Chapter 1

A REVIEW OF PARAMETRIC HIGH-RESOLUTION METHODS

Qi Cheng, *University of Western Sydney*
Yingbo Hua, *University of California*

1.1 Introduction

High-resolution methods are generally defined to be high-performance methods for estimating and/or detecting the desired and/or undesired signal components present in a given set of data. The term "high-resolution" also implies a good ability to resolve very "similar" signal components. One of the most common problems in signal processing is known as frequency estimation. In frequency estimation, "high-resolution" often refers to a good ability to resolve two or more closely located frequencies in the given data. There are two groups of high-resolution methods. One is parametric methods, and the other non-parametric methods. The parametric high-resolution methods result from ingenious exploitations of known data structures. The non-parametric high-resolution methods maximize the output of some desired information with little knowledge of the data structure. The choice between parametric methods and non-parametric methods largely depends on one's confidence in the assumed data model. In this chapter, we expose the readers to a range of existing parametric high-resolution methods.

In Section 1.2, we present several frequency estimation techniques using algebraic principles. They are linear prediction, matrix pencil, and iterative quadratic maximum likelihood. The linear prediction method and the matrix pencil method can achieve near-optimal accuracy of estimation without the local convergence issues associated with the optimal methods. The computational complexities of the two methods are among the most efficient. The iterative quadratic maximum likelihood method is an approximation of the (exact) maximum likelihood method. Under some condition (e.g., high SNR), this approximation achieves the optimal accuracy. Concepts like forward-and-backward averaging, total least square, and (joint) singular value decomposition of orthonormal matrices will also be discussed.

In Section 1.3, we present methods that exploit large sample theorems in statistics. In particular, we focus on data of multiple independent measurements. The key data structure is captured by the dominant (principal) subspace of the data matrix or the dominant eigenvectors of the data covariance matrix. The principal subspace is referred

to as signal subspace. The orthogonal complement of the signal subspace is referred to as noise subspace. Exploitation of the orthogonality between the signal subspace and the noise subspace results in a method known as MUSIC. For some data models such as the superimposed exponentials, the signal subspace possesses an invariance structure that can be exploited in a matrix pencil fashion also known as ESPRIT. A number of variations of MUSIC and ESPRIT are described there. We also present the maximum likelihood methods and their asymptotic (large sample) performances. The bilinear nature of the sensor array model is exploited in the conditional maximum likelihood method, unconditional maximum likelihood method and the so-called MODE method. The situation of coherent signals is also discussed.

In Section 1.4, several detection methods using a single data measurement are presented. Several effective detection methods using a single data measurement require the knowledge of noise variance. The noise variance is a key to determine the thresholds used in the effective singular values method, the noise significance level method, and the least squares data fitting method. But the noise variance is not required in the information theoretic criteria. The noise level is captured by the maximum log-likelihood function. With a proper choice of the penalty term, the information theoretic criteria have attractive asymptotic (long sample) properties. (Note that a long single data measurement is treated differently from a large number of measurements.)

In Section 1.5, we consider multiple data measurements for signal detection. The Akaike information criterion (AIC) and the minimum description length (MDL) methods are reviewed. Detection methods based on the eigenvalues of data covariance matrix are presented. The Bayesian methods for detection are also discussed.

1.1.1 Data Model

In the following, we define the domain of our problems by describing the data model under consideration. The reader will see that our data model is also known as sensor array model. An important specialization of this model is known as superimposed exponentials. The problem of frequency estimation and detection is largely associated with this model, which is also the main focus of this chapter.

The data model under consideration is:

$$\mathbf{y}(n) = \mathbf{A}\mathbf{s}(n) + \mathbf{w}(n), \quad n = 1, ..., N \tag{1.1.1}$$

where

$$\mathbf{y}(n) = [y_0(n), y_1(n), ..., y_{M-1}(n)]^T \tag{1.1.2}$$
$$\mathbf{s}(n) = [s_1(n), ..., s_I(n)]^T \tag{1.1.3}$$
$$\mathbf{w}(n) = [w_0(n), w_1(n), ..., w_{M-1}(n)]^T. \tag{1.1.4}$$

$\mathbf{s}(n)$ is the signal vector, and $\mathbf{w}(n)$ the noise vector. We refer to M as the size of each measurement, and N as the number of independent measurements. The matrix $\mathbf{A}$ is generally parameterized by an unknown vector:

$$\boldsymbol{\theta} = [\theta_1, ..., \theta_I]^T, \tag{1.1.5}$$

and sometimes denoted by $\mathbf{A}(\boldsymbol{\theta})$. For notional convenience, this dependence on frequencies is often suppressed throughout this book. Each element of the vector (1.1.5) could also be a vector in some multidimensional problems, as in other chapters. But for simple presentation, in this chapter, we will mainly consider the one-dimensional case. Some references on the multidimensional case will be mentioned.

The structure of the matrix $\mathbf{A}$ provides the key for all high-resolution methods. In particular, the following is an important specialization of (1.1.1):

$$\mathbf{A} = [\mathbf{a}(\theta_1), \mathbf{a}(\theta_2), ..., \mathbf{a}(\theta_I)] \tag{1.1.6}$$

$$\mathbf{a}(\theta) = [1, e^{j\theta}, ..., e^{j(M-1)\theta}]^T. \tag{1.1.7}$$

In other words, the noisy data

$$y_m(n) = \sum_{i=1}^{I} s_i(n) e^{j(n-1)\theta_i} + w_m(n) \stackrel{\triangle}{=} x_m(n) + w_m(n), \tag{1.1.8}$$
$$m = 0, 1, ..., M-1, \ n = 1, ..., N.$$

This data model is referred to as superimposed exponentials. Here, we call θ_i the frequency of the i-th exponential (or signal), $s_i(n)$ the corresponding amplitude (e.g., envelop of an incoming wave), and I the number of exponentials. $x_m(n)$ is the noise-free component of the data. In the context of sensor arrays, we may refer to n as the temporal dimension and m as the spatial dimension. But the actual meanings of n and m are not important in this chapter. It is the model (1.1.8) that many of the high-resolution methods are designed for. Some of the methods only need a single measurement of the data (1.1.8).

Signal amplitudes can be stochastic or deterministic. In the stochastic case, the signal amplitudes are assumed to be stationary circular Gaussian with zero mean and covariance matrix

$$E\{\mathbf{s}(n)\mathbf{s}^H(l)\} = \mathbf{R}_s \delta_{n,l}, \quad E\{\mathbf{s}(n)\mathbf{s}^T(l)\} = \mathbf{0}, \quad \forall\, n, l \tag{1.1.9}$$

where E stands for the statistical expectation, $\mathbf{R}_s$ is called the signal covariance matrix which is unknown, $\delta_{k,l}$ is the Kronecker delta with $\delta_{n,l} = 1$ for $n = l$, and $\delta_{n,l} = 0$ for $n \neq l$. The first expression of (1.1.9) implies the independence of signal amplitudes at distinct instants, which is called temporally independent. The second one is a consequence of the assumption that the real and imaginary parts of the signal amplitudes are identical (but not necessarily independent) joint Gaussian and satisfy the property $E\{\Re e\{\mathbf{s}(n)\}(\Im m\{\mathbf{s}(n)\})^T\} = -E\{\Im m\{\mathbf{s}(n)\}(\Re e\{\mathbf{s}(n)\})^T\}$. Unless specified otherwise, $\mathbf{R}_s$ will be assumed to be of full rank, in which case we say that the signals are incoherent. If $\mathbf{R}_s$ does not have a full rank, we say that the signals are coherent.

In the deterministic case, the signal amplitudes are also assumed to have the following property

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{s}(n)\mathbf{s}^H(n) = \mathbf{R}_s. \tag{1.1.10}$$

This assumption is merely needed to establish the optimal performance of the maximum likelihood methods in this case.

The statistical property of the noise also has a major impact on algorithm development and analysis. We will assume that the noise is stationary circular Gaussian with zero mean and the covariance matrix

$$E\{\mathbf{w}(n)\mathbf{w}^H(l)\} = \mathbf{R}_n\delta_{n,l}, \quad E\{\mathbf{w}(n)\mathbf{w}^T(l)\} = \mathbf{0}, \quad \forall\, n, l \tag{1.1.11}$$

where $\mathbf{R}_n$ can be an unknown non-singular matrix. Unless otherwise specified, the noise is spatially white (independent) ($\mathbf{R}_n = \sigma^2\mathbf{I}_M$).

The noise in (1.1.11) is also temporally independent. Temporal independence is a fundamental assumption used for signals and noise in all chapters but Chapter 6. A temporally dependent noise model will be considered therein.

The signal and noise are assumed to be independent of each other, i.e.,

$$E\{\mathbf{w}(n)\mathbf{s}^H(l)\} = \mathbf{0}, \quad E\{\mathbf{w}(n)\mathbf{s}^T(l)\} = \mathbf{0}, \quad \forall\, n, l. \tag{1.1.12}$$

1.2 Estimation Techniques Using Algebraic Principles

In this section, we discuss several frequency estimation techniques using algebraic principles. These techniques exploit the algebraic structure within each measurement. For convenience of presentation, we will focus on a single measurement of data. For multiple measurements, the same principles can be applied to a matrix stacked with multiple measurements.

Some of the key structural information within each measurement of data can be exploited via the following data matrix:

$$\mathbf{X} = \begin{bmatrix} x_0 & x_1 & \cdots & x_L \\ x_1 & x_2 & \cdots & x_{L+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M-L-1} & x_{M-L-2} & \cdots & x_{M-1} \end{bmatrix} \tag{1.2.1}$$

where L is the length of a sliding window used to divide the original full length data into overlapped segments, and we have dropped the index n. According to (1.1.8), we can write that,

$$\mathbf{X} = \mathbf{Z}_{M-L}\mathbf{S}\mathbf{Z}_{L+1}^T \tag{1.2.2}$$

where

$$\mathbf{Z}_m = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_I \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{m-1} & z_2^{m-1} & \cdots & z_I^{m-1} \end{bmatrix} \tag{1.2.3}$$

$$\mathbf{S} = diag[s_1, s_2, ..., s_I] \tag{1.2.4}$$

with $z_i = e^{j\theta_i}$. z_i is called the i-th signal pole. It is easy to show from the above decomposition of the matrix $\mathbf{X}$ that this matrix has a constant rank equal to the number of signals provided that the number of rows and the number of columns are larger than or equal to the number of signals while the original (single) measurement vector $\mathbf{x} = [x_0, ..., x_{M-1}]^T$ only has a rank equal to 1. The same idea will also be used in Subsections 1.3.4, 1.4.1, and Section 4.3.

1.2.1 Linear Prediction

This technique models the data sequence by a set of linear prediction equations. The linear prediction coefficients are used to retrieve the signal poles.

a. Linear Prediction Equation

Let the parameters $(b_1, b_2, \cdots, b_L)$ be the coefficients of the the following polynomial [1]

$$P(z) = 1 + \sum_{i=1}^{L} b_i z^i \tag{1.2.5}$$

where $L \geq I$. According to the theory of polynomials, there exist such coefficients $(b_1, b_2, \cdots, b_L)$ that the signal poles $z_1, z_2, \cdots, z_I$ are I roots of the above L-degree polynomial. It is then easy to verify that

$$x_l + \sum_{i=1}^{L} b_i x_{l+i} = 0. \tag{1.2.6}$$

Let

$$\mathbf{b} = \begin{bmatrix} 1 & b_1 & \cdots & b_L \end{bmatrix}^T. \tag{1.2.7}$$

Then the matrix form of (1.2.6) is

$$\mathbf{X}\mathbf{b} = \mathbf{0}. \tag{1.2.8}$$

The equation (1.2.8) is known as the linear prediction equation.
Case (1): $L = I$: Provided that $m \geq I$, $\mathbf{Z}_m$ has the full column rank I if and only if the poles $\{z_i, i = 1, ..., I\}$ are distinct of each other. Considering (1.2.2), we see that if $M - I \geq I$ (i.e., $M \geq 2I$), then $\mathbf{X}$ has rank I. Furthermore, we see that $\mathbf{Z}_{I+1}^T$ has a one-dimensional (right) null space, and hence so does $\mathbf{X}$. Therefore, (1.2.8) has a unique solution for $\mathbf{b}$ up to a scalar.

[1] Note that in (1.2.5), b_0 is intentionally chosen to be 1. This choice enforces no zero roots for the polynomial (1.2.5). Another equivalent choice is $b_L = 1$. Although one is allowed to normalize any coefficient to be 1, numerical ill-conditioning may occur if the coefficient in consideration is small or even zero.

Case (2): $L > I$: It is easy to verify that provided $M - L \geq I$, $\mathbf{X}$ has an $(L - I + 1)$-dimensional right null space. This is because $\mathbf{Z}_{L+1}^T$ has an $(L-I+1)$-dimensional right null space, and $\mathbf{Z}_{M-L}$ has the full column rank given $M - L \geq I$. In other words, all solutions of (1.2.8) are the solutions of

$$\mathbf{Z}_{L+1}^T \mathbf{b} = \mathbf{0} \tag{1.2.9}$$

and vice versa. From (1.2.9), we know that the I poles $\{z_i, i = 1, 2, ..., I\}$ are among the L roots of the polynomial (1.2.5). Therefore, any solution of (1.2.8) has this desired property. But there are also $L-I$ additional roots associated with any solution of (1.2.8). These additional roots are called the extraneous roots. Next, we show how to select a solution of (1.2.8) such that the extraneous roots are separable from the desired roots.

b. Unique Determination of Signal Poles

Let us rewrite (1.2.5) as

$$P(z) = P_{1,I}(z) P_{2,L-I}(z) \tag{1.2.10}$$

where $P_{1,I}(z)$ and $P_{2,L-I}(z)$ are two polynomials of degrees I and $L - I$ respectively. Without loss of generality, we assume that the I poles $\{z_i, i = 1, 2, ..., I\}$ are the I roots of $P_{1,I}(z)$, and the other extraneous roots of $P_L(z)$ are the $L - I$ roots of $P_{2,L-I}(z)$. Among all the solutions of (1.2.8), there is the minimum-norm solution, i.e., $\|\mathbf{b}\|$ is minimized subject to that the first element of $\mathbf{b}$ is one. Let $P_{\mathrm{MN}}(z)$ be the polynomial associated with this minimum-norm solution. Then, we have

$$P_{\mathrm{MN}}(z) = P_{1,I}(z) P_{2,L-I,\mathrm{MN}}(z) \tag{1.2.11}$$

where the $L - I$ roots of $P_{2,L-I,\mathrm{MN}}(z)$ have a special property. Experiments suggest that the roots of $P_{2,L-I,\mathrm{MN}}(z)$ are all outside the unit circle in the complex plane. This property is also supported by the following analysis. Let a root of $P_{2,L-I,\mathrm{MN}}(z)$ be denoted by $z_0 = e^{\alpha_0 + j\omega_0}$. Then, $P_{\mathrm{MN}}(z) = P_{1,0}(z) P_{2,0}(z)$ where $P_{2,0}(z) = 1 + b_{2,0} z$ and

$$1 + b_{2,0} z_0 = 0. \tag{1.2.12}$$

Based on the well known Parseval's Theorem, we have:

$$\|\mathbf{b}\|^2 = \frac{1}{2\pi} \int_0^{2\pi} |P_{1,0}(e^{j\omega})|^2 |P_{2,0}(e^{j\omega})|^2 d\omega. \tag{1.2.13}$$

We now consider a small perturbation of the root z_0 moving away from the unit circle (along the radial direction). Then, (1.2.12) implies

$$\Delta b_{2,0} z_0 + b_{2,0} \Delta z_0 = 0.$$

But $\Delta z_0 = z_0 \Delta\alpha_0$. Therefore

$$\Delta b_{2,0} = -b_{2,0}\Delta\alpha_0. \tag{1.2.14}$$

Then, provided that $\mathbf{b}$ is a solution of (1.2.8), the perturbation of the norm-square of $\mathbf{b}$ due to the perturbation of the root z_0 away from the unit circle is given below:

$$\begin{aligned}\Delta\|\mathbf{b}\|^2 &= \frac{1}{2\pi}\int_0^{2\pi} |P_{1,0}(e^{j\omega})|^2 \Delta|P_{2,0}(e^{j\omega})|^2 d\omega \\ &= \frac{1}{2\pi}\int_0^{2\pi} |P_{1,0}(e^{j\omega})|^2 2\Re e\{P_{2,0}^*(e^{j\omega})\Delta P_{2,0}(e^{j\omega})\} d\omega \\ &= \frac{1}{2\pi}\int_0^{2\pi} |P_{1,0}(e^{j\omega})|^2 2\Re e\{(1+b_{2,0}^* e^{-j\omega})\Delta b_{2,0}(e^{j\omega})\} d\omega. \end{aligned} \tag{1.2.15}$$

Using (1.2.14) in (1.2.15) yields

$$\begin{aligned}\Delta\|\mathbf{b}\|^2 &= -\Delta\alpha_0 \frac{1}{2\pi}\int_0^{2\pi} |P_{1,0}(e^{j\omega})|^2 2\Re e\{(1+b_{2,0}^* e^{-j\omega}) b_{2,0}(e^{j\omega})\} d\omega \\ &= -\Delta\alpha_0 \frac{1}{2\pi}\int_0^{2\pi} |P_{1,0}(e^{j\omega})|^2 2(1-cos(\omega-\omega_0)) d\omega. \end{aligned} \tag{1.2.16}$$

Note that we used $b_{2,0} = -e^{-j\omega_0}$ in (1.2.16). It is now clear that as the root moves away from the unit circle, the norm of $\mathbf{b}$ decreases. By examining (1.2.13), we see that the extraneous root $z_0 = e^{\alpha_0 + j\omega_0}$ should also be close to the unit circle to keep the norm of $\mathbf{b}$ small. Furthermore, it is not hard to see that if there are multiple extraneous roots in $P(z)$, these roots should be evenly distributed near the unit circle to keep the norm of $\mathbf{b}$ small. Therefore, it is a valid conjecture that all the extraneous roots of the polynomial $P(z)$ constructed from the minimum-norm solution $\mathbf{b}$ of (1.2.8) (where the first element of $\mathbf{b}$ is equal to one) are evenly distributed near but outside the unit circle. Although supported by simulation and analysis, this conjecture is however not yet proven in a strict mathematical sense. This property and the above analysis were originally shown in [55].

With the above property of the extraneous roots, we can easily separate the desired roots from the extraneous roots if all the desired roots are inside or on the unit circle. Note that if the data sequence x_n is an impulse response of a stable system, all the desired poles (i.e., the desired roots of $P(z)$) are inside the unit circle.

We now rewrite (1.2.5) as

$$P^{'}(z) = z^{-L}\sum_{i=0}^{L} b_i z^i = \sum_{i=0}^{L} b_{L-i} z^{-i}. \tag{1.2.17}$$

We can do a similar analysis of this polynomial under the constraint that the last element of $\mathbf{b}$ is one. The conclusion is that the extraneous roots of $P^{'}(z)$ should be evenly distributed near but inside the unit circle. Note that the above property implies

that the minimum-norm $\mathbf{b}$ with $b_0 = 1$ is not a scaled version of the minimum-norm $\mathbf{b}$ with $b_L = 1$. Furthermore, if the desired poles can be anywhere in the complex plane, one can use both $P^{'}(z)$ and $P(z)$ to separate the desired roots from the extraneous roots [80] .

c. Estimation of Linear Prediction Coefficients

When the data are noisy, the ideal minimum-norm vector can only be estimated. Let $\mathbf{Y}$ be defined in the same way as $\mathbf{X}$ in (1.2.1) using the noisy data $y_m, m = 0, 1, ..., M-1$. For the case where $b_0 = 1$, we define

$$\mathbf{Y} = \left[\begin{array}{cc} \mathbf{y}_0 & \mathbf{Y}_0 \end{array} \right] \tag{1.2.18}$$

$$\mathbf{b} = \left[\begin{array}{c} 1 \\ \mathbf{b}_0 \end{array} \right] \tag{1.2.19}$$

where $\mathbf{y}_0$ is the first column of $\mathbf{Y}$ and $\mathbf{b}_0$ consists of the last L elements of $\mathbf{b}$. Note that (1.2.8) is no longer satisfied when $\mathbf{X}$ is replaced by $\mathbf{Y}$. In this case, $\mathbf{b}_0$ can be determined from the minimization problem

$$\min_{\mathbf{b}_0} \|\mathbf{Y}_0\mathbf{b}_0 + \mathbf{y}_0\|^2 \tag{1.2.20}$$

where $\|\cdot\|$ denotes the 2-norm. According to [25], [26], this problem has two choices of minimum-norm solutions for $\mathbf{b}_0$. The two solutions are called the least squares (LS) solution and the total least squares (TLS) solution. The determination of the two solutions is described in the following.

Since $\mathbf{Y}_0$ has a rank I without noise, we can replace it by a rank-I approximation. According to Theorem 6.7 of [94] (originally due to Eckart and Young in 1936), the most accurate approximation can be obtained via SVD. This approximation is also called the rank-I truncation of $\mathbf{Y}_0$.

The rank-I truncation of a full rank noisy matrix $\mathbf{Y}$, is defined as the truncated singular value decomposition (SVD) [2] of $\mathbf{Y}$, given by

$$[\hat{\mathbf{Y}}]_T = \sum_{i=1}^{I} \sigma_i \mathbf{u}_i \mathbf{v}_i^H \tag{1.2.21}$$

where the (full) singular value decomposition (SVD) of $\hat{\mathbf{Y}}$ is

$$\hat{\mathbf{Y}} = \sum_{i=1}^{K} \sigma_i \mathbf{u}_i \mathbf{v}_i^H, \quad \sigma_1 \geq \sigma_1 \geq \cdots \geq \sigma_K. \tag{1.2.22}$$

[2] Since only the I principal singular vectors/values are required, the rank-I truncation can be more efficiently computed using fast techniques suggested in [109], [120]. The rank-I truncation can also be obtained even without SVD [35].

With the rank-I truncation of $\mathbf{Y}_0$, the LS minimum-norm solution of (1.2.20) for $\mathbf{b}_0$ is given by

$$(\mathbf{b}_0)_{\mathrm{MN-LS}} = -([\mathbf{Y}_0]_T)^\dagger \mathbf{y}_0 \tag{1.2.23}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudo inverse of a matrix. Given the SVD, it can be determined simply by inverting the non-zero singular values of the matrix, e.g., $([\mathbf{Y}_0]_T)^\dagger = \sum_{i=1}^{I} \sigma_i^{-1} \mathbf{u}_i \mathbf{v}_i^H$.

We now denote the rank-I truncation of $\mathbf{Y}$: $[\mathbf{Y}]_T = [\mathbf{h}_0, \mathbf{Q}]$ where $\mathbf{Q}$ is also a rank-I matrix. Then the TLS minimum-norm solution of (1.2.20) for $\mathbf{b}_0$ is given by

$$(\mathbf{b}_0)_{\mathrm{MN-TLS}} = -(\mathbf{Q})^\dagger \mathbf{h}_0. \tag{1.2.24}$$

Note that the TLS solution requires two SVDs (one for rank-I truncation and one for pseudo inverse) while the LS solution requires only one SVD (where one SVD is used for both rank-I truncation and pseudo inverse). The solution (1.2.23) was proposed in [108], [57], [55] as the SVD-Prony approach, and the solution (1.2.24) in [81], [55] as the TLS-LP approach [3]. The SVD-Prony and the TLS-LP have been shown in [40] to yield the identical solution to the first-order approximation, i.e., both approaches have the same performance at high SNR. At a medium SNR (or in the threshold region of SNR), the TLS-LP performs better than the SVD-Prony, because the TLS-LP entails an additional noise cleaning procedure on $\mathbf{y}_0$. An improvement on TLS-LP for known noise covariance matrix was shown in [41]. When $L = I$, the SVD-Prony coincides with the Prony approach [29] and the TLS-LP with the Pisarenko approach [79]. The choice of $L > I$ yields a better accuracy.

