the matrix equation $\mathbf{Bc} = \mathbf{0}$.
3. Estimate the number of sinusiodal components, $L$, by the number of the "principal" singular values of $\mathbf{B}$.

4. Approximate the null space from the right singular vectors of $\mathbf{B}$, (31). Obtain the minimum norm solution $\hat{\mathbf{c}}$, (33).

5. Find the zeros of $\psi(z)$, (24). Identify the signal zeros which should have unit norm and the extraneous zeros which should have norms less than 1. There should be $L$ signal zeros. The frequencies of the sinusoids are given by $\{\omega_i : e^{j\omega_i} = z_i\}_{i=1}^{L}$ where $z_i$'s are the signal zeros.

6. Once the frequencies are known, the amplitudes $\{p_i\}_{i=1}^{L}$ can be found by solving a set of linear equations as prescribed by the Prony method.

## Comparison between the two approaches

Let $\bar{\mathbf{c}} = [c_M, \ldots, c_1]^T$ and let $\sigma_i, \mathbf{u}_i, \mathbf{v}_i$ be the singular values and left and right singular vectors of the matrix $\mathbf{A}$, (23). Rahman and Yu [11] show that $\bar{\mathbf{c}}$ obtained in the least squares (LS) and total least squares approaches can be expressed as:

$$\hat{\mathbf{c}}_{LS} = \sum_{i=1}^{L} \frac{1}{\sigma_i}(\mathbf{u}_i^H \mathbf{b})\mathbf{v}_i \tag{34}$$

$$\hat{\mathbf{c}}_{TLS} \simeq \sum_{i=1}^{L} \frac{1}{\sigma_i - \frac{\eta^2}{\sigma_i}}(\mathbf{u}_i^H \mathbf{b})\mathbf{v}_i, \tag{35}$$

where $\eta$ is the near-zero singular values corresponding with the noise subspace of B:

$$\eta \equiv \sigma_{L+1}^B \approx \cdots \approx \sigma_{M+1}^B$$

where $\sigma_i^B$'s are the singular values of $\mathbf{B}$. The assumption that these singular values are approximately equal is acceptable especially when the noise is white.

According to (34) and (35), each principal singular value of $\mathbf{A}$ is corrected by $\eta^2/\sigma_i$ in TLS. These correction terms in TLS must be the result of the inclusion of $\mathbf{b}$ in the lower rank approximation. Evidently, the additional information provided by $\mathbf{b}$ indicates that the prinicipal singular values of $\mathbf{A}$ are perturbed by these amounts. Note that the larger the singular values, the smaller the perturbation is. This is in accord with the theme behind the noise reduction technique based on the dimensionality of data matrices.

## Experimental comparison

Rahman and Yu [11] have conducted experiments to compare the performance of the least squares and total least squares approaches. The signal in the experiment is composed of equi-magnitude sinusoids with close frequencies: $2\pi(0.52)$ and $2\pi(0.50)$. The signal is corrupted by an additive complex white Gaussian noise by signal to noise ratios of: 12, 7, 3, 0.5 dB. The performance is measured by the average of the sums of square differences between the estimated and true frequencies. The results are summarized below, note that $L \leq M \leq N - L$:

1. TLS is consistently better than LS. The margin of performance is especially large when $M$ is low or when signal to noise ratio is high.

2. Both TLS and LS perform better as $M$ increases. (We have discussed previously that the resolution of the estimator should improve as the order of the underlying LP model increases.)

3. TLS and LS perform the same when $M$ is maximum. ($M = N - L$ is called the "Kumaresan-Prony" case, and $\mathbf{c}$ can be solved for without singular value decomposition. Use $\mathbf{A}^{-P} = (\mathbf{A}^H\mathbf{A})^{-P}\mathbf{A}^H$ in (25) [14].)

4. At lower signal to noise ratios, both TLS and LS achieve the Cramer-Rao bound for a maximum likelihood frequency estimator.

5. In TLS some extraneous zeros of $\psi(z)$ have magnitudes greater than one, making frequency identification difficult.

These experimental results are rather disappointing because TLS is not so much better than LS in important practical cases: high resolution ($\sim M$) and low signal to noise ratio. Also, the erratic zero locations in TLS (item 5 above) are perplexing. Rahman and Yu did not give explanation to this observation.

# Pisarenko Harmonic Decomposition (PHD)

Pisarenko [10] has shown that any finite sequence of correlations, $r_k \equiv E(x_{n+k}x_n^*)$, can be decomposed into a linear combination of sinusoids plus a white noise:

$$r_k = \sum_{i=1}^{L} \overline{p_i}^2 e^{j\omega_i k} + \eta\delta_k. \tag{36}$$

The corresponding PSD is a line spectrum in a background white noise:

$$P(\omega) = \eta + \sum_{i=1}^{L} \overline{p_i}^2 \delta(\omega - \omega_i).$$

Grenander and Szego [3] show (in the proof for the Caratheodory's theorem) that a finite sequence with the Hermitian property, $r_{-k} = r_k^*$, can always be decomposed as in (36). This served as the basis for the original derivation of the Pisarenko's harmonic decomposition method. In this section we analyze PHD from the perspective of Prony's formulation, which allows us to compare PHD with TLS.

