Advanced Macromodeling Algorithm for Sampled Time/Frequency Domain Measured/Tabulated Data

by

Seyed-Behzad Nouri, B.E.,

A thesis submitted to the Faculty of Graduate Studies and Research
In partial fulfillment of the requirements for the degree of
Master of Applied Sciences

Ottawa-Carleton Institute for Electrical Engineering
Department of Electronics
Faculty of Engineering
Carleton University
Ottawa, Ontario, Canada

Your file Votre référence
ISBN: 978-0-494-36830-5
Our file Notre référence
ISBN: 978-0-494-36830-5

Abstract

In deep-submicron VLSI design, as signal rise times drop into the sub-nanosecond range, signal integrity analysis consistently demands efficient modeling and simulation of passive structures, such as packages and interconnects. Accordingly, efficient algorithms allowing accurate and compact (low-order) macromodeling based on the measured or simulated port-to-port response have recently attracted more attention.

The frequency-domain vector fitting (VF) techniques have become a popular tool for system identification, when frequency-domain observed data is available from measurements or EM simulators. Similarly, time-domain VF techniques have also been developed to create macromodels by utilizing the available time-domain sampled data.

In this thesis, an advanced macromodeling tool based on z-domain orthonormal vector fitting (ZD-OVF) is proposed. This novel multiport vector fitting method provides accurate and compact models for linear sub-networks using either frequency-domain or time-domain tabulated data. Utilizing the new z-domain orthonormal basis functions remarkably improves the numerical conditioning and robustness of resulting system equations. The proposed approach leads to significantly better-conditioned equations even when the initial choice of starting poles is not optimal. The new algorithm was applied to many application examples, and the results confirm that ZD-OVF exhibits efficient computation and produces accurate approximants.

Dedicated:

To my parents, my wife, and my son.

Acknowledgments

I would sincerely like to express my appreciation to my supervisors, Professors Michel Nakhla and Ram Achar, for introducing me to the world of CAD, wherein the exactness of mathematics, as pure science, combines with the practicality of engineering applications to make a glorious area worth discovering. I have learnt from Professor Nakhla many aspects of science and life. In particularly, learning research methodology has proved an invaluable experience. I am also grateful to Professor Achar, for his helpful suggestions and guidance, which was crucial in many stages of the research for this thesis. Most of all, I wish to thank him for his motivation and encouragement.

I would like to thank Dr. D. Saraswat for all the useful discussions on vector fitting and my fellow colleagues who were always readily available for some friendly deliberations. This made my graduate life most enjoyable. I will always fondly remember their support and friendship.

I am thankful towards the staff of the Department of Electronics in Carleton University for having been so helpful, supportive, and resourceful.

I am eternally indebted to my wife and my son for their unconditional, invaluable and relentless support, encouragement, patience and respect. My final thoughts are with my parents and other family members whose love and understanding have been an excellent motivation. I believe that I could not have achieved this without their unlimited sacrifice. This is for them.

Thank you,

Table of Contents

List of Figures

Glossary of Terms

TD	Discrete Time Domain
FD	Continuous-Time Frequency-Domain
ZD	z-Domain (or Z transforms Domain)
SD	s-Domain (or Laplace Domain)
TF	Transfer Function
RBF	Rational Basis Function(s)
OBF	Orthonormal Basis Function(s)
ROBF	Rational Orthonormal Basis Function(s)
ZD-OBF	z-Domain (Rational) Orthonormal Basis Function(s)
SS	State Space form
LTI	Linear Time Invariant (Dynamical System)
IP	Initial Poles
VF	Vector Fitting
OVF	Orthonormal Vector Fitting (VF using orthonormal bases)
ZD-OVF	z-Domain Orthonormal Vector Fitting
FD-OVF	Continuous-Time Frequency-Domain OVF
SK	Sanathanan and Koerner
LLS	Linear Least Square
SVD	Singular Value Decomposition
NE	Normal Equations Method
CAD	Computer Aided Design
VLSI	Very Large Scale Integrated Circuit
CPU	Central Processing Unit
SISO	Single Input and Single Output system
MIMO	Multi Input and Multi Output (multiport) system

Notation

$\mathbf{A}$	A matrix $\mathbf{A}=[a_{ij}]$, composed of elements a_{ij} in i-th row and j-th column
$\mathbf{A}^T$	The transpose of matrix $\mathbf{A}=[a_{ij}]$ defined as: $\mathbf{A}^T=[a_{ji}]$
$\mathbf{A}^{-1}$	The inverse of matrix $\mathbf{A}$
$\mathbf{A}^*$	The complex conjugate of complex $\mathbf{A}=[a_{ij}]$, defined as: $\mathbf{A}^*=[\overline{a}_{ij}]$
$\mathbf{A}^H$	The conjugate transpose of complex matrix $\mathbf{A}=[a_{ij}]$, defined as: $\mathbf{A}^*=[\overline{a}_{ji}]$
$\mathbf{I}_{m\times m}$	An $m\times m$ identity matrix $\mathbf{I}=\left[\mathfrak{I}_{ij}\right]$, wherein $\mathfrak{I}_{ij}=1$, for $i=j$ and $\mathfrak{I}_{ij}=0$, for $i\neq j$
$\mathbb{R}$	The field of real numbers
$\mathbb{R}^{\mathrm{N}}$	The set of real vectors of size N
$\mathbb{R}^{\mathrm{N}\times\mathrm{N}}$	The set of real matrices of size $N\times N$
$\mathbb{C}$	The field of complex numbers, e.g.: s-plane
$\mathbb{C}^+$	The open right half plane in the complex s-plane; $\mathbb{C}^+=\{s\in\mathbb{C}:\ \mathrm{Re}(s)>0\}$
$\mathbb{C}^-$	The open left half plane in the complex s-plane; $\mathbb{C}^-=\{s\in\mathbb{C}:\ \mathrm{Re}(s)<0\}$
$\overline{\mathbb{C}}$	The closed right half plane in s-plane; $\overline{\mathbb{C}}=\{s\in\mathbb{C}:\ \mathrm{Re}(s)\geq 0\}$
$\mathbb{E}$	The exterior of the unit disk in the complex z-plane including ∞; $\{z\in\mathbb{C}:\ \lvert z\rvert>1\}$
$\mathbb{T}$	The unit circle in the complex z-plane; $\{z\in\mathbb{C}:\ \lvert z\rvert=1\}$
$\mathbb{D}$	The open unit disk in the complex z-plane; $\{z\in\mathbb{C}:\ \lvert z\rvert<1\}$
$\overline{\mathbb{D}}$	The closed unit disk in the complex z-plane; $\overline{\mathbb{D}}=\mathbb{D}\cup\mathbb{T}$
$\mathbb{C}^N$	The set of complex vectors of size N
$\mathbb{C}^{N\times N}$	The set of complex matrices of size $N\times N$

$\bar{a}$ *or* a^*	The complex conjugate of a complex number $a \in \mathbb{C}$
s	$(s = \alpha + j\omega)$ presents the complex frequency (Laplace variable) as a point in complex s-plane. Considering $s = j\omega$ converts the Laplace transform to the Fourier transform. This property is instrumental when tackling the steady state response in frequency domain.
z	The complex value as variable in z domain presenting a general point in the z-plane, $z \triangleq e^{sT_s}$, when T_s is the sampling period.
$\mathcal{H}_p(\mathbb{E})$	'Hardy space' denotes the set of all integrable functions f on T, analytical in the E region. that are such that: $\|f\|_p^p = \frac{1}{2\pi}\sup_{r>1}\int_0^{2\pi}\left\|f(re^{j\theta})\right\|^p d\theta < \infty \quad ;Where\ 0 < p < \infty$ $\|f\|_\infty = \sup_{z\in E}\left\|f(re^{j\theta})\right\| < \infty$
$\mathcal{RH}_2^{p\times m}$	Set of $p \times m$ matrices with rational functions as entries that are analytic for $\|z\| \geq 1$ and squared integrable on the unit circle.

CHAPTER 1. Introduction

1.1 Background and Motivation

The high-speed design of communication, digital, and microwave electronic systems, in contrast to the design at low speeds, requires accurate characterization of passive circuit elements. These passive elements may include the wires, vias, circuit board traces, integrated-circuit packages, on-chip passive components, and high-frequency microwave devices that make up a product.

At higher speeds, the behavior of passive circuit elements such as transmission lines, formed from packaging and interconnect, can have a direct impact on system performance [1]. Accordingly, a significant amount of research has been done to include (accurate models of) non-ideal transmission lines in circuit simulation [2]-[3].

The complexity of modeling interconnects depends on the operation frequency and physical structure. At high frequencies, the length of interconnects become a significant fraction of operating wavelength, hence distributed models are required. The parasitic effects such as ringing, signal delay, distortion, reflection and crosstalk that were not significant at lower frequencies should be indispensably considered [4], [1]-[5]. Skin and proximity effects also become prominent, requiring the per-unit length (p.u.l.) parameters to be functions of frequency [4], [5], [6], [7]-[8]. The ever-rising operating frequencies have posed serious challenges to simulators; due to the emergence of the

above-mentioned high frequency effects besides interconnect dispersion and mutual couplings [9].

One of the highly focused fields in the analysis of high-speed modules is the modeling and characterization based on tabulated data [2], [10]. It is rapidly becoming an integral part of signal validations, due to the diverse and complex nature of high-speed modules in microwave and integrated designs. Tabulated data can be obtained from simulations or measurements. For instance, the behavior of strip-line and micro-strip printed circuit board traces, inter-chip connections on multichip modules, and coaxial cable connections are most easily represented by frequency-dependent scattering parameters. A vector/matrix of tabulated (S) parameters data can be derived directly from measurements or a proper simulation method such as using full-wave electromagnetic (EM) analysis and detailed finite-element simulation, or even analytic formulas.

A general approach to including transmission lines in circuit simulators is the macromodeling of the passive linear subnetworks based on sampled data, for which the most straightforward approach is to approximate the frequency-domain representation with a rational function. This has been a topic of intense research during the recent years. Several methods have been proposed to generate rational representations of distribute systems. These methods include asymptotic waveform evaluation, complex frequency hopping method and its variants that are based on moment matching at single or multiple points as well as other methods that are based on Krylov subspace techniques such as Lanczos and Arnoldi methods [10], [11]- [12].

Following the same line, a frequency-domain system identification technique, called vector fitting (VF), was introduced for fitting frequency-dependent data through high-order rational function approximation over wide frequency ranges [13]. It is a formulation based on the Levi's linearized least square error estimator [14] using partial fraction bases. This iterative pole-relocation process starts from prescribed initial poles for the partial fractions and is successively repeated until minimizing the weighted error LLS estimator. The VF technique that was originally introduced for power system and transmission line modeling, subsequently, has become a method of choice for many other applications in microwave engineering, interconnect, and package macromodeling [9]. The robustness of the method is mainly due to the use of rational bases instead of polynomials, which are numerically advantageous if prescribed poles are properly chosen [15].

The numerical quality of the system equations in standard VF is sensitive to the choice of the location of starting poles in s-plane. This fact explicitly reveals that the accuracy and convergence of the method is significantly affected by the selection of starting poles. The standard methods for choosing the orders and optimum initial poles are in the most part heuristic [13]- [16].

To address this problem, recently, orthogonal vector fitting (OVF) has been introduced in [15], to build macromodel based on frequency-domain data samples. In this method, orthonormal rational functions are utilized to improve the numerical stability of the method and reduce the numerical sensitivity of the system equations to the choice of starting poles.

These pure frequency-domain methods are associated with the following issues:

- They operate entirely in frequency-domain, providing rational approximations of transfer functions using frequency-Domain samples. They are limited to carry out the macromodeling task for 'continuous-time' LTI systems.

- In frequency-domain methods it is indispensable to scale the frequency axis (and hence also the parameters) to guarantee the numerical stability [17] and to minimize the error and optimize the convergence speed. It also requires that the parameters for the final macromodel to be scaled back to support the original data. The s-domain OVF algorithm often faces rank deficient equations, when identifying the final model by orthonormal bases. Thus, the final model would be mostly limited to the partial fraction form.

- In some cases, the time-domain measurement or simulation of transient excitations and responses at the ports is more feasible, and then the network is characterized by transient input/output responses [69], [70]. Thus, a technique is required to produce rational approximations using the time-domain vector/matrix data. A specific application, where this approach is convenient, is the equivalent circuit extraction for three-dimensional interconnect structures (e.g., electronic packages or connectors), that are characterized using a rigorous full-wave electromagnetic solver, based on (e.g.) the Finite-Difference Time-Domain (FDTD) method [18], [19]. Due to the high computational cost, it is desirable to terminate the full-wave analysis before all transients have extinguished, thus obtaining truncated time responses [19].

To tackle the above concerns, recently, a formulation of standard vector fitting method in the z-domain was proposed. Compared to the standard s-domain VF method, it has an

advantage of better numerical stability and better error in the model [11]. However, the algorithm has inherited the same sensitivity to the initial poles from its VF derivation.

In order to address the above issues, which are highly challenging the design automation and signal integrity community, the contributions of this thesis are given in the next section.

1.2 Contributions

By introducing the z-domain orthonormal vector fitting (ZD-OVF) technique, in this thesis, a robust and accurate macromodeling algorithm is developed. This novel algorithm is a powerful tool for modeling and simulation of multiport high-speed networks characterized by tabulated data. It is capable of accurately handling the sampled data in time-domain as well as frequency-domain. The developed technique also carries out the macromodeling task for both 'continuous-time' and 'discrete-time' stable LTI systems. The higher accuracy level obtainable from the proposed method, even when the starting poles are not optimally selected, is another remarkable advantage. Unlike the s-domain methods, the proposed ZD-OVF supports the original frequency data without degrading the numerical quality of the system equations or the accuracy of the results. The specific contributions are listed below.

1) A novel formation for z-domain orthonormal basis functions (ZD-OBF) has been introduced. Deviated from Takenaka-Malmquist functions, the proposed ZD-OBF associated with complex conjugate poles ensure the identification of physical systems in the form of a linear span of the proposed bases with real coefficient. By using the

proposed basis functions, an accurate realization is accomplished, while ensuring the real-valued time-domain impulse response for the identified model.

2) An efficient formulation has been developed to construct the minimal state-space realization for a system that has been identified in the form of a linear expansion of the proposed ZD-OBF.

3) An efficient multiport algorithm is developed for macromodeling of broadband passive networks (such as high-speed interconnect circuits). The proposed algorithm directly generates z-domain macromodel in the intermediate stage of modeling. This facilitates fast and accurate time-domain (transient) simulations.

4) An implementation of the 'Rank Revealing QR', a highly accurate solution method for the linear least square problems, is proposed. Utilizing this method enhances the accuracy level of the LLS solutions particularly when the system equations numerically are ill-conditioned. Thus, it ensures the sufficient accuracy for the final model.

5) Three multiport algorithms along with the Sanathanan-Koerner weighted LLS error estimation mechanism have been developed for the existing poles-relocation techniques listed below:

- s-domain conventional VF (1998)
- s-domain OVF (2005)
- z-domain conventional VF (2007)

Providing a comparable style and uniform program structure with those in the proposed method makes a trustworthy peer-to-peer comparison between all methods feasible. A comprehensive performance methodological comparison has been performed and presented in this work.

The proposed method in this thesis is an efficient tool to handle broadband high-speed passive networks (such as interconnect circuits) described with sampled data, in industrial grade simulators.

1.3 Organization of the Thesis

This thesis has been organized as follows. In chapter 2, a survey of fundamental concepts and definitions in discrete-domain is presented. As essential concepts for ZD-OVF, this chapter also discuses the transformation between counterpart domains and the feasible forms for the z-domain transfer function for physical systems. Chapter 3 provides theoretical background regarding the orthonormal functions from the applied perspective, with emphasis on the discrete-domain structures. Chapter 4 presents a novel ZD-OBF, to ensure the real-value time-domain impulse response for LTI systems. In chapter 5, for a system, identified in the form of a linear expansion of the proposed ZD-OBF, an efficient formulation to construct the minimal state-space realization is presented. This is an essential step in ZD-OVF process. Chapter 6 reviews the 'Error Estimator and Optimization' methods for the poles-relocation algorithms. It also provides a background on the linear least square (LLS) problems, constructed in the vector fitting process. It reviews the existing LLS solution methods, and proposes a formulation for RRQR as the most accurate solution. Chapter 7 presents the formulation for the proposed ZD-OVF algorithm and reviews a number of important issuses such as: selection of the initial poles, convergence, pole-refinement strategy, bounding the angular frequencies of poles. Chapter 8 presents the numerical results from examining the proposed algorithm with practical data. Moreover, It presents a performance

methodological comparison between the existing vector fitting techniques in both s and z-domains. Chapter 9 contains concise conclusions of this thesis work and discuses the directions for the future work.

CHAPTER 2. Background on Signals and Systems in Discrete Time-Domain

Identification of continuous-time LTI systems is possible in either time-domain or frequency-domain. Time-domain methods usually determine a discrete-time model of the system, while continuous-time frequency-domain methods can identify either a discrete-domain or a continuous-time model. This fact reveals the importance of the discrete-domain macromodeling techniques as a common tool in both-domains.

This chapter establishes an analytical framework and notation suitable to discrete-time systems. It considers the fundamental concepts and definitions in discrete-domain. It emphasizes on z-transform that is the most commonly used tool for the analysis of sampled-data systems. This chapter also reviews a number of essential concepts such as Nyquist theorem, Aliasing phenomenon and Zero-Order-Hold. Following this line, the definitions and forms of transfer function, impulse response, and stability will be declared in the z-domain. In addition, it is shown how the transformation between counterpart-domains is possible.

In this thesis, rather than attempt to force parallel results and conclusions from continues-time system theory into a discrete-time framework, they are directly derived in discrete-domain.

2.1 z-Transform in Discrete-domain

This section provides a mathematical approach to the concept of z-transform in discrete time-domain.

2.1.1 Definition Based on Inner Product

Consider an ordered set (sequence) of real numbers as:

$$\left\{f_n^{(k)}\right\} = \left\{\ldots, f_{-2}^{(k)}, f_{-1}^{(k)}, f_0^{(k)}, f_1^{(k)}, f_2^{(k)}, \ldots\right\} \tag{2.1}$$

in which k is an index on the set. Also, assume that a countable collection of such sequences is possible. A dot product of one sequence with another is defined like:

$$\left\{f_n^{(m)}\right\} \cdot \left\{f_n^{(n)}\right\} = \sum_n f_n^{(m)} f_n^{(n)} \tag{2.2}$$

i) The dot product of these real sequence with the complex sequence $\left\{e^{j\omega nT}\right\}$ yields Fourier series with period $\frac{2\pi}{T}$:

$$F^{(i)}(\omega) = \sum_n f_n^{(i)} e^{j\omega nT}, \tag{2.3}$$

where $\left\{e^{j\omega nT}\right\}$ satisfying (2.4) is an orthonormal set of bases.

$$\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} e^{-j(m-n)T\omega} d\omega = \begin{cases} 1, & m = n \\ 0, & m \neq n \end{cases} \tag{2.4}$$

Only those sequences for which the dot product is convergent ($< \infty$) is of interest in this context and will be further considered.

ii) Instead of forming the dot products of a real sequence with $\left\{e^{j\omega nT}\right\}$ we take their dot products with $\left\{z^{-n}\right\}$ as:

$$F^{(k)}(z)=\sum_{n} f_n^{(k)} z^{-n}. \tag{2.5}$$

This form is known as z-transform of a process and the real sequence $\left\{f_n^{(k)}\right\}$ presents the samples of a time signal in equal time intervals.

Similar to what was explained for Fourier series in continues-time-domain, a set of functions as $\left\{z^0,\ z^{-1},\ z^{-2},\ldots, z^{-n}\right\}=\left\{z^{-n}\right\}$, $n=0, 1,\ldots, N$ are orthonormal in terms of the inner product in $\mathcal{H}_2(\mathbb{E})$.

$$\left\langle z^{-n}, z^{-m}\right\rangle=\frac{1}{2\pi i}\oint_{\mathbb{T}} z^{-n} z^{+m}\frac{dz}{z}=\begin{cases}1, & m=n\\ 0, & m\neq n\end{cases} \tag{2.6}$$

Instead of the integral in (2.4), we have a contour integral taken counterclock-wise around the unit circle.

2.1.2 Definition based on Laplace Transforms

In this section, z-Transform is defined based on its continuous time-domain counterpart (Laplace Transforms). The ideal unit impulse signal, mathematically considered as "Dirac's delta function" shown in (2.7) is used for sampling the signal at time instants.

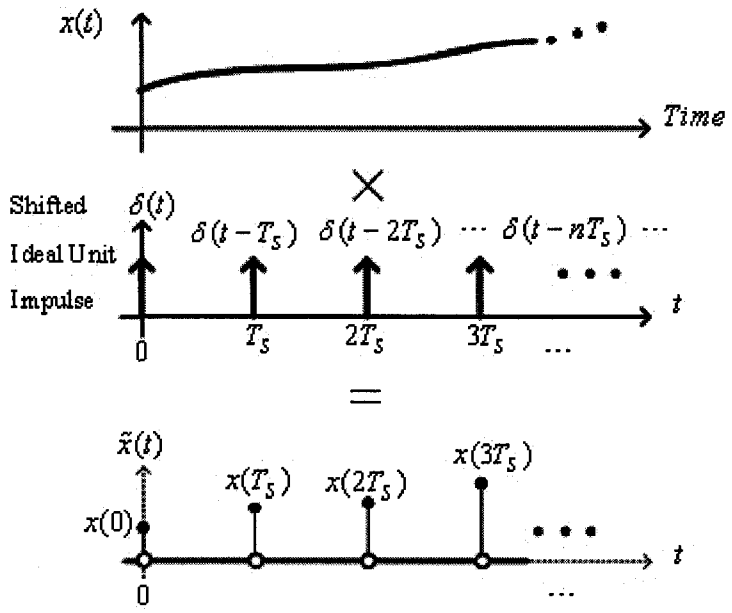


Figure 2-1: Ideal sampling scheme of a continuous-time signal

$$\delta(t) \triangleq \int_{-\infty}^{+\infty} \delta(t)dt = \int_{0^-}^{0^+} \delta(t)dt = 1 \tag{2.7}$$

The samples of a continuous transient are defined as those values of the continuous process attained at discrete times $t = 0, T_S, 2T_S, \cdots$. Thus, a discrete time-domain signal can be mathematically defined in the form of a train of weighted-and-shifted impulses as:

$$\tilde{x}(t) = \sum_{n=0}^{\infty} x(nT_S)\delta(t - nT_S) \tag{2.8}$$

All finite energy transients from real systems are aperiodic and zero for all values of the argument that are less than an arbitrary starting time, denoted as $t = 0$. This is true for either continuous transitions or their sampled counterparts [20][1].