For the case where $b_L = 1$, we define

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 & \mathbf{y}_1 \end{bmatrix} \tag{1.2.25}$$

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ 1 \end{bmatrix} \tag{1.2.26}$$

where $\mathbf{y}_1$ is the last column of $\mathbf{Y}$ and $\mathbf{b}_1$ consists of the first L-element of $\mathbf{b}$. Similar to the development for the case of $b_0 = 1$, the LS minimum-norm solution for $\mathbf{b}_1$ is given by

$$(\mathbf{b}_1)_{\mathrm{MN-LS}} = -([\mathbf{Y}_1]_T)^\dagger \mathbf{y}_1. \tag{1.2.27}$$

Denote $[\mathbf{Y}]_T = [\mathbf{Q}_1, \mathbf{h}_1]$ where $\mathbf{Q}_1$ is also a rank-I matrix. The TLS minimum-norm solution for $\mathbf{b}_1$ can be given by

$$(\mathbf{b}_1)_{\mathrm{MN-TLS}} = -(\mathbf{Q}_1)^\dagger \mathbf{h}_1 \tag{1.2.28}$$

[3] The approach in [55] was originally called the improved Pisarenko approach which turned out to have the same formulation as the approach called TLS-LP in [81]. The equivalence was proven in [20], [4]. The minimum-norm approach in [58] was originally proposed for multiple measurement case. An extension of the minimum-norm approach to single measurement case was considered in [20] and shown to have the identical solution as the TLS-LP approach.

Note that in noise-free case,

$$\mathbf{y}_0 = -\mathbf{Y}_0\mathbf{b}_0 \tag{1.2.29}$$

$$\mathbf{y}_1 = -\mathbf{Y}_1\mathbf{b}_1. \tag{1.2.30}$$

Because of the equations (1.2.30) and (1.2.29), $\mathbf{b}_1$ and $\mathbf{b}_0$ can be called the forward-prediction vector and backward-prediction vector, respectively. We know that the desired I roots of $P(z)$ associated with $\mathbf{b}_0$ are the same as the desired I roots of $P^{'}(z)$ associated with $\mathbf{b}_1$. But we also know that extraneous roots of $\mathbf{b}_0$ are outside the unit circle, and the extraneous roots of $\mathbf{b}_1$ are inside the unit circle. Therefore, the desired I roots can be detected by the common roots of $\mathbf{b}_0$ and $\mathbf{b}_1$. These desired roots can be anywhere in the complex plane. This idea was first reported in [80].

d. Forward-backward Averaging

If all the poles are on the unit circle, the data matrix $\mathbf{X}$ has the following property:

$$\mathbf{X}^*\mathbf{\Pi}_{L+1} = \mathbf{Z}^*_{M-L}\mathbf{S}^*(\mathbf{Z}^*_{L+1})^T\mathbf{\Pi}_{L+1} = \mathbf{Z}^*_{M-L}\mathbf{S}^*\mathbf{\Theta}^{-L}\mathbf{Z}^T_{L+1} \tag{1.2.31}$$

where the superscript * denotes complex conjugate,

$$\mathbf{\Theta} = diag[e^{j\theta_1}, ..., e^{j\theta_I}], \tag{1.2.32}$$

and $\mathbf{\Pi}_{L+1}$ is the $(L+1)\times(L+1)$ reverse permutation (exchange) matrix.

An $n \times n$ reverse permutation matrix is defined as

$$\mathbf{\Pi}_n = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix}_{n\times n}. \tag{1.2.33}$$

It is easy to verify that the following equation has the same solution as (1.2.9) and hence (1.2.8):

$$\mathbf{X}^*\mathbf{\Pi}_{L+1}\mathbf{b} = \mathbf{0} \tag{1.2.34}$$

Alternatively, we can combine (1.2.8) and (1.2.34) into

$$\mathbf{X}_{fb}\mathbf{b} = \mathbf{0} \tag{1.2.35}$$

where

$$\mathbf{X}_{fb} = \begin{bmatrix} \mathbf{X} \\ \mathbf{X}^*\mathbf{\Pi}_{L+1} \end{bmatrix}. \tag{1.2.36}$$

With $\mathbf{X}_{fb}$ (1.2.35) replaced by $\mathbf{Y}_{fb}$, one can compute the minimum-norm solution with $b_0 = 1$ or $b_1 = 1$. The desired roots of the polynomial $P(z)$ are all on the unit circle,

and the extraneous roots are all outside the unit circle (if $b_0 = 1$) or all inside the unit circle (if $b_L = 1$). Since (1.2.35) contains more independent equations than (1.2.8) or equivalently (1.2.34), the solution of (1.2.35) in the noisy case is less sensitive to noise. This combination of singular-value-decomposition and forward-backward linear prediction was first proposed in [108] and analyzed in detail in [37].

1.2.2 Matrix Pencil

A matrix pencil is defined as the matrix $\mathbf{Y}_2 - z\mathbf{Y}_1$ for an arbitrary matrix pair $[\mathbf{Y}_2, \mathbf{Y}_1]$ of the same dimensions. The techniques to be discussed in this subsection yield the signal poles $z_i, i = 1, 2, ..., I$, as the rank reducing numbers of a matrix pencil constructed from data directly or indirectly. The accuracy of the matrix pencil methods is about the same as the SVD linear prediction based techniques. But the computational complexity of the matrix pencil methods is generally lower than the SVD linear prediction methods.

a. Underlying Principle

The relationship (1.2.8) can be exploited in another way without the extraneous roots that are encountered in linear prediction. The idea here is to exploit the matrix pencil inherent in the matrix $\mathbf{X}$. Let $\mathbf{X}_0$ and $\mathbf{X}_1$ be the noise-free versions of $\mathbf{Y}_0$ in (1.2.18) and $\mathbf{Y}_1$ in (1.2.25). It follows that

$$\mathbf{X}_0 - z\mathbf{X}_1 = \mathbf{Z}_{M-L}\mathbf{S}(\mathbf{\Theta} - z\mathbf{I}_I)\mathbf{Z}_L^T \tag{1.2.37}$$

where $\mathbf{\Theta}$ is defined in (1.2.32), and refer to (1.2.3) and (1.2.4) for the definitions of the other matrices in the above equation. From (1.2.37), it is clear that if $\text{rank}(\mathbf{Z}_{M-L}) = \text{rank}(\mathbf{Z}_L) = I$, then each element of the poles $\{z_i, i = 1, ..., I\}$ is a rank-reducing number of the matrix pencil $\mathbf{X}_0 - z\mathbf{X}_1$, i.e.,

$$\text{rank}(\mathbf{X}_0 - z\mathbf{X}_1) = 1 + \text{rank}(\mathbf{X}_0 - z_i\mathbf{X}_1) \tag{1.2.38}$$

where $z \notin \{z_i, i = 1, 2, ..., I\}$. Note that $\text{rank}(\mathbf{Z}_l) = I$ if and only if $l \geq I$.

If all the poles are on the unit circle, an FB (forward-backward) matrix pencil can be formed based on $\mathbf{X}_{fb}$ (1.2.36). We define

$$\mathbf{X}_{fb} = \begin{bmatrix} \mathbf{x}_{0,fb} & \mathbf{X}_{0,fb} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{1,fb} & \mathbf{x}_{1,fb} \end{bmatrix} \tag{1.2.39}$$

where $\mathbf{X}_{0,fb}$ consists of the last L columns of $\mathbf{X}_{fb}$, and $\mathbf{X}_{1,fb}$ consists of the first L columns of $\mathbf{X}_{fb}$. Then

$$\mathbf{X}_{0,fb} - z\mathbf{X}_{1,fb} = \begin{bmatrix} \mathbf{Z}_{M-L}\mathbf{S} \\ \mathbf{Z}_{M-L}^*\mathbf{S}^*\mathbf{\Theta}^{-L} \end{bmatrix} (\mathbf{\Theta} - z\mathbf{I}_I)\mathbf{Z}_L^T \tag{1.2.40}$$

and

$$\text{rank}(\mathbf{X}_{0,fb} - z\mathbf{X}_{1,fb}) = 1 + \text{rank}(\mathbf{X}_{0,fb} - z_i\mathbf{X}_{1,fb}) \tag{1.2.41}$$

where $z \notin \{z_i, i = 1, 2, , ..., I\}$.

Since $\mathbf{X}_0$ and $\mathbf{X}_1$ are of rank I and range$(\mathbf{X}_0)$ = range$(\mathbf{X}_1)$, one can let $\mathbf{X}_0 = \mathbf{Q}\mathbf{R}_0^H$ and $\mathbf{X}_1 = \mathbf{Q}\mathbf{R}_1^H$ where $\mathbf{Q}$ is an $(M-L) \times I$ full column rank matrix, $\mathbf{R}_0$ and $\mathbf{R}_1$ are $L \times I$ full column rank matrices. To compute the rank reducing numbers of the matrix pencil $\mathbf{X}_0 - z\mathbf{X}_1$, the following property is useful:

$$\begin{aligned} \text{rank}(\mathbf{X}_0 - z\mathbf{X}_1) &= \text{rank}(\mathbf{R}_0^H - z\mathbf{R}_1^H) = \text{rank}((\mathbf{R}_1 - z\mathbf{R}_0) \\ &= \text{rank}((\mathbf{R}_0^H\mathbf{R}_0)^{-1}\mathbf{R}_0^H\mathbf{R}_1 - z\mathbf{I}_I). \end{aligned} \tag{1.2.42}$$

This property implies that if $\min(M-L, L) \geq I$, the poles can be obtained by the I rank reducing numbers of the matrix pencil $\mathbf{X}_0 - z\mathbf{X}_1$ or equivalently by the I eigenvalues of the $I \times I$ matrix $(\mathbf{R}_0^H\mathbf{R}_0)^{-1}\mathbf{R}_0^H\mathbf{R}_1$.

A similar procedure applies to the FB matrix pencil $\mathbf{X}_{1,fb} - z\mathbf{X}_{0,fb}$ if all the poles are on the unit circle.

Note that there is no such problem of extraneous roots in the matrix pencil method. Finding the I eigenvalues of a matrix is in general easier than finding the L roots of an L-degree polynomial when $L > I$. The results of noise analysis also favour the matrix pencil method over the linear prediction method. With a proper selection of the parameter L, the estimation variance of the matrix pencil method is close to the Cramér-Rao bound. The optimal choice of L is about one third of the data length. For details of the analysis of the matrix pencil method, see [39].

The matrix pencil method can be further generalized as follows. Let

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \cdots & \mathbf{x}_L \end{bmatrix}. \tag{1.2.43}$$

Then define

$$\mathbf{X}_{(i)} = \begin{bmatrix} \mathbf{x}_i & \mathbf{x}_{i+1} & \cdots & \mathbf{x}_{i+L-n} \end{bmatrix} \tag{1.2.44}$$

where $i = 0, 1, ..., n$ and

$$\bar{\mathbf{X}} = \sum_{i=0}^{n} c_i \mathbf{X}_{(i)}. \tag{1.2.45}$$

It can be shown that if $L - n + 1 \geq I$ and $M - L \geq I$, then

$$\text{rank}(\bar{\mathbf{X}}) = I - g \tag{1.2.46}$$

if and only if the roots of the polynomial $P_n(z) = \sum_{i=0}^{n} c_i z^i$ contain exactly g elements of $\{z_1, z_2, \cdots, z_I\}$. If we choose the coefficients c_i such that rank$(\bar{\mathbf{X}}) = 0$ (i.e., $\bar{\mathbf{X}} = 0$), then these coefficients are the same solution as to the linear prediction equation (1.2.8) with $L = n$. The two equations (1.2.45) and (1.2.46) constitute a notion called subspace linear prediction originally proposed in [38]. Subspace linear prediction unifies the notions of linear prediction and matrix pencil. But the numerical implementation of subspace linear prediction is not as straightforward as that of matrix pencil (where

$n = 1$ and $I \geq \text{rank}(\bar{\mathbf{X}}) \geq I - 1$) or that of linear prediction (where $n \geq I$ and $I \geq \text{rank}(\bar{\mathbf{X}}) \geq 0$). For matrix pencil, we have a generalized eigenvalue problem, which can be converted to a standard eigenvalue problem. For linear prediction, we invert a set of linear equations in the least squares sense and then compute the roots of a polynomial. Due to the computational reason, the subspace linear prediction method was not further investigated after the work [38].

Another advantage of the matrix pencil method over the linear prediction method is that for multidimensional frequency estimation, the required extension of the matrix pencil method is straightforward. For details of the multidimensional case, see [31], [32].

b. Practical Considerations

In the noisy case, the matrix pencil $\mathbf{X}_0 - z\mathbf{X}_1$ in (1.2.42) is replaced by

$$\mathbf{Y}_0 - z\mathbf{Y}_1, \tag{1.2.47}$$

where $\mathbf{Y}_0$ and $\mathbf{Y}_1$ are the noisy versions of $\mathbf{X}_0$ and $\mathbf{X}_1$ respectively. Note that $\mathbf{Y}_0 - z\mathbf{Y}_1$ is likely to have more than I rank-reducing numbers. To estimate the desired I rank-reducing numbers (the signal poles), there are a number of ways. One way is to replace $\mathbf{Y}_0$ and $\mathbf{Y}_1$ by their rank-I truncations $[\mathbf{Y}_0]_T$ and $[\mathbf{Y}_1]_T$ respectively. After the truncation, the matrix pencil $[\mathbf{Y}_0]_T - z[\mathbf{Y}_1]_T$ has only I rank-reducing numbers, which also equal to the signal poles in the noise-free case. The rank truncation can be done in several different ways.

We first let the rank-I truncations of $\mathbf{Y}_0$ and $\mathbf{Y}_1$ be defined via separate SVDs:

$$[\mathbf{Y}_0]_T = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^H \tag{1.2.48}$$

$$[\mathbf{Y}_1]_T = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^H. \tag{1.2.49}$$

The I rank-reducing numbers of the matrix pencil $[\mathbf{Y}_0]_T - z[\mathbf{Y}_1]_T$ can be obtained from the eigenvalues of the $I \times I$ matrix

$$\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^H\mathbf{U}_0\mathbf{\Sigma}_0. \tag{1.2.50}$$

This technique is the original matrix pencil approach [39].

We now define the rank-I truncation of $[\ \mathbf{Y}_0 \quad \mathbf{Y}_1\ ]$ via a joint SVD as follows

$$[\ \mathbf{Y}_0 \quad \mathbf{Y}_1\ ]_T = \mathbf{U}_x\mathbf{\Sigma}_x[\ \mathbf{V}_{x0}^H \quad \mathbf{V}_{x1}^H\ ]. \tag{1.2.51}$$

The matrix pencil $\mathbf{Y}_0 - z\mathbf{Y}_1$ can now be approximated by $\mathbf{U}_x\mathbf{\Sigma}_x(\mathbf{V}_{x0}^H - z\mathbf{V}_{x1}^H)$, of which the I rank-reducing numbers are the eigenvalues of the matrix

$$(\mathbf{V}_{x0}^H\mathbf{V}_{x1})(\mathbf{V}_{x1}^H\mathbf{V}_{x1})^{-1}. \tag{1.2.52}$$

This method is referred to as the LS (least squares) ESPRIT [90].

A further step of noise filtering can be carried out as shown here. Since $[\mathbf{V}_{x0} \quad \mathbf{V}_{x1}]$ has a rank I in the noise-free case, it can be replaced by the following rank-I truncation

$$[\ \mathbf{V}_{x0} \quad \mathbf{V}_{x1}\ ]_T = \mathbf{U}_v\mathbf{\Sigma}_v[\ \mathbf{V}_{v0}^H \quad \mathbf{V}_{v1}^H\ ] \tag{1.2.53}$$

where $\mathbf{V}_{v0}$ and $\mathbf{V}_{v1}$ are $I \times I$ matrices.

Note that since both $\mathbf{V}_{x0}$ and $\mathbf{V}_{x1}$ are orthogonal matrices, each of $\sqrt{2}\mathbf{V}_{v0}$ and $\sqrt{2}\mathbf{V}_{v1}$ can be shown to be unitary. This follows from a more general theorem on the property of $\mathbf{V}_{x0}$ and $\mathbf{V}_{x1}$:

Theorem 1.1: *Let $\mathbf{V}_{x0}$ and $\mathbf{V}_{x1}$ be two orthonormal matrices of I columns each. Let the following be the joint SVD of the two matrices:*

$$\begin{bmatrix} \mathbf{V}_{x0} & \mathbf{V}_{x1} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_v & \mathbf{U'}_v \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_v & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_v{}' \end{bmatrix} \begin{bmatrix} \mathbf{V}_{v0} & \mathbf{V}'_{v0} \\ \mathbf{V}_{v1} & \mathbf{V}'_{v1} \end{bmatrix}^H \tag{1.2.54}$$

where $\mathbf{\Sigma}_v > \mathbf{\Sigma}_v{}'$ ($\mathbf{A} > \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is positive-definite). Then, we have

$$\begin{bmatrix} \mathbf{V}_{v0} & \mathbf{V}'_{v0} \\ \mathbf{V}_{v1} & \mathbf{V}'_{v1} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{G}_0\mathbf{H}_0{}^H & \mathbf{G}_0\mathbf{H}_1{}^H \\ \mathbf{G}_1\mathbf{H}_0{}^H & -\mathbf{G}_1\mathbf{H}_1{}^H \end{bmatrix} \tag{1.2.55}$$

where $\mathbf{G}_0$, $\mathbf{H}_0$, $\mathbf{G}_1$ and $\mathbf{H}_1$ are all unitary matrices.

Proof: See proof of (2.11) in [42].

With (1.2.53), the matrix pencil $\mathbf{Y}_0 - z\mathbf{Y}_1$ is approximated by $\mathbf{U}_x\mathbf{\Sigma}_x(\mathbf{V}_{v0} - z\mathbf{V}_{v1})\mathbf{\Sigma}_v\mathbf{U}_v^H$, which has the same I rank-reducing numbers as those of the matrix pencil

$$\mathbf{V}_{v0} - z\mathbf{V}_{v1}. \tag{1.2.56}$$

These rank-reducing numbers are equal to the eigenvalues of the matrix

$$\mathbf{V}_{v0}\mathbf{V}_{v1}^{-1} \tag{1.2.57}$$

or, equivalently, the eigenvalues of the matrix $\mathbf{V}_{v1}^{-1}\mathbf{V}_{v0}$. The eigenvalues of (1.2.57) are also the same as the eigenvalues of the following matrix

$$-(\mathbf{V}'^H_{v0})^{-1}\mathbf{V}'^H_{v1} \tag{1.2.58}$$

where the range of $[\mathbf{V}'^T_{v1}, \mathbf{V}'^T_{v0}]^T$ is the orthogonal complement of the range of $[\mathbf{V}^T_{v1}, \mathbf{V}^T_{v0}]^T$. They are all by-products of the SVD used in (1.2.53). Note that

$$\mathbf{V}'^H_{v1}\mathbf{V}_{v1} + \mathbf{V}'^H_{v0}\mathbf{V}_{v0} = \mathbf{0} \rightarrow (\mathbf{V}'^H_{v0})^{-1}\mathbf{V}'^H_{v1} + \mathbf{V}_{v0}\mathbf{V}_{v1}^{-1} = \mathbf{0}. \tag{1.2.59}$$

The method based on (1.2.57) is known as the TLS (total least squares) ESPRIT [90].

Another way to generate the joint rank-I truncations of $\mathbf{Y}_0$ and $\mathbf{Y}_1$ follows the state space method in [61]. Let the rank-I truncation of $\mathbf{Y}$ be

$$[\mathbf{Y}]_T \ = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_s^H \tag{1.2.60}$$

and define two submatrices of $\mathbf{V}_s$

$$\mathbf{V}_{s1} = \text{the first } (L-1) \text{ rows of } \mathbf{V}_s \tag{1.2.61}$$

$$\mathbf{V}_{s0} = \text{the last } (L-1) \text{ rows of } \mathbf{V}_s. \tag{1.2.62}$$

Then $[\mathbf{Y}_0]_T = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_{s0}^H$ and $[\mathbf{Y}_1]_T = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_{s1}^H$. Hence, the signal poles can be computed from the rank-reducing numbers of the matrix pencil

$$\mathbf{V}_{s1}^H - z\mathbf{V}_{s0}^H, \tag{1.2.63}$$

or equivalently, from the eigenvalues of the following square matrix

$$\mathbf{V}_{s1}^H\mathbf{V}_{s0}(\mathbf{V}_{s0}^H\mathbf{V}_{s0})^{-1}. \tag{1.2.64}$$

In the noise-free case, the I eigenvalues are the same as the signal poles (provided, of course, $\min(M-L, L-1) \geq I$).

Similar to the TLS-ESPRIT, we can introduce an additional step of filtering for the state-space method, i.e., let

$$[\mathbf{V}_{s1} \quad \mathbf{V}_{s0}]_T = \mathbf{U}_{sv}\mathbf{\Sigma}_{sv}[\mathbf{V}_{sv1}^H \quad \mathbf{V}_{sv0}^H]. \tag{1.2.65}$$

Then $\mathbf{V}_{s1}^H - z\mathbf{V}_{s0}^H$ can be approximated by $(\mathbf{V}_{sv1} - z\mathbf{V}_{sv0})\mathbf{\Sigma}_{sv}\mathbf{U}_{sv}^H$, and the I rank-reducing numbers are the eigenvalues of

$$\mathbf{V}_{sv1}\mathbf{V}_{sv0}^{-1}. \tag{1.2.66}$$

The solution of (1.2.66) was proposed in [42] as the TLS state space method.

It is shown in [42] that the estimates by the above variations of the matrix pencil approach have the same variance at high SNR. Around a medium or threshold SNR, the TLS estimates are more accurate than the LS estimates.

For the case of $I = 1$, it is shown in [39] that at high SNR,

$$\min_L var_{\text{MP}}(\theta_1) = \min_L var_{\text{LP}}(\theta_1) = \frac{27}{4M^3\text{SNR}} \tag{1.2.67}$$

where $var_{\text{MP}}(\cdot)$ is the estimation variance for the matrix pencil and $var_{\text{LP}}(\cdot)$ for the linear prediction. The corresponding CRB (under white Gaussian noise model) is known to be

$$\text{CRB} = \frac{6}{M(M^2-1)\text{SNR}} \tag{1.2.68}$$

Then, the optimal statistical efficiency of both linear prediction and matrix pencil is

$$\text{efficiency} = \frac{\text{Variance}}{\text{CRB}} = \frac{27M(M^2-1)}{24M^3} \approx 1.125. \tag{1.2.69}$$

It is this near optimality and the computational efficiency that make the linear prediction method and the matrix pencil method practical choices. Between the two near-optimal methods, the matrix pencil method is simpler in computation.

It should be noted here that the matrix pencil method first shown in [36] has its roots coming from many sources. One of them is the pencil-of-functions method originally developed in [48], [49]. The pencil-of-functions method deals with continuous signals while the matrix pencil method is designed for discrete-time signals. The development of the matrix pencil method coincided with and was also influenced by the development of ESPRIT first shown in [73].