## The principle

Computationally, Pisarenko decomposition is an application of Prony's method. We have discussed that the data sequence $\{x_n\}$ may not be able to be fit into a linear combination of sinusoids. But since we are dealing with the correlation sequence, we are expressly assuming that $\{x_n\}$ is wide-sense stationary. Papoulis [9] shows that if a wide-sense stationary signal fits a finite LP model exactly then its correlation sequence can be described as a finite sum of sinusoids:

$$r_k = \sum_{i=1}^{L} \overline{p_i}^2 e^{j\omega_i k}. \tag{37}$$

Applying Prony's method, we see that the coefficients $\{c_i\}$ of $\psi(z)$, whose zeros are $e^{j\omega_i}$, are also the coefficients of the LP model similar to (19):

$$\sum_{i=0}^{L} c_i r_{k-i} = 0, \qquad c_0 = 1, \tag{38}$$

which yields the matrix equation:

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_L \\ r_{-1} & r_0 & \cdots & r_{L-1} \\ \vdots & \vdots & & \vdots \\ r_{-L} & r_{-L+1} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} c_L \\ c_{L-1} \\ \vdots \\ c_0 \end{bmatrix} = \mathbf{0}, \tag{39}$$

or

$$\mathbf{Rc} = \mathbf{0}. \tag{40}$$

$\mathbf{R}$ is given by

$$\mathbf{R} = \left[E(\mathbf{x}\mathbf{x}^H)\right]^T = E(\mathbf{x}^*\mathbf{x}^T) \tag{41}$$

To obtain a non-trivial estimate for $\mathbf{c}$, the $(L+1) \times (L+1)$ matrix $\mathbf{R}$ must be singular. But in practice $\mathbf{R}$ tends to be nonsingular because of the noise in data as well as possible modeling error in (37). One way to deal with this situation is to approximate $\mathbf{R}$ by a singular matrix $\mathbf{R}'$. In PHD,

$$\mathbf{R}' \equiv \mathbf{R} - \sigma\mathbf{I} \tag{42}$$

where $\sigma$ is an eigenvalue of $\mathbf{R}$. Note that $\mathbf{R}'$ is singular; $det(\mathbf{R}') = det(\mathbf{R} - \sigma\mathbf{I}) = 0$. We can assess the "closeness" of $\mathbf{R}'$ to $\mathbf{R}$ by the Frobenius norm:

$$\|\mathbf{R} - \mathbf{R}'\|_F^2 = \|\sigma\mathbf{I}\|_F^2 = (L+1)\sigma^2 \tag{43}$$

The smallest eigenvalue, $\sigma_{min}$, is used in (42) so that $\mathbf{R}'$ is as close as possible to $\mathbf{R}$. In fact, the smallest eigenvalue *must* be used to keep $\mathbf{R}'$ positive semi-definite. $\mathbf{R}'$ must be positive-definite because it approximates the *true* correlation matrix. By taking the similarity transformation (equivalent with the singular value decomposition in this case) of $\mathbf{R}$ in (42) we can see that every eigenvalue of $\mathbf{R}'$ can be obtained by subtracting $\sigma$ from each eigenvalue of $\mathbf{R}$. If $\sigma > \sigma_{min}$, then we would have at least one negative eigenvalue of $\mathbf{R}'$.

Using $\mathbf{R}'$ we can proceed with Prony's method to decompose $\{r'_k\}$:

$$r'_k = \sum_{i=1}^{L} \overline{p_i}^2 e^{j\omega_i k}. \tag{44}$$

Note that (42) with $\sigma = \sigma_{min}$ suggests that

$$r'_k = r_k - \sigma_{min}\delta_k;$$

thus, we have the desired decomposition of $\{r_k\}$:

$$r_k = \sum_{i=1}^{L} \overline{p_i}^2 e^{j\omega_i k} + \sigma_{min}\delta_k. \tag{45}$$

Comparing this with (36) we have $\eta = \sigma_{min}$.

## Choosing the model order and an unique $c$

The harmonic decomposition method described above requires prior knowledge of $L$, the number of the sinusoids in the signal. In many applications $L$ is not known and is to be estimated from the data. Suppose that we have $N$ data samples from which we can form a $K \times K$ correlation matrix $\mathbf{R}_K$, where the subscript denotes the size of the matrix and $L < K \leq N$. From (37) we can see that only $L$ columns of $\mathbf{R}_K$ are linearly independent in the ideal case. Hence, $\mathbf{R}_K$ must have $K - L$ zero eigenvalues. In practice, $\mathbf{R}_K$ is perturbed, but $K - L$ of its eigenvalues are nevertheless almost zero. Thus, $L$ can be estimated by the number of "principal" eigenvalues in $\mathbf{R}_K$.