By taking Laplace transform from the sampled signal:

[1] Concerning this fact, the z- transform concepts and definitions in the unilateral format (as opposed to bi-lateral) can be outlined. The discussion of the unilateral z-transform closely parallels the discussion of the unilateral Laplace transform in the literature.

$$\tilde{X}(s) = \mathcal{L}\{\tilde{x}(t)\} \triangleq \int_{-\infty}^{+\infty} \tilde{x}(t)e^{-st}dt = \int_{0}^{+\infty} \sum_{n=0}^{\infty} x(nT)\,\delta(t-nT)e^{-st}dt =$$

$$\sum_{n=0}^{\infty} x(nT) \int_{0}^{+\infty} \delta(t-nT)e^{-st}dt = \sum_{n=0}^{\infty} x(nT)e^{-(sT)n} \tag{2.9}$$

A simple variable change with defining z as shown in (2.10) will transform (map) the Laplace-domain (s-domain) to a new-domain, in which the independent variable is z, and hence the names z-domains and z-transform.

$$z \triangleq e^{sT} \tag{2.10}$$

where Laplace variable $s = \alpha + j\beta$ is 'complex frequency' and T presents the sampling interval (period). In a similar fashion, z is cited as 'discrete frequency'.

$$\underbrace{\tilde{X}(s) = \sum_{n=0}^{\infty} x(nT)e^{-(sT)n}}_{\text{s-Domain}} \xrightarrow{\overbrace{z \triangleq e^{sT}}^{\text{Original NonLinear Transformation}}} \underbrace{X(z) = \sum_{n=0}^{\infty} x(nT)z^{-n}}_{\text{z-Domain}} \ ^{\dagger} \tag{2.11}$$

This conceptual nonlinear transformation between domains will be used to investigate and clarify some fundamental concepts.

2.2 Impulse Response and System Identification in Discrete-Domain

The mathematical definition of z-transform, outlined in the previous section, was a subjective approach. It is an understandable fact that the impulse signal, the sampling

[†] According to footnote 1, the summation that is carried out only over nonnegative values of n is called unilateral z-transform. Chapter 10 in reference [21] in the bibliography can be referred to for a detailed account regarding the bilateral and unilateral concepts.

scheme based on it, as well as an infinite series of the time functions cannot have physical realization in the engineering fields. To satisfy the system identification applications the required physical qualities should be attributed to the above abstract concepts.

Assume a linear time-invariant stable discrete-time process represented by its proper transfer function H(z) in the Hilbert space, $\mathcal{H}_2(\mathbb{E})$. To be exact, H(z) is analytic outside and on the unit circle, $|z| \geq 1$. A general and common representation of H(z) is possible (convergent) in terms of its Laurent expansion (in z^{-1}) around $z = 0$, as shown in below:

$$H(z) = \sum_{n=0}^{\infty} h_n z^{-n} \ddagger, \tag{2.12}$$

where $\{h_k\}_{k=0,1,\ldots}$ is the sequence of Markov parameters[2] [22].

Exploiting orthogonal property, as shown in (2.6), the impulse response coefficients are given by:

$$h_n = \left\langle H, z^{-n} \right\rangle = \frac{1}{2\pi j} \oint_{\mathbb{T}} H(z) z^n \frac{dz}{z} \tag{2.13}$$

---

‡ Considering the general format of Laurent series as shown in below:

$$X(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n + \sum_{n=1}^{\infty} b_n (z - z_0)^{-n} = \sum_{n=-\infty}^{+\infty} c_n (z - z_0)^{-n} \quad ; \left(R_1 < |z - z_0| < R_2\right)$$

When X(z) is analytic throughout $1 \leq |z| < \infty$; if the coefficients a_n are worked out along the positively oriented unit circle $\mathbb{T}$, there will be $a_0 = x(0)$, and $a_n = 0$ for $n > 0$. For more relevant mathematical details regarding Laurent's theorem the references [23, pp.841-847] and [24, pp.190-195] can be referred to.

[2] See Appendix A for more details

The practical approximation will be carried out by expressing the given function H(z) with finite expansion.

$$H(z) \approx \hat{H}(z) = \sum_{n=0}^{N} h_n z^{-n} \tag{2.14}$$

Theorem: The mean-squared error between the original system and approximated model in the form of a truncated series shown in(2.14) for an arbitrary value of N is minimum when its coefficients are decided by (2.13). Hence, these coefficients are the optimum coefficients for the truncated series.

PROOF:

The mathematical proof outlined in below is the most general approach while the function H(z) is assumed to be a mathematical arbitrary function, which is analytic within $1 \le |z| < \infty$. For system identification applications wherein the original system is assumed to have a real impulse response, more constraints should be ensured throughout the proof. It will be proved and justified in the next section with adequate details.

Let the absolute error between the original function and approximated model be denoted like: $\mathcal{E} = H(z) - \hat{H}(z)$.

The mean-squared error can be obtained as:

$$\|\mathcal{E}\|_2^2 = \langle \mathcal{E}, \mathcal{E} \rangle \triangleq \frac{1}{2\pi j} \oint_{\mathbb{T}} \mathcal{E}(z,N)\mathcal{E}^*\left(1/z^*, N\right)\frac{dz}{z} =$$

$$= \frac{1}{2\pi j} \oint_{\mathbb{T}} \mathcal{E}(z,N)\mathcal{E}^*(z,N)\frac{dz}{z} = \frac{1}{2\pi j} \oint_{\mathbb{T}} \left[H(z) - \hat{H}(z)\right]\left[H^*(z) - \hat{H}^*(z)\right]\frac{dz}{z} =$$

$$\frac{1}{2\pi j} \oint_{\mathbb{T}} H(z)H^*(z)\frac{dz}{z} - \frac{1}{2\pi j} \oint_{\mathbb{T}} H(z)\hat{H}^*(z)\frac{dz}{z} - \frac{1}{2\pi j} \oint_{\mathbb{T}} \hat{H}(z)H^*(z)\frac{dz}{z} +$$

$$\frac{1}{2\pi j}\oint_{\mathbb{T}}\hat{H}(z)\hat{H}^*(z)\frac{dz}{z}=$$

$$=\|H(z)\|_2^2-\frac{1}{2\pi j}\oint_{\mathbb{T}}H(z)\sum_{n=0}^{N}h_n^*z^n\frac{dz}{z}-\frac{1}{2\pi j}\oint_{\mathbb{T}}\sum_{n=0}^{N}h_nz^{-n}H^*(z)\frac{dz}{z}+\sum_{n=0}^{N}h_nh_n^*$$

thus far, it is:

$$\|\mathcal{E}\|^2=\|H(z)\|_2^2-\sum_{n=0}^{N}h_n^*\left(\frac{1}{2\pi j}\oint_{\mathbb{T}}H(z)z^n\frac{dz}{z}\right)-\sum_{n=0}^{N}h_n\left(\frac{1}{2\pi j}\oint_{\mathbb{T}}H^*(z)z^{-n}\frac{dz}{z}\right)+\sum_{n=0}^{N}h_nh_n^* \tag{2.15}$$

To minimize the mean-squared error, $\|\mathcal{E}\|^2$ should be differentiated with respect to each unknown coefficient, h_n and make it equal to zero as: $\frac{\partial\|\mathcal{E}\|^2}{\partial h_n}=0$

$$\frac{\partial\|\mathcal{E}\|^2}{\partial h_n}=0-\sum_{n=0}^{N}\frac{\partial h_n^*}{\partial h_n}\times\left(\frac{1}{2\pi j}\oint_{\mathbb{T}}H(z)z^n\frac{dz}{z}\right)-\sum_{n=0}^{N}\left(\frac{1}{2\pi j}\oint_{\mathbb{T}}H^*(z)z^{-n}\frac{dz}{z}\right)+\sum_{n=0}^{N}h_n^*+$$

$$+\sum_{n=1}^{N}\frac{\partial h_n^*}{\partial h_n}\times h_n=0$$

Consequently:

$$\frac{\partial\|\mathcal{E}\|^2}{\partial h_n}=\sum_{n=0}^{N}\frac{\partial h_n^*}{\partial h_n}\times\left[h_n-\left(\frac{1}{2\pi j}\oint_{\mathbb{T}}H(z)z^n\frac{dz}{z}\right)\right]+\sum_{n=0}^{N}\left[h_n^*-\left(\frac{1}{2\pi j}\oint_{\mathbb{T}}H^*(z)z^{-n}\frac{dz}{z}\right)\right]=0$$

To ensure the above equality for any arbitrary value of h_n and N two following equations should be enforced simultaneously.

$$\begin{cases} h_n - \left(\dfrac{1}{2\pi j} \oint_{\mathbb{T}} H(z) z^n \dfrac{dz}{z} \right) = 0 & (i) \\ \text{and} \\ h_n^* - \left(\dfrac{1}{2\pi j} \oint_{\mathbb{T}} H^*(z) z^{-n} \dfrac{dz}{z} \right) = 0 & (ii) \end{cases}$$

The first condition resulted from (i) drops to the form of:

$$h_n = \frac{1}{2\pi j} \oint_{\mathbb{T}} H(z) z^n \frac{dz}{z} \tag{2.16}$$

Comparing (2.13) and (2.16) justifies the hypothesis and proves that the coefficients decided by (2.13) minimizes the absolute error for any arbitrary truncated series.

To continue, starting from the assumption of (2.16), it is shown in below that (ii) also will hold automatically and no extra condition is introduced by (ii).

$$h_n^* = \frac{1}{-2\pi j} \oint_{\mathbb{T}} H^*(z) z^{-n} \frac{d\left(\frac{1}{z}\right)}{z^{-1}} = \frac{1}{2\pi j} \oint_{\mathbb{T}} H^*(z) z^{-n} \frac{dz}{z}$$

$$\Rightarrow \; h_n^* - \left(\frac{1}{2\pi j} \oint_{\mathbb{T}} H^*(z) z^{-n} \frac{dz}{z} \right) = 0$$

Although the physical systems with real impulse response in a mathematical sense can be considered as a special case for this general proof, justifying the same fact while enforcing the compulsory constraints for physical systems bears a high importance. Based on the background provided in above, an authentic proof will be presented in the next chapter.

2.2.1 Real Impulse Response

The time-domain realization for a transfer function of physical systems, $z^{-1}\{H(z)\}$, appears in the form of a real sequence satisfying:[3]

$$H(z) = H^*(z^*) \tag{2.17}$$

This enforces all impulse response coefficients for transfer functions of the form (2.14) to be real-valued. According to this, we can verify the above theorem in its special form for the real-valued coefficients as follows.

PROOF:

Equation (2.15) will resemble:

$$\|\mathcal{E}\|^2 = \|H(z)\|_2^2 - \sum_{n=0}^{N} h_n \left(\frac{1}{2\pi j} \oint_{\mathbb{T}} H(z) z^n \frac{dz}{z} \right) - \sum_{n=0}^{N} h_n \underbrace{\left(\frac{1}{2\pi j} \oint_{\mathbb{T}} H^*(z) z^{-n} \frac{dz}{z} \right)}_{❶} + \sum_{n=0}^{N} h_n^2 \tag{2.18}$$

(2.17) and ❶:

$$❶ = \frac{1}{2\pi j} \oint_{\mathbb{T}} H(z^*) z^{-n} \frac{dz}{z} = \frac{1}{2\pi j} \oint_{\mathbb{T}} H(z^{-1}) z^{-n} \frac{dz}{z} \tag{2.19}$$

With a variable change like $z^{-1} = \mathsf{z}$, it is seen that the boundary of integral $(\mathbb{T})$ is not affected except integration instead of performing counterclockwise $[-\pi, \pi)$ is taken in reverse direction. To correct this we have to consider a negative sign, as shown below.

$$(2.19) \Rightarrow \frac{-1}{2\pi j} \oint_{\mathbb{T}} H(\mathsf{z}) \mathsf{z}^{n+1} d\left(\frac{1}{\mathsf{z}}\right) = \frac{-1}{2\pi j} \oint_{\mathbb{T}} H(\mathsf{z}) \mathsf{z}^{n+1} \times \frac{-d\mathsf{z}}{\mathsf{z}^2} = \frac{1}{2\pi j} \oint_{\mathbb{T}} H(\mathsf{z}) \mathsf{z}^n \times \frac{d\mathsf{z}}{\mathsf{z}}$$

Superficially, we can change the ' z ' to a more familiar notation as 'z':

[3] This will be discussed in this work later, with more details.

$$\textcircled{1} = \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^n \times \frac{dz}{z}$$

by plugging it back in (2.18):

$$\|\mathcal{E}\|^2 = \|H(z)\|_2^2 - \sum_{n=0}^{N} h_n \left(\frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^n \frac{dz}{z} \right) - \sum_{n=0}^{N} h_n \left(\frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^n \times \frac{dz}{z} \right) + \sum_{n=0}^{N} h_n^2$$

$$\|\mathcal{E}\|^2 = \|H(z)\|_2^2 - 2\sum_{n=0}^{N} h_n \left(\frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^n \frac{dz}{z} \right) + \sum_{n=0}^{N} h_n^2$$

$$\frac{\partial \|\mathcal{E}\|^2}{\partial h_n} = -2\sum_{n=0}^{N} \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^n \frac{dz}{z} + 2\sum_{n=0}^{N} h_n = 0$$

Since it is expected that the coefficients h_n do not depend upon the number of terms, N:

$$N=1: \quad h_1 = \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z \ \frac{dz}{z} \tag{2.20}$$

$$N=2: \quad h_1 + h_2 = \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z \ \frac{dz}{z} + \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^2 \frac{dz}{z} \tag{2.21}$$

(2.20) and (2.21):

$$h_2 = \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^2 \frac{dz}{z} \tag{2.22}$$

$$N=3: \quad h_1 + h_2 + h_3 = \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z \ \frac{dz}{z} + \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^2 \frac{dz}{z} + \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^3 \frac{dz}{z}$$

(2.23)

(2.21) and (2.23):

$$h_3 = \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z)z^3 \frac{dz}{z} \tag{2.24}$$

Generalizing the idea outlined within (2.20), (2.22), and (2.24) with intuition proves:

$$h_n = \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z) z^n \frac{dz}{z} \tag{2.25}$$

2.2.2 The z-Domain Response in the Form of a Rational Function

Based on what thus far has been explained, the z-domain response of any LTI passive physical network could be presented in the form of a finite series as:

$$\begin{cases} H(z) \approx \sum_{n=0}^{N} h_n z^{-n} \\ h_n = \frac{1}{2\pi j}\oint_{\mathbb{T}} H(z) z^n \frac{dz}{z} \end{cases} \tag{2.26}$$

For an N^{th}-order system response, it resembles:

$$H(z) \approx h_0 + h_1\left(z^{-1}\right) + h_2\left(z^{-1}\right)^2 + \cdots + h_N\left(z^{-1}\right)^N \tag{2.27}$$

Before proceeding further, recalling the similar case in continuous time s-domain may help to declare the concept better.

The s-domain response of any LTI dynamic system, using (Maclaurin-) Taylor series can be presented in the form of power series when the coefficients are the s-domain moments of the function.

$$H(s) \approx m_0 + m_1 s + m_2 s^2 + \cdots + m_N s^N \tag{2.28}$$

It is an established fact in literature that functions represented by convergent Taylor series, e.g. the one in the form of (2.28), can be well approximated by a proper rational function. In particular, when a function contains poles and has singularities, the use of rational function is even superior and represents the original system better. There are

established algorithms available in the literature to approximate a power series as a ratio of two polynomials with determining both the numerator and denominator coefficients. Accordingly, a function of the form (2.27) is well approximated as rational functions of two polynomials as:

$$H(z) \approx \frac{\sum_{k=0}^{N} a_k \left(z^{-1}\right)^k}{\sum_{k=0}^{D} b_k \left(z^{-1}\right)^k} \tag{2.29}$$

2.2.3 Alternative Approach to the Rational Transfer Function

Consider a LTI system for which the input and output satisfy a linear constant-coefficient difference equation of the form

$$\sum_{k=0}^{D} b_k y[n-k] = \sum_{k=0}^{N} a_k x[n-k] \tag{2.30}$$

Then:

$$\sum_{k=0}^{D} b_k z^{-k} Y(z) = \sum_{k=0}^{N} a_k z^{-k} X(z) \xrightarrow{or} Y(z) \sum_{k=0}^{D} b_k z^{-k} = X(z) \sum_{k=0}^{N} a_k z^{-k} \tag{2.31}$$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^{N} a_k z^{-k}}{\sum_{k=0}^{D} b_k z^{-k}} \tag{2.32}^{\dagger}$$

We note in particular that the system function for a system satisfying a linear constant-coefficient difference equation is always rational [21]. Such z-transforms arise

† (2.29) and (2.32) are identical results.

frequently in the study of linear time-invariant systems [25]. The constraint such as causality[4] or stability of the system serves to specify the region of convergence for H(z).

2.2.4 On the Forms of Rational Transfer Functions

The algebraic expression H(z) in (2.32) can, of course, be written so that the polynomials are expressed as powers of z rather than of z^{-1}. In the control and the signals processing contexts, it is common practice not to do so, whereas in the vector fitting formulation wherein we concern the poles, the equivalent expression in the following form, is considered.

$$H(z) = \frac{z^D \times \sum_{k=0}^{N} a_k z^{N-k}}{z^N \times \sum_{k=0}^{D} b_k z^{D-k}} = \left(z^{D-N}\right) \frac{\sum_{k=0}^{N} a_k z^{N-k}}{\sum_{k=0}^{D} b_k z^{D-k}} \tag{2.33}$$

Equation (2.33) explicitly shows that for such a function there will be N zeros and D poles at nonzero locations in the z-plane.

One may think of the following three circumstances:

i) When $D < N$:

From (2.33) it is[†]:

[4] ***Causality property for LTI systems***: a system is causal if the output at any time depends only on values of the input at the present time and in the past [26].

[†] For this case, in the Discrete-time signal processing context e.g. in [25], the general form for the equivalent expression is considered as:

$$H(z)=\frac{1}{z^{N-D}}\times\frac{\sum_{k=0}^{N}a_k z^{N-k}}{\sum_{k=0}^{D}b_k z^{D-k}}=\frac{\sum_{k=0}^{N}a_k z^{N-k}}{\sum_{k=0}^{D}b_k z^{N-k}} \tag{2.34}$$

The order of polynomials in numerator and denominator of this rational transfer function is equal to N. Therefore, it can be considered as a special case of the proper transfer function with z^{N-D} poles at the origin of the z-plane as shown below.

$$H(z)=\frac{a_0 z^N+\cdots+a_N}{b_0 z^N+\cdots+b_D z^{N-D}}=\frac{\hat{a}_0 z^{N-1}+\cdots+\hat{a}_{N-1}}{b_0 z^N+\cdots+b_D z^{N-D}}+d \tag{2.35}$$

Having co-incident (multiple-order) poles at the origin, the transfer function in the form of (2.35) is not an interesting or even a general case in the system identification.

ii) when $D=N$:

When $D=N$, the transfer function (2.32) is proper with respect to $\left(z^{-1}\right)$. A similar discussion with above would lead to a general form that resembles:

$$H(z)=\sum_{r=0}^{N-D}A_k\left(z^{-1}\right)^r+\underbrace{\frac{\sum_{k=0}^{D-1}\hat{a}_k\left(z^{-1}\right)^k}{\sum_{k=0}^{D}b_k\left(z^{-1}\right)^k}}_{\text{Strictly proper TF w.r.t. } z^{-1}}$$

This is the equivalent expression, in which the real coefficients A_k can be obtained by long division of the numerator by the denominator. The division process terminates when reaching the reminder with lower degree than the denominator.

$$H(z)=\frac{\sum_{k=0}^{D} a_k z^{D-k}}{\sum_{k=0}^{D} b_k z^{D-k}}=\frac{a_0 z^D+\cdots+a_D}{b_0 z^D+\cdots+b_D}=\frac{\hat{a}_0 z^{D-1}+\cdots+\hat{a}_{D-1}}{b_0 z^D+\cdots+b_D}+d \tag{2.36}$$

We can consider(2.36) as a general form, which includes the (2.35) case too.

iii) when $D>N$:

The alternative condition arises when $D>N$. Having $D>N$, makes (2.32) to be known as strictly proper w.r.t. $\left(z^{-1}\right)$.

A strictly proper transfer function in the most general form resembles:

$$H(z)=\frac{a_0\left(z^{-1}\right)^{N-1}+\cdots+a_{N-2}\left(z^{-1}\right)^2+a_{N-1}}{b_0\left(z^{-1}\right)^N+\cdots+b_{N-1}\left(z^{-1}\right)+b_N} \tag{2.37}$$

Then;

$$H(z)=z\underbrace{\left(\frac{a_{N-1}z^{N-1}+\cdots+a_1 z+a_0}{b_N z^N+\cdots+b_1 z+b_0}\right)}_{\tilde{H}(z)} \tag{2.38}$$

$$\tilde{H}(z)=\frac{H(z)}{z} \tag{2.39}$$

According to (2.39), the-domain of the $H(z)$ should not enclose the point $z=0$ (origin in the complex z-plane) which is the case for any $H(z)\in\mathcal{H}_2\left(\mathbb{E}\right)$. Thus, $\tilde{H}(z)$ falls in strictly proper format that can be expressed in terms of the sum of either partial fraction or orthonormal bases.

***Time Shifting Theorem*[5]:** If x(t) = 0 for t < 0 and x(t) has z-transform X(z)

[5] Also called real translation theorem

$$\mathcal{Z}\left[x(t-nT)\right]=z^{-n}X(z), \tag{2.40}$$

and for the delayed function:

$$\mathcal{Z}\left[x(t+nT)\right]=z^{n}\left[X(z)-\sum_{k=0}^{n-1}x(kT)z^{-k}\right], \tag{2.41}$$

where $n\geq 0$ [21], [27].

Considering the above theorem it is understood when the time-domain signal for *H(z)* is available as *h(t)* , the corresponding time-domain signal for $\tilde{H}(z)$ would be of the form *h(t-T)*. Simply, $\tilde{H}(z)$ is a strictly proper transfer function that represents a delayed time-domain signal. When the time-domain measurement results are available, $\tilde{H}(z)$ is attained by applying a delay as much as one sampling period (right shift in time axis) to the observed TD data.

It is important to remember that all $(a_i$ and $b_i)$ coefficients should be real numbers to assure a real time-domain response signal (e.g. impulse response).