1.2.3 Iterative Quadratic Maximum Likelihood

For an optimal accuracy of estimation, the maximum likelihood (ML) method is the classical choice. But the computational burden of the ML method can be prohibitive. Both the matrix pencil methods and the linear prediction methods can achieve a near optimal accuracy at a moderate computational cost. To compete with the sub-optimal methods in terms of computations, the maximum likelihood method needs to be implemented with some clever approximations. One of these clever approximations of the ML method is known as the iterative quadratic maximum likelihood (IQML) method. In this section, we show the principle of the IQML method.

a. Least Squares versus Maximum Likelihood

For white Gaussian noise ($\mathbf{R}_n = \sigma^2 \mathbf{I}_M$) and deterministic signal amplitudes with a single measurement, maximizing the likelihood function is equivalent to minimizing the following least squares (LS) function

$$\|\mathbf{y} - \mathbf{A}\mathbf{s}\|^2 \tag{1.2.70}$$

where $\|\cdot\|$ denotes the 2-norm and the measurement index n has been dropped for notational convenience. In the above formulation, the data is fit onto the given model in the LS sense. The method based on (1.2.70) is known as the LS method. Hence, for white Gaussian noise, the ML method is equivalent to the LS method. The ML method is known to have the least estimation variance asymptotically. By "asymptotically", we mean that either SNR or the data length approaches to infinity.

The main drawback of the ML method is its computation. It is often the most expensive in computation compared to other alternatives. For the data model shown in (1.1.8), it is possible to reduce the complexity to some extent. The matrix $\mathbf{A}$ in (1.2.70) is known as the Vandermonde matrix, which has independent columns if and only if the poles are distinct (and $I \leq M$). If the poles are given, i.e., the matrix $\mathbf{A}$ is given, then the ML estimate of the amplitude vector $\mathbf{s}$ is given by

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \|\mathbf{y} - \mathbf{A}\mathbf{s}\|^2 = \mathbf{A}^{\dagger}\mathbf{y} = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\mathbf{y} \tag{1.2.71}$$

where $\mathbf{A}^{\dagger} = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$ denotes the pseudo-inverse of $\mathbf{A}$, and $\mathbf{A}^H$ denotes the conjugate transpose of $\mathbf{A}$. Note that given any estimates of the poles, the LS estimates of the amplitudes can be obtained in one step by using (1.2.71). To find the ML

estimates of the poles, we substitute (1.2.71) into (1.2.70) to yield

$$\min_{\{\theta_i\}} \|\mathbf{y} - \mathbf{A}\mathbf{A}^{\dagger}\mathbf{y}\|^2 = \min_{\{\theta_i\}} \|\mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{y}\|^2 \tag{1.2.72}$$

where $\mathbf{P}_{\mathbf{A}}^{\perp} = \mathbf{I}_M - \mathbf{A}\mathbf{A}^{\dagger}$ is the projection matrix onto a space which is orthogonal to the space spanned by the columns of $\mathbf{A}$. This matrix is called the orthogonal projection matrix. Given a matrix $\mathbf{X}$ with full column rank, the orthogonal projection matrix can be defined as

$$\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{I} - \mathbf{X}(\mathbf{X}^H\mathbf{X})^{-1}\mathbf{X}^H. \tag{1.2.73}$$

The minimization problem (1.2.72) involves a smaller set of unknowns than the original one (1.2.70). Unfortunately, (1.2.72) is still generally a hard computational problem as the cost function is a highly nonlinear function of the unknowns. Only in some special cases, the computation of (1.2.72) can be reduced significantly. For example, if $\frac{1}{M}\mathbf{A}^H\mathbf{A}$ is constant, then (1.2.72) becomes a quadratic function of the poles.

b. IQML

The minimization of (1.2.72) can be approximated in another fashion. Define an $(M-I)\times M$ matrix as follows:

$$\mathbf{B}^H = \begin{bmatrix} b_0 & b_1 & \cdots & b_I & & \\ & b_0 & b_1 & \cdots & b_I & \\ & & \ddots & \ddots & \ddots & \ddots \\ & & & b_0 & b_1 & \cdots & b_I \end{bmatrix}_{(M-I)\times M} \tag{1.2.74}$$

where the parameters $(b_0, b_1, b_2, \cdots, b_I)$ are related to the poles z_i via the following polynomial equation:

$$P(z) = b_0 + \sum_{i=1}^{I} b_i z^i = 0 \quad \text{if and only if} \quad z \in (z_1, z_2, \cdots, z_I). \tag{1.2.75}$$

It follows that

$$\text{range}(\mathbf{B}) = M - I \tag{1.2.76}$$

$$\mathbf{B}^H\mathbf{A} = \mathbf{0}. \tag{1.2.77}$$

These two equations imply that range($\mathbf{B}$) is the orthogonal complement of range($\mathbf{A}$), and hence

$$\mathbf{P}_{\mathbf{B}} = \mathbf{P}_{\mathbf{A}}^{\perp} \tag{1.2.78}$$

where

$$\mathbf{P}_{\mathbf{B}} = \mathbf{B}\mathbf{B}^{\dagger} = \mathbf{B}(\mathbf{B}^H\mathbf{B})^{-1}\mathbf{B}^H \tag{1.2.79}$$

which is the projection matrix onto the column space of $\mathbf{B}$. Then, the minimization of (1.2.72) can be rewritten as

$$\begin{aligned}\min_{\{z_i\}} \|\mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{y}\|^2 &= \min_{\{b_i\}} \|\mathbf{P}_{\mathbf{B}}\mathbf{y}\|^2 = \min_{\{b_i\}} \mathbf{y}^H\mathbf{P}_{\mathbf{B}}\mathbf{y} \\ &= \min_{\{b_i\}} \mathbf{y}^H\mathbf{B}(\mathbf{B}^H\mathbf{B})^{-1}\mathbf{B}^H\mathbf{y} = \min_{\{b_i\}} \bar{\mathbf{b}}^H\bar{\mathbf{Y}}^H(\mathbf{B}^H\mathbf{B})^{-1}\bar{\mathbf{Y}}\bar{\mathbf{b}} \end{aligned} \tag{1.2.80}$$

where

$$\bar{\mathbf{Y}} = \begin{bmatrix} y_0 & y_1 & \cdots & y_I \\ y_1 & y_2 & \cdots & y_{I+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{M-I-1} & y_{M-I} & \cdots & y_{M-1} \end{bmatrix} \tag{1.2.81}$$

$$\bar{\mathbf{b}} = \begin{bmatrix} b_0 & b_1 & \cdots & b_I \end{bmatrix}^T. \tag{1.2.82}$$

The cost function in (1.2.80) is still highly nonlinear of the parameters $b_0, b_1, \cdots, b_I$. However, the last expression in (1.2.80) is useful. From the property of (1.2.75), it is easy to verify that in the absence of noise,

$$\bar{\mathbf{Y}}\bar{\mathbf{b}} = \mathbf{0}. \tag{1.2.83}$$

It has been shown in Subsection 1.2.1 that given a noise-free $\bar{\mathbf{Y}}$ ($= \mathbf{Y}$ when $L = I$), (1.2.83) implies a unique $\bar{\mathbf{b}}$ up to a scalar. To avoid trivial solution of (1.2.83), a constraint such that $\bar{\mathbf{b}} \neq 0$ is needed. Some linear nontriviality constraints have been in use, such as $b_0 = 1$ in [6] and $b_{I/2} = 1$ in [88]. However, as noted in [101], the above two simple constraints can not represent certain (though very few) practical scenarios. On the other hand, the quadratic nontriviality constraint

$$\|\bar{\mathbf{b}}\| = 1 \tag{1.2.84}$$

appears to be the most natural choice as it applies an equal weight on each element of the vector $\mathbf{b}$.

In the noise-free case, if b_0 is set to one, an equivalent equation of (1.2.83) is

$$\bar{\mathbf{Y}}_0\bar{\mathbf{b}}_0 = -\bar{\mathbf{y}}_0 \tag{1.2.85}$$

where $\bar{\mathbf{Y}} = \begin{bmatrix} \bar{\mathbf{y}}_0 & \bar{\mathbf{Y}}_0 \end{bmatrix}$ and $\bar{\mathbf{b}} = \begin{bmatrix} 1 & \bar{\mathbf{b}}_0^T \end{bmatrix}^T$. In the noisy case, (1.2.83) does not hold, and one can estimate $(b_0, b_1, \cdots, b_I)$ in two different ways. The first approach is to compute the least squares solution of (1.2.83), i.e.,

$$\arg\min_{\bar{\mathbf{b}}} \|\bar{\mathbf{Y}}\bar{\mathbf{b}}\|^2 = \arg\min_{\bar{\mathbf{b}}} \bar{\mathbf{b}}^H\bar{\mathbf{Y}}^H\bar{\mathbf{Y}}\bar{\mathbf{b}} \tag{1.2.86}$$

where $\|\bar{\mathbf{b}}\| = 1$. The solution of (1.2.86) is simply the eigenvector of $\bar{\mathbf{Y}}^H\bar{\mathbf{Y}}$ associated with the smallest eigenvalue. This solution is known as Pisarenko's method as mentioned in Subsection 1.2.1. It is also known as the "total least squares" solution of (1.2.85). The second approach is to compute the least squares solution of (1.2.85), i.e.,

$$\arg\min_{\bar{\mathbf{b}}_0} \|\bar{\mathbf{Y}}_0\bar{\mathbf{b}}_0 + \bar{\mathbf{y}}_0\|^2. \tag{1.2.87}$$

The solution of (1.2.87) is given by

$$\bar{\mathbf{b}}_{0,\mathrm{LS}} = -\bar{\mathbf{Y}}_0^{\dagger}\bar{\mathbf{y}}_0 \tag{1.2.88}$$

where $\bar{\mathbf{Y}}_0^{\dagger}$ is the pseudo-inverse of $\bar{\mathbf{Y}}_0$. This solution is known as Prony's method discussed in Subsection 1.2.1.

Neither Prony's method nor Pisarenko's method is optimal in the least squares sense given by (1.2.80). However, if the error in the estimate obtained by (1.2.86) or (1.2.87) is small, (1.2.80) can be approximated as follows:

$$\min_{\mathbf{b}} \bar{\mathbf{b}}^H \bar{\mathbf{Y}}^H (\mathbf{B}_E^H \mathbf{B}_E)^{-1} \bar{\mathbf{Y}} \bar{\mathbf{b}} \tag{1.2.89}$$

where $\mathbf{B}_E$ is determined by the estimates from (1.2.86) or (1.2.87). The minimization of (1.2.89) is quadratic, and hence the solution is given by the eigenvector of the matrix $\bar{\mathbf{Y}}^H(\mathbf{B}_E^H\mathbf{B}_E)^{-1}\bar{\mathbf{Y}}$ associated with the least eigenvalue. This minimization can be repeated once $\mathbf{B}_E$ is renewed by the previous estimate of $\bar{\mathbf{b}}$. This iterative procedure based on (1.2.89) is called the iterative quadratic maximum likelihood (IQML) method [6]. The reference of ML is because that if the noise in (1.2.70) is white Gaussian, the least squares criterion (1.2.80) is equivalent to the maximum likelihood criterion.

The IQML procedure proposed in [69] is the same as the one outlined above. The IQML procedure proposed in [14] exploits the knowledge of the noise covariance via the constraint $\bar{\mathbf{b}}^H \Re e\{\mathbf{R}_n\}\bar{\mathbf{b}} = 1$ where $\mathbf{R}_n$ is the noise covariance matrix defined in (1.1.11).

For multidimensional (M-D) models such as those in (e.g., [31], [32]), direct application of IQML is not possible because there does not exist a single M-D polynomial which has a one-to-one relationship with M-D frequencies. However, by reformulating the M-D models (with an added amplitude redundancy), multiple inter-related single-variate polynomials as in [16] and [65] can be formed. The M-D frequencies are uniquely related to those single-variate polynomials. Reparameterization of the likelihood function with the coefficients of the single-variate polynomials leads to quadratic functions of the same form as in (1.2.89). This allows the application of the IQML to the M-D case.

If a polynomial has all its roots located on the unit circle, its coefficients will be conjugate symmetric. More specifically, the coefficients of the polynomial (1.2.75) should satisfy the property:

$$b_i = b_{I-i}^*. \tag{1.2.90}$$

The estimation accuracy can be improved by imposing this constraint on $\mathbf{b}$. The incorporation of (1.2.90) and (1.2.84) into (1.2.89) can be implemented using the idea suggested in [101].

c. Implementation of IQML

Next, we show a fast implementation of the IQML method. This implementation follows from [33] where the unit circle constraint is not applied.

Step 1: Compute an initial estimate of $\bar{\mathbf{b}}$ by $\arg\min_{\bar{\mathbf{b}}} \bar{\mathbf{b}}^H \mathbf{Y}^H \mathbf{Y} \bar{\mathbf{b}}$. Each element of the $(I+1)\times(I+1)$ matrix $\mathbf{Y}^H\mathbf{Y}$ can be obtained by

$$(\mathbf{Y}^H\mathbf{Y})_{i,j} = \sum_{l=0}^{M-I-1} y_{i+L+1}^* y_{j+l-1} \tag{1.2.91}$$

where $1 \le i,j \le I+1$. Note that $(\mathbf{Y}^H\mathbf{Y})_{i,j} = (\mathbf{Y}^H\mathbf{Y})_{j,i}^*$. Hence, the floating point operations (flops) required are $(M-I)(I+1)(I+2)/2$. The least eigenvector of $\mathbf{Y}^H\mathbf{Y}$ can be obtained using an order of I^2 flops [25].

Step 2: Use the previous estimate of $\bar{\mathbf{b}}$ to compute $\mathbf{B}_E^H\mathbf{B}_E$ as follows:

$$(\mathbf{B}_E^H\mathbf{B}_E)_{i,j} = \begin{cases} \sum_{i=0}^{I-(j-i)} b_{j-i+1} b_l^*, & I \ge j-i \ge 0 \\ \sum_{i=0}^{I-(i-j)} b_l b_{l+i-j}^*, & I \ge i-j \ge 0 \\ 0, & |i-j| > I \end{cases} . \tag{1.2.92}$$

Note that this matrix $\mathbf{B}_E^H\mathbf{B}_E$ is Hermitian, Toeplitz and band-limited. Computing this matrix product only takes $(I+1)(I+2)/2$ flops (or $I(I+1)/2$ flops if $b_0 = 1$). The storage of this matrix requires only $I+1$ complex numbers.

Step 3: Compute the Cholesky decomposition of the $(M-I)\times(M-I)$ matrix $\mathbf{B}_E^H\mathbf{B}_E$, which is positive definite, Hermitian, and band-limited. Using Algorithm 5.3-5 in [25], we have

$$\mathbf{B}_E^H\mathbf{B}_E = \mathbf{R}^H\mathbf{R} \tag{1.2.93}$$

where $\mathbf{R}$ is an $(M-I)\times(M-I)$ upper triangular matrix with bandwidth $I+1$. This requires $(M-I)[I^2/2 + 3I/2] - I^3/3 - 3I^2/2$ flops [[25], p.76.].

Step 4: Compute the $(M-I)\times(I+1)$ matrix $\mathbf{R}^{-H}\mathbf{Y}$ by a simple modification of Algorithm 5.3-2 or 5.3-3 in [25]. This requires $(M-I)(I+1)^2 - I^2(I+1)/2$ flops.

Step 5: Compute the $(I+1)\times(I+1)$ Hermitian matrix $\mathbf{Y}^H(\mathbf{B}_E^H\mathbf{B}_E)^{-1}\mathbf{Y}$ by computing $(\mathbf{R}^{-H}\mathbf{Y})^H\mathbf{R}^{-H}\mathbf{Y}$. This requires $(M-I)(I+1)(I+2)/2$ flops.

Step 6: Compute a new estimate of $\bar{\mathbf{b}}$ by computing the least eigenvector of $\mathbf{Y}^H(\mathbf{B}_E^H\mathbf{B}_E)^{-1}\mathbf{Y}$. This requires an order of I^2 flops.

Step 7: Go to Step 2 until convergence.

The above implementation of IQML only requires an order of flops proportional to M (assuming $M >> I$). More precisely, assuming $M >> I >> 1$, the first step of the above implementation requires $I^2M/2$ flops, and each additional iteration from Step 2 to Step 6 requires $2I^2M$ flops. Other implementations such as given in [56] and [17] require a flop order proportional to M^2.

1.3 Estimation Techniques Using Large-Sample Theorems

Signal amplitude (statistical) and noise are further assumed to be realizations of ergodic processes. This assumption ensures that the time average converges to the statistical average in almost sure sense[4] :

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}(n)\mathbf{y}^H(n) = \mathbf{A}\mathbf{R}_s\mathbf{A}^H + \mathbf{R}_n. \tag{1.3.1}$$

For the finite deterministic amplitudes, by (1.1.10) and the strong law of the large number theorem, the above property also holds. Therefore with a large number of measurements, the following covariance matrix (for both stochastic and deterministic amplitudes):

$$\mathbf{R} \triangleq E\{\mathbf{y}(n)\mathbf{y}^H(n)\} = \mathbf{A}\mathbf{R}_s\mathbf{A}^H + \mathbf{R}_n \tag{1.3.2}$$

can then be estimated accurately where E denotes the statistical expectation. The structure shown in this expression is the key for all methods using a large number of measurements.

1.3.1 Subspace Rotation Invariance - ESPRIT

The principle of subspace rotation invariance is closely related to the principle of matrix pencil. Let $\mathbf{U}_s$ contain the unitary eigenvectors of the covariance matrix $\mathbf{R}$ in (1.3.2) corresponding to the I largest eigenvalues. When $\mathbf{R}_s$ is of full rank and $\mathbf{R}_n = \sigma^2\mathbf{I}_M$, it is easy to verify that $\mathbf{U}_s$ and $\mathbf{A}$ (defined in (1.1.6)) have the same range, i.e.,

$$\mathbf{U}_s = \mathbf{A}\mathbf{T} \tag{1.3.3}$$

where $\mathbf{T}$ is an $I \times I$ nonsingular matrix. Due to the relationship (1.3.3), the range space of $\mathbf{U}_s$ is called the signal subspace[5]. Let $\mathbf{U}_1$ be the submatrix of $\mathbf{U}_s$ with the last row deleted, and $\mathbf{U}_2$ the submatrix of $\mathbf{U}_s$ with the first row deleted. Then, one has:

$$\mathbf{U}_2 = \mathbf{U}_1\mathbf{\Psi}. \tag{1.3.4}$$

where $\mathbf{\Psi} = \mathbf{T}^{-1}\mathbf{\Theta}\mathbf{T}$ with $\mathbf{T}$ being a nonsingular $I \times I$ matrix. Eigenvalues of $\mathbf{\Psi}$ are equal to the desired signal poles. The relationship of (1.3.4) is called the subspace rotation invariance that forms the basis of the ESPRIT [90] and its many variations. It is important to note that this invariance property not only holds for the model

[4] Almost sure convergence of a sequence $\{a_n\}$ to a is defined by $P(\lim_{n\to\infty} |a_n - a| < \epsilon) = 1$ where ϵ is an arbitrarily small positive number. Other terms such as strong convergence and convergence with probability 1 are also used in the literature.

[5] For $\mathbf{X}$ in (1.2.81) used in the linear prediction and matrix pencil methods, a signal subspace can be similarly defined.

assumed in (1.1.6) but also for a larger class of models. The model corresponds to (1.1.6) is associated with a uniform linear array. But the subspace rotation invariance property in (1.3.4) is ensured by any array that consists of two identical subarrays that are spatially translated from each other. The two subarrays may have overlapped elements. The waves impinging on the arrays are assumed to be planar.

Given a finite sample estimate of the covariance matrix as defined below

$$\hat{\mathbf{R}} \triangleq \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}(n)\mathbf{y}^H(n), \tag{1.3.5}$$

$\mathbf{U}_s$ should be replaced by its estimate $\hat{\mathbf{U}}_s$ which consists of the eigenvectors of $\hat{\mathbf{R}}$ corresponding to the I largest eigenvalues. In this case, (1.3.4) does not hold exactly, but $\mathbf{\Psi}$ can be estimated from $\hat{\mathbf{U}}_s$ in a number of different ways as discussed next.

LS (least squares)-ESPRIT: This approach is to solve (1.3.4) through the following minimization:

$$\min_{\mathbf{\Psi}} \|\hat{\mathbf{U}}_2 - \hat{\mathbf{U}}_1\mathbf{\Psi}\|_F. \tag{1.3.6}$$

The solution for $\mathbf{\Psi}$ is given by

$$\hat{\mathbf{\Psi}}_{\mathrm{LS}} = (\hat{\mathbf{U}}_1^H\hat{\mathbf{U}}_1)^{-1}\hat{\mathbf{U}}_1^H\hat{\mathbf{U}}_2. \tag{1.3.7}$$

The LS-ESPRIT method estimates the signal poles by computing the eigenvalues of $\hat{\mathbf{\Psi}}_{\mathrm{LS}}$.

TLS (total least squares) -ESPRIT: The equation (1.3.4) can be alternatively solved in a total least squares (TLS) fashion. The corresponding solution is given as follows. Let the rank-I truncation of $[\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2]$ be expressed as $[\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2]_T = \hat{\mathbf{U}}_u\hat{\mathbf{\Sigma}}_u[\hat{\mathbf{V}}_{u1}^H, \hat{\mathbf{V}}_{u2}^H]$. Then the TLS solution is given by

$$\hat{\mathbf{\Psi}}_{\mathrm{TLS}} = (\hat{\mathbf{V}}_{u2}^H\hat{\mathbf{V}}_{u2})^{-1}\hat{\mathbf{V}}_{u2}^H\hat{\mathbf{V}}_{u1}. \tag{1.3.8}$$

The TLS-ESPRIT method estimates the signal poles by computing the eigenvalues of $\hat{\mathbf{\Psi}}_{\mathrm{TLS}}$. A detailed analysis of the TLS-ESPRIT is presented in [71]. The TLS-ESPRIT has a similar performance to the LS-ESPRIT when the number of measurements is sufficiently large [85]. The reason for this is also supported by the first order equivalence of various types of SVD truncations [42]. The difference between LS-ESPRIT and TLS-ESPRIT is nothing but in the way of SVD truncations.

W (weighted) -ESPRIT: Another variation of the ESPRIT is to introduce a weighting matrix as follows:

$$\hat{\mathbf{\Psi}}_{\mathrm{WESP}} = (\hat{\mathbf{U}}_1^H\boldsymbol{\mathcal{W}}\hat{\mathbf{U}}_1)^{-1}\hat{\mathbf{U}}_1^H\boldsymbol{\mathcal{W}}\hat{\mathbf{U}}_2 \tag{1.3.9}$$

where $\boldsymbol{\mathcal{W}}$ is an $(M-1)\times(M-1)$ Hermitian symmetric, positive semi-definite matrix. This corresponds to the so called W-ESPRIT method. The LS-ESPRIT is a special case of the W-ESPRIT when $\boldsymbol{\mathcal{W}} = \mathbf{I}_{M-1}$. Assuming that the signal vectors $\mathbf{s}(n)$ and the noise vectors $\mathbf{w}(n)$ are all *i.i.d.* Gaussian, a performance analysis of the W-ESPRIT yields the following theorem:

Theorem 1.2: *Let $\boldsymbol{\rho}_i^H$ be the i-th row of $(\mathbf{A}_-^H\boldsymbol{\mathcal{W}}\mathbf{A}_-)^{-1}\mathbf{A}_-^H\boldsymbol{\mathcal{W}}(\mathbf{I}_{M+} - z_i\mathbf{I}_{M-})$ where $\mathbf{G}_-$ is a submatrix of $\mathbf{G}$ without the last row, and $\mathbf{G}_+$ is a submatrix of $\mathbf{G}$ without the first row. If $N >> 1$, the weighted ESPRIT frequency estimation error vector $\hat{\boldsymbol{\theta}}^{WESP} - \boldsymbol{\theta}$ is Gaussian with zero mean and the covariance matrix*

$$\mathbf{C}_{\mathrm{WESP}}(\boldsymbol{\theta}) = \frac{\sigma^2}{2N}\Re e\{(\boldsymbol{\rho}^H\boldsymbol{\rho})\odot\boldsymbol{\mathcal{W}}_0^T\} \tag{1.3.10}$$

where

$$\boldsymbol{\rho} = [\boldsymbol{\rho}_1, \cdots, \boldsymbol{\rho}_I] \tag{1.3.11}$$

$$\boldsymbol{\mathcal{W}}_0 = \mathbf{R}_s^{-1} + \sigma^2\mathbf{R}_s^{-1}(\mathbf{A}^H\mathbf{A})^{-1}\mathbf{R}_s^{-1} \tag{1.3.12}$$

and $\mathbf{R}_s$ is defined in (1.1.9)

Proof: See [100].