To find an unique $\mathbf{c}$ in the null space of the correlation matrix, PHD forces the null space to be one-dimensional by forming $\mathbf{R}_{L+1}$. A single eigenvector corresponding with $\sigma_{min}$ of $\mathbf{R}_{L+1}$ spans the null space of $\mathbf{R}'_{L+1} = \mathbf{R}_{L+1} - \sigma_{min}\mathbf{I}_{L+1}$. $\mathbf{c}$ is obtained by scaling this vector so that the last element is 1. The Pisarenko harmonic decomposition method is summarized as below:

1. Estimate $\mathbf{R}_K$ from the data $\{x_n\}$ and find its eigenvalues $\{\sigma_i\}$, $i = 1, \ldots, K$.

2. Estimate L, the number of exponential components, by the size of the eigenvalues, i.e., $\sigma_1 \geq \cdots \geq \sigma_L > \sigma_{L+1} \approx \cdots \approx \sigma_K \approx 0$. If none of the eigenvalues is close to zero, more data samples and a larger $K$ are desirable.

3. Form $\mathbf{R}_{L+1}$ from the data. Find an eigenvector corresponding with its smallest eigenvalue. This eigenvector, scaled appropriately so that $c_0 = 1$, is the estimate for $\mathbf{c}$. The smallest eigenvalue is the estimate for $\eta$.

4. Proceed with the Prony's method to solove for $\{\omega_i\}$ and $\{p_i\}$.

## Interpretation of the noise strength $\eta$

Like TLS, PHD has an underlying LP model. Let us introduce a zero-mean white prediction error $\{n_n\}$:

$$\begin{aligned} x_n &= \sum_{i=1}^{L}(-c_i)x_{n-i} + n_n \\ r_k &= E(x_{n+k}x_n^*) \\ &= \sum_{i=1}^{L}(-c_i)E(x_{n+k-i}x_n^*) + E(n_{n+k}x_n^*) \\ &= \sum_{i=1}^{L}(-c_i)r_{k-i} + \nu\delta_k, \end{aligned} \tag{46}$$

where $\nu$ is the variance of the white noise. Note that (46) is an autoregressive (AR) model of the data sequence. We now have a variation of (38) incorporating an explicit noise model:

$$\sum_{i=0}^{L} c_i r_{k-i} = \nu\delta_k, \qquad c_0 = 1, k \geq 0. \tag{47}$$

This is called the Yule-Walker equations. Writing them for $k = L,\ L-1,\ \ldots,\ 0$ in a matrix form we have

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_L \\ r_{-1} & r_0 & \cdots & r_{L-1} \\ \vdots & \vdots & & \vdots \\ r_{-L} & r_{-L+1} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} c_L \\ c_{L-1} \\ \vdots \\ c_0 \end{bmatrix} \approx \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \nu \end{bmatrix}, \tag{48}$$

or

$$\mathbf{Rc} = \mathbf{h} \tag{49}$$

For a non-singular $\mathbf{R}$, we can apply the similarity transform to it and solve for $\nu$:

$$\begin{aligned} \mathbf{R} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H \\ \mathbf{\Lambda} &= diag(\sigma_1, \ldots, \sigma_{L+1}) \end{aligned}$$

$$\sigma_1 \geq \cdots \geq \sigma_{L+1} > 0$$

$$\begin{aligned} \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{c} &= \mathbf{h} \\ \mathbf{\Lambda}\mathbf{U}^H\mathbf{c} &= \mathbf{U}^H\mathbf{h} \end{aligned}$$

By equating the last elements from both sides of the last equation,

$$\begin{aligned} \sigma_{L+1}u^*_{L+1,L+1} &= u^*_{L+1,L+1}\nu \\ \sigma_{L+1} &= \nu \end{aligned}$$

$$\nu = \sigma_{L+1} = \sigma_{min} = \eta. \tag{50}$$

In another words, the magnitude of the noise in the Pisarenko decomposition (36) can be interpreted as the variance of the white noise in the $L^{th}$ order AR model (46).

## Pisarenko method is suboptimum

An important feature of PHD is that the decomposition yeilds a noise term which is *physically interpretable* as the white nosie in the AR model. This is the result of the lower rank approximation $\mathbf{R}'_{L+1} = \mathbf{R}_{L+1} - \sigma_{min}\mathbf{I}_{L+1}$, (42). Since for any rank deficient matrix $\mathbf{R}'$

$$\|\mathbf{R} - \mathbf{R}'\|_F \geq \sigma_{min},$$

(43) indicates that $\mathbf{R}'_{L+1}$ is not the optimal lower rank approximation of $\mathbf{R}_{L+1}$ in the sense of the Frobenius norm. If TLS is applied to the matrix $\mathbf{R}_{L+1}$, the decomposition will be (44). That is, PHD produces the same $\{\omega_i\}$ and $\{p_i\}$ as an optimal method does. It is the noise term $\eta\delta_k$ that makes Pisarenko decomposition suboptimal.