2.2.5 Strictly Proper Transfer Functions

According to the above discussion in the engineering applications, the most general form for the LTI systems transfer functions with respect to z is obtainable when $D\geq N$ in (2.32). It is worth noting that considering either one of these most general forms shown in (2.36) or (2.39); The transfer function finally falls in the form of the strictly proper form of function.

It is often assumed for strictly proper transfer function:

$$\lim_{|z|\to\infty} H(z)=0 \tag{2.42}$$

Although realizing the $|z| \to \infty$ is rarely possible, the proper transfer function is the one of most interest. Moreover, the rational functions of this type are capable of being expanded by the orthonormal bases in a convergent manner.

It is apparent that proper form of the transfer functions can be easily obtained by adding the "direct coupling constant" or "asymptote" terms to the expanded strictly proper form.

2.3 Sampling Theorem

Under certain conditions, a continuous-time signal can be completely represented by and recoverable from knowledge of its instantaneous values or equally spaced time samples. If a signal is band-limited[6] and if the samples are taken sufficiently close together in relation to the highest frequency present in the signal, the samples uniquely specify the signal and we can reconstruct it perfectly [21].

Sampling Theorem: In order for a band-limited baseband signal (i.e., one with a zero power spectrum for frequencies $f > f_{\max} > 0$) to be reconstructed fully, it must be sampled at a rate $f_S \geq 2f_{\max}$.

A signal sampled at $f_S = 2f_{\max}$ is said to be Nyquist sampled, and $f = 2f_{\max}$ is called the Nyquist frequency. No information is lost if a signal is sampled at the Nyquist frequency, and no additional information is gained by sampling faster than this rate (oversampling) [25], [28], [67].

[6] The "Band-Limited (BL)" signal assumption means that the measured signals contain no energy above a certain specified maximum frequency.

The importance of the sampling theorem lies in its role as a bridge between continuous-time signals and discrete-time signals [66].

2.3.1 The Effect of Under-Sampling: Aliasing

Assuming that the sampling frequency was sufficiently high, with a sampling rate greater than two times of the higher frequency component present in the spectrum of the signal (with no negligible energy) then the spectrum of the sampled signal consists of exact replication of the spectrum of the original time-domain signal.

It may happen that the condition of the sampling frequency outlined above is not met. This occurs when the sampling rate technically cannot be adequately high and "the signal is sampled without band-limiting it by an anti-aliasing filter" [29].

Then in the frequency spectra of an impulse-sampled signal $x(nT_S)$, where $\omega_S < \omega_{\max}$, there is found an arbitrary frequency points (e.g. ω_X) that falls in the region of the overlap of the frequency spectra. As a result, the frequency spectra at $\omega = \omega_X$ comprises two components $\left|X\left(j\omega_X\right)\right|$ and $\left|X\left(j(\omega_S - \omega_X)\right)\right|$.

This phenomenon that the frequency component $\omega_S - \omega_X$ (in general, $n\omega_S \pm \omega_X$, where n is an integer) shows up at frequency ω_X, when the signal $x(t)$ is sampled, is called ***aliasing***. This frequency $\omega_S - \omega_X$ (in general, $n\omega_S \pm \omega_X$) is called ***alias*** of ω_X [27].

2.3.2 Sampling with Zero-Order-Hold (ZOH)[7]

In first figuration of this chapter, a mathematical model of sampling process is illustrated. It shows how a band-limited signal can be represented by its samples based on impulse-train sampling scheme.

The mathematical exactness in presenting continuous signals in the form of a summation of weighted-and-shifted impulse signals practically cannot be achieved. Practically, the narrow large-amplitude pulses, approximating impulses, are relatively difficult to generate and transmit. It is more practical that the one samples the signal at a given sampling instant and holds that value until the succeeding sampling instant. The resulted staircase waveforms from a sample and hold process is referred as Zero-Order-Hold (ZOH).

2.4 Space Transforming

This section shows how one method of identification for any dynamic system in one space can be transformed to the other while preserving the main characteristics of the original physical system.

As an established fact by sampling theorem, it is remarked that band-limited continuous-time signal can be uniquely represented by its samples, when the measurements meet the certain conditions "that are different for z-domain and s-domain [29]". Exploiting a proper sampling scheme to ensure a unique representation of the signal will be the principal assumption in the rest of this chapter.

[7] Step invariant or constant (zero order) inter sample signal.

2.4.1 Domain Conversion in Complex Analysis

Mapping: Consider two complex variables as $z(\alpha_z, \beta_z)$ and $s(\alpha_s, \beta_s)$ in the complex z and s planes respectively. When a function f can be defined from s-plane to z-plane such that $z = f(s)$, it is often referred to as a *mapping* or *transformation*. In general, the inverse image of a 'z' point may contain just one point, many points or none at all [24].

2.4.2 Mapping Between s-Plane and z-Plane

As It was explained in section 2.1.2, when impulse sampling is incorporated into the process, the complex variables z, and s are related by the equation in (2.10). Using a transformation as (2.10) the points $s = \sigma + j\omega$ in the s-plane can be mapped to the z-plane as:

$$z = e^{sT_s} = \left(e^{\sigma T_s}\right)e^{j\omega T_s} = \left(e^{T\sigma}\right)e^{\left(j\left(\frac{\omega}{\omega_s}\right)2\pi\right)} \tag{2.43}$$

The equation (2.43) points to a fact of high importance.

When there are distinct points such as s_1 and s_2 for which the angular frequencies are $\omega_1 < \omega_s$ and $\omega_2 = n\omega_s + \omega_1$ respectively, these points in s-plane are mapped into the same location in the z-plane. This violates the mandatory criteria for isomorphism[8] and conformal[9] transformation.

---

[8] A transformation as f is isomorphism when it is one-to-one and for any $b \in \boldsymbol{B}$, there exists an $a \in \boldsymbol{A}$ for which $b=f(a)$. The latter is sometimes referred to as being "onto." See [23] and [24] for further details.

[9] A mapping in the plane is said to be a conformal mapping if it preserves angles (magnitude and direction) between oriented curves.

In general, this means that there are infinitely many points in s-plane, which can be mapped to a single point in z-plane (there is no one-to-one correspondence). While in transforming one space to other, ensuring the one to one relationship between correspondent elements in two-domains, is obligatory to perform inverse transformation. To ensure isomorphism, assume that a choice of sufficiently high $\omega_{sampling}$ causes all the angular frequencies (ω) for the s points of interest in s-plane to fall lower than or equal to $\left(\frac{1}{2}\omega_{samplig}\right)$. The left hand side region of this strip in s-plane is known as ***Primary strip*** as shown in Figure 2-2.

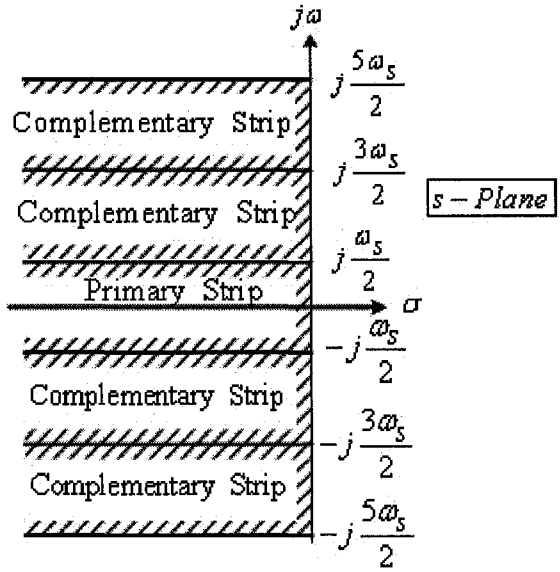


Figure 2-2: Periodic strips in the s-plane

Theorem (Conformal mapping): the mapping defined by an analytic function as $f(z)$ is conformal, except at critical points, at which the derivative $f'(z)$ is zero [23].

The "primary strip in S-plane" lies along the negative real axel and horizontally bounded to $[-j\omega_{\max}, +j\omega_{\max}] = \left[-j\frac{1}{2}\omega_s, +j\frac{1}{2}\omega_s\right] = \left[-j(2\pi F_{\max}), j(2\pi F_{\max})\right]$.

The minimum value for F_{max} is bounded[10] to the maximum frequency presents in the spectrum of the signal after which the amplitude of the harmonics is negligible.

Table 2-1: Mapping the primary strip in s-plane into the z-plane

#	$s \in$ s-Plane	Corresponding $z \in$ z-Plane
1	**Points at the RHP:** (including unstable poles) $s = \sigma \mp j\omega$; when $\sigma, \omega > 0$	**Corresponding points lie within the exterior of the unit circle $(\mathbb{E})$:** (including unstable z-domain poles) $z = \left(e^{\sigma T_s}\right)e^{j\left(\frac{\mp\omega}{\omega_s}\right)2\pi} = \rho e^{\mp j\theta}$; where $\sigma T_s > 0$; Thus $\lvert\rho\rvert > 1$, $\theta \in [-\pi, 0] \cup (0, \pi]$
2	**Points on the positive real axis:** (including pure real unstable poles) $s = \sigma$; when $\sigma > 0$, $\omega = 0$	**The counterpart points lie on the positive real axis at the exterior of the $\mathbb{D}$:** (including pure real unstable poles) $z = \left(e^{\sigma T_s}\right)e^{j0} = \rho$; Thus $\lvert\rho\rvert > 1$, $\theta = 0$
3	**Point on the $\mp j\omega$ axis :** $s = \mp j\omega$; when $\sigma = 0$ and $\omega > 0$	**The corresponding points lie on the unit circle $(\mathbb{T})$:** $z = e^{j\left(\frac{\mp\omega}{\omega_s}\right)2\pi} = \rho e^{\mp j\theta}$; *Thus* $\lvert\rho\rvert = 1$, $\theta \in [-\pi, 0] \cup (0, \pi]$

[10] Although according to the "Sampling Theorem", no additional information is gained by sampling faster than this maximum frequency, with bounding the primary strip to an arbitrary F_{max} in ZD-Vector fit algorithm, practically the angular frequency of the intended poles will be bound to this frequency. It may not suit the vector fitting case always. It will be explained in continue.

4	**Origin:** $s=0$	**Lies at point with modulus 1 on the real axis:** $z=1=1\angle 0$
5	**Points at the LHP:** (including stable poles) $s=-\sigma \mp j\omega$;when $\sigma,\omega>0$	**Corresponding points lie within the interior of the unit circle $(\mathbb{D})$:** (including z-domain stable poles) $z=\left(e^{-\sigma T_S}\right)e^{j\left(\frac{\mp\omega}{\omega_S}\right)2\pi}=\rho e^{\mp j\theta}$;where $\sigma T_S<0$;Thus $\lvert\rho\rvert<1$, $\theta\in[-\pi,0]\cup(0,\pi]$
6	**Points on the negative real axis:** (including pure real stable poles) $s=-\sigma$;when $\sigma>0$, $\omega=0$	**The counterpart points lie on the positive real axis in the $\mathbb{D}$:** (including pure real stable poles) $z=\left(e^{-\sigma T_S}\right)e^{j0}=\rho$;Thus $\lvert\rho\rvert<1$, $\theta=0$
7	**Points with infinite negative real part:** (Including the poles presenting the rapidly decaying exponentials) $s=-\sigma \mp j\omega$;when $\sigma\to+\infty$	**Origin of the z-plane:** $z=\left(e^{-\sigma T_S}\right)e^{j\theta}=\rho e^{j\theta}$;Thus $\rho\to 0$

Inside the primary region (not upper and lower borders) can be mapped into the z-plane such that each point as $s=\sigma+j\omega,\ \sigma\in(-\infty,0)\ and\ \omega\in(-\omega_{\max},\omega_{\max})$ to be mapped to a unique point in the unit disk centered at the origin of the z-plane excluding negative real axis $[-1,0)$. Similarly, all points on the $j\omega$ axis are mapped to the perimeter of the unit circle $(\mathbb{T})$.

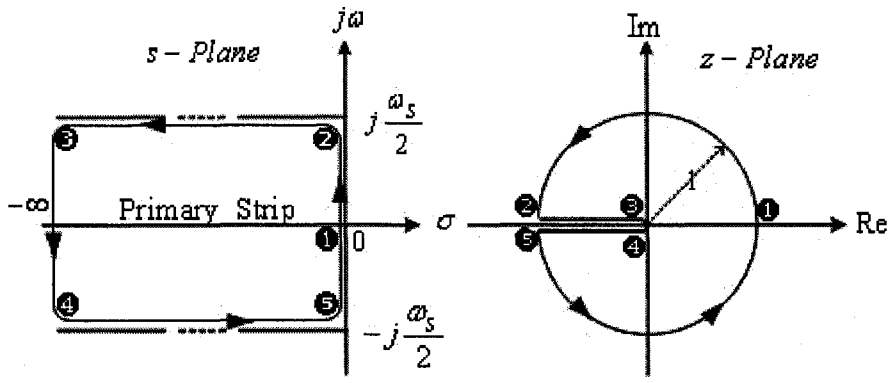


Figure 2-3: Transformation between s-domain and z-domain

As shown in the above figure, the special condition happens for the s points for which $\omega = \frac{1}{2}\omega_{sampling}$.

8	**Points at the left half plane lie on the upper and lower borders of the primary region:** $s = -\sigma \mp j\frac{1}{2}\omega_s$; when $\sigma, \omega > 0$	**Corresponding points lie on the Negative real axis in the $\mathbb{D}$:** (including stable poles) $z = \left(e^{-\sigma T_s}\right)e^{j\left(\frac{\mp\frac{1}{2}\omega_s}{\omega_s}\right)2\pi} = \rho e^{\mp j\pi}$; where $\sigma T_s < 0$; Thus $\lvert\rho\rvert < 1$, $\theta = \mp\pi \Rightarrow z \in [-1,0)$

Each point on the negative real axis inside the unit disk when being inversely mapped from z-plane to s-plane, can be transformed into two possible (and mathematically many[11]) points. This violates the one to one correspondence criterion between two-domains.

---

[11] $z_{pole} = -\alpha,\ z_{pole} = e^{\left(s_{poles}T_s\right)} \Rightarrow s_{poles} = \frac{1}{T_s}\ln(-\alpha)$

$T_s \times s_{poles} = \ln(z) = \ln|z| + j\arg(z) = \ln(\alpha) + j(1 \pm 2k)\pi,\ k = 0,1,2,...$

Even if we consider the z related to the primary strip for which there is $\arg(z) \in [-\pi, 0] \cup (0, \pi]$

We will get at least two answers and the transformation would not be unique.

Consequently, when there is a negative real pole in the resulted z-domain model, the process of $z \rightarrow s$ transformation will be cumbersome.

This is the reason why it is a common conviction that negative real poles cannot be transformed, in such cases there is no s-domain equivalent [27], [30], unless there is a negative real pole pair [31]. The latter also is not a case in this work because the underlying assumption is that there are no co-incident poles.

Since for negative real pole the unique *z*-to-*s*-domain transform is not possible, in vector-fitting algorithm, the situation will be avoided by widening the primary strip adequately.

2.5 Generating the z-Domain Response Data

The fact that there is a duality between the time-domain and frequency-domain for continuous-time signals indicates the possibility of identification of continuous-time linear systems either in time-domain, or in frequency-domain equivalently.

The discrete-time equivalent representation of a continuous-time system is a discrete-time system. Whose output in some way approximates the sampled response of the continuous-time system when a given signal is the input to the continuous-time system and the sampled version of that same signal is the input to the discrete system [28], [32].

Time-domain methods usually determine a discrete-time (z-domain) model; on the other hand, frequency-domain methods can identify either a discrete-time (z-domain) or a continuous-time (s-domain) method [33].

2.5.1 z-Domain Response Data from Time-Domain Observed Data

When a TD measurement has been performed, observed data is available in the form of the evenly spaced samples of the signal in the consecutive time instants. The effort of converting the data to the z-domain is limited to working out the values of the function in (2.11) at different z points. A set of equally spaced z points should be selected along an adequate contour in z-plane. This set at most has the length of the number of time instants in the experimented data.

One of the well-established methods in the signal processing literature based on the above theme is known as "chirp z-transform". It is performed by computing the z-transform of time-domain response along a specified spiral contour [26]. The chirp contour of interest is identified by the ratio between consecutive points and the complex starting point in z-plane. An interesting and useful spiral set can be defined when M evenly spaced z points are selected along the unit circle.

This particular chirp is specified by [34]:

- Complex starting point on z-plane contour = 1
- Uniform space between z samples on the contour = $e^{(j2\pi/M)}$, where M defines the length of the transform.

Mathematically, this is equivalent to the discrete Fourier transform of time response or $\mathrm{fft}(x)$.

2.5.2 z-Domain Response Data from Frequency-Domain Observed Data

When the measurement results are available in frequency-domain, the z-domain response data can be generated by either applying a proper (e.g. bilinear) transformation

to map s-domain to z-domain or using Inverse Fourier Transform (IFFT) to obtain time-domain data first. The z-domain response data can be extracted using the calculated time-domain data by employing the method explained in the previous section.

2.5.3 Bilinear Transformation

Converting from s-domain to the z-domain theoretically can be performed by exploiting direct definition of z variable as an exponential function, $z = e^{s \times T_{sampling}}$. Practically, a bilinear approximation of above function is utilized to convert data and systems from s-to-z domain (and vice versa). It is the ratio of the linear terms in Maclaurin series as shown below.

$$z \triangleq e^{s \times T_{sampling}} = \frac{e^{s \times \frac{T_s}{2}}}{e^{-s \times \frac{T_s}{2}}} = \frac{1 + \frac{T_s}{2} s}{1 - \frac{T_s}{2} s}$$

$$s \mapsto \; z = \frac{1 + (T_s/2)s}{1 - (T_s/2)s} \tag{2.44}$$

This is a "linear fractional transformation" known as bilinear (or Mobius Transformation [23]).

Property 1: Bilinear Transformation of the form in (2.44) conformally maps the left half-plane Π^- onto the unit disk $\mathbb{D}$, the right half-plane Π^+ onto the exterior of the unit disk $\mathbb{E}$, and the imaginary axis onto the unit circle $\mathbb{T}$.

Proof:

Given $s=\sigma+j\omega$: $z=\dfrac{1+(T_s/2)s}{1-(T_s/2)s}=\dfrac{1+(T_s/2)\sigma+(T_s/2)j\omega}{1-(T_s/2)\sigma-(T_s/2)j\omega}$

i) If $s\in\Pi^-$ then $\sigma<0$, $\Rightarrow$ $|z|<1$ for any ω so, $z\in\mathbb{D}$

ii) If $s\in\Pi^+$ then $\sigma>0$, $\Rightarrow$ $|z|>1$ for any ω so, $z\in\mathbb{E}$

iii) If s on $j\omega$ axis, then $\sigma=0$, $\Rightarrow$ $|z|=1$ for any ω so, $z\in\mathbb{T}$ maps onto the unit circle.

This conclusion complies with the details explained in the section 2.4.2.

Reversely, the bilinear transformation of the following form maps z-domain back to the s--domain.

(2.44) $\Rightarrow$ $z\left(1-(T_s/2)s\right)=1+(T_s/2)s,\quad \left((T_s/2)z+(T_s/2)\right)s=z-1$

Then: $$z\mapsto\ s=(2/T_s)\frac{z-1}{z+1} \qquad (2.45)$$

Bilinear transformation introduced above is one of the most common ways to form a discrete-time equivalent system. It is carried out by application of the bilinear transformation to the transfer function of a continuous-time system.

For example, assume that in the frequency-domain, the transfer function of a linear, continuous-time, time-invariant, finite-dimensional system is known in its closed form as H(s). By utilizing the equality in (2.45) a z-transform analytical representation for the discrete-time equivalent system denoted as H(z), can be attained simply by substitution of the quantity $(2/T_s)\dfrac{z-1}{z+1}$ for the complex frequency variable s in H(s). Mathematically, the bilinear transformation yields a set of difference equations based on

the numerical integration of the continuous-time system differential equations using the trapezoidal rule [35].

Corollary: Bilinear transformation of the forms in (2.44) and (2.45) preserve the stability property of the transfer function.

Hence, bilinear transformation coupled with Routh stability criterion [27] is a method frequently used in stability analysis of discrete time control systems.

2.5.4 Complementary Notes

I. Impulse- Invariant Transformation

The approach in section 2.1.2, outlined for theoretically clarifying the discrete time-domain concepts, is referred as *"impulse-invariant transformation"*. By definition, an impulse invariant discrete-time equivalent system produces an output that is the same as the sampled output of the continuous-time system. When the input to the discrete-time equivalent system is a unit pulse[12] and the input to the continuous-time system is the unit impulse.

II. Generalized Bilinear-Transform

[12] The "Unit Pulse" in discrete time-domain is defined as: $\{\delta(n)\}_{n=0}^{\infty} = \{1,0,0,\ldots\}$ so only the first term of this sequence is non-zero. As a time signal it is $x(0)=1$, and $x(n)=0$, for all $n>0$. Z-transform of unit pulse can be worked out using the definition of the z-transform in (2.11).

$$\mathcal{Z}\left\{\{\delta(n)\}_{n=0}^{\infty}\right\} = 1\times\frac{1}{z^0} = 1$$

If the form of the bilinear transformation $(z \mapsto s)$ in (2.45) is generalized in the following form by defining parameter α as a coefficient, one degree of freedom will be attributed to the transformation.

$$z \mapsto \quad s = \alpha \frac{z-1}{z+1}, \quad \alpha > 0 \tag{2.46}$$

Consequently, an explicit isomorphism would be feasible between $\mathcal{H}_2(\mathbb{E})$ and $\mathcal{H}_2(\Pi^+)$ by utilizing(2.46), which means this bilinear transformation preserves orthonormality of the functions [36], [37].

III. Negative Real Poles in the z-domain

It has been investigated and illustrated in section 2.4.1 that by using impulse invariant inverse transformation from s-to-z-domain, a complex conjugate pair of poles located at Nyquist frequency boundaries in s-domain is associated to a negative real pole in z-domain, which is seemingly a violation of isomorphism between-domains. In the case of the $z \mapsto s$ bilinear transformation, a z-domain negative real pole is associated to a negative real point in s-domain. Although this superficially ensures the isomorphism, the fact reveals that the method is unable in capturing the original complex conjugate poles.