In (1.3.10), $\odot$ denotes the Schur-Hadamard product (element-wise product) defined as below

$$\mathbf{X}_{m\times n}\odot\mathbf{Y}_{m\times n} = \begin{bmatrix} x_{1,1}y_{1,1} & \cdots & x_{1,n}y_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1}y_{m,1} & \cdots & x_{m,n}y_{m,n} \end{bmatrix}_{m\times n}. \tag{1.3.13}$$

The analysis result in Theorem 1.2 is based on the perturbation of the signal subspace [52] assuming that the number of measurements is sufficiently large. When the number of measurements is small, another (high SNR) perturbation analysis of the ESPRIT can be found in [64].

OW (optimally weighted) -ESPRIT: For the W-ESPRIT, it is of interest to find a weighting matrix which gives rise to a minimum estimation variance. For a minimum estimation variance of the i-th frequency, the optimal weighting matrix should be chosen as [100]

$$\boldsymbol{\mathcal{W}}_{opt}^i = (\mathbf{T}_i^H\mathbf{T}_i)^{-1} \tag{1.3.14}$$

where $\mathbf{T}_i^H = \mathbf{I}_{M+} - z_i\mathbf{I}_{M-}$. When $\boldsymbol{\mathcal{W}}$ is substituted by $\boldsymbol{\mathcal{W}}_{opt}^i$ in (1.3.9), the resulting method is called the OW-ESPRIT. However one should note that when $\boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{W}}_{opt}^i$ in (1.3.9), only the estimation variance of the i-th frequency is guaranteed to reach a

minimum, while the estimation variances of other frequencies using the same weighting matrix may not be reduced. In (1.3.14), the optimal weighting matrix $\boldsymbol{\mathcal{W}}_{opt}^{i}$ requires matrix inversion. But since $\mathbf{T}_i^H\mathbf{T}_i$ is a banded matrix, an analytic inversion is available and shown in the following lemma:

Lemma 1.1:

$$\boldsymbol{\mathcal{W}}_{opt}^{i} = \mathbf{Z}^H(\mathbf{I}_{M-1} - \frac{1}{M}\mathbf{z}\mathbf{z}^H)\mathbf{Z} \tag{1.3.15}$$

where

$$\mathbf{Z} = \begin{bmatrix} e^{-j\theta_i} & e^{-j2\theta_i} & \cdots & e^{-j(M-1)\theta_i} \\ & \ddots & \ddots & \vdots \\ & & \ddots & e^{-j2\theta_i} \\ & & & e^{-j\theta_i} \end{bmatrix}_{(M-1)\times(M-1)} \tag{1.3.16}$$

is upper triangular and $\mathbf{z}$ *is the last column of* $\mathbf{Z}$.

Proof: See Appendix D of [100].

Computation of $\boldsymbol{\mathcal{W}}_{opt}^{i}$ using (1.3.15) is more efficient than direct inversion. But the i-th signal pole z_i is required in computing $\boldsymbol{\mathcal{W}}_{opt}^{i}$. In practice, this knowledge is not available and an estimated value of z_i is used to construct an approximated weighting matrix $\hat{\boldsymbol{\mathcal{W}}}_{opt}^{i}$. The OW-ESPRIT is summarized in the following.

Procedure of OW-ESPRIT

Step 1: Determine estimates $\hat{z}_i^0$ of z_i, $i = 1, ..., I$ from eigenvalues of $(\hat{\mathbf{U}}_1^H\hat{\mathbf{U}}_1)^{-1}\hat{\mathbf{U}}_1^H\hat{\mathbf{U}}_2$.

Step 2: For $i = 1, 2, ..., I$, use $\hat{z}_i^0$ to construct the weighting matrix $\hat{\boldsymbol{\mathcal{W}}}_{opt}^{i}$ and determine an updated estimate $\hat{z}_i^{opt}$ of z_i as one eigenvalue of the matrix $(\hat{\mathbf{U}}_1^H\hat{\boldsymbol{\mathcal{W}}}_{opt}^{i}\hat{\mathbf{U}}_1)^{-1}\hat{\mathbf{U}}_1^H\hat{\boldsymbol{\mathcal{W}}}_{opt}^{i}\hat{\mathbf{U}}_2$.

A comparison between the OW-ESPRIT and the LS-ESPRIT is given in the following. In the single signal case,

$$\frac{var_{\text{OWESP}}(\theta_1)}{var_{\text{ESP}}(\theta_1)} = \frac{(M-1)^2}{24M(M^2-1)} \tag{1.3.17}$$

where $var_{\text{OWESP}}(\cdot)$ is the estimation variance for the OW-ESPRIT and $var_{\text{ESP}}(\cdot)$ for the ESPRIT. See Examples 4.1 and 5.1 of [100] for a proof. The conclusion is that the OW-ESPRIT can offer much better accuracy than the LS-ESPRIT when M is large.

FB-TLS-ESPRIT: Assume that all the signal poles are on the unit circle. Define the forward-backward averaging covariance matrix [6]

$$\hat{\mathbf{R}}_{fb} = \frac{1}{2}[\hat{\mathbf{R}} + \mathbf{\Pi}_M \hat{\mathbf{R}}^* \mathbf{\Pi}_M] \tag{1.3.18}$$

where $\mathbf{\Pi}_M$ is defined in (1.2.33). Let $\hat{\mathbf{U}}_{s,fb}$ contain the unitary eigenvectors of $\hat{\mathbf{R}}_{fb}$ corresponding to the I largest eigenvalues, $\hat{\mathbf{U}}_1$ the submatrix of $\hat{\mathbf{U}}_{s,fb}$ without the last row, and $\hat{\mathbf{U}}_2$ the submatrix of $\hat{\mathbf{U}}_{s,fb}$ without the first row. The TLS solution requires an additional SVD of $[\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2]$, the rank-$I$ truncation of which is denoted by

$$[\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2] = \hat{\mathbf{U}}_{fb} \hat{\mathbf{\Sigma}}_{fb} \hat{\mathbf{V}}_{fb}^H \tag{1.3.19}$$

where $\hat{\mathbf{\Sigma}}_{fb}$ is a diagonal matrix containing the singular values in non-increasing order, $\hat{\mathbf{U}}_{fb}$ and $\hat{\mathbf{V}}_{fb}$ contain the corresponding left- and right-unitary singular vectors respectively. Partition $\hat{\mathbf{V}}_{fb}$ into

$$\hat{\mathbf{V}}_{fb} = \begin{bmatrix} \hat{\mathbf{V}}_{fb,11} & \hat{\mathbf{V}}_{fb,12} \\ \hat{\mathbf{V}}_{fb,21} & \hat{\mathbf{V}}_{fb,22} \end{bmatrix} \tag{1.3.20}$$

where each submatrix is of dimensions $I \times I$. The FB-TLS-ESPRIT pole estimates are chosen as the eigenvalues of the matrix

$$-\hat{\mathbf{V}}_{fb,12} \hat{\mathbf{V}}_{fb,22}^{-1}. \tag{1.3.21}$$

Unitary-ESPRIT: Let

$$\mathbf{Q} = \begin{cases} \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_{M/2} & j\mathbf{I}_{M/2} \\ \mathbf{\Pi}_{M/2} & -j\mathbf{\Pi}_{M/2} \end{bmatrix}, & M \text{ even} \\ \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_{(M-1)/2} & 0 & j\mathbf{I}_{(M-1)/2} \\ \mathbf{0}^T & j\sqrt{2} & \mathbf{0}^T \\ \mathbf{\Pi}_{(M-1)/2} & 0 & -j\mathbf{\Pi}_{(M-1)/2} \end{bmatrix}, & M \text{ odd.} \end{cases} \tag{1.3.22}$$

It is easy to see that $\mathbf{Q}$ satisfies: $\mathbf{\Pi}_M \mathbf{Q} = \mathbf{Q}^*$. From this property, one can obtain a real covariance matrix

$$\mathbf{Q}^{-1} \hat{\mathbf{R}}_{fb} \mathbf{Q} = \Re e\{\mathbf{Q}^{-1} \hat{\mathbf{R}} \mathbf{Q}\}. \tag{1.3.23}$$

Let $\hat{\mathbf{U}}_r$ be the real unitary eigenvectors of this matrix, then $\hat{\mathbf{U}}_{fb}$ can be constructed as

$$\hat{\mathbf{U}}_{fb} = \mathbf{Q} \hat{\mathbf{U}}_r. \tag{1.3.24}$$

[6] Improved performance for the LS-ESPRIT was proved in [87]. When $\mathbf{R}_s$ is diagonal, it was further shown [47] that $\hat{\mathbf{R}}_{fb}$ is a more accurate estimate of $\mathbf{R}$ than $\hat{\mathbf{R}}$.

Now it is easy to prove that

$$\hat{\mathbf{U}}_{fb} = \mathbf{\Pi}_M \hat{\mathbf{U}}_{fb}^*, \tag{1.3.25}$$

which means that each column of $\hat{\mathbf{U}}_{fb}$ is conjugate symmetric. This property [7] is essential for signal pole estimates to lie on the unit circle.

Since $\hat{\mathbf{U}}_1 = \mathbf{\Pi}_{M-1}\hat{\mathbf{U}}_2^*$ according to (1.3.25), then $\hat{\mathbf{V}}_{fb}$ is equivalent to the unitary eigenvectors of the matrix

$$\mathbf{B} = \begin{bmatrix} \hat{\mathbf{U}}_1^H \\ \hat{\mathbf{U}}_1^T \mathbf{\Pi}_{M-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{U}}_1 & \mathbf{\Pi}_{M-1}\hat{\mathbf{U}}_1^* \end{bmatrix}. \tag{1.3.26}$$

Let $\mathbf{P} = diag[\mathbf{I}_I, -\mathbf{\Pi}_I] = \mathbf{P}^{-1}$, it can be seen that $\mathbf{P}^{-1}\mathbf{B}\mathbf{P}$ is a centro-hermitian matrix. Since $\mathbf{P}^{-1}\hat{\mathbf{V}}_{fb}$ are the unitary eigenvectors of this matrix, if they are put in a conjugate symmetric form, we can have $\hat{\mathbf{V}}_{fb,12} = \mathbf{\Pi}_I(-\mathbf{\Pi}_I\hat{\mathbf{V}}_{fb,22}^*) = -\hat{\mathbf{V}}_{fb,22}^*$. Then (1.3.21) can be rewritten as

$$\hat{\mathbf{V}}_{fb,12}(\hat{\mathbf{V}}_{fb,12}^*)^{-1}. \tag{1.3.27}$$

The matrix (1.3.27) has a property that if μ_i is an eigenvalue, so is $1/\mu_i^*$. When the number of measurements is large, eigenvalues will have distinct phases and thus $|\mu_i| = 1$. The resulting method is the Unitary-ESPRIT, proposed independently in [3], [28].

M-D ESPRIT: For multidimensional subspace invariance, see [107]. In the M-D case, the subspace invariance property can be exploited for each dimension of interest. Each of the previously discussed techniques can be applied to the M-D case. However, the estimated parameters from all dimensions need to be grouped together properly. Various techniques of grouping or association can be found in [34]. Performance analysis techniques for M-D parameters are presented in [9], [68]. Detailed discussion of M-D Unitary ESPRIT can be found in Chapter 5.

1.3.2 Subspace Fitting - MUSIC

The subspace rotation invariance method discussed in the previous section exploits a partial structure in the data. As a consequence, it is tolerant to some errors in the knowledge of the data structure. In particular, the exact knowledge of each of the two subarrays in the total array is not required as long as the two subarrays have an identical structure. But if the complete structure of the data is known, one can exploit more than just the subspace rotation invariance. Exploiting the complete

[7] This property seems to have been proved in many references. To our knowledge, an early proof can be traced to [24]. Proof for special choices of $\mathbf{Q}$ were given in [7]. This property has been used in [67], [28], [44] for reducing computation of eigendecomposition.

knowledge of the subspace structure can be done via the so-called MUSIC method. In this subsection, we will discuss the MUSIC method and its variants.

a. Spectral and Root MUSIC

Let $\mathbf{U}_n$ contain the unitary eigenvectors of the covariance matrix in (1.3.2) corresponding to the $M-I$ smallest eigenvalues, i.e., $\mathbf{U}_s^H\mathbf{U}_n = \mathbf{0}$. The fact that $\mathbf{U}_s$ and $\mathbf{A}$ (defined in (1.1.6)) share the same range (when $\mathbf{R}_n = \sigma^2\mathbf{I}_M$), leads to

$$\mathbf{a}^H(\theta_i)\mathbf{U}_n = \mathbf{0},\ i = 1, 2, ..., I. \tag{1.3.28}$$

where $\mathbf{a}(\theta)$ is defined in (1.1.7). It can be shown that the following function

$$f(\theta) = \mathbf{a}^H(\theta)\mathbf{U}_n\mathbf{U}_n^H\mathbf{a}(\theta) \tag{1.3.29}$$

takes on the minimum values (equal to zero) only at $\theta = \theta_i, i = 1, 2, ..., I$. $\mathbf{U}_n$ is called the noise subspace matrix because it contains all eigenvectors corresponding to (noise) eigenvalues σ^2. With eigenvectors estimated from a finite sample covariance matrix (1.3.5), the function (1.3.29) has the following form

$$\hat{f}(\theta) = \mathbf{a}^H(\theta)\hat{\mathbf{U}}_n\hat{\mathbf{U}}_n^H\mathbf{a}(\theta). \tag{1.3.30}$$

where $\hat{\mathbf{U}}_n$ contains the unitary eigenvectors of the finite sample covariance matrix in (1.3.5), corresponding to the $M-I$ smallest eigenvalues. The MUSIC frequency estimates are determined as the minimizing arguments of $\hat{f}(\theta)$ corresponding to the I smallest values. Note that $\hat{f}(\theta)$ does not necessarily have minimum values equal to zero.

Depending on how the minimizing arguments of $\hat{f}(\theta)$ are searched for, there exist two kinds of MUSIC implementations—spectral MUSIC and root-MUSIC. The spectral MUSIC finds the minimizing arguments by a one-dimensional search algorithm, and the root-MUSIC finds the minimizing arguments by polynomial rooting. Note that the function in (1.3.30) is actually a polynomial of order $2M-1$ for the case where the data are superimposed exponentials. But the MUSIC approach [92] can handle more general models and in general requires a spectrum searching procedure. For the data model considered in this chapter, a polynomial rooting procedure is applicable. The root-MUSIC was first advocated in [5]. In terms of computational load, the root-MUSIC is preferable. With regard to estimation accuracy, it has been found in [84] that the two versions are equivalent for a large number of measurements ($N >> 1$) but the root-MUSIC is more accurate for a moderate number of measurements.

Theorem 1.3: *If $N >> 1$ and $\mathbf{R}_n = \sigma^2\mathbf{I}_M$, then for both the Spectral MUSIC and the Root MUSIC, the frequency estimation error vector $\hat{\boldsymbol{\theta}}^{MU} - \boldsymbol{\theta}$ is Gaussian with zero mean and the covariance matrix:*

$$\mathbf{C}_{\text{MUSIC}}(\boldsymbol{\theta}) = \frac{\sigma^2}{2N}(\mathbf{H}\odot\mathbf{I}_M)^{-1}\Re e\{\mathbf{H}\odot\boldsymbol{\mathcal{W}}_0^T\}(\mathbf{H}\odot\mathbf{I}_M)^{-1} \tag{1.3.31}$$

where $\boldsymbol{\mathcal{W}}_0$ *in (1.3.12), and*

$$\mathbf{H} = \mathbf{D}\mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{D}^H \tag{1.3.32}$$

$$\mathbf{D} = [\frac{\partial \mathbf{a}(\theta_1)}{\partial \theta_1}, ..., \frac{\partial \mathbf{a}(\theta_I)}{\partial \theta_I}] \tag{1.3.33}$$

and $\mathbf{P}_{\mathbf{A}}^{\perp}$ *being defined according to (1.2.73).*

Proof: See [97].

b. Resolution of Spectral MUSIC

The resolution is a measure of the ability of the MUSIC to resolve two closely spaced frequencies. One resolution measure was proposed in [52] using the MUSIC spectrum $\hat{f}(\theta)$. According to that resolution measure, the MUSIC is considered to be able to distinguish the two closely spaced frequencies if two peaks are shown, i.e., $E\{1/\hat{f}(\theta_{1,2})] > E\{1/\hat{f}(\theta_m)]$ where $\theta_{1,2}$ denotes θ_1 or θ_2, and $\theta_m = (\theta_1 + \theta_2)/2$. A minimum signal-to-noise ratio required for the MUSIC resolvability can be solved for by solving the equation

$$E\{1/\hat{f}(\theta_{1,2})\} = E\{1/\hat{f}(\theta_m)\} \tag{1.3.34}$$

An explicit expression does not seem to exist in general. For the case of two signals with equal power P, if we define the total signal-to-noise ratio $\mathrm{SNR}_t = MP/\sigma^2$, the analytical result is [52]

$$\mathrm{SNR}_t = \frac{1}{N}\left\{\frac{2880(M-2)}{M^4}\Delta^{-4}\left[1 + \sqrt{1 + \frac{NM}{120(M-2)}\Delta^2}\right]\right\}. \tag{1.3.35}$$

The above expression explicitly shows the minimum value of SNR_t in terms of the array size M, the number N of measurements, and the frequency separation $\Delta = \theta_1 - \theta_2$. In [52], simulation examples were conducted to demonstrate the validity of this theoretical prediction. For a wide range of scenarios, the simulation result agrees well with the theoretical prediction.

The difference between the peak value $E\{1/\hat{f}(\theta_{1,2})\}$ and the dip $E\{1/\hat{f}(\theta_m)\}$ of the MUSIC spectrum was used as a resolution measure in [126] as well. The effect of frequency separation on the difference was investigated and a direct relationship was obtained. The equivalence between the results in [52] and [126] was revealed in [126].

A formula for calculating the probability of resolution was derived in [124], where two signals are said to be resolvable if

$$\hat{f}(\theta_m) < 0.5(\hat{f}(\theta_1) + \hat{f}(\theta_2)). \tag{1.3.36}$$

This resolution criterion (1.3.36) is based on the large sample peak values rather than the statistically averaged peak values as used in (1.3.34). The calculated

probability is therefore valid for a wide range of situations. A related work was presented in [62]. As shown in [124], the prediction by [62] has a larger mismatch than that by [124]. The MUSIC threshold is also investigated in [122] based on bias analysis.

c. Weighted MUSIC for Coloured Noise

Until now, we have assumed that the noise is white Gaussian. In some cases, ambient noise can be highly correlated [117]. Correlated noise is called coloured noise. Examples of coloured noise include distant shipping noise, wind-generated noise, flow noise and platform noise. For coloured Gaussian noise, the MUSIC in (1.3.30) can be reformulated as

$$\hat{f}_{WMU}(\theta) = \mathbf{a}^H(\theta)\hat{\mathbf{U}}_n\boldsymbol{\mathcal{W}}\hat{\mathbf{U}}_n^H\mathbf{a}(\theta) \tag{1.3.37}$$

where $\boldsymbol{\mathcal{W}}$ is an $M \times M$ positive definite weighting matrix to take into account the coloured noise fields. The matrix $\boldsymbol{\mathcal{W}}$ is determined such that the estimation variance reaches minimum. Such an optimal choice appears to be $\boldsymbol{\mathcal{W}} = \mathbf{R}_n^{-1}$ according to simulations in [50], [82]. This observation has been proven for white Gaussian noise in [99], but remains to be an open problem for coloured noise. Note that when the noise covariance is known, a prewhitening of data can also improve performance.

The approach in [82] was found to be identical to that in [50], by the authors of [105]. Weighted MUSIC was also suggested in [121], [51] aiming at reducing bias of frequency estimates. Reduced bias results in an improved threshold.

1.3.3 Maximum Likelihood Methods

If one can afford more computational cost to achieve optimal performance, one should consider maximum likelihood methods. Maximum likelihood methods exploit the full statistical model of the given data to achieve the optimality. The statistical properties of the signal amplitudes $\mathbf{s}(n)$ are critical for the development of maximum likelihood methods. There are two common choices for $\mathbf{s}(n)$: conditional and unconditional. In the former, the signal amplitudes are assumed to be deterministic unknowns. In the latter, the signal amplitudes are assumed to be Gaussian random variables with zero mean and covariance matrix equal to $\mathbf{R}_s$ (defined in (1.1.9)).

a. Cramér-Rao Bound (CRB) Expressions

In the later part of this subsection, CRB is needed to compare the performance of the various kinds of the maximum likelihood methods. To begin with, we will show CRB expressions for the conditional and unconditional amplitudes. Let $\boldsymbol{\Omega} = [\omega_1, ..., \omega_D]^T$ be the parameter vector consisting of D free real unknowns and

$$L(\boldsymbol{\Omega}) = \log f(\mathbf{y}(1), ..., \mathbf{y}(N)|\boldsymbol{\Omega}) \tag{1.3.38}$$

be the log-likelihood function. Then the so-called Fisher information matrix (FIM) can

be calculated from

$$\mathbf{F} = E\{\frac{\partial L}{\partial \mathbf{\Omega}} \frac{\partial L}{\partial \mathbf{\Omega}^T}\} \tag{1.3.39}$$

where $\frac{\partial L}{\partial \mathbf{\Omega}} = [\frac{\partial L}{\partial \omega_1}, ..., \frac{\partial L}{\partial \omega_D}]^T$. The Cramér-Rao lower bound (CRB) on the covariance matrix of the estimated unknowns is simply $\mathbf{CRB}(\mathbf{\Omega}) = \mathbf{F}^{-1}$. The (i,i)-th element of $\mathbf{F}^{-1}$ is the lower bound on the estimation variance of ω_i. Depending on whether the data model is conditional or unconditional, we will have a different log-likelihood function and hence a different CRB.

Conditional Model

Under the conditional model, the unknown vector $\mathbf{\Omega}$ consists of (independent parameters in) the noise covariance matrix $\mathbf{R}_n$, the amplitudes $\mathbf{s}(1), ..., \mathbf{s}(N)$, and the frequencies

$$\boldsymbol{\theta} = [\theta_1, ..., \theta_I]^T. \tag{1.3.40}$$

According to the assumption of signals and noise in Subsection 1.1.1, $\mathbf{y}(n)$ is Gaussian distributed with mean $\mathbf{A}\mathbf{s}(n)$ and covariance matrix $\mathbf{R}_n$. Then the likelihood function of the measurement vectors $\mathbf{y}(1), ..., \mathbf{y}(N)$, can be written as

$$\begin{aligned} &f_c(\mathbf{y}(1), ..., \mathbf{y}(N) | \mathbf{R}_n, \mathbf{s}(1), ..., \mathbf{s}(N), \boldsymbol{\theta}) \\ &\quad = \frac{1}{|\pi \mathbf{R}_n|^T} \exp(-\sum_{n=1}^{N} [\mathbf{y}(n) - \mathbf{A}\mathbf{s}(n)]^H \mathbf{R}_n^{-1} [\mathbf{y}(n) - \mathbf{A}\mathbf{s}(n)]) \end{aligned} \tag{1.3.41}$$

where $\mathbf{A}$ is defined in (1.1.6) and $\mathbf{s}(n)$ in (1.1.3).