# The MUSIC Method

MUSIC (MUltiple SIgnal Classification) was originally developed for the identification of the directions and amplitudes of the signals arriving at an array of antennas [12]. The "angles of arrival" in the direction finding problem can be treated as the spectral frequencies, and MUSIC is directly applicable to the high resolution spcetral estimation problem. Compared with TLS and PHD, MUSIC formulates the problem in a more general framework, hence its applicability to a wider range of estimation problems. It also requires more prior information about the signal—some second order statistics of the noise process and a set of "calibration" data detailed below.

## MUSIC in direction finding problems

Let us contrast the formulation of MUSIC with that of Prony's method on which TLS and PHD are based. In the direction finding problem which motivated the development of MUSIC, an array of antennas receives the linear combination of sinusoidal signals arriving from various directions. The problem is to find the arrival angle and complex amplitude of each sinusoid.

- In the direction finding problem, the unknown angles of arrivals are *multi-dimensional* (e.g., latitudes and longitudes). These unknowns are treated as the spectral frequencies $\omega$ when MUSIC is appliced to the high resolution spectral estimation problem. We denote the unknowns with $\theta$'s instead of $\omega$'s to signify the multi-dimentionality, but we still refer to them as "frequencies."

- The data set $\{x_n\}_{n=1}^{K}$ is the set of complex numbers registered at the $K$ receptors of the antenna array at a given time (or averages over time). Because of the arbitrary geometrical configuration of the receptor array, $\{x_n\}$ are *no longer equally spaced* samples.

- In the spectral estimation problem the signal subspace is spanned by vectors of the form $\mathbf{s}(\omega) = [e^{j\omega}, e^{2j\omega}, \cdots]^T$. The equivalent expression for the signal vectors in the direction finding problem is difficult to determine because of

the multi-dimensionality of the frequencies and arbitrary spacing between data samples. It must be obtained specifically for a given problem setting, e.g., for a particular arrangement of the receptor array, and it is referred to as the set of "*calibration data.*" We denote it as $\mathbf{s}(\theta)$ which is a $K$ dimensional vector function of $\theta$. The ideal signal is a linear combination of several samples of $\mathbf{s}(\theta)$.

- The second order statistics of the noise process $\{n_n\}$ must be known. The noise is assumed to be zero-mean but not white. The non-white assumption is appropriate because the noise components in the data from receptors located closely are likely to be correlated (e.g., the receivers may be under the same noisy interferrance.).

## The general principle

We have the ideal signal model $\mathbf{x} = \mathbf{Sp}$ from (12). We now introduce an additive noise $\mathbf{n} = [n_1, \ldots, n_K]^T$:

$$\mathbf{x} = \mathbf{Sp} + \mathbf{n} \tag{51}$$

Given this signal model and and the set of calibration data $\{\mathbf{s}(\theta), \forall\theta\}$, MUSIC provides us with a search procedure for $\mathbf{s}_i$, $i = 1, \ldots, L$. Ideally, each of the $L$ sinusoidal vectors $\{\mathbf{s}_i\}_{i=1}^{L}$ must satisfy the following conditions:

- It is in the signal subspace.
- It is in the set of the calibration data, i.e., $\exists\theta_i$ such that $\mathbf{s}_i = \mathbf{s}(\theta_i)$.

We thus need to identify the intersection between the signal subspace and the set $\mathbf{s}(\theta)$.

Like PHD, MUSIC tracks the problem from the singularity property of the ideal correlation matrix. Let the $K \times K$ data correlation matrix defined as follows:

$$\begin{aligned} \mathbf{Q} &= E(\mathbf{x}\mathbf{x}^H) \\ &= \mathbf{S}\,E(\mathbf{p}\mathbf{p}^H)\,\mathbf{S}^H + E(\mathbf{n}\mathbf{n}^H) \\ &= \mathbf{S}\mathbf{P}\mathbf{S}^H + \eta\mathbf{H} \end{aligned} \tag{52}$$

Note that we have assumed that each $p_i$ is uncorrelated with each $n_n$. Also, we assume that $\mathbf{P}$ is non-singular. $\eta$ is the noise strength, and $\mathbf{H}$ is the noise correlation matrix. $\mathbf{H}$ is an identity matrix if the noise turns out to be white.