IV. ZOH -Transformation Method

Beside the bilinear and impulse invariant transforms outlined above, there is the ***step-invariant*** or ***ZOH-transformation*** method in the literature. A step invariant discrete-time equivalent system produces an output that is the same as the sampled output of a continuous-time system when the input to each system is the unit step

function [35]. It uniquely determines the corresponding z-domain model from the s-domain rational transfer function of the LTI continuous-time system subject to fulfillment of specific assumptions [27], [32], [33], [38].

$$H(z) = ZOH\{H(s)\} = \left(1 - z^{-1}\right) \times \mathcal{Z}\left\{\mathcal{L}^{-1}\left\{\frac{H(s)}{s}\right\}\right\} \tag{2.47}$$

$\mathcal{Z}\{\bullet\}$ and $\mathcal{L}^{-1}\{\bullet\}$ denote the z- and inverse Laplace transforms respectively. It is proved that the transform in (2.47) is linear. By utilizing ZOH-transform on a first order system, it is shown that the s-domain pole $s_1 = -\alpha$ is transferred to the z-domain pole $z_1 = e^{s_1 T_s}$. Transforming the negative z-domain pole is still a place of further concern. A generalization of this method introduced in [33] maps the negative real z-domain poles to the poles on the aliasing margin. Precise writing for each negative pole in z-domain this method assigns a complex s-domain pole pair at Nyquist frequency to the system. Then the order of the s-domain system will be higher than that of the z-transform one, by the number of the negative real poles [33].

V. Bilinear Transformation is Preferable

It was showed thus far that the bilinear transformation preserves the stability characteristics when mapping between complex s and z -planes. This follows from the fact that stable and unstable poles for the continuous-time system are mapped respectively to stable and unstable poles in the discrete-time system, and vice versa. The second advantage for bilinear transformation defined of the form in (2.46) is due to a one degree of freedom (provided by α). By exploiting the bilinear transformation the

frequency-domain response function, $\hat{H}(s)$, obtained from the z-domain system function, $H(z)$, can preserve the magnitude characteristic of original transfer function. These main advantages are highlighted as the reasons for selecting bilinear transformation over the two other methods.

VI. Sampling Period for Transformation

In the above conversions choosing a proper value for sampling period ($T_{sampling}$) has a great place of importance to preserve the accuracy and avoiding **aliasing** effect. According to the Nyquist–Shannon sampling theory, to assure an accurate reconstruction of original signal from the sampled data, the sampling rate should be lower bounded to the Nyquist rate. It is two times the highest frequency presents in the spectrum of the signal.

Thus: $$\frac{1}{T_{sampling}} = f_S \geq 2 \times F_{\max}\Big|_{\text{in the spectrum of the signal}}$$

CHAPTER 3. Background on Rational Orthonormal Functions

It was pointed out in the first chapter that decomposing the transfer functions of the Linear Time Invariant (LTI) dynamics in terms of 'rational bases' has been dealt as a fundamental idea in various areas of applied mathematics, control theory, signal processing and system analysis. The construction, analysis and application of 'rational orthonormal bases (ROB)' suitable for describing LTI system characterizations also recently has been a subject of renewed interest with the goal of improving the accuracy. "This approach is of significant utility when accurate system descriptions are achieved with only a small number of basis functions" [39]. Exploiting the (z-domain) orthonormal bases for LTI systems identification is the core idea in this thesis.

Due to the large amount of the theoretical work on the orthogonal rational functions over the past several decades, it has been a robust concept in mathematics. In the system theoretic context, the applications of the orthonormal bases have also been manifold, but have concentrated mainly on the discrete time setting. Accordingly, this chapter has been devoted to investigate the orthonormal functions from the theoretical perspective with emphasis on discrete-domain structures.

3.1 Mathematical Concepts and Definitions[1]

First, the concept of orthogonality in the 'real value analysis' is reviewed. Next, the idea of orthonormality will be generalized for the complex analysis. It is the case for system identification applications.

3.1.1 Orthogonality of Functions

Definition 1: Let $\varphi_m(x)$ and $\varphi_n(x)$ be two "real-valued" functions that are defined on an interval $a \le x \le b$ and are such that the integral of product $'\varphi_m(x)\varphi_n(x)'$ over the interval $a \le x \le b$ exists. As a general standard notation, ***the inner product of two functions*** denoted as $\langle \varphi_m, \varphi_n \rangle$ is an integral as shown below:

$$\langle \varphi_m, \varphi_n \rangle = \int_a^b \varphi_m(x)\varphi_n(x)dx \tag{3.1}$$

Definition 2: Given $\mathbf{\Phi}$ as a collection of (here real-valued) functions as: $\mathbf{\Phi} = \{\varphi_1(x), \cdots, \varphi_i(x), \cdots\}$; it is called an ***orthogonal set of functions*** with respect to the inner product on an interval $a \le x \le b$, if all these functions are defined on the interval of interest and all the inner product integrals (3.1) exist and are zero for all pairs of distinct functions in the set.

Definition 3: The non-negative square root of $\langle \varphi_i, \varphi_i \rangle$ is called ***norm of function*** $\varphi_i(x)$ generally denoted as in (3.2)

[1] For the outlined pure mathematical concepts in this section, the ref. [23] in the bibliography list has been mainly relied.

$$\|\varphi_i\| = \sqrt{\langle \varphi_i, \varphi_i \rangle} = \sqrt{\int_a^b \varphi_i^2(x)dx} \tag{3.2}$$

The norm of an arbitrary non-zero function $\varphi_i(x)$ is assumed to be non-zero and bounded.

Definition 4: A set of functions over a range of interest, $a \le x \le b$, is called ***orthonormal*** iff:

$$\langle \varphi_m, \varphi_n \rangle = \int_a^b \varphi_m(x)\varphi_n(x)dx = \delta_{mn} \quad \text{(Kronecker delta)}, \tag{3.3}$$

where $$\delta_{mn} = \begin{cases} 0 & m \neq n \\ 1 & m = n \end{cases} \; ; \; m,n = 1,2,\cdots$$

Definition 5: Let $f(x)$ be a given continuous function that on the interval $a \le x \le b$ can be presented with a linear combination of orthogonal functions in the form of ***convergent*** series,

$$f(x) = \sum_{i=1}^{\infty} a_i \varphi_i(x) \tag{3.4}$$

The series shown in (3.4) is called a ***generalized Fourier series*** of $f(x)$. Coefficients a_i are called the ***Fourier constants*** of $f(x)$ with respect to the orthogonal set of functions.

- Theoretically, this orthogonal set of functions consists of infinitely large number of possible φ_n for $n = 1,2,\cdots$.
- Similarly, the definition stays valid for the case of orthonormal set of functions.

At this point, it should be noted that, using an infinitely large number of basis to factorize the function $f(x)$ is a subjective mathematical concept. However, for

engineering applications, a set of "sufficiently many" basis functions is considered as a practically complete set. Accordingly, the problem of convergence of constructed series would lead to the concern of ***completeness*** for the exploited orthonormal sets.

Definition 6: Consider $\phi_n(x)$ denoting a linear combination of 'n' orthonormal functions as $\phi_n(x) = \sum_{i=1}^{n} a_i \varphi_i(x)$, where n can be "infinitely large" (as a mathematically abstracted approach). Based on the definition of the '***convergence in the norm***' (or mean square convergence) $\phi_n(x)$ is called ***convergent*** with the limit of f if:

$$\lim_{n\to\infty} \left\| \phi_n(x) - f \right\| = 0 \tag{3.5}$$

Considering the Definition 3, equation (3.5) would logically appear as:

$$\lim_{n\to\infty} \int_a^b \left(\phi_n(x) - f(x) \right)^2 dx = 0 \tag{3.6}$$

It means that function $f(x)$ can be "equivalently" presented by $\phi_n(x)$, an "infinite" series. In order to adapt this pure mathematical definition for the applied conditions, let the term "equivalently" pragmatically be modified as "reasonably approximated" and term "infinite" as "adequately large". Consequently, the concept of convergent would lead to another more practical definition outlined below.

Definition 7: Consider **F** to be a set of functions $f(x)$ defined on $a \le x \le b$ and there is an orthonormal set of functions like $S_n = \{\varphi_1, \varphi_2, \cdots, \varphi_n\}$ on the same interval. If every

$f \in \mathrm{F}$ can be "reasonably approximated" with a linear combination as $\phi_n = a_1\varphi_1 + a_2\varphi_2 + \cdots + a_n\varphi_n$, by definition, the orthonormal set S_n is ***complete***.

3.1.1.1 Complementary Explanation

An adequate theoretical background for orthogonal functions from applied perspective has been provided thus far. Proceeding further into the depth of mathematical concepts falls beyond the scope of this thesis.

Most references peruse the discussion beyond this point by defining 'Bessel's inequality' by considering the case of $n \to \infty$. This subsequence will also subjectively yield far beyond the practical concerns by emphasizing on the condition of $n \to \infty$ in 'Parseval's equality'[2] and its corollary 'completeness theorem'.

The weight of these theoretical concepts in understanding the ***convergence*** and ***completeness,*** and ***uniqueness*** properties for the orthogonal realizations is not questionable. However, to provide an applied approach only the concept of 'completeness' was reviewed in above without providing more mathematical details.

3.2 Discrete-Time Orthonormal Rational Functions

Within the previous section, fundamental concepts of orthogonality were illustrated in the simplest possible format for real-valued functions. These concepts are also valid in the complex analysis although they appear in more involved mathematical form of equations. To continue, and in compliance with the above fundamentals, specific features of complex-value orthonormal basis functions are shown.

[2] Appendix A in [20] can be referred for derivation of Parseval's Theorem.

Definition 8: Given $X(z), Y(z) \in \mathcal{H}_2(E)$ that do not have singularities[3] on $\mathbb{T}$, the inner product is defined as the following contour integral [36],

$$\langle X, Y \rangle \triangleq \frac{1}{2\pi j} \oint_{\mathbb{T}} X(z) Y^*(1/z^*) \frac{dz}{z} \tag{3.7}$$

In this work, the application of orthonormal components in actual signals and systems decomposition is the topic of interest. Thus, the form of (3.7) can be further simplified considering $Y(z)$ as a transfer function of a physical system, bearing a particular property.

If $Y(z) = \sum_k y_k z^{-k}$ is a transfer function of a physical system, all the coefficients in rational format of the (discrete z-domain) transfer function necessarily are real[4]. This obligatory property for the transfer functions of actual systems guarantees that y_k also appears as real scalar, $y_k \in \mathbb{R}$.

Property: For $Y(z)$, a transfer function of a physical system, we have:[5]

$$Y^*\left(1/z^*\right) = Y\left(1/z\right) \tag{3.8}$$

Proof:

---

[3] A singularity is a point at which an equation, surface, etc., blows up or becomes degenerate. Singularities are often also called singular points and are extremely important in complex analysis, where they characterize the possible behaviors of analytic functions. Complex singularities are points z_0 in the-domain of a function, where it fails to be analytic [46].

[4] To hold the above criteria, the dynamic modes (poles) of the system are not allowed to happen in the exterior and on the unit circle, which is the case for stable systems. The poles should be enforced into $\mathbb{D}$ in the pole relocation algorithm.

[5] For an idea of proof see (3.17)

$$Y^*(1/z^*) = \left(\sum_k y_k z^{-k}\right)^*\Bigg|_{z\Rightarrow 1/z^*} = \left(\sum_k y_k^* \left(z^*\right)^{-k}\right)\Bigg|_{z\Rightarrow 1/z^*} = \sum_k y_k \left(\left(1/z^*\right)^*\right)^{-k} =$$

$$= \sum_k y_k \left(1/z\right)^{-k} = Y\left(\frac{1}{z}\right)$$

It is a provable that, for any orthogonal (independent) basis functions (such as $X(z)$ and $Y(z)$ in (3.7)), adapted for system identification in discrete time-domain, the property shown in (3.8) still holds.

Another formation of the equation (3.7) can be derived considering the fact that, for the contour integral in (3.7), the variable z is constrained to vary along the unit circle, $z \in \mathbb{T}$. Thus:

$$z \in T \Rightarrow |z| = 1, \quad z = |z| e^{j\angle z} = e^{j\theta} \quad ; -\pi \le \theta \le \pi$$

$$\therefore \quad z^* = e^{-j\angle z} = \frac{1}{z} \quad \Rightarrow \quad z = \frac{1}{z^*} \tag{3.9}$$

According to (3.9), the definition of the "inner product using these boundary values [36] " on the positively oriented unit circle would fall in a shortened form as followings:

for (3.7) and (3.8) we can write:

$$\langle X, Y \rangle = \frac{1}{2\pi j} \oint_{\mathbb{T}} X(z) Y(1/z) \frac{dz}{z} \quad ‡ \tag{3.10}$$

for (3.7) and (3.9), we have:

‡ An interesting method for extracting this equation directly from Parseval's Theorem can be found in appendix A in [20].

$$\boxed{\langle X,Y\rangle = \frac{1}{2\pi j}\oint_{\mathbb{T}:|z|=1} X(z)Y^*(z)\,\frac{dz}{z}} \tag{3.11}$$

Note: Equations (3.10) and (3.11) are equivalent format for the inner product of complex-valued X(z) and Y(z), when they denote arbitrary analytical functions identifying stable physical systems in discrete time-domain.

Depending upon the problem, any one of the above three equivalent forms in (3.7), (3.10) or (3.11) can be used.

3.3 Norm and Orthonormality

Defining the ***Norm*** and ***Orthonormality*** in the complex analysis (in either discreet or continuous-domain) is manageable in the same fashion with previous section.

The ***norm***[6] of $X(z) \in \mathcal{H}_2(\mathbb{E})$ denoted by $\|X\|$ and is equal to:

$$\|X\| \triangleq \sqrt{\langle X,X\rangle} \tag{3.12}$$

Definition 9: Two functions $X_m(z)$ *and* $X_n(z)$ are called ***Orthonormal*** if the following conditions are hold [36].

$$\langle X_m, X_n\rangle = 0, \quad \text{when } m \neq n \text{ and } \|X_m\| = \|X_n\| = 1$$

3.4 Constructing the ROBF for Discrete Time-domain

Consider the linearly independent functions with the general form of $\phi_k(z) = \frac{1}{z-a_k}$, for $k = 0, 1, \cdots, n$ that are not orthogonal in the space $\mathcal{H}_2(\mathbb{E})$. Direct

[6] The "norm" refers to the 2-norm which is also denoted as $\|X\|_2$

corollary from the linear independency property is that, there are no (common) multiple poles. Special class of orthonormal basis functions is constructed utilizing the Gram-Schmidt procedure [36] on the (linearly independent) partial fractions as shown in above.

To obtain the first orthonormal basis, a function is chosen and normalized by its 2-norm, $\phi_1^{\perp} = \frac{\phi_1}{\|\phi_1\|}$. Then a second orthonormal basis is constructed by projecting the second un-orthonormal function ϕ_2 by applying an appropriate transforms Π_2. The resulting function would be $(\Pi_2 \times \phi_2)$. This function should be orthogonal[7] to the first basis in the inner product sense. This fact can be investigated by checking $\langle \phi_1, \Pi_2 \phi_2 \rangle = 0$. Then, it should be normalized, $\phi_2^{\perp} = \frac{\Pi_2 \phi_2}{\|\Pi_2 \phi_2\|}$. The third orthonormal basis as well is constructed based on the third un-orthonormal function in a similar manner by applying appropriate transforms Π_3. The resulting function $(\Pi_3 \phi_3)$ should be orthogonal[8] to both $\phi_1^{\perp}$ and $\phi_2^{\perp}$. The orthonormal base would be obtained by normalizing it like

[7] The inner product integral in below is zero.

$$\left\langle \phi_1^{\perp}(z), \Pi_2 \phi_2(z) \right\rangle \triangleq \frac{1}{2\pi j} \oint_T \left[\phi_1^{\perp}(z) \times \Pi_2 \phi_2(1/z) \right] \frac{dz}{z} = 0$$

As sufficient conditions for above, it is required that the product function $\left[\phi_1^{\perp}(z) \times \Pi_2 \phi_2(1/z) \right]$ to be analytic in the exterior of the unit circle and the degree of the denominator in z to exceed that of the numerator.

[8] Meaning that both $\left[\phi_1^{\perp}(z) \times \Pi_3 \phi_3(1/z) \right]$ and $\left[\phi_2^{\perp}(z) \times \Pi_3 \phi_3(1/z) \right]$ should bear the same property explained in the previous footnote.

$\phi_3^{\perp} = \frac{\Pi_3 \phi_3}{\|\Pi_3 \phi_3\|}$. This orthogonalization process can continue until the entire set is constructed.

For the general case with possibly several complex poles, this procedure of orthogonalization will result in the so-called ***Takenaka-Malmquist*** functions.

$$F_k(z) = \prod_{i=1}^{k-1}\left[\frac{1-p_i^* z}{z-p_i}\right] \times \frac{\sqrt{1-|p_k|^2}}{z-p_k}, \; k = 1,2,\ldots, \quad p_i \in \mathbb{C}, \; |p_i| < 1 \qquad (3.13)$$

In the signal-processing context, functions (3.13) are called Kautz functions (filters). They inherit their name and a specific way of deduction from a method proposed by Kautz to orthonormalize a set of continuous-time exponential components [40], [41]. The discrete-time version can be attributed to Broome [20], [41].

Property 1: In (3.13) the sub-function $G(z) = \prod_{i=1}^{k-1}\left[\frac{1-p_i^* z}{z-p_i}\right]$ has the numerator and the denominator of the same order for which the corresponding poles and zeros are interrelated as $Z_i = \frac{1}{p_i^*}$. These attributions indicate function $G(z)$ as an all-pass transfer function.

Property 2: The projection property of all-pass structures is particularly utilized to obtain orthogonal functions from the original partial fraction bases.

A declaration of the fact can be induced in an envision sense, as below.

$$F_n^{\perp}(z) = \underbrace{\sqrt{1-|p_n|^2}}_{\substack{\text{Normalization} \\ \text{factor}}} \times \underbrace{\prod_{i=1}^{n-1}\left[\frac{1-p_i^* z}{z-p_i}\right]}_{\substack{\text{Projector function} \\ \text{to orthogionalize} \\ \text{the partial fractional Basis} \\ \text{shown to the right}}} \times \underbrace{\frac{1}{z-p_n}}_{\substack{\text{a general} \\ \text{partial} \\ \text{fraction} \\ \text{basis}}}$$

The projection property of all-pass function $G(z) = \prod_{i=1}^{k-1}\left[\frac{1-p_i^* z}{z-p_i}\right]$ leads to the following:

$$\left\langle F_j(z),\ G_k(z)F_k(z)\right\rangle = 0,\quad j \le k-1,\ \forall F_j(z)\ \&F_k(z)\ \in \mathcal{H}_2(\mathbb{E}) \qquad (3.14)$$

It is known in a somewhat more abstract setting as the Beurling-Lax theorem [36].

Property 3: The roots of denominators in (3.13) are limited to the unit disk, $p_i \in \mathbb{D}$. This reveals that p_i preserves the main stability condition expected for the poles in transfer function of physical passive stable linear system. It is analytical (non-singular thus) throughout the $\mathbb{E}$. Having poles over unit disk makes the Takenaka-Malmquist functions an appropriate candidate for the orthonormal basis in the area of the actual passive systems identification.

Property 4: The Takenaka-Malmquist functions, in their general form shown in (3.13), will have complex impulse response [36].

In contrary to Property 3, this characteristic makes the general form of these functions unfortunate for physical systems characterizations.

3.5 LTI Dynamical System Identification [9]

The use of orthogonal basis functions for the Hilbert space of stable systems has a long track in the modeling and identification of dynamical systems [15, 20, 36, 39, 40, 41, 42, 43]. Referring to the classical work of Lee [44], every stable system has a unique series of expansion in terms of pre-chosen basis functions. Assume that a z-domain transfer function of a stable physical system $H(z)$ can be adequately approximated with a finite-length series expansion using orthonormal basis functions,

$$H(z) \approx \Phi(z) = \varpi_1 \phi_1^{\perp}(z) + \varpi_2 \phi_2^{\perp}(z) + \cdots + \varpi_N \phi_N^{\perp}(z) \ . \qquad (3.15)$$

Here, N represents the required number of the poles or the order of the intended model. The coefficients of the series expansion can be estimated from the measured /tabulated data.

The above (macro-) model representation can serve as an approximated model if it follows the behavior of the original system (at the ports) closely. In addition, the model should preserve the major properties of the physical system.

The response (output signal), from actual systems with respect to any arbitrary physical excitations, is realistically expected to be a real-value in time-domain. With the property of convolution integral in mind, this fundamental property is held when the impulse response of the system is ensured to be real-value in time-domain.

Property 5: The intrinsic property for this approximated model is that when it is converted to the time-domain by applying inverse z-transforms the result is a real-valued function with respect to time.

[9] In this section the main consideration is focused on the discrete time-domain; however the outlined concepts are not merely limited to this-domain or to orthonormal functions. A most general form of concepts are also valid in continues-domain.

3.6 Real Impulse Response

It is remarked that, in system theoretical context, the transfer function of a LTI system is defined as the response of system when it is excited by an impulse source. Therefore, the corresponding time functions or inverse z-transforms [41] of z-domain transfer function in the general form of $\left\{h_n(n)\right\}_{n=0}^{\infty}$ is impulse response of the system. Hence, the ***real-valued time-domain impulse response*** of an arbitrary LTI dynamical system identified in the form of a linear span of independent orthonormal basis functions presented in (3.15) can be accurately shown as $\varphi(n) = \mathcal{Z}^{-1}\{\phi(z)\} \in \mathbb{R}$ where $\mathcal{Z}^{-1}$ denotes the inverse z-transform,

$$\varphi(k) = \mathcal{Z}^{-1}\{\phi(z)\} = \mathcal{Z}^{-1}\left\{ \varpi_1\phi_1^{\perp}(z) + \varpi_2\phi_2^{\perp}(z) + \cdots + \varpi_N\phi_N^{\perp}(z) \right\} =$$

$$\varpi_1\mathcal{Z}^{-1}\left\{\phi_1^{\perp}(z)\right\} + \varpi_2\mathcal{Z}^{-1}\left\{\phi_2^{\perp}(z)\right\} + \cdots + \varpi_N\mathcal{Z}^{-1}\left\{\phi_N^{\perp}(z)\right\} = \sum_{i=1}^{N} \varpi_i\varphi_i\left(kT_S\right) \quad (3.16)$$

N : presents the practical order of the model (theoretically $N = \infty$),
k : is the count of the time instances,
T_S : shows sampling period

As shown in (3.16), the total time-domain signal is constructed from a linear combination of inverse z-transforms of the basis functions. The following conditions assure the real sequence, shown in (3.16) as a time-domain signal.