We now define a real parameter vector for the case of white Gaussian noise as

$$\mathbf{\Omega}_c = [\sigma^2, \Re e\{\mathbf{s}(1)\}^T, \Im m[\mathbf{s}(1)]^T, ..., \Re e\{\mathbf{s}(N)\}^T, \Im m\{\mathbf{s}(N)\}^T, \boldsymbol{\theta}^T]^T \tag{1.3.42}$$

where σ^2 is the noise variance. The CRB matrix under the conditional likelihood function (1.3.41) is denoted by $\mathbf{CRB}_c(\mathbf{\Omega}_c)$ and described by the following [102]:

$$\mathbf{CRB}_c(\mathbf{\Omega}_c) = \begin{bmatrix} [\mathrm{CRB}(\sigma^2)]^{-1} & & & & \\ \mathbf{0} & \mathbf{Q} & & \mathbf{0} & \boldsymbol{\Delta}_1 \\ & & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & & \mathbf{Q} & \boldsymbol{\Delta}_N \\ \mathbf{0} & \boldsymbol{\Delta}_1^T & \cdots & \boldsymbol{\Delta}_N^T & \boldsymbol{\Gamma} \end{bmatrix}^{-1} \tag{1.3.43}$$

where

$$\mathrm{CRB}(\sigma^2) = \frac{\sigma^2}{NM} \tag{1.3.44}$$

$$\mathbf{Q} = \frac{2}{\sigma^2}\mathbf{A}^H\mathbf{A} \tag{1.3.45}$$

$$\boldsymbol{\Delta}_n = \frac{2}{\sigma^2}[\Re e\{\mathbf{A}^H\mathbf{D}\mathbf{S}_n\}^T, \Im m\{\mathbf{A}^H\mathbf{D}\mathbf{S}_n]^T\}^T \tag{1.3.46}$$

$$\mathbf{S}_n = diag[s_1(n), ..., s_I(n)] \tag{1.3.47}$$

$$\boldsymbol{\Gamma} = \frac{2}{\sigma^2}\sum_{n=1}^{N}\Re e\{\mathbf{S}_n^H\mathbf{D}^H\mathbf{D}\mathbf{S}_n\}. \tag{1.3.48}$$

Using the matrix inversion lemma, the CRB matrix for the estimated frequencies can be obtained as

$$\mathbf{CRB}_c(\boldsymbol{\theta}) = \frac{\sigma^2}{2N}\left(\sum_{n=1}^{N}\Re e\{\mathbf{S}_n^H\mathbf{H}\mathbf{S}_n\}\right)^{-1} \tag{1.3.49}$$

where $\mathbf{H}$ is defined in (1.3.32). The CRB on the estimation variance of $\theta_i, i \in 1, ..., I$, is given by the (i,i)-th element of the CRB matrix in (1.3.49). Note that the CRB expression in (1.3.49) holds for any value of N. Under the assumption of (1.1.10), the asymptotic ($N \to \infty$) CRB expression is given below

$$\mathbf{CRB}_c^{as}(\boldsymbol{\theta}) = \lim_{N\to\infty}\mathbf{CRB}_c(\boldsymbol{\theta}) = \frac{\sigma^2}{2}\left(\Re e\{\mathbf{H}\odot\mathbf{R}_s^T\}\right)^{-1} \tag{1.3.50}$$

where $\odot$ denotes the Schur-Hadamard (element-wise) product defined in (1.3.13) and $\mathbf{H}$ is defined in (1.3.32). The conditional CRB for frequencies is achievable as SNR or M increases.

For the signal model of superimposed exponentials, a more explicit expression of the conditional CRB under white Gaussian noise is given in [39]. For coloured noise, a CRB expression [8] was recently developed in [75].

Unconditional Model

Under the unconditional model, $\mathbf{s}(n)$ and $\mathbf{w}(n)$ are two independent zero mean Gaussian vectors, and $\mathbf{y}(n)$ is then a Gaussian vector with zero mean and the covariance matrix $\mathbf{R}$ (defined in (1.3.2)). The unknown vector $\boldsymbol{\Omega}$ now consists of (independent parameters in) the noise covariance matrix $\mathbf{R}_n$, (independent parameters in) the signal covariance matrix $\mathbf{R}_s$, and the frequencies $\boldsymbol{\theta}$ (defined in (1.3.40)). In this case, $\mathbf{y}(n)$ is also Gaussian distributed but with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$. Then the

[8] Though white but non-uniform noise is assumed throughout, the derived CRB expression in [75] also applies to coloured noise.

likelihood function of the independent measurements $\mathbf{y}(1), ..., \mathbf{y}(N)$ can be written as

$$\begin{aligned} & f_u(\mathbf{y}(1), ..., \mathbf{y}(N) | \mathbf{R}_n, \mathbf{R}_s, \boldsymbol{\theta}) \\ &= \frac{1}{|\pi \mathbf{R}|^T} \exp(-\sum_{n=1}^{N} \mathbf{y}(n)^H \mathbf{R}^{-1} \mathbf{y}(n)) = \frac{1}{|\pi \mathbf{R}|^T} \exp(-N \mathrm{tr}[\mathbf{R}^{-1} \hat{\mathbf{R}}]) \end{aligned} \quad (1.3.51)$$

where tr[·] stands for the trace operation, and $\hat{\mathbf{R}}$ is defined by (1.3.5) which is called the sample covariance matrix.

The CRB matrix for the unconditional likelihood function (1.3.51) is denoted by $\mathbf{CRB}_u(\boldsymbol{\Omega}_u)$ and described by

$$[(\mathbf{CRB}_u(\boldsymbol{\Omega}_u))^{-1}]_{i,j} = N \mathrm{tr}[\mathbf{R}^{-1}(\boldsymbol{\Omega}_u) \frac{\partial \mathbf{R}(\boldsymbol{\Omega}_u)}{\partial \omega_i} \mathbf{R}^{-1}(\boldsymbol{\Omega}_u) \frac{\partial \mathbf{R}(\boldsymbol{\Omega}_u)}{\partial \omega_j}] \quad (1.3.52)$$

where $\mathbf{R}(\boldsymbol{\Omega}_u)$ is defined as in (1.3.2). To find an explicit expression of the above CRB for frequencies had been a challenging task for a decade and was recently solved in [95].

The unconditional CRB for frequencies is achievable as SNR, M or N increases. A comparison of the CRBs under the two different models is shown in [43].

An expression of the unconditional CRB for frequencies in coloured Gaussian noise was found in [23], with result for (special) non-uniform white noise (where $\mathbf{R}_n \neq \sigma^2 \mathbf{I}_M$ but is still diagonal) presented in [75]. As shown in [23], [22], even with the Hermitianity of $\mathbf{R}_n$, there would be too many parameters in $\mathbf{R}_n$ to be identified. Additional prior information about $\mathbf{R}_n$ is required to reduce parameters for identifiability.

Next we will describe the maximum likelihood methods under the conditional and unconditional models.

b. The Conditional Maximum Likelihood (CML) Method

Ignoring the constant terms, the log-likelihood function (1.3.41) is equal to

$$L(\boldsymbol{\Omega}_c) = -N \log |\pi \mathbf{R}_n| - \sum_{n=1}^{N} [(\mathbf{y}(n) - \mathbf{A}\mathbf{s}(n))^H \mathbf{R}_n^{-1} (\mathbf{y}(n) - \mathbf{A}\mathbf{s}(n))] \quad (1.3.53)$$

For an arbitrary but identifiable $\mathbf{R}_n$, a multi-dimensional search procedure is normally required.

When $\mathbf{R}_n = \sigma^2 \mathbf{I}_M$, the procedure can be significantly simplified to a much lower dimensional search! In this case, the log-likelihood function (1.3.53) becomes

$$L(\boldsymbol{\Omega}_c) = -MN \log \sigma^2 - \frac{1}{\sigma^2} \sum_{n=1}^{N} [(\mathbf{y}(n) - \mathbf{A}\mathbf{s}(n))^H (\mathbf{y}(n) - \mathbf{A}\mathbf{s}(n))]. \quad (1.3.54)$$

Similar to the arguments used in the derivation of IQML in Subsection 1.2.3, the signal

amplitudes and the noise variance can be determined from

$$\hat{\mathbf{s}}(n) = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\mathbf{y}(n) \tag{1.3.55}$$

$$\hat{\sigma}^2 = \frac{1}{MN}\sum_{n=1}^{N}(\mathbf{y}(n) - \mathbf{A}\hat{\mathbf{s}}(n))^H(\mathbf{y}(n) - \mathbf{A}\hat{\mathbf{s}}(n)) = \frac{1}{M}\mathrm{tr}\{\mathbf{P}_{\mathbf{A}}^{\perp}\hat{\mathbf{R}}\} \tag{1.3.56}$$

where $\mathbf{P}_{\mathbf{A}}^{\perp}$ is defined in (1.2.73) and $\hat{\mathbf{R}}$ in (1.3.5). Substituting into (1.3.54) the above expressions transforms (1.3.54) into the following concentrated function containing only I frequencies as unknowns

$$L_c(\boldsymbol{\theta}) = -MN\log\frac{1}{M}\mathrm{tr}\{\mathbf{P}_{\mathbf{A}}^{\perp}\hat{\mathbf{R}}\} \tag{1.3.57}$$

where $\boldsymbol{\theta}$ is defined in (1.3.40). The maximization of the function (1.3.57) is equivalent to the minimization of the function

$$f_c(\boldsymbol{\theta}) = \mathrm{tr}\{\mathbf{P}_{\mathbf{A}}^{\perp}\hat{\mathbf{R}}\}. \tag{1.3.58}$$

Thus the estimates of $\theta_i, i = 1, ..., I$ can be determined as the minimizing argument of the function (1.3.58) by an I-dimensional search procedure. The resulting method is called the conditional ML (CML) method. Note that the CML method can be used for any given number of measurements.

Next we look at the consistency of the frequency estimates given by the CML method. From (1.3.1), $f_c(\boldsymbol{\theta})$ approaches

$$f_c^{as}(\boldsymbol{\theta}) = \lim_{N\to\infty} f_c(\boldsymbol{\theta}) = \mathrm{tr}[\mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{R}]. \tag{1.3.59}$$

Let $\tilde{\mathbf{A}}$ be a matrix with the same structure as $\mathbf{A}$, but constructed using the frequency estimates $\tilde{\theta}_i, i = 1, ..., I$. Then

$$\mathrm{tr}[\mathbf{P}_{\tilde{\mathbf{A}}}^{\perp}\mathbf{R}] = \mathrm{tr}[\mathbf{P}_{\tilde{\mathbf{A}}}^{\perp}(\mathbf{A}\mathbf{R}_s\mathbf{A}^H + \sigma^2\mathbf{I}_M)] = \mathrm{tr}[\mathbf{P}_{\tilde{\mathbf{A}}}^{\perp}\mathbf{A}\mathbf{R}_s\mathbf{A}^H] + \sigma^2(M - I). \tag{1.3.60}$$

(1.3.60) indicates that $\tilde{\mathbf{A}} = \mathbf{A}$ is the only minimizer of (1.3.60) and hence the frequency estimates given by $f_c(\boldsymbol{\theta})$ are asymptotically consistent. Their large sample properties are given in the following theorem.

Theorem 1.4: *If $N >> 1$, the CML frequency estimation error vector $\hat{\boldsymbol{\theta}}^{CML} - \boldsymbol{\theta}$ is Gaussian with zero mean and the covariance matrix*

$$\mathbf{C}_c(\boldsymbol{\theta}) = \frac{\sigma^2}{2N}[\Re e\{\mathbf{H}\odot\mathbf{R}_s^T\}]^{-1}\Re e\{\mathbf{H}\odot(\mathbf{R}_s\boldsymbol{\mathcal{W}}_0\mathbf{R}_s)\}[\Re e\{\mathbf{H}\odot\mathbf{R}_s^T\}]^{-1} \tag{1.3.61}$$

where $\mathbf{H}$ is defined in (1.3.32) and $\boldsymbol{\mathcal{W}}_0$ is defined in (1.3.12).

Proof: See [99].

But the estimates of the amplitudes and noise variance are not asymptotically consistent, when M is finite. It can be shown that

$$\lim_{N\to\infty} \hat{\mathbf{s}}(n) = \mathbf{s}(n) + (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\mathbf{w}(n) \tag{1.3.62}$$

$$\lim_{N\to\infty} \hat{\sigma}^2 = \frac{1}{M}\mathrm{tr}[\mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{R}] = \frac{M-I}{M}\sigma^2. \tag{1.3.63}$$

This fact reveals that the CML frequency estimation variance will not achieve the CRB as the estimator is not asymptotically consistent for all parameters. The following example (originally given in [105]) provides a further explanation.

Example 1.1: Consider a one-signal situation where $\mathbf{R}_s = \sigma_s^2\mathbf{I}_I$ and M is finite. One can obtain that

$$\mathbf{C}_c(\theta_1) = \frac{6\sigma^2}{N}\frac{M\sigma_s^2+\sigma^2}{M^2(M^2-1)\sigma_s^4}$$

$$\mathbf{CRB}_c(\theta_1) = \frac{6\sigma^2}{N}\frac{1}{M(M^2-1)\sigma_s^2}$$

and

$$\frac{\mathbf{C}_c(\theta_1)}{\mathbf{CRB}_c(\theta_1)} = 1 + \frac{\sigma^2}{M\sigma_s^2}$$

which can be greater than 1 if M is finite.

We now assume that both N and M tend to infinity. Then the consistency of frequencies and noise variance follows from (1.3.60) and (1.3.63). The bias term of the amplitude vector in (1.3.62) can be shown to be a Gaussian vector with mean zero and the covariance matrix,

$$E[(\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\mathbf{w}(n)\mathbf{w}(n)^H\mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}] = \sigma^2(\mathbf{A}^H\mathbf{A})^{-1} \to 0. \tag{1.3.64}$$

The signal amplitude estimates are therefore asymptotically consistent in this case.

The I-dimensional minimization of the CML function (1.3.58) can also be reduced to a series of quadratic minimizations similar to that shown in Subsection 1.2.3. Using the polynomial coefficients b_i as defined in (1.2.86), the conditional likelihood function (1.3.58) can be reparameterized as

$$f_{\mathrm{IQ}}(\mathbf{b}) = \mathrm{tr}\{\mathbf{B}(\hat{\mathbf{B}}^H\hat{\mathbf{B}})^{-1}\mathbf{B}^H\hat{\mathbf{R}}\} \tag{1.3.65}$$

where $\hat{\mathbf{R}}$ is defined in (1.3.5) and $\hat{\mathbf{B}}$ is the matrix $\mathbf{B}$ with the polynomial coefficients replaced by available estimates. The fast implementation used for the single measurement IQML approach (1.2.89) can be applied here. However as shown in [66], the minimizing argument of the function (1.3.65) may be a local minimum (i.e., estimates

are not consistent) even for the limiting case of $N = \infty$. In other words, the original IQML is not large-sample consistent (although it is high-SNR consistent). Later, we will show that with a simple modification, the IQML can be made to be large-sample consistent as well.

c. Unconditional Maximum Likelihood (UML) Method

It is known that the UML does not achieve the CRB for finite M. We now look at frequency estimation using the unconditional likelihood function (1.3.51). Direct maximization of this function involves all unknown parameters. The dimension of search can be reduced by exploiting the fact that some of the unknowns can be represented by the estimates of others. If $\mathbf{R}_n = \sigma^2\mathbf{I}_M$, the estimates of the signal covariance matrix $\mathbf{R}_s$ and the noise variance σ^2 can be represented in terms of the frequency estimates

$$\hat{\sigma}^2 = \mathrm{tr}\{\mathbf{P}_{\mathbf{A}}^{\perp}\hat{\mathbf{R}}\}/(M-I) \tag{1.3.66}$$

$$\hat{\mathbf{R}}_s = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\hat{\mathbf{R}}\mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1} - \hat{\sigma}^2(\mathbf{A}^H\mathbf{A})^{-1} \tag{1.3.67}$$

where $\hat{\mathbf{R}}$ is defined in (1.3.5). Substituting σ^2 and $\mathbf{R}_s$ in (1.3.51) with $\hat{\sigma}^2$ and $\hat{\mathbf{R}}_s$, leads to the following concentrated function parameterized using frequencies

$$f_u(\boldsymbol{\theta}) = |\mathbf{A}\hat{\mathbf{R}}_s\mathbf{A}^H + \hat{\sigma}^2\mathbf{I}_M|. \tag{1.3.68}$$

The frequency estimates can be determined as the maximizing argument of (1.3.68). The resulting method is called the unconditional maximum likelihood (UML) method. The large sample properties of the UML frequency estimates are given in the following theorem.

Theorem 1.5: *If $N >> 1$, the UML frequency estimation error vector $\hat{\boldsymbol{\theta}}^{UML} - \boldsymbol{\theta}$ is Gaussian with zero mean, and the covariance matrix*

$$\mathbf{C}_u(\boldsymbol{\theta}) = \frac{\sigma^2}{2N}[\Re e\{\mathbf{H} \odot (\mathbf{R}_s\mathbf{A}^H\mathbf{R}^{-1}\mathbf{A}\mathbf{R}_s)^T\}]^{-1} \tag{1.3.69}$$

where $\mathbf{H}$ is defined as in (1.3.32), and $\mathbf{R}$ as in (1.3.2).

Proof: See [98].

Note that the unknowns (under the unconditional model) include the frequencies, $\mathbf{R}_s$ and σ^2. The UML frequency estimates were shown to be consistent [96]. It is then easy to establish that the UML estimates of σ^2 and $\mathbf{R}_s$ given by (1.3.66) and (1.3.67) are also consistent. The UML method is an ML estimator (under the unconditional model) and the performance of such a consistent ML estimator achieves the Cramér-Rao bound according to [70], i.e., $\mathbf{C}_u(\boldsymbol{\theta}) = \mathbf{CRB}_u(\boldsymbol{\theta})$.

Another ML algorithm termed as MODE was proposed in [101]. The MODE estimates the frequencies as the minimizing argument of the following function

$$f_{\mathrm{MODE}}(\boldsymbol{\theta}) = \mathrm{tr}[\mathbf{A}^H\hat{\mathbf{U}}_n\hat{\mathbf{U}}_n^H\mathbf{A}\hat{\boldsymbol{\mathcal{W}}}_0^{-1}] \tag{1.3.70}$$

where $\hat{\boldsymbol{\mathcal{W}}}_0$ is a consistent estimate of $\boldsymbol{\mathcal{W}}_0$ as defined in (1.3.12). The minimization of (1.3.70) requires a multidimensional search. From the implementation point of view, another (large N equivalent) MODE function is preferred

$$f_{\text{MODE1}}(\boldsymbol{\theta}) = \text{tr}[\hat{\mathbf{U}}_s^H \mathbf{B}(\hat{\mathbf{B}}^H\hat{\mathbf{B}})^{-1}\mathbf{B}^H\hat{\mathbf{U}}_s\hat{\boldsymbol{\Sigma}}_s^{-1}(\hat{\boldsymbol{\Sigma}}_s - \hat{\sigma}^2)^2] \tag{1.3.71}$$

where $\hat{\boldsymbol{\Sigma}}_s$ is a diagonal matrix containing the I-largest eigenvalues of the finite sample covariance matrix (1.3.5), $\hat{\mathbf{U}}_s$ contains the corresponding unitary eigenvectors, $\hat{\sigma}^2 = \sum_{i=I+1}^{M}\hat{\sigma}_i/(M-I)$ is the average of the remaining eigenvalues, and $\hat{\mathbf{B}}$ is constructed using currently available estimates of $b_i, i = 1, ..., I$.

As shown in [102], for a sufficiently large N, $f_{\text{MODE}}(\boldsymbol{\theta})$ and $f_{\text{MODE1}}(\boldsymbol{\theta})$ are the first-order equivalent. Function (1.3.71) can be implemented by a two-step quadratic minimization procedure as outlined in [101]. Initial frequency estimates (for the construction of $\hat{\mathbf{B}}$) can be obtained by other methods or from the minimizing argument of the following function

$$f(\boldsymbol{\theta}) = \text{tr}[\hat{\mathbf{U}}_s^H \mathbf{B}\mathbf{B}^H\hat{\mathbf{U}}_s\hat{\boldsymbol{\Sigma}}_s^{-1}(\hat{\boldsymbol{\Sigma}}_s - \hat{\sigma}^2)^2]. \tag{1.3.72}$$

Theorem 1.6: *If $N >> 1$, the MODE frequency estimation error vector $\hat{\boldsymbol{\theta}}^{\text{MODE}} - \boldsymbol{\theta}$ is Gaussian with zero mean, and the covariance matrix*

$$\mathbf{C}_{\text{MODE}}(\boldsymbol{\theta}) = \frac{\sigma^2}{2N}[\Re e\{\mathbf{H} \odot \boldsymbol{\mathcal{W}}_0^{-T}\}]^{-1} \tag{1.3.73}$$

where $\mathbf{H}$ is defined in (1.3.32) and $\boldsymbol{\mathcal{W}}_0$ is defined in (1.3.12).

Proof: See [102].

The same result as Theorem 1.6 was independently developed in [112]. From (3.20) in [98], we know $\boldsymbol{\mathcal{W}}_0 = (\mathbf{R}_s\mathbf{A}^H\mathbf{R}^{-1}\mathbf{A}\mathbf{R}_s)^{-1}$, and then $\mathbf{C}_u(\boldsymbol{\theta}) = \mathbf{C}_{\text{MODE}}(\boldsymbol{\theta})$. This result implies that the UML method is a large sample realization of the UML method. Minimization of the UML function (1.3.68) is much involved, while minimization of the MODE function (1.3.71) only involves a much appealing (iterative) quadratic minimization procedure. When $N >> 1$, the minimization of the UML function only needs two iterations of quadratic minimization without significantly reducing accuracy. The MODE method is also called the weighted subspace fitting (WSF) approach and its properties have been independently studied in [112], [72].

It is also shown in Appendix A of [98] that $\mathbf{C}_c(\boldsymbol{\theta}) \geq \mathbf{C}_u(\boldsymbol{\theta}) = \mathbf{C}_{\text{MODE}}(\boldsymbol{\theta})$ where $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semi-definite. The $\geq$ part is not surprising since the conditional model has more unknowns than the unconditional model. The following summarizes the relative performances of the UML, UML, and MODE methods.

Lemma 1.2: *Asymptotically ($N >> 1$), $\mathbf{C}_c(\boldsymbol{\theta}) \geq \mathbf{C}_{\text{MODE}}(\boldsymbol{\theta}) = \mathbf{C}_u(\boldsymbol{\theta}) = \mathbf{CRB}_u(\boldsymbol{\theta})$.*

Proof: See [98].

Lemma 1.3: *For finite M,* $\mathbf{CRB}_u(\boldsymbol{\theta}) > \mathbf{CRB}_c(\boldsymbol{\theta})$.

Proof: See [98].