The $K \times K$ matrix $\mathbf{SPS}^H$ is of rank $L$ because $rank(\mathbf{S}) = L$, $rank(\mathbf{P}) = K$, and $K > L$. Its column vectors belong to the signal subspace. To identify the signal and noise subspace we start by requiring

$$det(\mathbf{Q} - \eta\mathbf{H}) = det(\mathbf{SPS}^H) = 0$$

since $\mathbf{SPS}^H$ is singular. We see that $\eta$ must be a *generalized eigenvalue* corresponding with the *pencil* $\mathbf{Q} - \eta\mathbf{H}$. As we show later in the section, since $\mathbf{SPS}^H$ has $K - L$ zero eigenvalues, the minimum eigenvalue of $\mathbf{Q} - \eta\mathbf{H}$ has the multiplicity of $K - L$. The corresponding right (generalized) eigenvectors $\{\mathbf{u}_i\}_{i=L+1}^{K}$ span the noise subspace. Now let us consider the Euclidean ("2-norm") distance $d(\theta)$ from a calibration vector $\mathbf{s}(\theta)$ to the signal subspace. $d(\theta)$ is equal to the norm of the projection of $\mathbf{s}(\theta)$ onto the noise subspace:

$$d(\theta) = \|\mathbf{s}^H(\theta)\mathbf{U}_{ns}\|$$

$$\mathbf{U}_{ns} = [\mathbf{u}_{L+1}|\cdots|\mathbf{u}_K] \tag{53}$$

In principle $d(\theta)$ would have at least $L$ zeros corresponding with the desired frequencies $\{\theta_i\}_{i=1}^{L}$. In practice we expect the noise to perturb $d(\theta)$. MUSIC uses the function

$$g(\theta) = \frac{1}{\|\mathbf{s}^H(\theta)\mathbf{U}_{ns}\mathbf{U}_{ns}^H\mathbf{s}(\theta)\|} \tag{54}$$

on which the locations of $\{\theta_i\}_{i=1}^{L}$ are identified as sharp positive peaks. Schmidt [12] does not describe what to do when the number of peaks is not $L$, even though he mentions such a possibility. It seems that the peaks must be chosen with some qualitative judgements, and the value of $L$ may have to be adjusted according to the form of $g(\theta)$.

Once the "frequencies" $\{\theta_i\}_{i=1}^{L}$ are known, $\mathbf{S}$ can be evaluated and the elements of $\mathbf{P}$ can be obtained from (52). The diagonal elements are the average powers of

the sinusoidal components. Only the *relative* phase between a pair of sinusoids can be identified from the cross-correlation term. If a cross-correlation term is zero, the phase relation between the corresponding pair of sinusoids cannot be determined directly.

## Summary of the MUSIC method

1. Estimate the $K \times K$ correlation matrix $\mathbf{Q}$ from the data sequence.
2. Find the multiplicity of the minimum generalized eigenvalue of the pencil $\mathbf{Q} - \eta\mathbf{H}$. $K$ minus the multiplicity is the estimate for the number of signal components $L$.
3. Find the $K - L$ eigenvectors corresponding with the minimum eigenvalue. Form $\mathbf{U}_{ns}$ as in (53).
4. Find the $L$ tallest peaks in $g(\theta)$, from which $\{\theta_i\}_{i=1}^{L}$ are identified.
5. Evaluate $\mathbf{P}$ and identify from its elements the average powers of the signal components as well as the relative phases.

## MUSIC in high resolution harmonic decompositions

In spectral estimation, white noise and equally spaced samples are reasonable assumption. We have:

$$E(\mathbf{n}\mathbf{n}^H) = \eta\mathbf{I}$$

$$\mathbf{s}(\theta) = [e^{j\theta}, \cdots, e^{jK\theta}]^T$$

where $\theta$ is now one-dimensional. From (52) we have

$$\mathbf{S}\mathbf{P}\mathbf{S}^H = \mathbf{Q} - \eta\mathbf{I} \tag{55}$$

which is analogous to the rank $L$ approximation in PHD, (42).

MUSIC then performs the search for the spectral frequencies using the eigenvectors associated with the noise subspace. Contrastingly, PHD reforms a smaller correlation matrix to force the noise subspace to be one-dimensional. The noise

subspace yields the coefficients of $\psi(z)$ whose $L$ roots are the spectral frequencies; thus, in principle PHD does not require a search. It seems that PHD has thrown away some available correlation data by restricting the size of the correlation matrix. However, one can argue that for a fixed number of data samples smaller correlation matrices are estimated more accurately because more data are averaged. Thus, there seems no definitive answer to which of the two methods should perform better.