Property: The z-transform of a real sequence satisfies the following condition [36],

$$\phi(z) = \phi^*(z^*) \quad (3.17)$$

3.7 System Identification Using Partial Fraction Bases

It is recalled that a real (physical) system can be identified (approximated) by the linear combination of the set of (conventional) partial fraction basis functions, $\phi_n = \frac{1}{z-p_n}$, as:

$$\phi(z) = \sum_{n=1}^{N} R_k \phi_n = R_1\left(\frac{1}{z-p_1}\right) + R_2\left(\frac{1}{z-p_2}\right) + \cdots + R_N\left(\frac{1}{z-p_N}\right) \qquad (3.18)$$

To ensure a real (time-domain) impulse response, the constructed transfer function $\phi(z)$ in (3.18) contains real poles and real corresponding residues as well as (more often) complex conjugate poles and residues. Consequently, the time-domain transform appears as real responses. It results from the basis of the real poles, as well as "complex conjugate time-domain transforms" from basis associated with the pairs of complex poles.

$$R_n\phi_n(z) + R_{n+1}\phi_{n+1}(z), \quad \text{where } p_{n+1} = p_n^* \text{ and } R_{n+1} = R_n^*$$

$$\mathcal{Z}^{-1}\left\{R_n\phi_n(z) + R_n^*\phi_{n+1}(z)\right\} = R_n\varphi_n(k) + R_n^*\varphi_{n+1}(k)$$

$$= \underbrace{\{h_n + h_{n+1}\}}_{\substack{\text{2 complex}\\ \text{conjugate}\\ \text{time-domain}\\ \text{transforms}}}(t) \quad \Rightarrow \text{ real-valu time domain signal}$$

h_n and h_{n+1}[10] are complex conjugate TD transforms (mathematical functions but not signals), conforming to a real TD response[11].

To preserve the above property, this time with real coefficients, $\omega_i \in \mathbb{R}$ the modified form of the partial fraction functions has been introduced in [13]. These bases are directly derived from partial fractions by a simple algebraic manipulation[12]. It ensures that the residues associated with complex conjugate poles come in perfect conjugate pairs while all the coefficients in linear combination of new bases (below) are being enforced as real values.

Consider the real and complex conjugate poles lying at

$$
\begin{aligned}
& p_i = \alpha_i; && \phi_i(z) = \frac{1}{z - p_i} \\
& p_{2i-1} = -\alpha_i - j\beta_i; && \phi_{2i-1}(z) = \frac{1}{z - p_{2i-1}} + \frac{1}{z - p_{2i}} \\
& p_{2i} = p^{*}_{2i-1} = -\alpha_i + j\beta_i; && \phi_{2i}(z) = \frac{j}{z - p_{2i-1}} - \frac{j}{z - p_{2i}},
\end{aligned}
\qquad (3.19)
$$

where $\alpha_i, \beta_i \in \mathbb{R}$

Since the application of the partial fraction is not limited to stable problems, in (3.19) it is not considered $|P_i| < 1$ for discrete-domain or equivalently, $\alpha_i > 0$ for continuous-domain. If unstable poles are allowed, one still can resort to the partial fraction bases. In

[10] The notation as $\{h_k + h_{k+1}\}(t)$ emphasizes that none of the h_k *and* h_{k+1} transforms presents a real time-domain signal on its own.

[11] Considering the inverse z-transform rules to obtain each sub-response can reveal the fact better.

[12] The analogous form for continues-time system is reviewed in the Appendix C, where instead of z-transforms, Laplace transforms is considered.

contrary, for the orthonormal basis functions, the assumption of the stable poles is essential.

3.8 ROBF for Continuous-Time LTI Systems Identification

This section focuses attention on the continuous time scenario by considering the set of basis functions defined by a choice of a set of stable poles on the open left half plane in the complex s-plane, $a_i \in \mathbb{C}^-$.

The form for the continuous time orthonormal functions is introduced in section 3.8.1 with adequate details. These functions are orthonormal on $\mathcal{H}_2(\mathbb{C}^+)$ with respect to the inner product. By replacing the exterior of the unit circle, $\mathbb{E}$, and unit circle $\mathbb{T}$ from discrete-domain with the right half of the complex plane $\mathbb{C}^+$ and the imaginary axis respectively, the scalar 'inner product for continues-domain as shown in (3.20)' [36] is obtained.

$$\langle X(s), Y(s) \rangle \triangleq \frac{1}{2\pi i} \int_{i\mathbb{R}} X(s) Y^*(-s^*) ds \,, \tag{3.20}$$

in the above equation, integration is performed over the imaginary axis.

3.8.1 Orthonormal Set for Continuous-Time

The general polynomials form for the continuous-time orthonormal basis function is:

$$\Phi_n(s) = \kappa_n \times \sqrt{-2p_n} \times \left(\prod_{i=1}^{n-1} \frac{s + p_i^*}{s - p_i} \right) \times \frac{1}{s - p_n} \tag{3.21}$$

In the mathematical level of generality, κ_i is an arbitrary unimodular[13] complex number. This base originates from the discrete-time Takenaka-Malmquist function in (3.13) by transforming it to the continuous time-domain.

- The special cases of the basis (3.21) wherein all poles are the same real number, $p_i = -\alpha \in \mathbb{R}$, is known as the '**Laguerre**' basis [39], [45, p.241].
- If all the poles are real, $p_i = -\alpha_i \in \mathbb{R}$ or where all poles are the in complex conjugate pairs and lie at $\left(p_{2\nu-1} = -\alpha_\nu - j\beta_\nu\right)$ and $\left(p_{2\nu} = -\alpha_\nu + j\beta_\nu\right)$ for $\left(\nu = 1, 2, \ldots, N/2\right)$ the function (3.21) would fall into a set of transforms known as **Kautz** bases (filters) [20], [40].

In the control and signal processing literature, the orthonormal bases mostly refer as Kautz filters. These bases, introduced in [40] were originally obtained from orthogonalizing the exponential sequences and then adapted by Kautz. Appendix B clarifies the Kautz functions in continuous time-domain and identifies them as a variation of the Takenaka-Malmquist functions with special conditions.

3.8.2 Generalized Orthonormal Bases

In appendix B, it is declared that the inverse Laplace transforms of an arbitrary transfer function, approximated with a linear combination of Kautz bases with real coefficients, presents a real-valued time-domain function. The Kautz functions have been adapted for specific conditions, when all the poles are either real or in complex

[13] $|\kappa_n| = 1$, Based on [15]

conjugate pairs. However, this is not always the case in linear dynamical system identification problems, wherein a combination of real and complex conjugate poles are mostly required to model the behavior of the system accurately. Therefore, the original Kautz functions cannot be considered satisfactorily general. To obtain the generalized form for the orthogonal basis [15] adopting the dynamical system representation, the idea of Laguerre functions and two-parameter Kautz functions can be expanded and generalized as outlined in the next chapter.

3.8.3 Orthonormal Basis Functions in Continuous-Domain [14]

Laguerre functions are especially appropriate for accurate modeling of systems with dominant first-order dynamics, whereas Kautz functions are directed toward systems with dominant second-order resonant dynamics [22]. The basis functions in this section suite the systems with a wide range of dominant dynamics. With a set of real and complex conjugate poles, exploiting the following generalized bases ensures the real coefficients in the resulting transfer function. It is worth noting that the assumptions of no coincident poles and negative real parts for all the poles are essentially important.

- **Orthonormal basis associated with a real pole, $p_n = -\alpha_n$:** [15]

$$\Phi_n(s) = \sqrt{-2p_n} \times \left(\prod_{i=1}^{n-1} \frac{s + p_i^*}{s - P_i} \right) \times \frac{1}{s - p_n} \tag{3.22}$$

And when $p_n = p_{n+1}^*$

[14] Based on [15]

[15] in the system identification problems to guarantee the real-valued coefficients for the resulting transfer function it should be strictly assumed $\kappa_i = 1$.

- **Orthonormal basis associated with a complex pole, $p_n = -\alpha_n - j\beta_n$:**

$$\Phi_n(s) = \sqrt{-2\,\mathrm{Re}(p_n)} \times \left(\prod_{i=1}^{n-1} \frac{s + p_i^*}{s - p_i} \right) \times \frac{s + |p_n|}{(s - p_n)(s - p_n^*)} \tag{3.23}$$

- **Orthonormal basis associated with the complex pole, which is the complex conjugate of p_n , $p_{n+1} = p_n^* = -\alpha_n + j\beta_n$:**

$$\Phi_{n+1}(s) = \sqrt{-2\,\mathrm{Re}(p_n)} \times \left(\prod_{i=1}^{n-1} \frac{s + p_i^*}{s - p_i} \right) \times \frac{s - |p_n|}{(s - p_n)(s - p_n^*)} \tag{3.24}$$

CHAPTER 4. Proposed z-Domain Orthonormal Basis Functions

As it was mentioned in Chapter 3, the original discrete-domain Takenaka-Malmquist functions of the form shown in (4.1), will not have real impulse response. They produce real-valued signal only in the case of real poles. Considering the fact that, realization of a physical passive system in most cases requires a combination of real and complex conjugate pairs of poles, it is revealed that Takenaka-Malmquist functions cannot be directly utilized for system identification purposes. To address the issue, this chapter is devoted to propose a novel formulation for the Rational Orthonormal Basis Functions (ROBF). The proposed basis functions provide the ***real-valued time-domain impulse response*** even for complex conjugate poles. To obtain the proposed functions the Takenaka-Malmquist bases have been extended for the real-valued time-domain impulse response of LTI systems.

$$F_k(z)=\prod_{i=1}^{k-1}\left[\frac{1-p_i^* z}{z-p_i}\right]\times\frac{\sqrt{1-|p_k|^2}}{z-p_k},\ k=1,2,\ldots,\quad p_i\in\mathbb{C},\ |p_i|<1 \qquad (4.1)$$

In system identification and in signal processing context, it is stated that if the poles are real or occurs in a complex conjugate pair, $p_{n+1}=p_n^*$, it is possible to form linear combinations of corresponding orthonormal bases, $\phi_n(z)$ and $\phi_{n+1}(z)$, to obtain two orthonormal functions with real impulse response, spanning the same space [15], [36].

There is no reason that a possible solution will be unique. The structure of sets of orthonormal sequences proposed by Broome in [20] can be mentioned as an instance of attempts in this direction. However, in order to arrange an efficient VF problem formulation, an intuitively simpler analytical form is necessary. Accordingly, a set of orthonormal sequences possessing the explained properties is proposed in this thesis.

4.1 The Compulsory Properties for Proposed Functions

The primary objective is to ensure the resulting transfer function, expressed in the form of a ratio of two polynomials, to have real-valued coefficients. The focus of this chapter is centered on constructing a sequence of orthonormal functions $\phi = \{\phi_n(z):\ n = 1,2,\ldots N\}$ by modifying the Takenaka-Malmquist functions to satisfy this constraint. The resulting basis functions should ensure the following essential properties.

A) The basis associated with the real stable poles, $p_{n=-\alpha_n}$, where $\alpha_n \in \mathbb{R}$ and $0 < \alpha_n < 1$,

$$\phi_n(z) = F_n(z) \tag{4.2}$$

The original format of Takenaka-Malmquist functions associated to the real poles can produce real output signal.

B) For the $\phi_n(z)$ and $\phi_{n+1}(z)$, new bases corresponding to the complex conjugate pair of the poles $\left(p_n,\ p_n^*\right)$, it is:

i) $\phi_n(z)$ and $\phi_{n+1}(z)$ are perpendicular to each other, as well as to every other elements of the sequence,

$$\langle \phi_n, \phi_{n+1} \rangle = 0$$

$$\langle \phi_i, \phi_n \rangle = 0 \quad ; i = 1,2,\ldots,N-1, \quad n > i$$

$$\langle \phi_i, \phi_{n+1} \rangle = 0 \quad ; i = 1,2,\ldots,N-2, \quad n+1 > i$$

ii) $\phi_n(z)$ *and* $\phi_{n+1}(z)$ are unimodular complex-valued functions,

$$\|\phi_n\|_2^2 = \langle \phi_n, \phi_n \rangle = 1$$

$$\|\phi_{n+1}\|_2^2 = \langle \phi_{n+1}, \phi_{n+1} \rangle = 1$$

This property should hold for the entire basis.

Hence, the properties (i) and (ii) prove that $\phi_n(z)$ and $\phi_{n+1}(z)$ are orthonormal functions.

C) The inverse z-transform of any linear combination of $\phi_n(z)$ *and* $\phi_{n+1}(z)$ with real coefficients should be a real function.

D) $\phi_n(z)$ *and* $\phi_{n+1}(z)$ should span the same space with $F_n(z)$ and $F_{n+1}(z)$.

4.2 The Formulation of Proposed Orthonormal Functions

Followings are the proposed z-domain rational orthonormal basis functions (ROBF) ensuring the real time-domain impulse response.

- For p_n as a real and stable pole, $p_n \in \mathbb{R}$, $|p_n| < 1$:

$$\phi_n(z) = \kappa_n \times \prod_{i=1}^{n-1}\left[\frac{1-p_i^* z}{z-p_i}\right] \times \frac{1}{z-p_n} \tag{4.3}$$

- For complex conjugate stable poles, $(p_n = -\alpha_n - j\beta_n)$ & $\left(p_{n+1} = p_n^*\right) \in \mathbb{C}$, $|p_n| < 1$

$$\phi_n(z) = \gamma_n \times \prod_{i=1}^{n-1}\left[\frac{1-p_i^* z}{z-p_i}\right] \times \frac{1-z}{\left(z-p_n\right)\left(z-p_n^*\right)} \tag{4.4}$$

$$\phi_{n+1}(z) = \gamma_{n+1} \times \prod_{i=1}^{n-1}\left[\frac{1-p_i^* z}{z-p_i}\right] \times \frac{1+z}{\left(z-p_n\right)\left(z-p_n^*\right)} \tag{4.5}$$

$$\begin{cases} \kappa_n = \sqrt{1-p_n p_n^*} = \sqrt{1-|p_n|^2} \\ \gamma_n = \left(\frac{\sqrt{2}}{2}|1+p_n|\right)\sqrt{1-|p_n|^2} \\ \gamma_{n+1} = \left(\frac{\sqrt{2}}{2}|1-p_n|\right)\sqrt{1-|p_n|^2} \end{cases} \tag{4.6}$$

Since a non-unique solution is expected, the problem of extracting the above bases from the mathematical derivation falls into heuristic classification. Then providing a theoretical justification to address whether the obtained solution is optimum is not feasible. The next section presents a mathematical investigation of the expected important properties in the suggested functions.

4.3 Investigation of Properties of the Proposed ZD-OBF

To continue, the essential properties of the proposed ZD-OFB will be precisely discussed. It will be investigated that these functions, fully satisfying the essential properties, can be considered a complete set of orthonormal bases.

I) Elementary property of the all-pass filter block[1] in the ZD-OFB:

Lemma 1: The all-pass function defined as $G(z)=\prod_{i=1}^{k-1}\left(\frac{1-p_i^* z}{z-p_i}\right)$, $z\in\mathbb{C}$, $p_i\in\mathbb{D}$

satisfies the condition. $$G(z)G^*(\frac{1}{z^*})=1 \tag{4.7}$$

PROOF:

$$G^*(\frac{1}{z^*})=\left(\prod_{i=1}^{k-1}\frac{1-p_i^* z}{z-p_i}\right)^*\Bigg|_{z\Rightarrow\left(\frac{1}{z^*}\right)}=\prod_{i=1}^{k-1}\left(\frac{1-p_i z^*}{z^*-p_i^*}\Bigg|_{z\Rightarrow\left(\frac{1}{z^*}\right)}\right)=\prod_{i=1}^{k-1}\frac{z-p_i}{1-p_i^* z},\quad z\neq 0 \tag{4.8}$$

substituting (4.8) in(4.7):

$$G(z)G^*(\frac{1}{z^*})=\prod_{i=1}^{k-1}\left[\frac{1-p_i^* z}{z-p_i}\times\frac{z-p_i}{1-p_i^* z}\right]=1 \tag{4.9}$$

Note: The property (4.7) is held only when the point $z=0$ is excluded from the range of the function $G(z)$ means:

$$G(z)G^*(\frac{1}{z^*})=1,\quad z\in\mathbb{C}-\{0\} \tag{4.10}$$

[1] This sub-function and some of its properties have been previously addressed.

Property 1: For every z on the unit circle there is:

$$G(z)G^*(z)=1, \quad z\in\mathbb{T}=\{z:\ |z|=1\} \tag{4.11}$$

PROOF:

By substituting $\frac{1}{z^*}=z$ for $z\in T:|z|=1$ as it was showed before, in (4.10).

Eq. (4.11) will be frequently use when proving the properties for the ZD-OBF in continue.

II) Cauchy's Integral Formula:

Cauchy's integral formula and its conditions of validity can be stated as follows. This formula will be utilized to evaluate complex contour integrals. The complex integration chapter in [23] can be referred for the details and proof.

Theorem 1: (Cauchy's integral formula) Let $f(z)$ be analytical[2] in a simply connected-domain $\tilde{D}$. Then for any point z_0 in $\tilde{D}$ and any simple closed path C in $\tilde{D}$ that encloses z_0 (Figure 4-1), we have:

$$\oint_C \frac{f(z)}{z-z_0}\,dz = 2\pi j \times f(z_0) \tag{4.12}$$

[2] **Definition (Analyticity)**
A function $f(z)$ is said to be analytic in a domain $\mathbb{D}$ if $f(z)$ is defined and differentiable at all points of $\mathbb{D}$. The function $f(z)$ is said to be analytic at a point $z=z_0$ in $\mathbb{D}$ if $f(z)$ is analytic in a neighborhood of z_0.

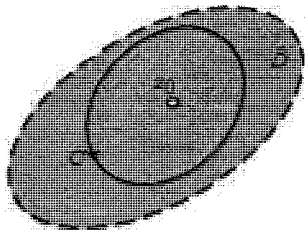


Figure 4-1: Cauchy's integral

Note: The integration is taken counterclockwise.

Lemma 2: A set of the proposed discrete basis functions $\{\phi_n(z)\}_{n>1}$ forms an orthonormal set.

Requirements for Orthogonality: A set of functions is an orthonormal set if the scalar inner product of any one with any other function of the set has a value of zero and the inner product of any one with itself has a value of unity.

The above Lemma is proved with a better clarity if it is broken down into followings.

Case *I*: If $\phi_m(z)$ is any of the bases in the set associated with real poles, it has a unity norm and is orthogonal to any other basis in the set corresponding to any other real pole (multiple poles are not allowed).

PROOF:

Since the Takenaka-Malmquist original functions are orthonormal, this is a fundamentally verified fact. However, a short account of the proof is clarified in below.

Let p_m and p_n be any arbitrary stable poles, p_m, $p_n \in \mathbb{D}$. Without loss of generality, it is assumed $n > m$.

For Orthogonality:

Thus, $\phi_m(z) = \kappa_m \times \underbrace{\prod_{i=1}^{m-1}\left[\frac{1-p_i^* z}{z-p_i}\right]}_{G_m(z)} \times \frac{1}{z-p_m}$, $\phi_n(z) = \kappa_n \times \prod_{j=1}^{n-1}\left[\frac{1-p_j^* z}{z-p_j}\right] \times \frac{1}{z-p_n}$

$$\phi_m(z) = \kappa_m \times G_m(z) \times \frac{1}{z-p_m} \quad \text{and} \quad \phi_n(z) = \kappa_n \times G_m(z) \times \prod_{j=m}^{n-1}\left[\frac{1-p_j^* z}{z-p_j}\right] \times \frac{1}{z-p_n}$$

$$\left\langle \phi_m, \phi_n \right\rangle = \frac{1}{2\pi j} \oint_T \phi_m(z)\phi_n^*(z) \frac{dz}{z} =$$

$$\underbrace{\frac{\kappa_n \kappa_m}{2\pi j} \int_{\substack{-\pi \\ |z|=1}}^{+\pi} G_m(z) \times \frac{1}{z-p_m} \times G_m^*(z) \times \prod_{j=m}^{n-1}\left[\frac{1-p_j z^*}{z^*-p_j^*}\right] \times \frac{1}{z^*-p_n^*} \times \frac{dz}{z}}_{\mathbf{I}}$$

by considering: $G(z)G^*(z) = 1$ for $z \in \mathbb{T} = \{ z: \ |z| = 1 \}$

$$\mathbf{I} = \int_{\substack{-\pi \\ |z|=1}}^{+\pi} \overbrace{\prod_{j=m}^{n-1}\left[\frac{z-p_j}{1-p_j^* z}\right]}^{F_{n-m}(z)} \times \frac{1}{1-p_n^* z} \times \frac{1}{z-p_m} \, dz$$

To preserve the analyticity of function $F_{n-m}(z)$ within the unit disk region, the point $z = p_m$ cannot be excluded from its-domain. This fact forces $F_{n-m}(z)$ to keep the zero at p_m [†].

[†] This point may superficially resemble a simple "pole-zero cancellations" [20]; however, it would be more mathematically precise, when stated as "having a zero at a pole" and not cancellation.
As a mater of fact, dividing the numerator of a fractional function by a factor is possible only when the root of the divisor is not enclosed in the range of the rational function. In other words, cancelling a term from the numerator of a function requires the divisor term not to be zero at any point along the range of the function.
Since the above integrand fractional function is analytical all over the unit disk, this condition is not hold. Thus, cancelling a zero from integrand function with a pole of integral kernel is not allowed here.

Then we can say the pole of integral kernel occurs at one of the zeros of the integrand function. This is the key point in the form of the Takenaka-Malmquist functions that guarantees the orthogonality property.

The general result logically anticipated from Cauchy integral formula is:

$$\left\langle \phi_m, \phi_n \right\rangle = \kappa_n \kappa_m \left(F_{n-m}(p_m) \right) = 0, \qquad \forall n-m>0$$

In a similar fashion, using the above argument for $\left\langle \phi_m, \phi_n \right\rangle$, $n=1,2,\ldots,m-1$, leads to the zero structure as $F_{n-m}(p_m)$, $\forall (n-m)>0$.

For the unity norm:

For any general term with the form of original Takenaka-Malmquist as:

$$\phi_n(z) = \kappa_n \times \prod_{j=1}^{n-1} \left[\frac{1-p_j^* z}{z-p_j} \right] \times \frac{1}{z-p_n}$$

It is:

$$\left\langle \phi_n, \phi_n \right\rangle = \frac{1}{2\pi j} \oint_T \phi_n(z) \phi_n^*(z) \frac{dz}{z} =$$

$$\frac{\kappa_n^2}{2\pi j} \int_{\substack{-\pi \\ |z|=1}}^{+\pi} G_n(z) \times \frac{1}{z-p_n} \times G_n^*(z) \times \frac{1}{z^*-p_n^*} \times \frac{dz}{z} = \frac{\kappa_n^2}{2\pi j} \int_{\substack{-\pi \\ |z|=1}}^{+\pi} \overbrace{\frac{1}{1-p_n^* z}}^{F(z)} \times \frac{1}{z-p_n} dz =$$

$$\left(1-p_n p_n^*\right) \times \frac{1}{1-p_n^* p_n} = 1$$

The case of the real stable poles $p_m, p_n \in \mathbb{D} \cap \mathbb{R}$ can be considered as a special form for the above general proof.