To show some details, one can verify that

$$(\mathbf{CRB}_c(\boldsymbol{\theta}))^{-1} - (\mathbf{CRB}_u(\boldsymbol{\theta}))^{-1} = \frac{2N}{\sigma^2}\Re e\{\mathbf{H} \odot (\mathbf{R}_s - \boldsymbol{\mathcal{W}}_0^{-1})^T\}. \tag{1.3.74}$$

and then

$$\begin{aligned} \mathbf{R}_s - \boldsymbol{\mathcal{W}}_0^{-1} &= \mathbf{R}_s - \mathbf{R}_s[\mathbf{I}_I + \sigma^2\mathbf{R}_s^{-1}(\mathbf{A}^H\mathbf{A})^{-1}]^{-1} \\ &= \sigma^2\mathbf{R}_s[\mathbf{I}_I + \sigma^2\mathbf{R}_s^{-1}(\mathbf{A}^H\mathbf{A})^{-1}]^{-1}\mathbf{R}_s^{-1}(\mathbf{A}^H\mathbf{A})^{-1} \\ &= \sigma^2\mathbf{R}_s[(\mathbf{A}^H\mathbf{A} + \sigma^2\mathbf{R}_s^{-1})\mathbf{R}_s]^{-1} \\ &= \sigma^2[\mathbf{A}^H\mathbf{A} + \sigma^2\mathbf{R}_s^{-1}]^{-1} \end{aligned} \tag{1.3.75}$$

It is clear that the above matrix is positive definite for any $M < \infty$ and hence the above lemma follows.

Observing both of the above lemmas, one can see that the CRB for the conditional model with a finite SNR is not achievable by increasing the sample size N. (This is actually obvious because for the conditional model, the number of unknowns (the signal amplitudes) increases as the sample size increases.) The CRB for the conditional model is only achievable by increasing SNR.

It has been shown [23] that Lemma 1.3 holds for coloured noise as well. However, there has been no report on whether the UML performance for coloured noise remains to achieve the unconditional CRB.

In M-D cases, discussions on the maximum likelihood techniques and their performances, will be provided for the applications of radar processing and biomedical processing in Chapters 6 and 7 respectively.

1.3.4 Smoothing for Coherent Signals

In Section 1.3, we have assumed that the signal covariance matrix $\mathbf{R}_s$ is non-singular. When this assumption does not hold, we say that the signals are coherent. Coherency among signals is caused by multipath propagation and smart jamming in radar processing and wireless communications. For coherent signals, the methods (except MODE/WSF) shown so far in Section 1.3 need some modifications.

As pointed out in [112], by results of [7], [115], the MODE/WSF formulation (1.3.70) can still provide asymptotically consistent estimates for coherent situations, if the signal subspace is replaced by a lower-dimensional one with the dimension equal to the rank of the signal covariance matrix. But the statistical efficiency of the coherent

MODE/WSF was not established, until [106]. A brief summary of the proof will be presented in Part b.

a. Suboptimal Algorithms

One key to handle the coherent signals is to use the so-called spatial smoothing as the first stage of processing. The idea of spatial smoothing scheme was advocated in [21], [93]. But in fact, it is also inherent in the methods of linear prediction and matrix pencil. Specifically, the matrix as given (1.2.1) exploits the idea of spatial smoothing. In order for the MUSIC and ESPRIT methods shown previously to work, we need first to construct a smoothed covariance matrix as shown below. This smoothed covariance matrix should have a rank equal to I in the absence of noise.

For each measurement vector $\mathbf{y}(n)$, construct a smoothed data matrix $\mathbf{Y}(n)$ similar to $\mathbf{X}$ in (1.2.1):

$$\mathbf{Y}(n) = \begin{bmatrix} y_0(n) & y_1(n) & \cdots & y_L(n) \\ y_1(n) & y_2(n) & \cdots & y_{L+1}(n) \\ \vdots & \vdots & \ddots & \vdots \\ y_{M-L-1}(n) & y_{M-L-2}(n) & \cdots & y_{M-1}(n) \end{bmatrix}. \tag{1.3.76}$$

A forward-smoothed covariance matrix is then given by

$$\mathbf{R}_f = \frac{1}{L+1} E\left\{\mathbf{Y}(n)\mathbf{Y}^H(n)\right\} = \mathbf{Z}_{M-L}\mathbf{R}_{s,f}\mathbf{Z}_{M-L}^H + \sigma^2\mathbf{I}_{M-L} \tag{1.3.77}$$

where $\mathbf{Z}_{M-L}$ is defined in (1.2.3),

$$\mathbf{R}_{s,f} = \frac{1}{L+1}\sum_{l=0}^{L}\mathbf{B}^l\mathbf{R}_s(\mathbf{B}^l)^H \tag{1.3.78}$$

$$\mathbf{B} = diag[e^{j\theta_1}, ..., e^{j\theta_I}]. \tag{1.3.79}$$

Because signal amplitudes have zero mean, the signal covariance matrix can be decomposed into two matrices:

$$\mathbf{R}_s = E[(\mathbf{s}(n) - \mathbf{1})(\mathbf{s}(n) - \mathbf{1})^H] + \mathbf{1}\mathbf{1}^T = \bar{\mathbf{R}}_s + \mathbf{1}\mathbf{1}^T \tag{1.3.80}$$

where $\mathbf{1}$ is an I-element column vector with all elements equal to 1. $\mathbf{R}_{s,f}$ can then be rewritten as

$$\mathbf{R}_{s,f} = \frac{1}{L+1}(\mathbf{Z}_{L+1}^T\mathbf{Z}_{L+1}^* + \sum_{l=0}^{L}\mathbf{B}^l\bar{\mathbf{R}}_s(\mathbf{B}^l)^H) \tag{1.3.81}$$

where the second matrix is obviously positive semi-definite. If

$$L + 1 \geq I, \tag{1.3.82}$$

$\mathbf{Z}_{L+1}$ is of full rank I and hence so is $\mathbf{R}_{s,f}$ (regardless of the coherence between signals). To ensure that the signal subspace has a dimension I and the noise subspace has a dimension larger than zero (e.g., as required by MUSIC), we should choose

$$M - L > I. \tag{1.3.83}$$

Similarly, a forward-backward (FB) smoothed covariance matrix can be defined as

$$\mathbf{R}_{fb} = \frac{1}{2}(\mathbf{R}_f + \mathbf{\Pi}_L \mathbf{R}_f^* \mathbf{\Pi}_L) \triangleq \mathbf{Z}_{M-L} \mathbf{R}_{s,fb} \mathbf{Z}_{M-L}^H + \sigma^2 \mathbf{I}_{M-L} \tag{1.3.84}$$

where

$$\mathbf{R}_{s,fb} = \frac{\mathbf{Z}_{L+1}^T \mathbf{Z}_{L+1}^* + \mathbf{B}^{-L} \mathbf{Z}_{L+1}^H \mathbf{Z}_{L+1} \mathbf{B}^L}{2(L+1)} + \bar{\mathbf{R}}_{s,fb}. \tag{1.3.85}$$

It is clear that $\mathbf{Z}_{L+1}^T \mathbf{Z}_{L+1}^*$ and $\bar{\mathbf{R}}_{s,fb}$ are both positive semi-definite. If conditions (1.3.82) and (1.3.83) are satisfied, $\mathbf{R}_{s,fb}$ is of the full rank I as desired. Several methods based on the above smoothed covariance matrix (1.3.84) are presented and analyzed in [103], [76], [78], [77].

It is felt that Condition (1.3.82) may be relaxed to $L + 1 > I/2$ for most practical situations. However, up to now, there has been no conditions which could be tested when frequencies and correlation coefficients are unknown. (A recent attempt was made in [13].)

b. Maximum Likelihood Method

The second approach is to follow the same principle of maximum likelihood estimation except for a step of reparameterization. This reparameterization results in a column-wise smoothing of the matrix $\mathbf{A}$. This (unconditional) maximum likelihood method was proposed in [106] for the coherent case. The eventual algorithm formulation for frequency estimation has a similar form as that for the non-coherent case. Next, we present a brief description of the statistical efficiency of this algorithm. Assuming that the signal covariance matrix $\mathbf{R}_s$ has a rank p ($p < I$), one can uniquely decompose it as

$$\mathbf{R}_s = \begin{bmatrix} \mathbf{I}_p \\ \mathbf{C} \end{bmatrix} \mathbf{S} \begin{bmatrix} \mathbf{I}_p & \mathbf{C} \end{bmatrix} \tag{1.3.86}$$

where $\mathbf{S}$ is a $p \times p$ matrix and $\mathbf{C}$ an $(n-p) \times p$ matrix. Then the covariance matrix in (1.3.2) can be written as

$$\mathbf{R} = \mathbf{A}_c \mathbf{S} \mathbf{A}_c^H + \sigma^2 \mathbf{I}_M \tag{1.3.87}$$

where

$$\mathbf{A}_c = \mathbf{A} \begin{bmatrix} \mathbf{I}_p \\ \mathbf{C} \end{bmatrix}. \tag{1.3.88}$$

This is what was referred to as "smoothing". Let $\boldsymbol{\alpha}$ denote the unknown parameters composed of $\theta_i, i = 1, ..., I$, and the real and imaginary parts of the elements in $\mathbf{C}$. With respect to the new parameter vector $\boldsymbol{\alpha}$, the concentrated function (1.3.68) now becomes

$$f_u(\boldsymbol{\alpha}) = |\mathbf{A}_c \hat{\mathbf{S}} \mathbf{A}_c^H + \hat{\sigma}^2 \mathbf{I}_M|. \tag{1.3.89}$$

where

$$\hat{\sigma}^2 = \mathrm{tr}\{\mathbf{P}^{\perp}_{\mathbf{A}_c} \hat{\mathbf{R}}\}/(M-p) \tag{1.3.90}$$

$$\hat{\mathbf{S}} = (\mathbf{A}_c^H \mathbf{A}_c)^{-1} \mathbf{A}_c^H (\hat{\mathbf{R}} - \hat{\sigma}^2 \mathbf{I}_M) \mathbf{A}_c (\mathbf{A}_c^H \mathbf{A}_c)^{-1} \tag{1.3.91}$$

where $\mathbf{P}^{\perp}_{\mathbf{A}_c}$ is the projection of $\mathbf{A}_c$ onto its orthogonal subspace (defined in (1.2.73)). Note that $\mathbf{A}_c$ is a function of $\boldsymbol{\alpha}$.

The UML estimate of $\boldsymbol{\alpha}$ is given by

$$\arg\min_{\boldsymbol{\alpha}} |\mathbf{A}_c \hat{\mathbf{S}} \mathbf{A}_c^H + \hat{\sigma}^2 \mathbf{I}_M|. \tag{1.3.92}$$

Just like the MODE in the non-coherent case, the minimization of (1.3.92) is asymptotically equivalent to

$$\arg\min_{\boldsymbol{\alpha}} \mathrm{tr}[\mathbf{P}^{\perp}_{\mathbf{A}_c} \boldsymbol{\mathcal{W}}] \tag{1.3.93}$$

where $\boldsymbol{\mathcal{W}} = \hat{\mathbf{U}}_s (\hat{\boldsymbol{\Lambda}} - \hat{\sigma}^2)^2 \hat{\boldsymbol{\Lambda}}^{-1} \hat{\mathbf{U}}_s^H$. Here, $\hat{\boldsymbol{\Lambda}}$ is a diagonal matrix which contains the p largest eigenvalues of $\hat{\mathbf{R}}$ on its diagonal, $\hat{\mathbf{U}}_s$ contains the corresponding unitary eigenvectors, and $\hat{\sigma}^2 = (\mathrm{tr}[\hat{\mathbf{R}}] - \mathrm{tr}[\hat{\boldsymbol{\Lambda}}])/(M-p)$. The above optimization involves additional unknowns in the signal covariance.

Interestingly, as shown in [106], a minimizing $\mathbf{C}$ of (1.3.93) can be fully represented using frequencies as $\hat{\mathbf{C}}(\boldsymbol{\theta}) = \hat{\mathbf{V}}_2(\boldsymbol{\theta}) \hat{\mathbf{V}}_1^{-1}(\boldsymbol{\theta})$ where $\hat{\mathbf{V}}_1(\boldsymbol{\theta})$ is $p \times p$, $\hat{\mathbf{V}}_2(\boldsymbol{\theta})$ is $(I-p) \times p$, and $[\hat{\mathbf{V}}_1^T(\boldsymbol{\theta}), \hat{\mathbf{V}}_2^T(\boldsymbol{\theta})]^T \triangleq (\mathbf{A}^H(\boldsymbol{\theta}) \mathbf{A}(\boldsymbol{\theta}))^{-1} \mathbf{A}^H(\boldsymbol{\theta}) \hat{\mathbf{U}}_s$. Replacing this estimate of $\mathbf{C}$ in (1.3.93) leads to the following minimization problem for estimating frequencies only

$$\arg\min_{\boldsymbol{\theta}} \mathrm{tr}[\mathbf{P}^{\perp}_{\mathbf{A}} \boldsymbol{\mathcal{W}}]. \tag{1.3.94}$$

The function (1.3.94) has an equivalent form to (1.3.70), except $\boldsymbol{\mathcal{W}}$ now has a lower rank. Under mild conditions on $\mathbf{A}(\boldsymbol{\theta})$ in [7], [115], the function (1.3.94) yields asymptotically consistent estimates. Then it readily follows that $\hat{\mathbf{C}}$, and $\hat{\sigma}^2$ in (1.3.90) and $\hat{\mathbf{S}}$ in (1.3.91) will all be asymptotically consistent. Therefore, the frequency estimates give by the minimization problem in (1.3.94) achieves the (coherent) CRB (derived in [106]).

However, in this case, the formulation (1.3.71) is no longer valid, and hence the computationally efficient two-step implementation in [101] is not applicable. A multi-dimensional search must be used instead. The initial estimates can not be efficiently determined from (1.3.72), and suboptimal approaches in Part a can be used to yield initial estimates.

1.4 Detection Techniques Using a Single Measurement

Detection techniques refer to the techniques of estimating the number I in our data model. This number is also referred to as the number of signals. All detection methods are to find a minimum value of I such that a prespecified condition is satisfied. This prespecified condition can be explicit or implicit. An explicit condition can be a threshold value on singular values, eigenvalues or some data matching errors. An implicit condition is often hidden in a penalized log-likelihood function. The penalty term in a penalized log-likelihood function provides a soft control over the minimum value of I that satisfies the implicit condition. It is the prespecified condition as well as the penalty term that have many possible forms and hence lead to different detection methods. Each method also depends on certain assumptions and hence often has its own unique properties.

The detection techniques can be grouped into two categories: techniques using a single measurement and techniques using multiple measurements. This section introduces the first category, and the next section the second category.

1.4.1 Effective Singular Values

Assuming a single measurement of data, the effective singular value (ESV) method [60] estimates the number of signals by determining a number $\hat{I}_{\text{ESV}}$ of the "effective singular values" (termed in [60]) from

$$\hat{I}_{\text{ESV}} = \{k|\lambda_k > \epsilon^{\text{ESV}} \geq \lambda_{k+1}\} \tag{1.4.1}$$

where $\lambda_1, \lambda_2, ..., \lambda_K$ are the singular values, in descending order, of the following $K \times (M-K+1)$ data matrix $\mathbf{Y}_e$:

$$\mathbf{Y}_e = \begin{bmatrix} y_0 & y_1 & \cdots & y_{M-K} \\ y_1 & y_2 & \cdots & y_{M-K+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{K-1} & y_K & \cdots & y_{M-1} \end{bmatrix}. \tag{1.4.2}$$

Clearly, the threshold ϵ^{ESV} is the most critical parameter for deciding the number of "effective" singular values. This threshold depends on the effect of the noise on the singular values. Although finding the statistical distribution of the singular values is a difficult problem, an analysis shown in [60] suggests the following selection of the threshold: [9]

$$\epsilon^{\text{ESV}} = \sqrt{c(M-K+1)}. \tag{1.4.3}$$

c is determined by $P\{S^2 \leq c\} = \alpha$ where $S^2 = |w_0|^2 + \cdots + |w_{K-1}|^2 = 0.5\sigma^2\chi^2_{2K}$, α is a user-chosen confidence level (normally close to 1), and σ^2 is the variance of the

[9] According to [60], this bound is the tightest of the three bounds derived therein for correlated noise in $\mathbf{Y}_e$ in (1.4.2).

white Gaussian noise in the data. The ESV method does not have a good performance at low SNR or for closely-spaced frequencies because in these situations the statistical bound (1.4.3) on the effective singular values tends to be too loose.

1.4.2 Noise Significance Level

A method called the noise significance level (NSL) method [110] also works on the singular value decomposition of the same data matrix $\mathbf{Y}_e$. Based on the perturbation analyses in [63], [54], it is shown in [110] that

$$\sum_{i=I+1}^{K} \lambda_i^2 = \dot{\mathbf{w}}^T \mathbf{Q}^H \mathbf{Q} \dot{\mathbf{w}} \tag{1.4.4}$$

where $\dot{\mathbf{w}} = [\Re e\{\mathbf{w}^T\}, \Im m\{\mathbf{w}^T\}]^T$, $\mathbf{w}$ is defined in (1.1.4) with the time index n dropped, and $\mathbf{Q}$ is a matrix depending on the singular vectors of $\mathbf{X}_e$ (noise-free version of $\mathbf{Y}_e$). If $\mathbf{Q}$ is available, a threshold can be setup based on the distribution of the weighted χ^2 variable $\sum_{i=I+1}^{K} \lambda_i^2$. Unfortunately, the noise-free data matrix $\mathbf{X}_e$ is unknown and neither are its singular vectors. In [110], it is suggested that $S_0 = \sum_{k=0}^{M-K-1} k(|w_k|^2 + |w_{K-k}|^2) + (M-K+1)\sum_{k=M-K}^{K-1} |w_k|^2$ be used as an approximation to $\sum_{i=I+1}^{K} \lambda_i^2$. It is later observed that this approximation results in a performance degradation at medium- and low-SNRs [11].

Let $\epsilon_k^{\mathrm{NSL}}$ be a threshold given by $P\{S_0 \leq \epsilon_k^{NLS}\} = \alpha$. Then the NLS method estimates the number of signals from

$$\hat{I}_{\mathrm{NSL}} = \min\{k | \sum_{i=k+1}^{K} \lambda_i^2 \leq \epsilon_k^{NLS}\}. \tag{1.4.5}$$

The forward-backward version of $\mathbf{Y}_e$, i.e., $[\mathbf{Y}_e, \mathbf{\Pi}_K \mathbf{Y}_e^*]$ can be used to reduce the noise eigenvalue deviation. In this case, the above test becomes

$$\hat{I}_{\mathrm{NSL}} = \min\{k | \sum_{i=k+1}^{K} \mu_i^2 \leq \epsilon_k^{NLS}\} \tag{1.4.6}$$

where $\mu_1, \mu_2, ..., \mu_K$ are the singular values of $[\mathbf{Y}_e, \mathbf{\Pi}_K \mathbf{Y}_e^*]$ in descending order.

Another method developed in [59] is based on a series of hypothesis testings on linear prediction errors. It is computationally expensive and performs worse than the NSL method.

1.4.3 Least Squares Data Fitting

The least squares (LS) method [11] employs a sequential hypothesis testing similar to the ESV and NSL methods. However, the LS method performs a series of tests on the least squares errors between the data and the reconstructed signal (as opposed to the

singular values). As is shown in [11], the least squares error asymptotically exhibits χ^2 distribution when the estimated number of signals is equal to the true one. Unlike the distribution of the singular values, the distribution of the least squares errors (at a given hypothesized number of signals) does not depend on the signal parameters. This fact leads to an accurate threshold setup for testing on the least squares errors. Though more computations are required, the least squares method performs much better than both the ESV and NSL methods [11]. The poor performance of the ESV and NSL methods is caused by a much less accurate threshold setup, as a result of the dependence of noise columns in $\mathbf{Y}_e$. The LS method is described below.

Let H_k be a hypothesis that the number of signals is equal to k. Under H_k, the least squares error of fitting the model to the measurement is defined by

$$\min_{\mathbf{s},\phi_1,...,\phi_k} \|\mathbf{y} - \mathbf{A}_k\mathbf{s}\|^2 \tag{1.4.7}$$

where $\mathbf{y}$ is defined in (1.1.2) with the time index n dropped,

$$\mathbf{s} = [s_1, \cdots, s_k]^T \tag{1.4.8}$$

$$\mathbf{A}_k \stackrel{\triangle}{=} [\mathbf{a}_1, \cdots, \mathbf{a}_k] \tag{1.4.9}$$

$$\mathbf{a}_r \stackrel{\triangle}{=} [1, e^{j\phi_r}, \cdots, e^{j(M-1)\phi_r}]^T. \tag{1.4.10}$$

with $r = 1, ..., k$. The parameters $\phi_1, ..., \phi_k$ are the hypothesized k unknown frequencies. In analogy to (1.2.71), the least squares amplitude estimates are given by

$$\hat{\mathbf{s}} = (\mathbf{A}_k^H\mathbf{A}_k)^{-1}\mathbf{A}_k^H\mathbf{y}. \tag{1.4.11}$$

Substituting (1.4.11) into (1.4.7), we obtain the following error function in terms of the frequencies,

$$L_k \stackrel{\triangle}{=} L(\phi_1, ..., \phi_k) \stackrel{\triangle}{=} \mathbf{y}^H\mathbf{P}^{\perp}_{\mathbf{A}_k}\mathbf{y} \tag{1.4.12}$$

$$L_0 \stackrel{\triangle}{=} \mathbf{y}^H\mathbf{y} \tag{1.4.13}$$

where $\mathbf{P}^{\perp}_{\mathbf{A}_k}$ is defined in (1.2.73). The least squares frequency estimates $\hat{\phi}_1^{(k)}, ..., \hat{\phi}_k^{(k)}$ are defined as the minimizing argument of the k-variate function L_k (1.4.12) and calculated via a k-dimensional search. The minimum value of the cost function L_k, is denoted by

$$\hat{L}_k \stackrel{\triangle}{=} L_k(\hat{\phi}_1^{(k)}, ..., \hat{\phi}_k^{(k)}), \tag{1.4.14}$$

which will be called the least-squares error. Note that $\hat{L}_k$ is a monotonically decreasing function of k.

According to Theorem 1 in [11], in the case of high SNR ($\stackrel{\triangle}{=} \min_i |s_i|^2/\sigma^2$) and $M > 3I/2$,

$$\hat{L}_I \approx \frac{\sigma^2}{2}\chi^2_{2M-3I} \tag{1.4.15}$$

where $\approx$ denotes the first order approximation (i.e., keeping the nonzero terms of the lowest order). The expression (1.4.15) simply says that the least squares error $\hat{L}_I$ is χ^2 distributed with the $2M-3I$ degrees of freedom. The number $2M$ is the number of (real) data points, and $3I$ is the number of (real) unknowns (complex amplitudes and real frequencies). According to Theorem 2 in [11], at high SNR,

$$\hat{L}_k = \textit{ a positive value with negligible noise} \tag{1.4.16}$$

for $k = 0, ..., I-1$. Therefore for small σ^2, $\hat{L}_I$ is separable from $\hat{L}_k$ for $k = 0, \cdots, I-1$. Based on this observation, the LS method for finding I can be derived as follows. Define a threshold ϵ_k^{LS} such that

$$P(\mu_k \leq \epsilon_k^{\mathrm{LS}}) = \alpha \tag{1.4.17}$$

where $\mu_k = \frac{\sigma^2}{2}\chi^2_{2M-3k}$ and α $(0 < \alpha < 1)$ is a user chosen value of "confidence". Then we estimate I by

$$\hat{I}_{\mathrm{LS}} = \min\{k | \hat{L}_k \leq \epsilon_k^{\mathrm{LS}}\}. \tag{1.4.18}$$

For the high SNR case (small σ^2), the probability of underestimation is very small because of (1.4.16) and the fact that ϵ_k^{LS} is proportional to σ^2, and the probability of correct detection

$$P(\hat{I}_{\mathrm{LS}} = I) \approx P(\hat{L}_I \leq \epsilon_I^{\mathrm{LS}}) = \alpha. \tag{1.4.19}$$

Note that ϵ_k^{LS} can be obtained (off-line) using standard MATLAB software or by just table-lookup [74]. In contrast, the computation of a threshold $\epsilon_k^{\mathrm{NLS}}$ used in the NSL method is a more tedious work because the distribution of a weighted sum of χ_2^2 random variables is required.