## The eigen-structure of $Q - \lambda H$

We sidetrack a moment to show that the minimum eigenvalue $\eta$ of the pencil $\mathbf{Q}-\lambda\mathbf{H}$ does repeat $K-L$ times and the corresponding eigenvectors span the noise subspace of $\mathbf{SPS}^H = \mathbf{Q} - \eta\mathbf{H}$. By definition *generalized* eigenvalues $\{\lambda_i\}$ and eigenvectors $\{\mathbf{u}_i\}$ for the pencil satisfy

$$det(\mathbf{Q} - \lambda_i\mathbf{H}) = 0$$
$$(\mathbf{Q} - \lambda_i\mathbf{H})\mathbf{u}_i = 0.$$

Let $\tilde{\mathbf{Q}} \equiv \mathbf{SPS}^H$ and its eigenvalues and eigenvectors be denoted $\{\tilde{\sigma}_i\}$ and $\{\tilde{\mathbf{u}}_i\}$, respectively, i.e.,

$$det(\tilde{\mathbf{Q}} - \tilde{\sigma}_i\mathbf{I}) = 0$$
$$(\tilde{\mathbf{Q}} - \tilde{\sigma}_i\mathbf{H})\tilde{\mathbf{u}}_i = 0.$$

Then, since $\tilde{\mathbf{Q}} = \mathbf{Q} - \eta\mathbf{H}$ we have

$$\tilde{\mathbf{Q}} - \tilde{\sigma}\mathbf{I} = (\mathbf{Q} + \tilde{\sigma}\mathbf{H}_1) - (\eta + \tilde{\sigma})\mathbf{H} \quad (56)$$
$$= (\mathbf{Q} + \tilde{\sigma}\mathbf{H}_1) - \lambda\mathbf{H} \quad (57)$$

where $\mathbf{H}_1 \equiv \mathbf{H} - \mathbf{I}$. For each $\tilde{\sigma}$ there are $K$ eigenvalues $\lambda$ for the pencil $(\mathbf{Q}+\tilde{\sigma}\mathbf{H}_1) - \lambda\mathbf{H}$ and we observe from (56) and (57) that at least one of them must equal $\eta + \tilde{\sigma}$. Note that the generalized eigenvector corresponding with this eigenvalue ($\lambda = \eta+\tilde{\sigma}$) is $\tilde{\mathbf{u}}_i$. Since $\{\tilde{\mathbf{u}}_i\}$ are distinct, we can set the index so that the $i^{th}$ $\lambda$ relates to $\tilde{\sigma}$:

$$\lambda_i(\tilde{\sigma}_i) = \eta + \tilde{\sigma}_i$$

Note that $\lambda_i(0)$ is the generalized eigenvalue of the pencil $\mathbf{Q}-\lambda\mathbf{H}$. Since $\{\tilde{\sigma}_i = 0\}_{i=L+1}^{K}$,

$$\lambda_i(0) = \eta, \qquad i = L+1, \ldots, K$$

Thus, $K - L$ of the eigenvalues of $\mathbf{Q} - \eta\mathbf{H}$ are equal to $\eta$.

We now show that $\eta$ is the minimum eigenvalue. From [1], there exists an invertible matrix $\mathbf{X}$ such that

$$\mathbf{Q} - \eta\mathbf{H} = \mathbf{X}^H diag\{q_i - \eta h_i\}\mathbf{X}$$

where $\{q_i/h_i,\ i = 1, \ldots, K\}$ are the eigenvalues. Note that the pencil is positive semi-definite: $\mathbf{Q} - \eta\mathbf{H} = \tilde{\mathbf{Q}} \geq 0$. Since $X$ is invertible, we must have

$$\begin{aligned} diag\{q_i - \eta h_i\} &\geq 0 \\ q_i - \eta h_i &\geq 0, \qquad i = 1, \ldots, K \\ \eta &\leq \frac{q_i}{h_i}, \qquad i = 1, \ldots, K \end{aligned}$$

Hence, $\eta$ must be the minimum eigenvalue.

# An Unified Viewpoint Revisited

## The signal and noise subspaces in Prony's method

In TLS and PHD, we have a data matrix whose *rows* are derived from the data vectors. From the singular value decomposition (13) we see that $\{\mathbf{v}_i^H\}$ span the row space of $\mathbf{X}$. The signal and noise subspaces are thus identified as

$$\mathcal{S}_s = span(\mathbf{v}_1^*, \ldots, \mathbf{v}_L^*) \tag{58}$$

$$\mathcal{N}_s = span(\mathbf{v}_{L+1}^*, \ldots, \mathbf{v}_\chi^*) \tag{59}$$

where $\{\mathbf{v}_i\}$ are the right singular vectors and $\chi = rank(\mathbf{X})$. The vector $\mathbf{c}$ in (30) and (40) is in the null space of the rank $L$ approximation $\mathbf{X}'$:

$$\mathbf{c} \in Null(\mathbf{X}') = span(\mathbf{v}_{L+1}, \ldots, \mathbf{v}_\chi)$$

Let

$$\gamma \equiv \mathbf{c}^*.$$

Then,

$$\gamma \in \mathcal{N}_s$$

The polynomial $\psi(z)$ in (24) is the inner product

$$\gamma^H \mathbf{z} = \mathbf{c}^T \mathbf{z} = \psi(z)$$

where $\mathbf{z} = [1, z, \ldots, z^{K-1}]^T$. If the noise subspace is estimated accurately, we have

$$\mathbf{s}_i \perp \mathcal{N}_s \qquad i = 1, \ldots, L. \tag{60}$$

Thus,

$$\gamma^H \mathbf{s}_i = \psi(e^{j\omega_i}) = 0, \qquad \gamma \in \mathcal{N}_s. \tag{61}$$

This shows that in the ideal case *any* vector in the noise subspace will yield Prony's polynomial $\psi(z)$, whose $L$ of $K-1$ roots have unit norm.