Case *II*: The proposed bases $\phi_n(z)$ and $\phi_{n+1}(z)$, shown below, corresponding to the stable complex conjugate pair of the poles $\left(p_n,\ p_n^*\right)$ are orthonormal respectively.

$$\phi_n(z)=\gamma_n\times\underbrace{\prod_{i=1}^{n-1}\left[\frac{1-p_i^*z}{z-p_i}\right]}_{G_n(z)}\times\frac{1-z}{\left(z-p_n\right)\left(z-p_n^*\right)}$$

and

$$\phi_{n+1}(z)=\gamma_{n+1}\times\prod_{i=1}^{n-1}\left[\frac{1-p_i^*z}{z-p_i}\right]\times\frac{1+z}{\left(z-p_n\right)\left(z-p_n^*\right)}$$

where $\left(p_{n+1}=p_n^*\right)\in\mathbb{C}$, $|p_n|<1$ and γ_n and γ_{n+1} are defined, given by $\left(\frac{\sqrt{2}}{2}|1+a_n|\right)\sqrt{1-|p_n|^2}$ and $\left(\frac{\sqrt{2}}{2}|1-a_n|\right)\sqrt{1-|p_n|^2}$ respectively.

PROOF:

First, the above two bases should be orthogonal to each other, hence it is:

$$\left\langle\phi_n,\phi_{n+1}\right\rangle=\frac{1}{2\pi j}\oint_T\phi_n(z)\phi_{n+1}^*(z)\ \frac{dz}{z}=$$

$$\frac{\gamma_n\gamma_{n+1}}{2\pi j}\int_{\substack{-\pi\\|z|=1}}^{+\pi}G_n(z)\times\frac{1-z}{\left(z-p_n\right)\left(z-p_n^*\right)}\times G_n^*(z)\times\frac{1+z}{\left(1-p_n^*z\right)\left(1-p_nz\right)}\times dz=$$

$$\frac{\gamma_n\gamma_{n+1}}{2\pi j}\int_{\substack{-\pi\\|z|=1}}^{+\pi}\overbrace{\frac{1-z^2}{\left(1-p_n^*z\right)\left(1-p_nz\right)}}^{\mathrm{F(z)}}\times\underbrace{\frac{1}{\left(z-p_n\right)\left(z-p_n^*\right)}}_{\text{❶}}dz=$$

By decomposing ❶ to the partial fraction forms:

$$\frac{\gamma_n\gamma_{n+1}}{2\pi j\left(p_n-p_n^*\right)}\int_{\substack{-\pi\\|z|=1}}^{+\pi}\frac{\mathrm{F(z)}}{\left(z-p_n\right)}-\frac{\mathrm{F(z)}}{\left(z-p_n^*\right)}dz=\frac{\gamma_n\gamma_{n+1}}{\left(p_n-p_n^*\right)}\left(\mathrm{F(p_n)}-\mathrm{F(p_n^*)}\right)=$$

$$\frac{\gamma_n\gamma_{n+1}}{\left(p_n-p_n^*\right)}\left[\frac{\left(1-p_n^2\right)}{\left(1-|p_n|^2\right)\left(1-p_n^2\right)}-\frac{\left(1-\left(p_n^*\right)^2\right)}{\left(1-\left(p_n^*\right)^2\right)\left(1-|p_n|^2\right)}\right]=0$$

Thus: $\left\langle\phi_n,\phi_{n+1}\right\rangle=0 \quad\Rightarrow\quad \phi_n\perp\phi_{n+1}$

Next, to prove the unity norm for the basis by definition, it should be:

$$\left\langle\phi_n,\phi_n\right\rangle=\left\langle\phi_{n+1},\phi_{n+1}\right\rangle=1$$

This can be investigated as below:

$$\left\langle\phi_n,\phi_n\right\rangle=\frac{1}{2\pi j}\oint_T\phi_n(z)\phi_n^*(z)\ \frac{dz}{z}=$$

$$\frac{-\gamma_n^2}{2\pi j}\int_{\substack{-\pi\\|z|=1}}^{+\pi}\overbrace{\frac{(1-z)^2}{\left(1-p_n^*z\right)\left(1-p_nz\right)}}^{\mathrm{F(z)}}\times\frac{1}{\left(z-p_n\right)\left(z-p_n^*\right)}\,dz=\frac{\gamma_n^2}{\left(p_n^*-p_n\right)}\left(\mathrm{F(p_n)}-\mathrm{F(p_n^*)}\right)$$

$$\frac{\frac{1}{2}|1+a_n|^2}{\left(p_n^*-p_n\right)}\times\left[\frac{(1-p_n)}{\left(1+p_n\right)}-\frac{\left(1-p_n^*\right)}{\left(1+p_n^*\right)}\right]=\frac{|1+a_n|^2}{2\left(p_n^*-p_n\right)}\times\frac{2\left(p_n^*-p_n\right)}{|1+a_n|^2}=1$$

By exploiting similar techniques, $\left\langle\phi_{n+1},\phi_{n+1}\right\rangle=1$ can be investigated too.

Case *III*: Assume a sequence comprised with $\left\{\phi_m(z),\phi_{m+1}(z)\right\}$ and $\left\{\phi_n(z),\phi_{n+1}(z)\right\}$, the proposed bases related to the two arbitrary stable complex

conjugate pairs, and $\{\phi_i(z)\}$ for any arbitrary real stable pole. The resulting set would be orthonormal with respect to the inner product.

PROOF:

The assumption of $n > m$ does not degrade the generality level of the problem. The convincing evidence has already been provided for:

$$\langle \phi_n, \phi_n \rangle = \langle \phi_{n+1}, \phi_{n+1} \rangle = 1, \quad \text{and} \quad \langle \phi_n, \phi_{n+1} \rangle = 0$$

also equally: $$\langle \phi_m, \phi_m \rangle = \langle \phi_{m+1}, \phi_{m+1} \rangle = 1, \quad \text{and} \quad \langle \phi_m, \phi_{m+1} \rangle = 0$$

In the light of all above explanations, with the fact in mind that $F_{n-m}(z)$ always has a zero at p_m[‡], applying the similar techniques to demonstrate the similar properties for any of the following pairs is possible by performing a few algebraic manipulations.

$$\{\phi_n, \phi_m\}, \{\phi_{n+1}, \phi_m\}, \{\phi_n, \phi_{m+1}\}, \{\phi_{n+1}, \phi_{m+1}\}, \{\phi_k, \phi_i\}_{k=n,\, n+1,\, m, \text{ and } m+1}$$

In fact, the minimum conditions on the integrand under which the inner product integral is zero can be listed as below.

(1) No poles exist on the unit circle; and either (2) The function is analytic in the interior of the unit circle, or (3) The function is analytic in the exterior of the unit circle and the degree of the denominator in z is at least two greater than that of numerator (not the case in this context). These are sufficient, but not necessary condition for integral to be zero [20].

‡ See Footnote †

Conclusion: By verifying all above hypotheses (cases), covering all possible cases in the most general form, It can be confirmed that,

"A set of the proposed discrete basis functions $\{\phi_n(z)\}_{n>1}$ is an orthonormal set."

Lemma 3: The all-pass structure $G(z)=\prod_{i=1}^{k}\left[\frac{1-p_i^* z}{z-p_i}\right]$ satisfies $G(z)=G^*(z^*)$ where

$z\in\mathbb{D}$ and $p_i\in P=\left\{p:\ p\in\mathbb{D},\ \forall\, p_i\notin\mathbb{R}\ \exists\ p_{i+1}=p_i^*\right\}$

PROOF:

$$G^*(z^*)=\left\{\prod_{i=1}^{k}\left[\frac{1-p_i^* z}{z-p_i}\right]\right\}^*\Bigg|_{z\to z^*}=\prod_{i=1}^{k}\left[\frac{1-p_i^* z}{z-p_i}\right]^*\Bigg|_{z\to z^*}=\prod_{i=1}^{k}\left[\frac{1-p_i z}{z-p_i^*}\right] \tag{4.13}$$

a) $\forall\ p_i$: Real, $\quad G_i^*(z^*)=\frac{1-p_i z}{z-p_i^*}=\frac{1-p_i z}{z-p_i}=G_i(z)$ (4.14)

b) $\forall\ p_i$: complex $\exists\, p_{i+1}=p_i^*$,

$$\left(G_i^*(z^*)\times G_{i+1}^*(z^*)\right)=\left(\frac{1-p_i z}{z-p_i^*}\times\frac{1-p_{i+1}z}{z-p_{i+1}^*}\right)$$

$$=\left(\frac{1-p_{i+1}^* z}{z-p_{i+1}}\times\frac{1-p_i^* z}{z-p_i}\right)=\left(G_i(z)\times G_{i+1}(z)\right) \tag{4.15}$$

considering (4.13) and (4.14), (4.15) results in:

$$G^*(z^*)=\prod_{i=1}^{k}G_i^*(z^*)=\prod_{i=1}^{k}G_i(z)=G(z)$$

Hence, when the stable poles are either real or in complex conjugate, it is:

$$G^*(z^*) = G(z) \tag{4.16}$$

Before proceeding further it is remarked that in compliance with what was discussed earlier, "if $\phi(z)$ is formed by the Z transforms of a sequence of real time-domain signals then it satisfies the $\phi(z) = \phi^*(z^*)$".

Lemma 4: When the poles are either real or occur in complex conjugate pairs, the proposed orthonormal basis functions as well as any linear combination of them with real coefficients present a real time-domain impulse response.

PROOF:

i) First, a brief review of the proof for basis functions associated with real poles is given in the following:

$$\phi_n(z) = \kappa_n \times G_{n-1}(z) \times \frac{1}{z - p_n}, \quad p_n \in \mathbb{D} \cap \mathbb{R}$$

$$\phi_n^*(z^*) = \kappa_n \times G_{n-1}^*(z^*) \times \frac{1}{z - p_n} = \kappa_n \times G_{n-1}(z) \times \frac{1}{z - p_n} = \phi_n(z)$$

$$\phi^*(z^*) = \phi(z), \qquad p_n \in \mathbb{D} \cap \mathbb{R} \text{ (real-stable pole)} \tag{4.17}$$

ii) For the basis functions related to a pair of complex conjugate stable poles it is:

$$\phi_n(z) = \left(\gamma_n \times G_{n-1}(z) \times \frac{1 - z}{\left(z - p_n\right)\left(z - p_n^*\right)} \right) \tag{4.18}$$

$$\phi_{n+1}(z) = \left(\gamma_{n+1} \times G_{n-1}(z) \times \frac{1 + z}{\left(z - p_n\right)\left(z - p_n^*\right)} \right) \tag{4.19}$$

$$\phi_n^*(z^*) = \left(\gamma_n \times G_{n-1}^*(z^*) \times \frac{1-z}{\left(z-p_n^*\right)\left(z-p_n\right)} \right) \tag{4.20}$$

$$\phi_{n+1}^*(z^*) = \left(\gamma_{n+1} \times G_{n-1}^*(z^*) \times \frac{1+z}{\left(z-p_n^*\right)\left(z-p_n\right)} \right) \tag{4.21}$$

(4.16), (4.18), and (4.20) similarly (4.16), (4.19), and (4.21) result in:

$$\phi_n^*(z^*) = \phi_n(z) \tag{4.22}$$

$$\phi_{n+1}^*(z^*) = \phi_{n+1}(z) \tag{4.23}$$

,where $p_{n+1} = p_n^*$ (stable complex conjugate pole).

According to the (4.17), (4.22), and (4.23) proved in (i) and (ii) we can attempt to prove the problem in the most general form as following.

$$\phi(z) = \sum_{i=1}^{N} \varpi_i \phi_i(z), \qquad \varpi \in \mathbb{R}$$

$$\phi^*(z^*) = \left(\sum_{i=1}^{N} \varpi_i \phi_i(z) \right)^* \Bigg|_{z \to z^*} = \sum_{i=1}^{N} \varpi_i \phi_i^*(z^*) = \sum_{i=1}^{N} \varpi_i \phi_i(z)$$

Hence, $$\phi^*(z^*) = \phi(z)$$

"Lemma 4" authenticates that, if a transfer function of physical system $H(z)$ is approximated with a series of proposed orthonormal basis functions with real coefficients, the inverse Z transform of the approximated function presents a real function in time. Bearing this inherent property for the proposed bases makes them suitable for the system identification applications.

CHAPTER 5. Real-Valued State-space Realization

The LTI finite-dimensional system can be represented by its rational transfer function. For every proper stable transfer function matrix, $\mathbf{H}(z) \in \mathcal{RH}_2^{p\times m}$ [†], there exists a State-Space realization (m presents the number of input excitations and p is the number of outputs).

Within the proposed z-domain orthonormal vector fitting (ZD-OVF) process, it is necessary to convert the obtained transfer function to the state-space representation. This process of transformation is repeatedly performed in every iteration, to obtain the relocated poles. In addition, when processing the 'time-domain' data by utilizing the ZD-OVF technique, converting the final macromodel into a state-space representation is necessary. This transformation serves the purpose of transient response simulation.

This necessity would be sensible when considering the fact that performing the numerical integration on the resulting system of the first order differential equations (in SS equation form) is more cost efficient and produces better results in comparison to computing the inverse Laplace or inverse z -transforms by numerical methods.

[†] Notation $\mathcal{RH}_2^{p\times m}$ presents a set of $p\times m$ matrices with rational functions as its entries that are analytic for $|z| \geq 1$ and squared integrable on the unit circle.

This chapter discusses an efficient formulation to construct the minimal state-space realization, for a system that has been identified in the form of a linear expansion of the propose ZD-OBF.

5.1 Background on Discrete-Domain State-Space Theory

5.1.1 Preliminaries

The state-space models with real matrices[1] that adopt real-valued impulse response for LTI dynamical systems are of interest in this thesis. Accordingly, the discrete-domain state-space realizations in general notation resembles:

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \tag{5.1}$$

$$\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k) \tag{5.2}$$

with respect to a known sampling period, T_S. Here, system's dynamic $\mathbf{A} \in \mathbb{R}^{N\times N}$, input vector $\mathbf{B} \in \mathbb{R}^{N\times m}$, output vector $\mathbf{C} \in \mathbb{R}^{p\times N}$, and vector consisting of the "direct feedthrough[‡]" constants $\mathbf{D} \in \mathbb{R}^{p\times m}$. Also, $\mathbf{u}(k) = \mathbf{u}(kT_S)$ is a vector including the input excitations and $\mathbf{x}(k) \in \mathbb{R}^N$ is the states vector.

A being a real matrix ensures that all eigenvalues of **A** (dynamic modes of the system) appears either as real or in complex conjugate pairs.

---

[1] Applying specific constraints that suit the physical systems, leads to degradation in the mathematical level of generality in problem formulation. In general, applied cases can be considered as special cases for the general mathematical problems.

[‡] This name has borrowed from "Model reduction by Kautz filters, authored by A.C. den Brinker". Citing the 'direct coupling constant term', *D*, as "direct feedthrough" sounds more descriptive. It infers that the *direct* effect of the source excitations is *fed through* the network of interest, and *directly* affects the output states.

In general, the state-space model in (5.1) and (5.2) is an n-dimensional realization of $\mathbf{H}(z)$ defined in below.

$$\mathbf{H}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \tag{5.3}$$

$\mathbf{H}(z)$ presents a matrix of complex-valued transfer functions from the inputs (ports) to the output states (at ports).

5.1.2 Review of Concepts

There is a noticeable diversity in the available approaches to the following concepts and terminologies. A review of the related concepts is given below. It is required to avoid any possibility of ambiguity especially when yielding the focus from continuous- to the discrete-domain.

5.1.2.1 Stability in Realization

A realization is stable if all eigenvalues of **A** (poles) lie strictly inside the unit circle in z-plane, which is corresponding to the left hand side half plane in Laplace-domain.

5.1.2.2 Controllability

i) Controllability of LTI Continuous-Time Systems

For continuous-time-domain system, presented as:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{5.4}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{5.5}$$

Definition: The state equation (5.4) or the pair $(\mathbf{A},\mathbf{B})$ is said to be controllable if for any initial state $\mathbf{x}(0)=\mathbf{x}_0$ and any final state $\mathbf{x}_1$, there exists an input that transfers $\mathbf{x}_0$ to $\mathbf{x}_1$ in a finite time. Otherwise, the system is said to be uncontrollable [47].

Theorem: For an n-dimensional controllable pair $(\mathbf{A},\mathbf{B})$, the following statements are equivalent.

1. The $n\times nm$ construability matrix in (5.6) has full row rank.

$$\mathbf{C}_{n\times nm}=\left[\mathbf{B},\ \mathbf{AB},\ \mathbf{A}^2\mathbf{B},\ \ldots,\mathbf{A}^{(n-1)}\mathbf{B}\right] \tag{5.6}$$

2. For stable realizations the "*Controllability Gramian*" defined as

$$\mathbf{W}_C=\int_0^\infty e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^T e^{\mathbf{A}^T\tau}d\tau \ ^{\dagger} \tag{5.7}$$

is an unique positive definite matrix satisfying the equation

$$\mathbf{A}\mathbf{W}_C+\mathbf{W}_C\mathbf{A}^T=-\mathbf{B}\mathbf{B}^T \tag{5.8}$$

ii) Complete State Controllability for LTI Discrete-Time Systems

Definition: The discrete-time control system given by (5.1) is said to be 'completely state controllable' or simply 'state controllable' if there exists a piecewise-constant control signal $u(k)$ defined over a finite number of sampling periods such that, starting from any initial state, the state $\mathbf{x}(k)$ can be transferred to the desired state $\mathbf{x}_F$ in at most n sampling periods [27].

---

† For a real matrix *M*, the complex conjugate (Hermitian) transpose denoted as $\left(\mathbf{M}^H\right)$ reduces to the simple transpose, shown as $\mathbf{M}^T$.

In discrete-domain controllability matrix is defined in the form resembled to (5.6) and it should be full rank for the controllable systems.

5.1.2.3 Observability

i) Observability of LTI Continuous-Time Systems

Definition: The state equations (5.4) and (5.5) or the pair (A,C) is said to be observable if for any unknown initial state $\mathbf{x}(0)=\mathbf{x}_0$, there exist a finite $t_1>0$ such that the knowledge of the input $\mathbf{u}(k)$ and the output state y over $[0,t_1]$ suffices to uniquely determine the initial state $\mathbf{x}_0$. Otherwise, the system is said to be unobservable [47].

Theorem: For an n-dimensional observable pair $(\mathbf{A},\mathbf{C})$ the following statements are equivalent.

1. The $np\times n$ observability matrix, in below, has full column rank.

$$\mathbf{O}_{np\times n}=\begin{bmatrix}\mathbf{C}\\ \mathbf{AC}\\ \vdots\\ \mathbf{A}^{(n-1)}\mathbf{C}\end{bmatrix} \tag{5.9}$$

2. For stable realizations the "*Observability Gramian*" defining as

$$\mathbf{W}_o=\int_0^{\infty} e^{\mathbf{A}^T\tau}\mathbf{C}^T\mathbf{C}e^{\mathbf{A}\tau}d\tau \tag{5.10}$$

is unique positive definite matrix satisfying the equation,

$$\mathbf{A}^T\mathbf{W}_o+\mathbf{W}_o\mathbf{A}=-\mathbf{C}^T\mathbf{C}. \tag{5.11}$$

It is worth noting that the format of both equations in (5.8) and (5.11) is known as Lyapunov equation in control contexts.

Corollary[†]: if a realization is stable, the controllability gramian $\mathbf{W}_c$ and observability gramian $\mathbf{W}_o$ are defined as the solutions to the Lyapunov equations of the forms $\mathbf{A}\mathbf{W}_c\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{W}_c$ and $\mathbf{A}^T\mathbf{W}_o\mathbf{A} + \mathbf{C}\mathbf{C}^T = \mathbf{W}_o$, respectively [22].

ii) **Complete Observability for LTI Discrete-Time Systems**

Definition: The system defined by (5.1) and (5.2) is completely observable if, given the output $y(k)$ over a finite number of sampling periods, it is possible to determine the initial state vector $\mathbf{x}(0)$ [27]. This requires that the $np \times n$ matrix as shown in (5.9) be of rank n.

5.1.2.4 Minimal State-Space Realization

Theorem: A state equation (**A**, **B**, **C**, D) is a minimal realization of a proper rational function, $H(z)$, if and only if $(\mathbf{A}, \mathbf{B})$ is controllable and $(\mathbf{A}, \mathbf{C})$ is observable [47] .

A noteworthy condition happens when the numerator polynomial has a divisor in common with the denominator (zeros-poles cancellation). In this case the system is either uncontrollable, or unobservable, or both. Consequently, any form of the State-space equation presenting these transfer functions cannot be considered as a minimal realization.

Definition: A stable minimal realization is called '**internally balanced realization**' if $\mathbf{W}_c = \mathbf{W}_o = \mathbf{\Sigma}$, with $\mathbf{\Sigma} = diag(\delta_1, \ldots, \delta_n)$, $\delta_1 \geq \ldots \geq \delta_n$, a diagonal matrix with positive Hankel singular values as diagonal elements [22].

[†] This important conclusion is frequently referred in the investigation of Orthogonality in the state-space presentation of a system (in matrix notation). An attempt to prove it however drags the line of this work to the depth of the control area. This may fall beyond the scope of this context.

5.1.3 Orthogonal Realization

Assume that the dynamics of an LTI SISO system is governed by a state-space equation as:

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}u(k) \tag{5.12}$$

From (5.12) the transfer functions from the input to the states, would have the form as:

$$\frac{\mathbf{X}(z)}{u(k)} = \mathbf{\Phi}(z) = \left[(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\right]$$

For a system of order n excited by a single source it is:

$$\begin{bmatrix} X_1(z) \\ X_2(z) \\ \vdots \\ X_N(z) \end{bmatrix} = \underbrace{\begin{bmatrix} \phi_1(z) \\ \phi_2(z) \\ \vdots \\ \phi_n(z) \end{bmatrix}}_{\mathbf{\Phi}(z)} \times u(k)$$

First, simply assume that by any means one could construct a structure in which the transfer functions from states to input fall in the form of the orthonormal basis functions.