Among the three techniques discussed so far, the LS method offers a much better SNR threshold although at a higher cost of computation. The computation required to obtain the least squares estimates of the frequencies and amplitudes can be reduced by exploiting some near-optimal methods. As shown in [11], the maximum likelihood estimates can be approximated by the matrix pencil estimates without sacrificing much of the performance. It should be noted that the determination of any of the thresholds ϵ^{ESV}, $\epsilon_k^{\mathrm{NSL}}$ and ϵ_k^{LS} requires the knowledge of the noise variance σ^2. If this information is not available, it has to be estimated before the detection of signals. For noise variance estimation, see [104] where a separate measurement of data is considered, or [110] where the same measurement is considered.

1.4.4 Variations of Information Theoretic Criteria

The AIC (Akaike Information Criterion) method and the MDL (Minimum Description Length) method are two popular detection methods. These two methods belong to a large group of methods that all use some penalized log-likelihood functions. For many practical situations, the data model is best expressed by a likelihood function

$f(\mathbf{Y}|\mathbf{\Omega}^{(k)})$ where $\mathbf{Y}$ is the (total) data vector and $\mathbf{\Omega}^{(k)}$ is the (total) parameter vector of the dimension $p(k)$. The integer k is often referred to as the hypothesized model order. In the context of frequency estimation, k is referred to as the hypothesized number of frequencies or signals. If the true value of k is known, the parameter vector that maximizes the likelihood function is the maximum likelihood estimation of the desired parameter vector. If the true value of k is unknown, the likelihood function also has a role to play in estimating (detecting) the desired value of k.

A penalized log-likelihood function has the following form:

$$d(k) = -\log f(\mathbf{Y}|\hat{\mathbf{\Omega}}^{(k)}) + p(k)C(M,N) \tag{1.4.20}$$

where $\hat{\mathbf{\Omega}}^{(k)}$ is the maximum likelihood estimate of the parameter vector of the assumed model order k (hypothesis H_k), and N is the total number of independent measurements in the data. The first term in $d(k)$ is simply the negative log-likelihood function with order k, and the second term $p(k)C(M,N)$ is the penalty term. Without the penalty term, $d(k)$ would be a decreasing function in general. So, with an increasing function of k as the penalty term, the minimum of $d(k)$ may correspond to the desired value of k. Among all methods in this group, it is the penalty term that makes each of the algorithms unique. Also, behind each of the different penalty terms, there is a theory and/or intuition.

The AIC and MDL methods were originally developed for multiple measurements. But for a single measurement, there exist some variations of the two. We now discuss the two cases: single measurement and multiple measurements, separately. For comparison purpose, we will also discuss two additional methods based on the eigenvalues of the data covariance matrix.

Given a single measurement of data (i.e., $N = 1$), the following choices of $d(k)$ have been discussed in the literature:

$$d_{\mathrm{BAY}(1)}(k) = (M-L)\log(\mathbf{y}^H\hat{\mathbf{P}}_k^{\perp}\mathbf{y}) + \frac{5k}{2}\log\frac{M}{L} \tag{1.4.21}$$

$$d_{\mathrm{EDC}}(k) = M\log\mathbf{y}^H\hat{\mathbf{P}}_k^{\perp}\mathbf{y} + \frac{3k}{2}C(M) \tag{1.4.22}$$

where M is the dimension of the measurement vector $\mathbf{y}$ which corresponds to a series of superimposed undamped complex exponentials, and $\hat{\mathbf{P}}_k^{\perp}$ is as defined in (1.2.73) except that k frequencies are assumed under H_k. The parameter L in (1.4.21) is a "user's choice" recommended in [19]. $d_{\mathrm{BAY}(1)}(k)$ can be derived by maximizing the following with respect to k:

$$f(\mathbf{y}_2|k,\mathbf{y}_1) = \frac{f(\mathbf{y}|k)}{f(\mathbf{y}_1|k)}. \tag{1.4.23}$$

where f is a corresponding probability density function, and $\mathbf{y}_1 = [y_0,...,y_{L-1}]^T$ and $\mathbf{y}_2 = [y_L,...,y_{M-1}]^T$ are the two subvectors of $\mathbf{y}$. The function $C(M)$ in (1.4.22)

provides a more flexible choice as recommended in [125]. In fact, there exists such a choice of $C(M)$ that the minimum of $d_{\mathrm{EDC}}(k)$ corresponds to the exact number I of frequencies when M is arbitrarily large. This is a consequence of the following result:

Theorem 1.7: *The $d_{\mathrm{EDC}}(k)$ is large-M strongly consistent*[10]*, if $C(M)$ satisfies the following conditions:*

$$\lim_{M\to\infty} \frac{C(M)}{M} = 0 \tag{1.4.24}$$

$$\lim_{M\to\infty} \frac{C(M)}{\log M} = \infty. \tag{1.4.25}$$

Proof: See [10].

The condition (1.4.24) ensures that I is not under-estimated when M is large. The condition (1.4.25) ensures that I is not over-estimated when M is large. It is easy to check that for $d_{\mathrm{BAY}(1)}(k)$, the second condition of the theorem is not satisfied, and hence $d_{\mathrm{BAY}(1)}(k)$ may not be large-M (strongly) consistent. But by choosing a smaller L, the chance of over-estimation is reduced.

The first term of $d(k)$ is often expensive in computation. To reduce the computation, some approximation is necessary. One simple approximation is to replace the maximum likelihood estimates of the parameters by some sub-optimal estimates such as the estimates by linear prediction, matrix pencil, MUSIC or ESPRIT. In [45], the notch periodogram is used to compute the frequency estimates.

1.5 Detection Techniques Using Multiple Measurements

When N is large (the large sample case), the unconditional model is the better choice for frequency estimation and detection. We will assume this model from now on.

1.5.1 Information Theoretic Criteria

For the large sample case, there are two popular detection schemes known as AIC (Akaike Information Criterion) and MDL (Minimum Description Length) methods:

$$d_{\mathrm{AIC}}(k) = -\log f(\mathbf{Y}|\hat{\mathbf{\Omega}}^{(k)}) + p(k) \tag{1.5.1}$$

$$d_{\mathrm{MDL}}(k) = -\log f(\mathbf{Y}|\hat{\mathbf{\Omega}}^{(k)}) + \frac{p(k)}{2}\log N. \tag{1.5.2}$$

[10] Estimates of I strongly converge to the true value as $M \to \infty$. The strong convergence is defined on p.21.

But a more general method is given previously in (1.4.20). It is shown in [125] that the general detection method is large-N strongly consistent[11], if

$$\lim_{N\to\infty} \frac{C(M,N)}{N} = 0, \tag{1.5.3}$$

$$\lim_{N\to\infty} \frac{C(M,N)}{\log\log N} = \infty. \tag{1.5.4}$$

It is clear that the AIC method does not meet the second condition above. Hence, the AIC method tends to over-estimate the model order. But the MDL method is strongly consistent.

The computation of $\log f(\mathbf{Y}|\hat{\mathbf{\Omega}}^{(k)})$ can be simplified if we assume that the covariance matrix $\mathbf{R}^{(k)}$ is parameterized as follows:

$$\mathbf{R}^{(k)} = E\{\mathbf{y}(n)\mathbf{y}^H(n)\} \triangleq \sum_{i=1}^{k}(\lambda_i - \sigma^2)\mathbf{u}_i\mathbf{u}_i^H + \sigma^2\mathbf{I} \tag{1.5.5}$$

where the vector of unknowns is defined as

$$\mathbf{\Omega}^{(k)} = [\lambda_1, ..., \lambda_k, \sigma^2, \mathbf{u}_1^T, ..., \mathbf{u}_k^T]^T. \tag{1.5.6}$$

Parameterized with $\mathbf{\Omega}^{(k)}$, the probability density function (the likelihood function) of the multiple measurements is given by

$$f(\mathbf{Y}|\mathbf{\Omega}_k) = \frac{1}{(\pi^M|\mathbf{R}^{(k)}|)^N}\exp(-\sum_{n=1}^{N}\mathbf{y}(n)^H(\mathbf{R}^{(k)})^{-1}\mathbf{y}(n)). \tag{1.5.7}$$

The maximum likelihood estimate of $\mathbf{\Omega}_k$ is given in [2] and cited below,

$$\hat{\lambda}_i = l_i, \quad i = 1, .., k \tag{1.5.8}$$

$$\hat{\sigma}^2 = \frac{1}{M-k}\sum_{i=k+1}^{M} l_i \tag{1.5.9}$$

$$\hat{\mathbf{u}}_i = \mathbf{c}_i, \quad i = 1, .., k \tag{1.5.10}$$

where $l_1 > l_2 > \cdots > l_M$ and $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_M$ are the eigenvalues and corresponding unitary eigenvectors of the sample covariance matrix $\hat{\mathbf{R}}$ in (1.3.5).

Substituting these ML estimates into (1.5.7), the maximum of the log-likelihood function is given by

$$\log f(\mathbf{Y}|\hat{\mathbf{\Omega}}^{(k)}) = -N\log|\hat{\mathbf{R}}^{(k)}| - \sum_{n=1}^{N}\mathbf{y}(n)^H(\hat{\mathbf{R}}^{(k)})^{-1}\mathbf{y}(n) \tag{1.5.11}$$

[11] Estimates of I strongly converge to the true value as $N \to \infty$. The strong convergence is defined on p.21.

where

$$\hat{\mathbf{R}}^{(k)} = \sum_{i=1}^{k}(l_i - \hat{\sigma}^2)\mathbf{u}_i\mathbf{u}_i^H + \hat{\sigma}^2\mathbf{I} \tag{1.5.12}$$

is the maximum likelihood estimate of $\mathbf{R}^{(k)}$ with the structure defined in (1.5.5). Using

$$|\hat{\mathbf{R}}^{(k)}| = (\prod_{i=1}^{k} l_i)(\hat{\sigma}^2)^{M-k} = (\prod_{i=1}^{M} l_i)(\hat{\sigma}^2)^{M-k}/(\prod_{i=k+1}^{M} l_i) \tag{1.5.13}$$

$$\sum_{n=1}^{N} \mathbf{y}(n)^H(\hat{\mathbf{R}}^{(k)})^{-1}\mathbf{y}(n) = NM, \tag{1.5.14}$$

(where the second equation holds for large N) and dropping the terms independent of k, we obtain the following approximation of the maximum log-likelihood function:

$$\log f(\mathbf{Y}|\hat{\mathbf{\Omega}}^{(k)}) = N(M-k)\log\left[(\prod_{i=k+1}^{M} l_i^{1/(M-k)})/(\frac{1}{M-k}\sum_{i=k+1}^{M} l_i)\right]. \tag{1.5.15}$$

Note that this function does not depend on the eigenvectors of the covariance matrix. With the above parameterization of $\mathbf{R}^{(k)}$ and the conjugate symmetry condition of $\mathbf{R}^{(k)}$, the number of free parameters is $p(k) = k+1+2Mk-k(k-1) = k(2M-k)+1$.

The random variable $(\prod_{i=k+1}^{M} l_i^{1/(M-k)})/(\frac{1}{M-k}\sum_{i=k+1}^{M} l_i)$ is known [2] to be χ^2 distributed. Based on this result, it is shown in [89], [125] that the MDL criterion is strongly consistent, and the AIC criterion tends to overestimate with a nonzero probability. Further analysis on the error probability of the AIC and MDL criteria is available in [113], [53].

If we consider the orthogonality among the eigenvectors of $\mathbf{R}^{(k)}$, the number of free parameters reduces from $k(2M-k)$ to $k(2M-k-1)+1$. In this case, a slight improvement was reported in [116]. In [119], it is suggested to use the forward-backward covariance matrix $\hat{\mathbf{R}}_{fb}$ (see (1.3.18)). Because $\hat{\mathbf{R}}_{fb}$ is a centro-Hermitian matrix, the number of free parameters in the eigenvectors is reduced to $p(k) = k(M+k+1)+1$. Also, since the eigenvalues of $\hat{\mathbf{R}}_{fb}$ are more reliable than $\hat{\mathbf{R}}$ with a given number of observations, the performance of the MDL method can be improved with the use of $\hat{\mathbf{R}}_{fb}$.

It is important to note that the parameterization of $\mathbf{R}^{(k)}$ is critical to the final outcome of the detection algorithm. There seems no optimal way to do the parameterization. Different parameterizations lead to different detection algorithms of varied performances. A different parameterization of $\mathbf{R}^{(k)}$ also implies a different level of complexity for computing the maximum log-likelihood function. Furthermore, it should be noted that almost all discussions shown in this chapter assume spatially white noise. But many of the principles can be extended to spatially colored noise. For examples, see [123].

1.5.2 Treating Eigenvalues as the Observations

In [118], a different strategy is applied. Since the probability density function $f(\mathbf{Y}|\mathbf{\Omega}^{(k)})$ only depends on the eigenvalues of $\mathbf{R}^{(k)}$ when N is large, one may consider these eigenvalues to be the sole observation for the detection purpose. Denote by $f(l_1, ..., l_M|\boldsymbol{\lambda}^{(k)})$ the probability density function of the eigenvalues $l_1, ..., l_M$ under H_k, where $\boldsymbol{\lambda}^{(k)} = [\lambda_1, \lambda_2, \cdots, \lambda_k, \sigma^2]^T$ is the parameter vector. An exact expression of the density function $f_k(\cdot)$ is however intractable due to the involvement of hypergeometric function. An asymptotic ($N >> 1$) expression of $f_k(\cdot)$ is developed by Chattopadhyay [18] and briefly described in [118]. This expression is

$$f(l_1, ..., l_M|\boldsymbol{\lambda}^{(k)}) = c_1 c_2 F_1 F_2 \tag{1.5.16}$$

where

$$c_1 = N^{MN-k(2M-k-1)/2}, \quad c_2 = \frac{\pi^{M(M-1)-k(2M-k-1)}}{\bar{\Gamma}_M(N)\bar{\Gamma}_{M-k}(M-k)} \tag{1.5.17}$$

$$F_1 = \exp\{-N[\sum_{i=1}^{k}\frac{l_i}{\lambda_i} + \sum_{i=k+1}^{M}\frac{l_i}{\sigma^2}]\} \cdot [\prod_{i=1}^{k}\frac{l_i^{N-M}}{\lambda_i^N}\prod_{i=k+1}^{M}\frac{l_i^{N-M}}{(\sigma^2)^N}] \tag{1.5.18}$$

$$F_2 = [\prod_{i=1,i<j}^{k}\alpha_{ij} \cdot \prod_{i=1}^{k}\prod_{j=k+1}^{M}\beta_{ij}] \cdot \prod_{i=1,i<j}^{M}(l_i - l_j)^2 \tag{1.5.19}$$

with $\alpha_{ij} = \lambda_i\lambda_j/[(\lambda_i - \lambda_j)(l_i - l_j)]$, $\beta_{ij} = \lambda_i\sigma^2/[(\lambda_i - \sigma^2)(l_i - l_j)]$ and $\bar{\Gamma}_m(k) = \pi^{m(m-1)/2}\prod_{i=1}^{m}\Gamma(n-i+1)$.

In the density function (1.5.16), there are $k+1$ free parameters. Then the AIC and MDL criteria become (with the constant term dropped)

$$AIC(k) = -\log f(l_1, ..., l_M|\hat{\boldsymbol{\lambda}}^{(k)}) + k \tag{1.5.20}$$

$$MDL(k) = -\log f(l_1, ..., l_M|\hat{\boldsymbol{\lambda}}^{(k)}) + (k/2)\log N \tag{1.5.21}$$

where $\hat{\boldsymbol{\lambda}}^{(k)}$ denotes the maximum likelihood estimate of $\boldsymbol{\lambda}^{(k)}$, which has to be found via a costly search procedure.

1.5.3 Thresholding Eigenvalues

An eigen-threshold approach is introduced in [8]. This approach is based on a large-sample property of the eigenvalues of the (large sample) covariance matrix. Define

$$\hat{\sigma}_k^2 = \frac{1}{M-k}\sum_{i=k+1}^{M} l_i. \tag{1.5.22}$$

According to [46], [27], for $k \geq k_0$ under H_{k_0}

$$\sqrt{N(M-k)}(\frac{\hat{\sigma}_k^2}{\sigma^2} - 1) \sim \mathcal{N}\{0, 1\} \tag{1.5.23}$$

where $\mathcal{N}\{m, r\}$ stands for a univariate complex normal (Gaussian) pdf with mean m and variance r. Let μ be a threshold such that $\int_{-\mu}^{\mu} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \alpha$ where α is a user specified confidence interval. Then, with the confidence α, for $k \geq k_0$ under H_{k_0},

$$1 - \frac{\mu}{\sqrt{N(M-k)}} \leq \frac{\hat{\sigma}_k^2}{\sigma^2} \leq 1 + \frac{\mu}{\sqrt{N(M-k)}}. \tag{1.5.24}$$

A simple detection method would be to decide on H_{k_0} if k_0 is the first integer between $[0, M]$ such that the above condition holds. But this method requires the knowledge of σ. To develop an alternative method, we now consider that for $k-1 \geq k_0$ under H_{k_0}, with the confidence α,

$$1 - \frac{\mu}{\sqrt{N(M-k+1)}} \leq \frac{\hat{\sigma}_{k-1}^2}{\sigma^2} \leq 1 + \frac{\mu}{\sqrt{N(M-k+1)}}, \tag{1.5.25}$$

Combining (1.5.24) and (1.5.25), we have that for $k-1 \geq k_0$ under H_{k_0}, with a confidence no less than α,

$$\frac{\hat{\sigma}_{k-1}^2}{\hat{\sigma}_k^2} \leq \frac{1 + \mu/\sqrt{N(M-k+1)}}{1 - \mu/\sqrt{N(M-k)}}. \tag{1.5.26}$$

From the definition of $\hat{\sigma}_k^2$, it is easy to obtain

$$\begin{aligned} \hat{\sigma}_{k-1}^2 = \frac{1}{M-k+1} \sum_{i=k}^{M} l_i &= \frac{1}{M-k+1} [\sum_{i=k+1}^{M} l_i + l_k] \\ &= \frac{M-k}{M-k+1} \hat{\sigma}_k^2 + \frac{l_k}{M-k+1}. \end{aligned} \tag{1.5.27}$$

Combining (1.5.27) and (1.5.26) leads to that for $k-1 \geq k_0$ under H_{k_0}, with a confidence no less than α,

$$l_k \leq \left[(M-k+1) \frac{1 + \mu/\sqrt{N(M-k+1)}}{1 - \mu/\sqrt{N(M-k)}} - (M-k) \right] \hat{\sigma}_k^2 \triangleq l_k^{upper}. \tag{1.5.28}$$

We can use this inequality to decide on H_{k-1} if k is the first integer between 0 and M such that the above inequality holds. This eigenvalue thresholding method generally outperform both the AIC and MDL methods [8].

1.5.4 Bayesian Approach

It is known as the Bayesian approach to estimate an unknown by maximizing its probability density/mass function conditioned upon the observations. If the prior probability density/mass function of the unknown is constant (or unknown), the Bayesian approach is equivalent to the maximum likelihood approach. But when the unknown is the model

order, the maximum likelihood estimate of the model order would be infinity in general as the likelihood function generally is an increasing function of the model order. To avoid this problem, a mixed likelihood function is developed in [12]. This mixed likelihood function is simply the ratio of the likelihood function of the total observation over the likelihood function of a partial observation. It is clear that this ratio is generally no longer a monotonically increasing function.

Specifically, it is proposed in [12] that the number of signals should be found as follows:

$$\arg \min_{k=0,1,...,M-1} -\log \frac{f(\mathbf{Y}|k)}{f(\mathbf{Y}_1|k)} \tag{1.5.29}$$

where $\mathbf{Y}_1 = [\mathbf{y}(1), ..., \mathbf{y}(L)]$ is the partial observation, $1 < L < N$, and the $f(\dot{|}k)$ functions represent the probability density functions given the order k. This is equivalent to

$$\arg \min_{k=0,1,...,M-1} -\log f(\mathbf{Y}_2|k, \mathbf{Y}_1) \tag{1.5.30}$$

where $\mathbf{Y}_2 = [\mathbf{y}(L+1), ..., \mathbf{y}(N)]$. The above density function is marginalized over all other parameters, i.e.,

$$f(\mathbf{Y}_2|k, \mathbf{Y}_1) = \frac{\int_{\boldsymbol{\theta}^{(k)}} \int_{\boldsymbol{\phi}^{(k)}} f(\mathbf{Y}|\boldsymbol{\theta}^{(k)}, \boldsymbol{\phi}^{(k)}, k) f(\boldsymbol{\phi}^{(k)}|k) d\boldsymbol{\phi}^{(k)} f(\boldsymbol{\theta}^{(k)}|k) d\boldsymbol{\theta}^{(k)}}{\int_{\boldsymbol{\theta}^{(k)}} \int_{\boldsymbol{\phi}^{(k)}} f(\mathbf{Y}_1|\boldsymbol{\theta}^{(k)}, \boldsymbol{\phi}^{(k)}, k) f(\boldsymbol{\phi}^{(k)}|k) d\boldsymbol{\phi}^{(k)} f(\boldsymbol{\theta}^{(k)}|k) d\boldsymbol{\theta}^{(k)}} \tag{1.5.31}$$

where $\boldsymbol{\phi}^{(k)}$ consists of σ^2 and the independent unknowns in $\mathbf{R}_s$, and k frequencies $\boldsymbol{\theta}^{(k)} = [\theta_1^{(k)}, ..., \theta_k^{(k)}]^T$.

With some approximations of the prior distributions of $\boldsymbol{\phi}^{(k)}$ and $\boldsymbol{\theta}^{(k)}$, it is shown in [12] that the cost function of the above detector becomes

$$d_{\mathrm{BAY}(2)}(k) = N \log f(\mathbf{Y}|\hat{\boldsymbol{\Omega}}^{(k)}) + P(k) \tag{1.5.32}$$

where

$$P(k) = \frac{N}{N-L} \log \frac{d(k, M, N)}{d(k, L, N)} \tag{1.5.33}$$

$$d(k, M, N) = \frac{\pi^{k(k-1)/2}}{2(\pi N)^{MN}} \frac{\Gamma(N(M-k)) \prod_{i=0}^{k-1} \Gamma(N-i)}{(M-k)^{N(M-k)}}. \tag{1.5.34}$$

It is recommended in [12] that L is chosen to be $M-1$. This is obviously the largest value that one can choose for L while preserving a dependence of the cost function on the model order k. In fact, this choice is also very meaningful in that given almost all the observation ($\mathbf{Y}_1$), the order I should be chosen such that its likelihood is maximum with respect to the additional observation ($\mathbf{Y}_2$). When N is large, $P(k) \to \frac{1}{2}k(k+1)\log N$ which coincides with the MDL method. Simulations suggest that the Bayesian method performs better than the MDL method generally.