The spectral estimation based on Prony's method can be restated as following:

1. Identify the signal and noise subspaces based on the principle that the "principal" singular values of the data matrix is much larger than the "near-zero" singular values.

2. Choose a vector $\gamma$ in the noise subspace. Find a set of basis vectors for the $(K-1)$ dimensional subspace orthogonal to $\gamma$, under the constraint that these vectors must be of the form $\mathbf{z} = [z, z^2, \ldots, z^K]^T$. The constraint is there because we know that the sinusoidal vectors $\{\mathbf{s}_i\}_{i=1}^L$ are vectors of the same form.

3. Using the fact that the sinusoidal vectors span the signal subspace, identify these vectors in the set of $(K-1)$ basis vectors. Then, find $\{\omega_i\}_{i=1}^L$.

4. Form and solve linear equations for $\{p_i\}_{i=1}^L$.

A distinguishing feature of Prony's method is that step 2 is achieved by finding the zeros of a polynomial. The corresponding operation in MUSIC is the search for peaks in $g(\theta)$, (54). Also, whereas MUSIC employs the entire noise subspace in the search for the spectral frequencies (53), Prony's method finds the frequencies based on only one vector in the subspace. In another words, some information contained in the noise subspace is not used in Prony's method. PHD is a special case to this because its noise subspace is one-dimensional.

In practice, since the signal and noise subspaces in (58) and (59) are not exact, (60) and (61) are not strictly true, implying that the signal zeros may not have unit norm. We have stated that with the "minimum norm" $\gamma$ (i.e., $\gamma$ = projection of $[0, \ldots, 0, 1]^T$ on $\mathcal{N}_s$) all the extraneous zeros should be found closer to the origin than the signal zeros. When signal to noise ratio is low the zeros are expected to be dislocated greatly, and there seems no good criteria to choose the signal zeros and the corresponding spectral frequencies.

## Concluding summary

The common features of the three spectral estimators discussed in this report are:

- They model the signal as a linear combination of complex sinusoids; thus, they produce line spectra instead of smooth waveforms.
- They estimate the number of sinusoids in the signal from the number of "principal" singular values of some data matrices.

The three methods also have a number of contrasting features:

1. The signal and noise subspaces.
   - In TLS and PHD, the row space of the data matrix $\mathbf{X}$ is associated with the space of the (transposed) data vectors; thus, the signal and noise subspaces are spanned by $\{\mathbf{v}^*_i\}_{i=1}^{L}$ and $\{\mathbf{v}^*_i\}_{i=L+1}^{N}$, respectively.
   - In MUSIC, the column space is associated with the data vectors. The subspaces are given by $\{\mathbf{u}_i\}_{i=1}^{L}$ and $\{\mathbf{u}_i\}_{i=L+1}^{N}$.
2. The lower rank approximations.
   - In TLS, the rank $L$ approximation $\mathbf{X}'$, given in (14), has the smallest Frobenius distance to $\mathbf{X}$.
   - PHD and MUSIC use the suboptimal approximation in (16). The particular form of the approximation allows them to describe the noise process as the white noise in an AR model (46), in the PSD (36), or in the signal model (51).
3. The search procedures.
   - With TLS, one must search for the signal zeros from the roots of $\psi(z)$. This might be difficult under a low signal to noise ratio. The search is required because a higher order LP model is used to improve frequency resolution.
   - PHD requires no search; all roots of $\psi(z)$ are signal zeros.
   - The peaks of $g(\theta)$, (54), must be found in order to identify the frequencies.
4. Other comments.

- TLS works with the raw samples, i.e., unlike the other two methods it does not require estimation of the correlations. To use TLS, one must make sure that the ideal signal is a linear combination of sinusoids, for TLS does not guarantee that any of the roots of $\psi(z)$ has unit norm.
- PHD is the only one that does not require any kind of a search procedure. It achieves this at the cost of forming the second correlation matrix.
- MUSIC formulates the problem in a general framework and has a wider range of applications such as the direction finding problems. It also requires more prior data, i.e., the set of calibration data.

# Reference


[1] G. H. Golub and C. F. Van Loan, *Matrix Computations*, sections 2.3, 8.6, 12.3. Baltimore, MD: Johns Hopkins Univ. Press, 1984.