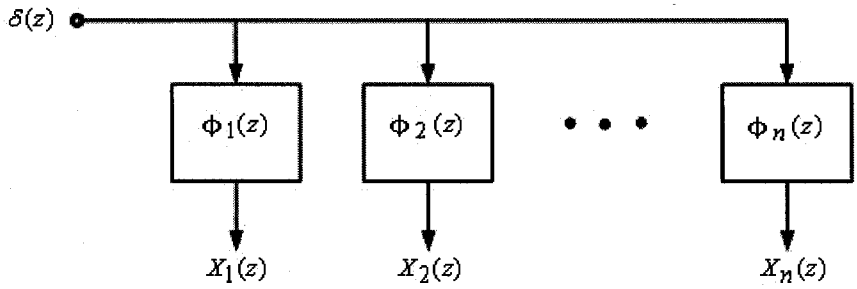


Figure 5-1: Parallel structures in which, TFs from states-to-input are orthonormal

According to the definition of orthogonality in matrix notations, it is:

$$\langle \mathbf{\Phi}, \mathbf{\Phi} \rangle = \frac{1}{2\pi j} \oint_{\mathbb{T}} \mathbf{\Phi}\mathbf{\Phi}^T (1/z) \frac{dz}{z} = \mathbf{I} \qquad \text{(the identity matrix)} \tag{5.13}$$

In the above block diagram, a general idea is illustrated. It shows that, among many possible presentations for a physical network, one(s) may be found, such that, in which each state presents an orthogonal basis when input excitation is the ideal unit impulse. Hence, the vector of states happens to be an orthogonal set, satisfying (5.13).

To continue, by extending this idea, a method will be outlined to obtain a SS realization for any arbitrary transfer function, expanded with orthonormal basis functions.

5.2 Real-Valued Minimal SS Realization with Proposed ZD-OBF

The representation of a transfer function in state-space form is obviously not unique. In this section, minimal state-space representation for the proposed ZD-OBF will be presented.

As shown in below a cascade (ladder) structure of the all-pass filters have been connected to every orthonormal basis. This all-pass section has a noticeable weight in the form and important impact in the properties of each function.

- For the real stable pole:

$$\phi_n(z) = \kappa_n \times \underbrace{\prod_{i=1}^{n-1}\left[\frac{1-a_i^* z}{z-a_i}\right]}_{\substack{\text{G(z):}\\ \text{all-pass TF}}} \times \underbrace{\frac{1}{z-a_n}}_{\substack{\text{first order}\\ \text{low-pass}\\ \text{filter}}}, \qquad a_n \in \mathbb{R},\ |a_n|<1 \tag{5.14}$$

- For complex conjugate stable poles:

$$\phi_n(z) = \gamma_n \times \underbrace{\prod_{i=1}^{n-1}\left[\frac{1-a_i^* z}{z-a_i}\right]}_{G(z)} \times \frac{1-z}{\left(z-a_n\right)\left(z-a_n^*\right)}, \qquad a_n = -\alpha_n - j\beta_n \ \&\ |a_n|<1 \tag{5.15}$$

$$\phi_{n+1}(z)=\gamma_{n+1}\times\underbrace{\prod_{i=1}^{n-1}\left[\frac{1-a_i^{*}z}{z-a_i}\right]}_{\mathrm{G(z)}}\times\underbrace{\frac{1+z}{\left(z-a_n\right)\left(z-a_n^{*}\right)}}_{\substack{\text{Second-order}\\ \text{low-pass filter}\\ \text{Section}}},\qquad a_{n+1}=a_n^{*}\tag{5.16}$$

5.2.1 All-Pass Transfer Functions Realization

5.2.1.1 Orthogonal All-Pass Transfer Function

Definition 1: Any arbitrary all-pass transfer function $H(z)$, of the order n, with the closed form shown below, holds following properties:

$$H(z)=\prod_{i=1}^{n}\left[\frac{1-a_i^{*}z}{z-a_i}\right]\tag{5.17}$$

i) Determined by a set of stable poles as $\{a_i\in\mathbb{C},\ |a_i|<1\}$,

ii) Asymptotically stable with real-valued impulse response,

iii) Satisfies $H(z)H(1/z)=H(z)H^{*}(z)=1$.

It is noted that for each pole, as a_i, there exists a corresponding zero located at the inverse of its conjugate as $\frac{1}{a^{*}}$.

Let (**A**, **B**, **C**, D)[†] be a minimal balanced state-space realization of (5.17). It is proved[2] that,

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & D\end{bmatrix}\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & D\end{bmatrix}^{T}=\mathbf{I}\tag{5.18}$$

[†] For a SISO system, D is scalar.

[2] One can prove it by using the all-pass network properties, controllability Gramian and observability matrix. For an idea in this regard, reference [36] can be referred to.

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & D \end{bmatrix}\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & D \end{bmatrix}^T = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & D \end{bmatrix}\begin{bmatrix} \mathbf{A}^T & \mathbf{C}^T \\ \mathbf{B}^T & D^T \end{bmatrix} = \mathbf{I}^{\dagger}$$

$$\begin{bmatrix} \mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T & \mathbf{A}\mathbf{C}^T + \mathbf{B}D^T \\ \mathbf{C}\mathbf{A}^T + D\mathbf{B}^T & \mathbf{C}\mathbf{C}^T + DD^T \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \tag{5.19}$$

Definition 2: State-space realization for $H(z)$ as shown in below, which satisfies (5.18) is called **orthogonal.**

$$\begin{bmatrix} \mathbf{x}(k+1) \\ y(k) \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & D \end{bmatrix}\begin{bmatrix} \mathbf{x}(k) \\ u(k) \end{bmatrix} \tag{5.20}$$

5.2.1.2 State-Space Realization for a Cascade All-Pass Filter Network

Consider an all-pass network in the form of:

$$H(z) = \underbrace{\left(\frac{1-a_1^* z}{z-a_1}\right)}_{H_1(z)} \times \underbrace{\left(\frac{1-a_2^* z}{z-a_2}\right)}_{H_2(z)} \times\cdots\times \underbrace{\left(\frac{1-a_i^* z}{z-a_i}\right)}_{H_i(z)} \times\cdots\times \underbrace{\left(\frac{1-a_n^* z}{z-a_n}\right)}_{H_n(z)} \tag{5.21}$$

$$Y(z) = H(z)\times U(z) \tag{5.22}$$

$$U(z) \rightarrow \overbrace{\underbrace{\left(\frac{1-a_1^* z}{z-a_1}\right)}_{H_1(z)} \xrightarrow{Y_1(z)} \underbrace{\left(\frac{1-a_2^* z}{z-a_2}\right)}_{H_2(z)} \xrightarrow{Y_2(z)} \cdots \xrightarrow{Y_{n-1}(z)} \underbrace{\left(\frac{1-a_n^* z}{z-a_n}\right)}_{H_n(z)}}^{H(z)} \rightarrow Y(z) \tag{5.23}$$

† For SISO cases, D is scalar and $D^T=D$. However, to preserve the generality in the form of equations, it is kept in transposed notation.

Each section of the above orthogonal all-pass filters, $H_i(z)$, presents a 'Single In Single Out' (SISO) network with the corresponding state as $X_i(z)$. Its state-space realization can be shown in the form of (5.20).

Theorem: Consider two orthogonal all-pass filters $H_1(z)$ and $H_2(z)$ with state vectors $\mathbf{x}_1(k)$ and $\mathbf{x}_2(k)$, respectively. Then the cascade (serial) connection of those two as $H_2(z)H_1(z)$ is also all-pass and orthogonal with the state vector, $\mathbf{x}(k) = \begin{bmatrix} \mathbf{x}_1(k) \\ \mathbf{x}_2(k) \end{bmatrix}$ [36].

The result can be generalized to any cascade structure of above all-pass filters as shown below:

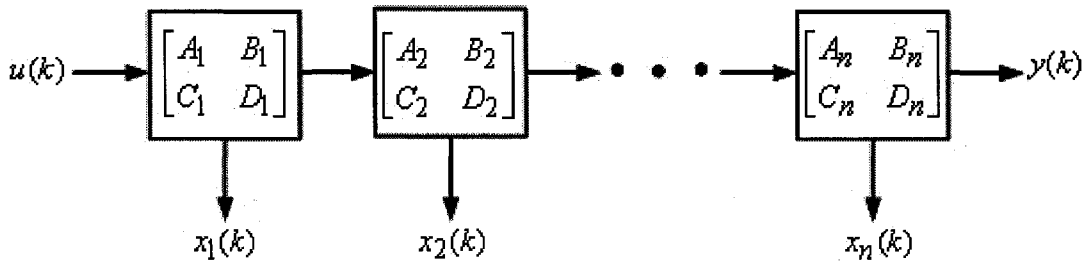


Figure 5-2: Block diagram of a cascade all-pass filters realization

The proof for this theorem, by itself is as an important mathematical assertion, it also results in the general form for the state-space realization for any order of cascade subsystems. Within the steps of proof, it is confirmed that the transfer functions from the source to the state of every stage are orthogonal. Moreover, as a corollary, it will be verified that the state-space realization for entire cascade structure, consists of n stages, is orthonormal.

PROOF:

I) **Realization for 2-Stage Network:**

- ***For the first Stage:***

$$\begin{bmatrix} x_1(k+1) \\ y_1(k) \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ u(k) \end{bmatrix} \tag{5.24}$$

Expanding the above equation in matrix notation by considering the $x_2(k) = x_2(k)$ as the third (and extra) equation would result in:

$$\begin{bmatrix} x_1(k+1) \\ x_2(k) \\ y_1(k) \end{bmatrix} = \begin{bmatrix} A_1 & 0 & B_1 \\ 0 & 1 & 0 \\ C_1 & 0 & D_1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ u(k) \end{bmatrix}, \tag{5.25}$$

- ***For the second Stage:***

$$\begin{bmatrix} x_2(k+1) \\ y_2(k+1) \end{bmatrix} = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} x_2(k) \\ y_1(k) \end{bmatrix} \tag{5.26}$$

similarly, expanding the above by considering the $x_1(k+1) = x_1(k+1)$ would result in:

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ y_2(k) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ A_2 & 0 & B_2 \\ 0 & C_2 & D_2 \end{bmatrix} \begin{bmatrix} x_1(k+1) \\ x_2(k) \\ y_1(k) \end{bmatrix} \tag{5.27}$$

by substituting the (5.25) in (5.27) :

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ y_2(k) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ A_2 & 0 & B_2 \\ 0 & C_2 & D_2 \end{bmatrix} \begin{bmatrix} A_1 & 0 & B_1 \\ 0 & 1 & 0 \\ C_1 & 0 & D_1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ u(k) \end{bmatrix} \tag{5.28}$$

$$\begin{bmatrix} \mathbf{A}^{(2)} & \mathbf{B}^{(2)} \\ \mathbf{C}^{(2)} & D^{(2)} \end{bmatrix} = \begin{bmatrix} A_1 & 0 & B_1 \\ B_2C_1 & A_2 & B_2D_1 \\ D_2C_1 & C_2 & D_2D_1 \end{bmatrix} \tag{5.29}$$

In (5.29) superscript (2) represents the number of cascade stages.

$$\begin{bmatrix} \mathbf{A}^{(2)} & \mathbf{B}^{(2)} \\ \mathbf{C}^{(2)} & D^{(2)} \end{bmatrix} = \underbrace{\begin{bmatrix} I & 0 & 0 \\ A_2 & 0 & B_2 \\ 0 & C_2 & D_2 \end{bmatrix}}_{\mathbf{M}_2} \overbrace{\begin{bmatrix} A_1 & 0 & B_1 \\ 0 & I & 0 \\ C_1 & 0 & D_1 \end{bmatrix}}^{\mathbf{M}_1} = \begin{bmatrix} A_1 & 0 & B_1 \\ B_2C_1 & A_2 & B_2D_1 \\ D_2C_1 & C_2 & D_2D_1 \end{bmatrix} \quad (5.30)$$

This is the total matrix known as state-space realization for 2-stage cascade network.

Property 1: Matrix $\mathbf{M}_1$ is orthonormal.

Proof:[3]

$$\mathbf{M}_1\mathbf{M}_1^T = \begin{bmatrix} A_1 & 0 & B_1 \\ 0 & 1 & 0 \\ C_1 & 0 & D_1 \end{bmatrix}\begin{bmatrix} A_1 & 0 & B_1 \\ 0 & 1 & 0 \\ C_1 & 0 & D_1 \end{bmatrix}^T = \begin{bmatrix} A_1 & 0 & B_1 \\ 0 & 1 & 0 \\ C_1 & 0 & D_1 \end{bmatrix}\begin{bmatrix} A_1' & 0 & C_1' \\ 0 & 1 & 0 \\ B_1' & 0 & D_1' \end{bmatrix} =$$

by considering (5.19), it is concluded:

$$\begin{bmatrix} A_1A_1'+B_1B_1' & 0 & A_1C_1'+B_1D_1' \\ 0 & 1 & 0 \\ C_1A_1'+D_1B_1' & 0 & C_1C_1'+D_1D_1' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = M_1M_1^T = I \quad (5.31)$$

Property 2: Similarly, matrix $\mathbf{M}_2$ is orthonormal too.

Property 3: The state-space realization for 2-stage structure is orthonormal.

Proof:

the product of two orthonormal matrices would results in an orthonormal matrix.

$$(\mathbf{M}_2\mathbf{M}_1)(\mathbf{M}_2\mathbf{M}_1)^T = \mathbf{M}_2\mathbf{M}_1\mathbf{M}_1^T\mathbf{M}_2^T = \mathbf{M}_2\mathbf{I}\mathbf{M}_2^T = \mathbf{I} \quad (5.32)$$

This idea can be expanded for higher number of stages in the similar fashion. After repeating the above procedure a few times (iterations), a regular pattern in the resulting

---

[3] To shorten the form of equations, the transpose of matrix $\mathbf{A} = [a_{ij}]$ defined as: $\mathbf{A}^T = [aji]$ occasionally is denoted as $\mathbf{A}'$.

matrix of state-space realization would be recognizable. Even in the case of 3-stage structures, the regularity in the outcome can be seen.

II) Realization for 3-stage network:

With adding the $x_3(k) = x_3(k)$ to equation (5.30)

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k) \\ y_2(k) \end{bmatrix} = \begin{bmatrix} A_1 & 0 & 0 & B_1 \\ B_2C_1 & A_2 & 0 & B_2D_1 \\ 0 & 0 & 1 & 0 \\ D_2C_1 & C_2 & 0 & D_2D_1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ u(k) \end{bmatrix} \tag{5.33}$$

Also, enclosing the two extended equations as $x_1(k+1) = x_1(k+1)$ and $x_2(k+1) = x_2(k+1)$ in the SS equations for the third stage:

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \\ y_3(k) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & A_3 & B_3 \\ 0 & 0 & C_3 & D_3 \end{bmatrix} \begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k) \\ y_2(k) \end{bmatrix} \tag{5.34}$$

Substituting (5.33) in (5.34) results in:

$$\begin{bmatrix} \mathbf{A}^{(3)} & \mathbf{B}^{(3)} \\ \mathbf{C}^{(3)} & D^{(3)} \end{bmatrix} = \begin{bmatrix} A_1 & 0 & 0 & B_1 \\ B_2C_1 & A_2 & 0 & B_2D_1 \\ B_3B_2C_1 & B_3C_2 & A_3 & B_3D_2D_1 \\ D_3D_2C_1 & D_3C_2 & C_3 & D_3D_2D_1 \end{bmatrix} \tag{5.35}$$

Property 4: As a product of two orthonormal matrices, this realization for a 3-stage cascade all-pass network also preserves the orthonormal characteristic.

Pushing this procedure further, we can obtain state-space representation for a given all-pass filter with any arbitrary number of stages.

III) State-Space Representation in General Form

This section shows how the idea, outlined in the previous section, can be generalized. Consider a cascade structure comprising of a number of all-pass filter networks in series with a first-order block at the last. This condition may be judged as the most common form of realization. For the reason that, to extract state-space representation for any arbitrary transfer functions, its all first-order sub-blocks corresponding to the first-order differential equations should be recognized.

Accordingly, the problem of converting the orthonormal basis functions of interest shown in (5.14), (5.15), or (5.16) would ultimately fall in the form of a cascade network as illustrated in below.

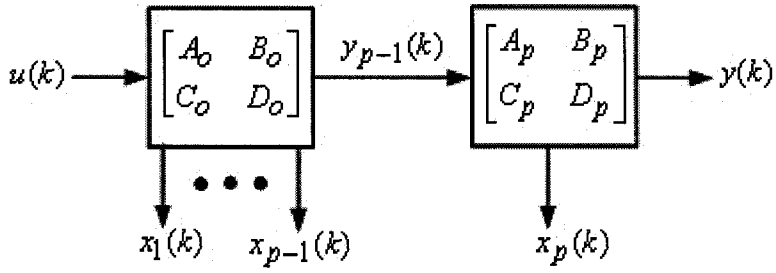


Figure 5-3: A general cascade network

$$\begin{bmatrix} x_1(k+1) \\ \vdots \\ x_{p-1}(k+1) \\ x_p(k) \\ y_{p-1}(k) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_o & \mathbf{0} & \mathbf{B}_o \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{C}_o & \mathbf{0} & D_o \end{bmatrix} \begin{bmatrix} x_1(k) \\ \vdots \\ x_{p-1}(k) \\ x_p(k) \\ u(k) \end{bmatrix} \tag{5.36}$$

For the last stage:

$$\begin{bmatrix} x_1(k+1) \\ \vdots \\ x_{p-1}(k+1) \\ x_p(k+1) \\ y(k) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & A_p & B_p \\ 0 & \cdots & 0 & C_p & D_p \end{bmatrix} \begin{bmatrix} x_1(k+1) \\ \vdots \\ x_{p-1}(k+1) \\ x_p(k) \\ y_{p-1}(k) \end{bmatrix} \tag{5.37}$$

Plugging the (5.36) in (5.37):

$$\begin{bmatrix} x_1(k+1) \\ \vdots \\ x_{p-1}(k+1) \\ x_p(k+1) \\ y(k) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & A_p & B_p \\ 0 & \cdots & 0 & C_p & D_p \end{bmatrix} \begin{bmatrix} \mathbf{A}_o & \mathbf{0} & \mathbf{B}_o \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{C}_o & \mathbf{0} & D_o \end{bmatrix} \begin{bmatrix} x_1(k) \\ \vdots \\ x_{p-1}(k) \\ x_p(k) \\ u(k) \end{bmatrix} \tag{5.38}$$

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & D \end{bmatrix} = \begin{bmatrix} \mathbf{A}_o & \mathbf{0} & \mathbf{B}_o \\ Bp\mathbf{C}_o & A_p & B_pD_o \\ D_p\mathbf{C}_o & C_p & D_pD_o \end{bmatrix} \tag{5.39}$$

In the most general approach, the orthonormality of the resulting SS realization would be decided by the orthonormal property at the last stage.

Continuing the extraction process, a general state-space realization for a p-stage network with the *(p-1)* first-order all-pass blocks and a first-order section at the end would result in defining the **A**, **B**, **C**, D matrices as shown in below.

$$\mathbf{A}_{P\times P} = \begin{bmatrix} A_1 & 0 & \vdots & 0 & 0 & 0 & 0 \\ B_2C_1 & A_2 & \vdots & 0 & 0 & 0 & 0 \\ B_3D_2C_1 & B_3C_2 & \vdots & 0 & 0 & 0 & 0 \\ B_4D_3D_2C_1 & B_4D_3C_2 & \vdots & 0 & 0 & 0 & 0 \\ B_5D_4D_3D_2C_1 & B_5D_4D_3C_2 & \vdots & 0 & 0 & 0 & 0 \\ \cdots & \cdots & \vdots & \cdots & \cdots & \cdots & \cdots \\ B_{P-3}D_{P-4}\cdots D_2C_1 & B_{P-3}D_{P-4}\cdots D_3C_2 & \vdots & A_{p-3} & 0 & 0 & 0 \\ B_{P-2}D_{P-3}\cdots D_2C_1 & B_{P-2}D_{P-3}\cdots D_3C_2 & \vdots & B_{P-2}C_{P-3} & A_{p-2} & 0 & 0 \\ B_{P-1}D_{P-2}\cdots D_2C_1 & B_{P-1}D_{P-2}\cdots D_3C_2 & \vdots & B_{P-1}D_{P-2}C_{P-3} & B_{P-1}C_{P-2} & A_{P-1} & 0 \\ B_PD_{P-1}\cdots D_2C_1 & B_PD_{P-1}\cdots D_3C_2 & \vdots & B_PD_{P-1}D_{P-2}C_{P-3} & B_PD_{P-1}C_{P-2} & B_PC_{P-1} & A_P \end{bmatrix} \tag{5.40}$$

$$\mathbf{B} = \begin{bmatrix} B_1 \\ B_2D_1 \\ B_3D_2D_1 \\ B_4D_3D_2D_1 \\ \cdots \\ B_PD_{P-1}\cdots D_2D_1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} D_PD_{p-1}\cdots D_2C_1 \\ D_PD_{p-1}\cdots D_3C_2 \\ \cdots \\ D_PD_{P-1}C_{P-2} \\ D_PC_{P-1} \\ C_P \end{bmatrix}^T \tag{5.41}$$

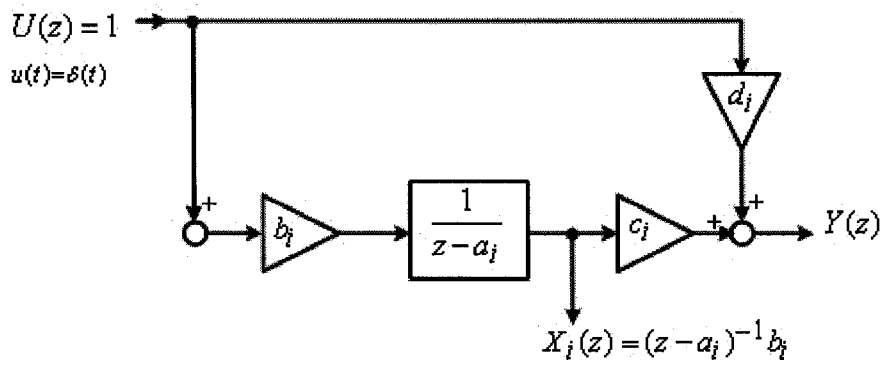
$$D = D_P\cdots D_1 \tag{5.42}$$

5.3 Minimal SS Representation for Takenaka-Malmquist OBF

So far, a general form for minimal state-space representation for a system consisting of serial first-order structures has been discussed. Following the above explanations, obtaining a state-space realization for the general discrete-time Takenaka-Malmquist orthonormal basis functions with any arbitrary order is presented. First attempt in this direction would be the forming matrices **A** and **B** in the equation $\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}u(k)$.

This point is emphasized as an essential criteria that the state matrix **A,** and input matrix **B,** must be formed such that every state $X_i(z)$ happens to be the un-normalized i[th] orthogonal basis function.