1.6 Conclusions

In this chapter, we have reviewed a range of estimation and detection methods for the sensor array model and in particular for the model of superimposed exponentials. Each of these methods has a different balance between computational complexity and performance. The choice of a particular method depends on the needs and constraints in practice. This chapter should also have given the readers a feel of how much has been done in this field. Some of the latest advances and open issues will be discussed in later chapters.

Acknowledgements

Qi Cheng's work was supported by the 2000 UWS Nepean Research Grant Scheme and 2001 UWS Study Leave Program.

REFERENCES

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *Proc. 2nd International Symposium on Information Theory*, pp. 267-281, Tsahkadsor, Armenia, USSR, 2-8 September 1971 (Imprint 1973).

[2] T. W. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Stat.*, vol. 34, pp. 122-148, 1963.

[3] R. Bachl, "The forward-backward averaging technique applied to TLS-ESPRIT processing," *IEEE Trans. Signal Processing*, vol. 43, no. 11, pp. 2691-2699, Nov. 1995.

[4] Z. Banjanin, J. R. Cruz, D.S. Zrnic, "A unified approach to three eigen-decomposition methods for frequency estimation," *IEEE Trans. Signal Processing*, vol. 40, no. 1, pp. 214-218, Jan. 1992.

[5] A. J. Barabell, "Improving the resolution performance of eigenstructure-based direction finding algorithms," *Proc. 1983 IEEE Intl. Conf. on Acoustics, Speech, Signal Processing*, vol. 4, pp. 336-339, 1983.

[6] Y. Bresler and A. Macovski, "Exact maximum likelihood parameter estimation of superimposed exponential signals in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 10, pp. 1081-1089, Oct. 1986.

[7] Y. Bresler and A. Macovski, "On the number of signals resolvable by a uniform linear array," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1361-1375, Dec. 1986.

[8] W. Chen, K. M. Wong and J. P. Reilly, "Detection of the number of signals: a predicted eigen-threshold approach," *IEEE Trans. Signal Processing*, vol. 39, no. 5, pp. 1088-1098, May 1991.

[9] Q. Cheng and Y. Hua, "Performance analysis of the MUSIC and Pencil-MUSIC algorithms," *IEEE Trans. Signal Processing*, vol. 42, no. 11, pp. 284-299, Nov. 1994.

[10] Q. Cheng and Y. Hua, "Consistency of two detection techniques using single measurement of data," *Signal Processing*, vol. 66, no. 1, pp. 61-77, April 1998.

[11] Q. Cheng and Y. Hua, "Detection of cisoids using least square error function," *IEEE Trans. Signal Processing*, vol. 45, no. 6, pp. 1584-1590, June 1997.

[12] C. M. V. Cho and P. M. Djurić, "Detection and estimation of DOA's of signals via Bayesian predictive densities," *IEEE Trans. Signal Processing*, vol. 42, no. 11, pp. 3051-3060, Nov. 1994.

[13] Y. H. Choi, "On conditions for the rank restoration in forward/backward spatial smoothing," *IEEE Trans. Signal Processing*, vol. 50, no. 11, pp. 2900-2901, Nov. 2002.

[14] N. Chotikakamthron and J. A. Chambers, "IQML algorithm for multiple signal parameter estimation," *IEE Proc. F - Radar, Sonar and Navigation*, vol. 144, no. 5, pp. 237-244, Oct. 1997.

[15] Y. S. Chow and H. Teicher, *Probability Theory*: New York: Springer-Verlag, 1988.

[16] M. P. Clark and L. L. Scharf, "Two-dimensional modal analysis based on maximum likelihood," *IEEE Trans. Signal Processing*, vol. 42, no. 6, pp. 1443-1452, June 1994.

[17] M. P. Clark and L. L. Scharf, "On the complexity of IQML algorithms," *IEEE Trans. Signal Processing*, vol. 40, no. 7, pp. 1811-1813, July 1992.

[18] A. K. Chattopadhyay, "An asymptotic distribution theory and applications in multivariate analysis," *Tech. Report 256*, Dept. of Statistics, Purdue University, 1971.

[19] P. M. Djurić, "Simultaneous detection and frequency estimation of sinusoidal signals," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Processing*, vol. IV, pp. 53-56, Minneapolis, Minnesota, 27-30 April 1993.

[20] E. M. Dowling and R. D. DeGroat, "The equivalence of the total least squares and minimum norm methods," *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1891-1892, Aug. 1991.

[21] J. E. Evans, J. R. Johnson and D. F. Sun, "High resolution angular spectrum estimation techniques for terrain scattering analysis and angle of arrival estimation," *Proc. 1st ASSP Workshop on Spectral Estimation*, pp. 134-139, Hamilton, Canada, 1981.

[22] B. Friedlander and A. J. Weiss, "Direction finding using noise covariance modeling," *IEEE Trans. Signal Processing*, vol. 43, no. 7, pp. 1557-1567, July 1995.

[23] A. Gershman, P. Stoica, M. Pesavento and E. Larsson, "Stochastic Cramér-Rao bound for direction estimation in unknown noise fields," *IEE Proc.–Radar, Sonar and Navigation*, vol. 149, no. 1, pp. 2-8, Feb. 2002.

[24] M. J. Goldstein, "Reduction of the eigenproblem for Hermitian persymmetric matrices," *Mathemat. Comput.*, vol. 28, pp. 237-238, Jan. 1974.

[25] G. H. Golub and C. F. Van Loan, *Matrix Computation*. Baltimore, MD: John Hopkins University Press, 1983.

[26] G. H. Golub and C. F. Van Loan, "An analysis of the total least squares problem," *SIAM Journal Numer. Anal.*, vol. 17, no.6, pp. 883-893, Dec. 1980.

[27] N. R. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction)," *Ann. Math. Stat.*, vol. 34, pp. 152-176, 1963.

[28] M. Haardt and J. A. Nossek, "Unitary ESPRIT: How to obtain increased estimation accuracy with a reduced computational burden," *IEEE Trans. Signal Processing*, Vol.43, no. 5, pp. 1232-1242, May 1995.

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

[30] Y. H. Hu and B. C. Phan, "Frequency estimation error in Pisarenko harmonic decomposition method," *Proc. IEEE*, vol. 76, no. 1, pp. 82-84, Jan. 1988.

[31] Y. Hua, "Estimating two-dimensional frequencies by matrix enhancement and matrix pencil," *IEEE Trans. Signal Processing*, vol. 40, no. 9, pp. 2267-2280, Sept. 1992.

[32] Y. Hua, "High resolution imaging of continuously moving object using steeped frequency radar," *Signal Processing*, vol. 35, no. 1, pp. 33-40, Jan. 1994.

[33] Y. Hua, "The most efficient implementation of the IQML algorithm," *IEEE Trans. Signal Processing*, Vol. 42, no. 8, pp. 2203-2204, Aug. 1994.

[34] Y. Hua and K. Abed-Meraim, "Techniques of eigenvalues estimation and association," *Digital Signal Processing*, vol. 7, pp. 253-259, 1997.

[35] Y. Hua, M. Nikpour and P. Stoica, "Optimal reduced rank estimaiton and filtering," *IEEE Trans. Signal Processing*, vol. 49, no. 3, pp. 457-469, March 2001.

[36] Y. Hua and T. K. Sarkar, "Statistical analysis of three high-resolution techniques for radio directional estimation," *Proc. IEEE Workshop on Spectral Estimation and Modeling*, pp. 161-164, Boston, MA, Nov 1986.

[37] Y. Hua and T. K. Sarkar, "Perturbation analysis of the TK method for harmonic retrieval problem," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 2, pp. 228-240, Feb. 1988.

[38] Y. Hua and T. K. Sarkar, "Subspace linear prediction approach to extracting poles," *Proc. IEEE ASSP 4th Workshop on Spectral Estimation and Modeling*, pp. 367-370, Minneapolis, MN, USA, Aug. 1988.

[39] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 5, pp. 814-824, May 1990.

[40] Y. Hua and T. K. Sarkar, "A perturbation property of the TLS-LP method," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 11, pp. 2004-2005, Nov. 1990.

[41] Y. Hua and T. K. Sarkar, "On the total least squares linear prediction method for frequency estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 12, pp. 2186-2189, Dec. 1990.

[42] Y. Hua and T. K. Sarkar, "On SVD for estimating generalized eigenvalues of singular matrix pencil in noise," *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 892-900, April 1991.

[43] Y. Hua and T. K. Sarkar, "A note on the Cramér-Rao bound for 2-D direction finding based on 2-D array," *IEEE Trans. Signal Processing*, vol. 39, no. 5, pp. 1215-1218, May 1991.

[44] K. C. Huarng and C. C. Yeh, "A unitary transformation method for angle of arrival estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, no. 4, pp. 975-977, April 1989.

[45] J. K. Hwang and Y. C. Chen, "A combined detection-estimation algorithm for harmonic-retrieval," *Signal Processing*, vol. 30, no. 2, pp. 177-197, Jan. 1993.

[46] A. T. James, "Distributions of matrix values and latent roots derived from normal samples," *Ann. Math. Stat.*, vol. 35, pp. 475-501, 1964.

[47] M. Jansson and P. Stoica, "Forward-only and forward-backward sample covariances-A comparative study," *Signal Processing*, vol. 77, pp. 235-245, 1999.

[48] V. K. Jain, "Filter analysis by use of pencil-of-functions: Part I," *IEEE Trans. Circuits and Syst.*, vol. 21, no. 5, pp. 574-579, Sept. 1974.

[49] V. K. Jain, "Filter analysis by use of pencil-of-functions: Part II," *IEEE Trans. Circuits and Syst.*, vol. 21, no. 5, pp. 580-583, Sept. 1974.

[50] D. H. Johnson and S. R. deGraff, "Improving resolution of bearing in passive sonar arrays by eigenvalue analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no. 8, pp. 638-647, Aug. 1982.

[51] M. Kaveh and A. Bassias, "Threshold extension based on a new paradigm for MUSIC-type estimation," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Processing*, vol. 5, pp. 2535-2538, 3-6 April 1990, Albuquerque, NM, USA.

[52] M. Kaveh and A. J. Barabell, "The statistical performance of the MUSIC and minimum-norm algorithms in resolving plane waves in noise," *IEEE Trans. Signal Processing*, vol. 34, no. 4, pp. 331-341, April 1986.

[53] M. Kaveh, H. Wang and H. Hung, "On the theoretical performance of a class of estimators of the number of narrow-band sources," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 9, pp. 1350-1352, Sept. 1987.

[54] I. Kirsteins, "Analysis of Reduced Rank Interference Cancellation," Ph.D. Thesis, University of Rhode Island, Kingston, RI, 1990.

[55] R. Kumaresan, "Estimating the parameters of exponentially damped/undamped sinusoidal signals in noise", Ph.D. dissertation, Univ. of Rhode Island, Kingston, RI, Aug. 1982.

[56] R. Kumaresan, L. L. Scharf and A. K. Shaw, "An algorithm for pole-zero modeling and spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.34, no. 6, pp. 637-640, June 1986.

[57] R. Kumaresan and D.W. Tufts, "Estimating parameters of exponentially damped/undamped sinusoidal signals in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no.12, pp.833-840, Dec. 1982.

[58] R. Kumaresan and D. W. Tufts, "Estimating angles of arrival of multiple plane waves," *IEEE Trans. Aerospace and Electron. Syst.*, vol. 19, no. 1, pp.134-149, Jan. 1983.

[59] R. Kumaresan, D. W. Tufts and L. L. Scharf, "A Prony method for noisy data: Choosing the signal components and selecting the order in exponential signal models," *Proc. IEEE*, vol. 72, no.2, pp. 230-233, Feb. 1984.

[60] K. Konstantinides and K. Yao, "Statistical analysis of effective singular values in matrix rank determination," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 5, pp. 757-763, May 1988.

[61] S. Y. Kung, K. S. Arun and B. D. Rao, "State space and singular value decomposition based methods for the harmonic retrieval problem," *J. Opt. Soc. Amer.*, vol. 73, no. 12, pp.1791-1811, Dec. 1983.

[62] H. B. Lee and M. S. Wengrovitz, "Statistical characterization of the MUSIC spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, no. 6, pp. 1333-1347, June 1991.

[63] F. Li, R. J. Vaccaro and D. W. Tufts, "Unified performance analysis of subspace estimation algorithms," *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Processing*, pp. 2575-2578, 1990.

[64] F. Li, R. J. Vaccaro and D. W. Tufts, "Performance analysis of the state-space realization (TAM) and ESPRIT algorithms for DOA estimation," *IEEE Trans. Antennas and Propagation*, vol. 39, no. 3, pp. 418-423, March 1991.

[65] J. Li, P. Stoica, D. Zheng, "An efficient algorithm for two-dimensional frequency estimation," *Multidimensional Systems and Signal Processing*, vol. 7, pp. 151-178, 1996.

[66] J. Li, P. Stoica, T. Soderstrom, "On the consistency of IQML," *Signal Processing*, vol. 56, no. 2, pp. 185-190, January 1997.

[67] D. A. Linebarger, R. D. DeGroat and E. M. Dowling, "Efficient direction finding methods employing forward-backward averaging," *IEEE Trans. Signal Processing*, vol. 43, no. 8, pp. 2136-2145, August 1995.

[68] C. P. Mathews and M. Zoltwoski, "Performance analysis of UCS-ESPRIT algorithm for circular ring arrays," *IEEE Trans. Signal Processing*, vol. 42, no. 9, pp. 2535-2539, Sept. 1995.

[69] V. Nagesha and S. Kay, "On frequency estimation with the IQML algorithm," *IEEE Trans. Signal Processing*, vol. 42, no. 9, pp. 2509-2513, Sept. 1994.

[70] B. Ottersten and L. Ljung, "Asymptotic results for sensor array processing," *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Processing*, vol. 37, no. 5, pp. 2266-2269, May 1989.

[71] B. Ottersten, M. Viberg and T. Kailath, "Performance analysis of the total least squares ESPRIT algorithm," *IEEE Trans. Signal Processing*, vol. 39, no. 5, pp. 1122-1135, May 1991.

[72] B. Ottersten, M. Viberg and T. Kailath, "Analysis of subspace fitting and ML techniques for parameter estimation from sensor array data," *IEEE Trans. Signal Processing*, vol. 40 , no.3, pp. 590-600, March 1992.

[73] A. Paulraj, R. Roy and T. Kailath, "Estimation of signal parameters via rotational invariance techniques - ESPRIT," in Proc. 19th Asilomar Conference on Circuits, Systems and Computers, Asilomar, CA, Nov. 1985.

[74] E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, vol. I. Cambridge University Press, 1962.

[75] M. Pesavento and A. Gershman, "Maximum-likelihood direction of arrival estimation stochastic Cramér-Rao bound for direction estimation in the presence of unknown noise," *IEEE Trans. Signal Processing*, vol. 49, no. 7, pp. 1310-1324, July 2001.

[76] S. U. Pillai and B. H. Kwon, "Forward-backward smoothing techniques for coherent signal identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, pp. 8-15, Jan. 1989.

[77] S. U. Pillai and Y. Lee, "Coherent signal classification using symmetry considerations Performance," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, pp. 135-138, Jan. 1989.

[78] S. U. Pillai and B. H. Kwon, "Performance analysis of MUSIC-type high resolution estimators for direction finding in correlated and coherent scenes," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 8, pp. 1176-1189, Aug. 1989.

[79] V. F. Pisarenko, "The retrieval of harmonics from covariance function," *Geophysics. J. Royal Astronom. Soc.*, vol. 33, pp. 347-366, 1976.

[80] B. Porat and B. Friedlander, "A modification of Kumaresan-Tufts method for estimating rational impulse responses," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 34, no. 5, pp. 1336-1338, Oct. 1986.

[81] M.A. Rahman and K. B. Yu, "Total least squares approach for frequency estimation using linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 10, pp.1440-1454, Oct. 1987.

[82] R. Rajagopal and P. Ramakrishna, "DOA estimation with unknown noise fields: a matrix decomposition method," *IEE Proc. F- Radar Signal Processing*, vol. 138, no. 5, pp.495-501, Oct. 1991.

[83] B. D. Rao and K. V. S. Hari, "Statistical performance analysis of the minimum-norm method," *IEE Proc. F - Radar Signal Processing*, vol. 136, no. 3, pp. 125-134, June 1989.

[84] B. D. Rao and K. V. S. Hari, "Performance analysis of Root-MUSIC," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 1939-1949, Dec. 1989.

[85] B. D. Rao and K. V. S. Hari, "Performance analysis of ESPRIT and TAM in determining the direction of arrival of plane waves in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 1990-1995, Dec. 1989.

[86] B. D. Rao, "Relationship between matrix pencil and state space based harmonic retrieval methods," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 1, pp. 177-179, Jan. 1990.

[87] B. D. Rao and K. V. S. Hari, "Weighted subspace methods and spatial smoothing: analysis and comparison," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 788-803, Feb. 1993.

[88] W. J. L. Read, "Improving threshold performance of the IQML algorithm," *IEEE Trans. Signal Processing*, vol. 48, no. 9, pp. 2662-2665, Sept. 2000.

[89] J. Rissanen, "Modelling by shortest data description," *Automatica*, vol. 14, pp. 465-471, 1978.

[90] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 7, pp. 984-995, July 1989.

[91] H. Sakai, "Statistical Analysis of Pisarenko's method for sinusoidal frequency estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 2, pp. 95-101, Feb. 1984.

[92] R. O. Schmidt, "Multiple emitter location Signal parameter estimation," in *Proc. RADC Spectral Estimation Workshop*, Rome, NY, pp. 243-258, 1979.

[93] T. J. Shan, M. Wax and T. Kailath,"On spatial smoothing for estimation of coherent signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 8, pp. 806-811, Aug. 1985.

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

[95] P. Stoica, E. G. Larsson and A. Gershman, "The stochastic CRB for array processing–A textbook derivation," *IEEE Signal Processing Letters*, vol. 8, no. 5, pp. 148-150, May 2001.

[96] P. Stoica and A. Nehorai, "MODE, maximum likelihood, and Cramér-Rao bound: conditional and unconditional results," Center Syst. Sci., Yale University, New Haven, CT, Rep. 8901, Jan. 1989.

[97] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramér-Rao bound," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 5, pp. 720-741, May 1989.

[98] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 10, pp. 1783-1795, Oct. 1990.

[99] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramér-Rao bound: further results and comparisons," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 12, pp. 2140-2150, Dec. 1990.

[100] P. Stoica and A. Nehorai, "Performance comparison of subspace rotation and MUSIC methods for direction estimation," *IEEE Trans. Signal Processing*, vol. 39, no. 2, pp. 446-453, Feb. 1991.

[101] P. Stoica and K. C. Sharman, "Novel eigenanalysis method for direction estimation," *IEE Proc. F - Radar Signal Processing*, vol. 137, no. 1, pp. 19-26, Feb. 1990.

[102] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 7, pp. 1132-1143, July 1990.

[103] P. Stoica and T. Söderström, "Statistical analysis of MUSIC and subspace rotation estimates of sinusoidal frequencies," *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1836 -1847, Aug. 1991

[104] P. Stoica, "On estimating the noise power in array processing," *Signal Processing*, vol. 26, pp. 205-220, 1992.

[105] P. Stoica, A. Eriksson, T. Beron and K. D. Ward, "DOA estimation method for unknown noise fields," *IEE Proc. F - Radar, Sonar and Navigation*, vol. 141, no. 3, pp. 149-150, June 1994.

[106] P. Stoica and B. Ottersten, M. Viberg and R. L. Moses, "Maximum likelihood array processing for stochastic coherent sources," *IEEE Trans. Signal Processing*, vol. 44, no. 1, pp. 96- 105, Jan. 1996.

[107] A. Swindlehurst and T. Kailath, "Azimuth/elevation direction finding using regular array geometries," *IEEE Trans. Aerospace and Electron. Syst.*, vol. 29, no. 1, pp. 145-155, Jan. 1993.

[108] D. W. Tufts and R. Kumaresan, "Frequency estimation of multiple sinusoids: Making linear prediction perform like maximum likelihood," *Proc. IEEE*, vol. 70, no. 9, pp. 975-989, Sept. 1982.

[109] D. W. Tufts and C. D. Melissinos, "Simple, effective computation of principal eigenvectors and their eigenvalues and application to high-resolution estimation of frequencies," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 5, no. 10, pp. 1046-1052, Oct. 1986.

[110] D. W. Tufts and A. A. Shah, "Rank determination in time series analysis," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Processing*, vol. 4, pp. 21-25, April 1994, Adelaide, Australia.

[111] R. J. Vaccaro and Y. Ding, "A new state-space approach for direction finding," *IEEE Trans. Signal Processing*, vol. 42, no. 11, pp.3234-3237, Nov. 1994.

[112] M. Viberg and B. Ottersten, "Sensor array processing based on subspace fitting," *IEEE Trans. Signal Processing*, vol. 39, no. 5, pp. 1110-1121, May 1991.

[113] H. Wang and M. Kaveh, "On the performance of signal subspace processing – Part I: Narrow band systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 5, pp. 1201-1209, May 1986.

[114] M. Wax and T. Kailath, "Detection of signals by information-theoretic-criterion," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 387-392, April 1985.

[115] M. Wax and I. Ziskind, "On unique localization of multiple sources by passive sensor arrays," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 7, pp. 996-1000, July 1989.

[116] D. B. Williams, "Counting the degrees of freedom when using AIC and MDL to detect signals," *IEEE Trans. Signal Processing*, vol. 42, no. 11, pp. 3282-3284, Nov. 1994.

[117] G. M. Wnnz, "Acoust. of ambient noise: spectra and sources," *J. Acoust. Soc. Am.*, vol. 30, pp. 1936- 1956, 1962.

[118] K. M. Wong, Q. T. Zhang, J. P. Reilly and P. C. Yip, "On information theoretic criteria for determining the number of signals in high resolution array processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 11, pp. 1959-1971, Nov. 1990.

[119] G. Xu, R. Roy and T. Kailath, "Detection of number of sources via exploitation of centro-symmetry property," *IEEE Trans. Signal Processing*, vol. 42, no. 1, pp. 102-112, Jan. 1994.

[120] G. Xu and T. Kailath, "Fast subspace decomposition," *IEEE Trans. Signal Processing*, vol. 42, no. 3, pp. 539-551, March 1994.

[121] W. Xu and M. Kaveh, "Design of two MUSIC-like estimators based on bias minimization," *IEEE Trans. Signal Processing*, vol. 44, no. 9, pp. 2284-2299, Sept. 1996.

[122] X. L. Xu and K. M. Buckley, "Bias analysis of the MUSIC location estimator," *IEEE Trans. Signal Processing*, vol. 40, no. 10, pp. 2559-2569, Oct. 1992.

[123] Q. T. Zhang and K. M. Wong, "Information theoretic criteria for the determination of the number of signals in spatially correlated noise," *IEEE Trans. Signal Processing*, vol. 41, no. 4, pp. 1652-1663, April 1993.

[124] Q. T. Zhang, "Probability of resolution of the MUSIC algorithm," *IEEE Trans. Signal Processing*, vol. 43, no. 4, pp. 978-987, April 1995.

[125] L. C. Zhao, P. R. Krishnaiah and Z. D. Bai, "On detection of the number of signals in presence of white noise," *J. Multivariate Analysis*, vol. 20, no. 1, pp. 1-25, Oct. 1986.

[126] C. Zhou, F. Haber and D. L. Jaggard, "A resolution measure for the MUSIC algorithm and its application to plane wave arrivals contaminated by coherent interference," *IEEE Trans. Signal Processing*, vol. 39, no. 2, pp. 454-463, Feb. 1991.