[2] G. H. Golub and C. F. Van Loan, "An analysis of the total least squares problem," *SIAM J. Numer. Anal.*, v.17(6), pp.883-893, Dec. 1980.

[3] O. Grenander and G. Szego, *Toeplitz Forms and Their Applications*, chapter 4. Berkeley, CA: Univ. California Press, 1958.

[4] S. Haykin, *Adaptive Filter Theory*, sections 7.12, 7.14. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[5] F. B. Hildebrand, *Introduction to Numerical Analysis*, section 9.4. New York: McGraw-Hill, 1956.

[6] S. M. Kay and S. L. Marple, "Spectrum Analysis—A Modern Perspective," *Proc. IEEE*, v.69(11), pp.1380-1419, Nov. 1981.

[7] R. Kumaresan, "On the zeros of the linear prediction-error filter for deterministic signals," *IEEE Trans. Acoust., Speech., Signal Processing*, v.ASSP-31(1), pp.217-220, Feb. 1983.

[8] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, section 5.5. Englewood Cliffs, NJ: Pretice-Hall, 1975.

[9] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, section 13.4. New York: McGraw-Hill, 1984.

[10] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophysical J. Royal Astronomical Soc.*, v.33, pp.347-366, 1973.

[11] M. A. Rahman and K. Yu, "Total least squares approach for frequency estimation using linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, v.ASSP-35(10), pp.1440-1454, Oct. 1987.

[12] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Proc. RADC Spectrum Estimation Workshop*, pp.243-258, 1979.

[13] D. W. Tufts and R. Kumaresan, "Singular value decomposition and improved frequency estimation using linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, v.ASSP-30(4), pp.671-675, Aug. 1982.

[14] D. W. Tufts and R. Kumaresan, "Frequency estimation of multiple sinusoids: making linear prediction perform like maximum likelihood," *Proc. IEEE*, v.70(9), pp.975-989, Sep. 1982.

[15] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, chapter 2. Oxford, England: Clarendon, 1965.

# After Thoughts (March, 1988)

So, which of the three methods is the best? We have revealed some differences among the three methods such as the ability to retain the phase information (p.4, pp.29-30, p.35) and so on. But the main issue is the estimation of the frequencies, and with this regard the three methods are pretty much the same. Their characteristic features are:

1. The use of some subsequences of the data sequence to form what we call the "data matrices" whose row or column spaces are ideally of rank $L$.
2. The estimation of $L$ from the sizes of the singular values of the data matrices.
3. The identification of $\{\omega_i\}$ using the fact that the sinusoidal vectors (p.7) are perpendicular to the noise subspace (p.29)(60).

The methods differ in accomplishing step 3. Whereas MUSIC uses a full set of the basis vectors of the noise subspace to estimate the frequencies (53)(54), the Prony-based methods (TLS, PHD) use only one vector in the subspace (61). It seems that Prony's approach is wasting some available information in the subspace, especially in TLS. However, there is no experimental evidence indicating that MUSIC estimates the frequencies more accurately than TLS [1].

Step 2 is a crucial step. The task of identifying the "near-zero" singular values becomes easier if their values are equal. We have shown that for the correlation matrices ($\mathbf{Q}$ and $\mathbf{R}$) such is the case (pp.31-32, p.23). Are MUSIC and PHD better than TLS then? Theoretically the answer may be yes, but practically it is no. Suppose for a given data sequence the matrix $\mathbf{B}$ (TLS) yields such singular values $\{\sigma_i^B\}$ that the estimation of $L$ is quite difficult. Let us try MUSIC then. The correlation matrix $\mathbf{Q}$ (10) is typically *estimated* by

$$\mathbf{Q} = \frac{1}{K}(\mathbf{B}\mathbf{B}^H)$$

[1] private correspondence with Bruce Musicus

whose singular values can be shown to be $\{\sigma_i^{B^2}\}$, unfortunately. Thus, practically speaking, if TLS has a problem identifying the "near-zero" singular values, so would MUSIC and PHD.

The interpretations of the noise in the three methods differ slightly. In TLS, the rank $L$ approximation to the data matrix is chosen to be the one which minimizes the Frobenius norm of the difference between the two (31). We have called this an "optimal" rank $L$ approximation (p.8). Is it really the right thing to do? Maybe not. The data samples $\{x_n\}$ are repeated in the data matrix $\mathbf{B}$, e.g., $x_2$ appears twice; however, the minimization of the Frobenius norm implies that each element of the matrix is perturbed randomly, ignoring the obvious correlations among the matrix elements.

To identify the signal subspace accurately the three methods must assume that the noise is additive and white (p.9). The "suboptimal" rank $L$ approximation (16) is shown to correspond with an additive white noise model (pp.25-26). In a sense this approximation is superior to the "optimal" one (14), because it correctly reflects the underlying assumption behind the noise reduction technique.