This important property can be illustrated within following block diagrams:



$$\text{Dynamics}: \quad y_i(k+1) = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} u(k)$$

$$\text{Xfer function}: \quad H_i(z) = c_i(z-a_i)^{-1}b_i + di$$

Figure 5-4: Illustration of the TF for the i[th] first-order sub-network when being disturbed by Dirac delta (impulse) function

A block diagram representation of an arbitrary function expanded as a linear span of the general discrete-time Takenaka-Malmquist orthonormal bases is shown in the Figure 5-5 , wherein:

- $X_1,\ldots,X_P(z)$: States, evaluated as un-normalized orthogonal functions
- $\phi_i(z)$: Represent normalize orthogonal basis functions
- γ_i : Real-valued coefficients as normalization factor
- ω_i : Coefficients of the orthonormal bases in expansion

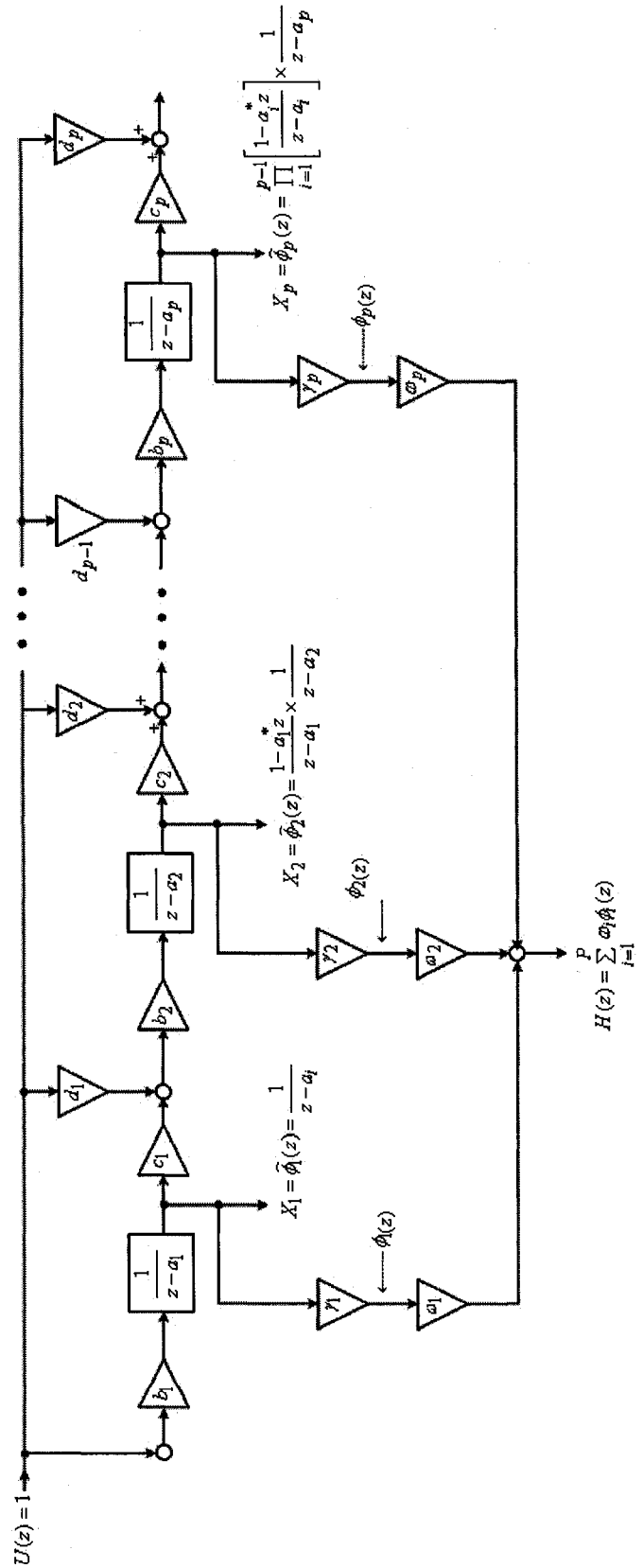


Figure 5-5: Block diagram of an arbitrary rational function, expanded as a linear span of discrete-time Takenaka-Malmquist orthonormal bases

In the above illustration, $H(z)$ can be composed as a linear combination of normalized orthogonal states of the system like:

$$H(s)=\mathbf{C}\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}+D \quad \Rightarrow \mathbf{C}=\begin{bmatrix} \omega_1\gamma_1 & \omega_2\gamma_2 & \cdots & \omega_P\gamma_P \end{bmatrix} \tag{5.43}$$

The vector **C** defined in (5.43) is utilized for realization of the intended transfer functions. In Vector Fitting algorithm when constructing the vector **C**, for any γ_i, corresponding normalization factor according to the type of the associated poles (real or complex pairs) should be substituted.

5.4 Realization of Sub-Structures in the Proposed ZD-OBF

In this section, the SS realization for each sub-block in the orthonormal basis structures shown below in (5.44), (5.45) and (5.46) is explained. The total **A** and **B** matrices would be formed by plugging (A_i, B_i, C_i, D_i) matrices from each sub-block in the general form for **A** and **B**.

I) Un-normalized Basis function corresponding to a real pole, a_p:

$$\underbrace{\left(\frac{1-a_1^* z}{z-a_1} \rightarrow \frac{1-a_2^* z}{z-a_2} \rightarrow \frac{1-a_3^* z}{z-a_3} \rightarrow \cdots \rightarrow \frac{1-a_{P-1}^* z}{z-a_{P-1}}\right)}_{G(z):\ \text{Cascade structure of first-order all-pass filter sections}} \rightarrow \underbrace{\frac{1}{z-a_P}}_{\substack{\text{First-order} \\ \text{low-pass filter} \\ \text{section}}} \tag{5.44}$$

II) Un-normalized orthogonal basis corresponding to the two subsequent complex conjugate pairs of poles $\left(a_{P+1}=a_P^*\right)$:

$$\underbrace{\left(\frac{1-a_1^* z}{z-a_1} \to \frac{1-a_2^* z}{z-a_2} \to \frac{1-a_3^* z}{z-a_3} \to \cdots \to \frac{1-a_{P-1}^* z}{z-a_{P-1}}\right)}_{G(z)} \to \underbrace{\frac{1-z}{(z-a_P)(z-a_P^*)}}_{\substack{\text{Second-order} \\ \text{low-pass filter} \\ \text{Section}}} \tag{5.45}$$

$$\underbrace{\left(\frac{1-a_1^* z}{z-a_1} \to \frac{1-a_2^* z}{z-a_2} \to \frac{1-a_3^* z}{z-a_3} \to \cdots \to \frac{1-a_{P-1}^* z}{z-a_{P-1}}\right)}_{G(z)} \to \underbrace{\frac{1+z}{(z-a_P)(z-a_P^*)}}_{\substack{\text{Second order} \\ \text{lowpass filter} \\ \text{section}}} \tag{5.46}$$

To practice a consistency in entire following text (without any loss of generality) it is assumed that:

$$a_n = -\alpha_n - j\beta_n \text{ and } a_{n+1} = -\alpha_n + j\beta_n \qquad \alpha_n, \beta_n > 0 \;\&\in \mathbb{R}$$

5.4.1 The First-Order Network

Assume the dynamic of a given first-order (asymptotical stable) SISO LTI subsystem is governed by: $\begin{cases} zX = aX + bU \\ Y = cX + dU \end{cases}$

Its transfer function has a general form as: $H(z) = \dfrac{dz + (bc - ad)}{z - a}$

5.4.2 First-Order Sections in the ZD-OBF

Comparing the transfer functions for the following first-order cases with the above general form leads to the results shown below.

i) SISO all-pass section:

$$H_n(z) = \frac{1-a_n^* z}{z-a_n}$$

$$\begin{cases} \boxed{A_n = a_n} \ (the\ pole) \\ \boxed{B_n = 1} \\ \boxed{C_n = 1 - |a_n|^2} \\ \boxed{D_n = a_n^*} \end{cases} \tag{5.47}$$

ii) SISO low-pass section:

$$H_n(z) = \frac{1}{z - a_n}$$

$$\begin{cases} \boxed{A_n = a_n} \\ \boxed{B_n = 1} \\ \boxed{C_n = 1} \\ \boxed{D_n = 0} \end{cases} \tag{5.48}$$

5.4.3 The Second-Order Network

Two complex conjugate poles can be dynamic modes for a second-order subsystem with a transfer function, resembling:

$$\frac{1 - a_n^* z}{z - a_n} \times \frac{1 - a_n z}{z - a_n^*} \tag{5.49}$$

According to the format of corresponding bases, the transfer function from input to the second-order section, to every state is in the following form:

$$\begin{bmatrix} x_1(z) \\ x_2(z) \end{bmatrix} = \begin{bmatrix} \dfrac{1 - z}{(z - a_P)(z - a_P^*)} \\ \dfrac{1 + z}{(z - a_P)(z - a_P^*)} \end{bmatrix} \tag{5.50}$$

state-space realization for the second-order system presented by transfer function in (5.49) requires the effort of extracting **A**, **B** matrices, such that the two states of the system should appear as orthogonal functions shown in (5.50).

More accurately, with respect to the system matrix and input vector, it should be:

$$(zI-A)^{-1}B=\begin{bmatrix}\dfrac{1-z}{(z-a_P)(z-a_P^*)}\\ \dfrac{1+z}{(z-a_P)(z-a_P^*)}\end{bmatrix} \tag{5.51}$$

In addition, the work should continue with extracting the matrices **C** and D such that the transfer function for the entire second-order block turns to appear in the following form:

$$\mathbf{C}(z\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+D=\frac{\left(1-a_n^*z\right)\left(1-a_nz\right)}{\left(z-a_n\right)\left(z-a_n^*\right)} \tag{5.52}$$

To have a better overview of the concept, for two orthonormal bases associated with the first two complex conjugate poles the illustration is shown in below.

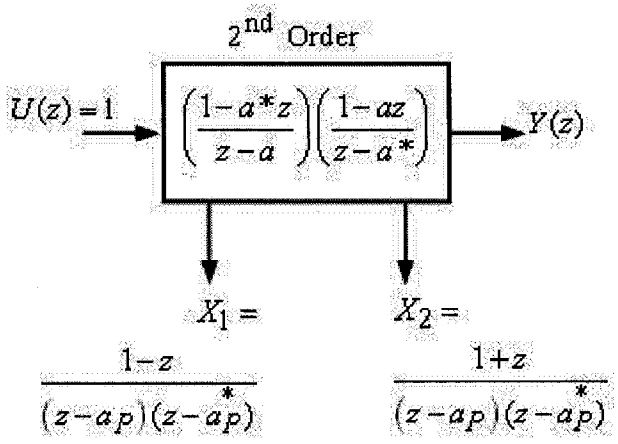
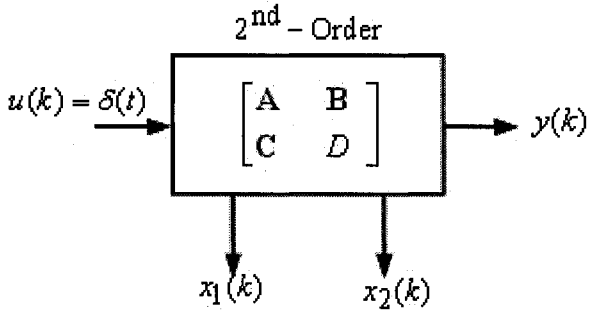


Figure 5-6: A second-order block, presenting un-normalized two orthogonal bases associated with the pair of first two complex conjugate poles



A, **B**	such that : $X_1(z) \perp X_2(z)$
and **C**, D	" " : $\mathbf{C}(z\mathbf{I}-\mathbf{A})\mathbf{B}+D=\left(\frac{1-a^*z}{z-a}\right)\left(\frac{1-az}{z-a^*}\right)$

Figure 5-7: An idea of how a structure shown in Figure 5-6 can be represented in SS form

Example: Given a system with order of four, whose dynamic modes lie at $a_1 = real$, $a_2 = complex$, $a_3 = a_2^*$, $a_4 = real$. Following the outlined idea in above, a block diagram presentation using the orthonormal bases is shown in below.

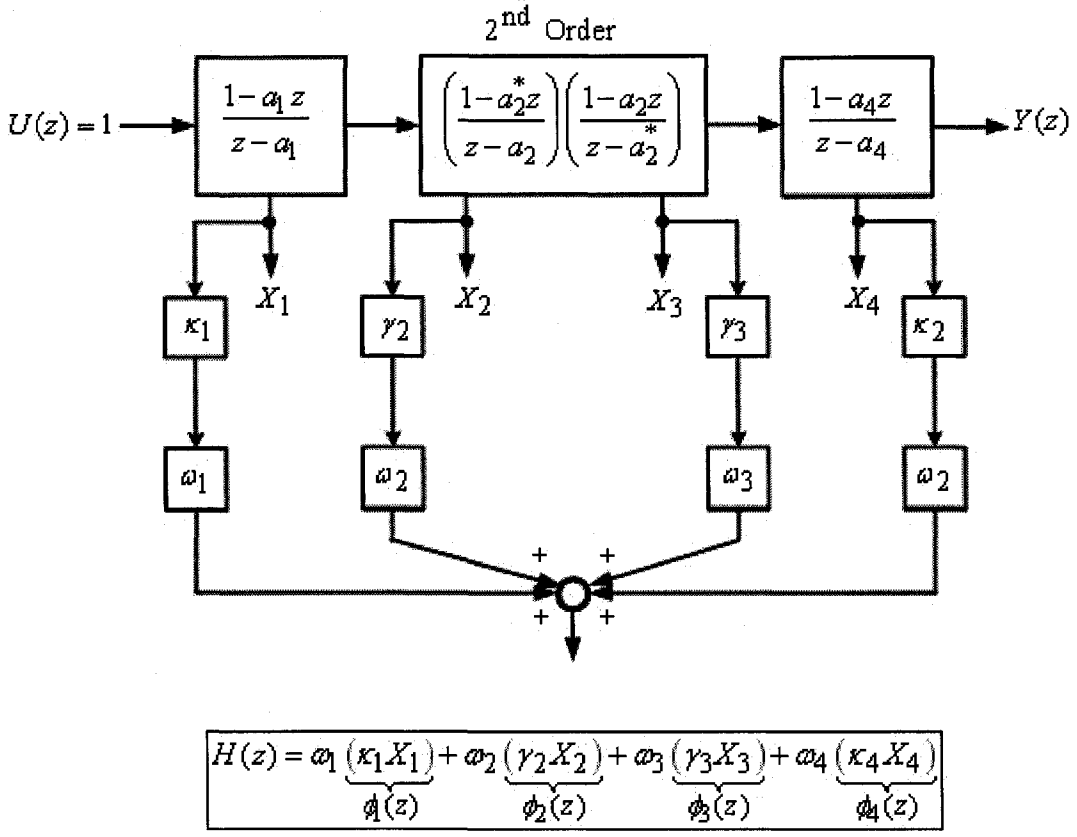


Figure 5-8: Block diagram for z-domain system identification using proposed orthonormal functions for the an example fourth-order system

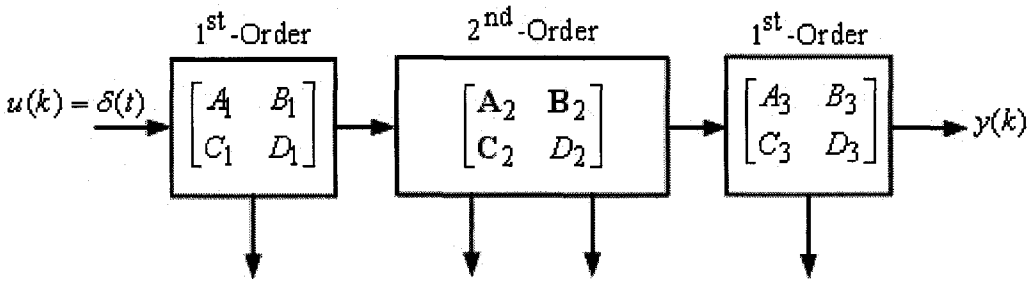


Figure 5-9: The total A and B matrices for this realization should provide states as orthogonal functions $X_i(z) \perp X_j(z) \quad ; i \neq j$

Where the un-normalized bases would be:

$$X_1 = \hat{\phi}_1(z) = \frac{1}{z - a_1},$$

$$X_2 = \hat{\phi}_2(z) = \left(\frac{1-a_1 z}{z-a_1} \right) \frac{1-z}{(z-a_2)(z-a_2^*)}$$

$$X_3 = \hat{\phi}_3(z) = \left(\frac{1-a_1 z}{z-a_1} \right) \frac{1+z}{(z-a_2)(z-a_2^*)}$$

$$X_4 = \hat{\phi}_4(z) = \left(\frac{1-a_1 z}{z-a_1} \times \frac{1-a_2^* z}{z-a_2} \times \frac{1-a_2 z}{z-a_2^*} \right) \frac{1}{z-a_4}$$

5.4.4 Second-Order Sections in the ZD-OBF

In this section, it is described, how a state-space realization, (**A**, **B**, **C**, D), for a transfer function in (5.53) can be obtained.

$$\left(\frac{1-a_n^* z}{z-a_n} \right) \left(\frac{1-a_n z}{z-a_n^*} \right) = \frac{|a_n|^2 z^2 - 2\mathrm{Re}(a_n) z + 1}{z^2 - 2\mathrm{Re}\left(a_n\right) z + \left|a_n\right|^2} \tag{5.53}$$

The outcomes should hold the following important constraints to guarantee the resulting model adopts real-valued impulse response.

- All entries of matrix **A** should be real. This ensures the complex eigenvalues of A occur in complex conjugate.
- Real **B** and **C** matrices ensure the zeros in complex conjugate pair
- Scalar D also should be real to ensure the real-valued response.

Step 1: Calculating **A** and **B**:

Let **A** and **B** be denoted as $\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$. Recalculating the left hand side in (5.51) with these two matrices:

$$(z\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}=\begin{bmatrix} z-A_{11} & -A_{12} \\ -A_{21} & z-A_{22} \end{bmatrix}^{-1}\mathbf{B}=$$

$$\left(\frac{1}{z^2-(A_{11}+A_{22})z+(A_{11}A_{22}-A_{12}A_{21})}\right)\begin{bmatrix} z-A_{22} & A_{12} \\ A_{21} & z-A_{11} \end{bmatrix}\begin{bmatrix} B_1 \\ B_2 \end{bmatrix}=$$

$$=\left(\frac{1}{z^2-(A_{11}+A_{22})z+(A_{11}A_{22}-A_{12}A_{21})}\right)\begin{bmatrix} B_1z+(A_{12}B_2-A_{22}B_1) \\ B_2z+(A_{21}B_1-A_{11}B_2) \end{bmatrix} \tag{5.54}$$

With comparing (5.51) and (5.54) :

$$\frac{B_1z+(A_{12}B_2-A_{22}B_1)}{z^2-(A_{11}+A_{22})z+(A_{11}A_{22}-A_{12}A_{21})}=\frac{1-z}{z^2-2\mathrm{Re}(a_n)z+|a_n|^2} \tag{5.55}$$

$$\frac{B_2z+(A_{21}B_1-A_{11}B_2)}{z^2-(A_{11}+A_{22})z+(A_{11}A_{22}-A_{12}A_{21})}=\frac{1+z}{z^2-2\mathrm{Re}(a_n)z+|a_n|^2} \tag{5.56}$$

From (5.55) and (5.56) it is concluded that:

$$\boxed{B_1=-1} \tag{5.57}$$

$$\boxed{B_2=1} \tag{5.58}$$

$$\begin{cases} A_{11}+A_{21}=-1 & \text{(i)} \\ A_{12}+A_{22}=1 & \text{(ii)} \\ A_{11}+A_{22}=2\mathrm{Re}(a_n) & \text{(iii)} \\ A_{11}A_{22}-A_{12}A_{21}=|a_n|^2 & \text{(iv)} \end{cases} \tag{5.59}$$

From (ii) in (5.59): $A_{22}=1-A_{12}$ (5.60)

by substituting (5.60) in (iv): $A_{11} - A_{12}(A_{11} + A_{21}) = |a_n|^2$ (5.61)

by substituting (i) in(5.61): $A_{11} + A_{12} = |a_n|^2$ (5.62)

from (ii) and (iii): $A_{12} - A_{11} = 1 - 2\mathrm{Re}(a_n)$ (5.63)

from (5.62) and (5.63): $\boxed{A_{12} = \frac{1}{2}\left[|a_n|^2 - 2\mathrm{Re}(a_n) + 1\right]}$ (5.64)

by plugging (5.64) in (5.60): $\boxed{A_{22} = \frac{1}{2}\left[-|a_n|^2 + 2\mathrm{Re}(a_n) + 1\right]}$ (5.65)

and (5.65) in (iii): $\boxed{A_{11} = \frac{1}{2}\left[|a_n|^2 + 2\mathrm{Re}(a_n) - 1\right]}$ (5.66)

also, (5.66) in (i): $\boxed{A_{21} = -\frac{1}{2}\left[|a_n|^2 + 2\mathrm{Re}(a_n) + 1\right]}$ (5.67)

Step 2: Calculating **C** and D:

Here, D is a scalar while Matrix C is a row vector as: $\mathbf{C} = [C_1 \quad C_2]$

$$\mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + D = \frac{|a_n|^2 z^2 - 2\mathrm{Re}(a_n)z + 1}{z^2 - 2\mathrm{Re}(a_n)z + |a_n|^2} = \frac{2\mathrm{Re}(a_n)\left(|a_n|^2 - 1\right)z - \left(|a_n|^4 - 1\right)}{z^2 - 2\mathrm{Re}(a_n)z + |a_n|^2} + |a_n|^2 \tag{5.68}$$

$$\boxed{D = |a_n|^2} \tag{5.69}$$

by substituting (5.51) in the left hand side of (5.68):

$$\mathbf{C}(z\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+D=$$

$$[C_1 \quad C_2]\begin{bmatrix}\dfrac{1-z}{z^2-2\mathrm{Re}(a_n)z+|a_n|^2}\\ \dfrac{1+z}{z^2-2\mathrm{Re}(a_n)z+|a_n|^2}\end{bmatrix}+D=\frac{2\mathrm{Re}(a_n)\left(|a_n|^2-1\right)z-\left(|a_n|^4-1\right)}{z^2-2\mathrm{Re}(a_n)z+|a_n|^2}+|a_n|^2 \tag{5.70}$$

$$C_1(1-z)+C_2(1+z)=2\mathrm{Re}(a_n)\left(|a_n|^2-1\right)z-\left(|a_n|^4-1\right) \tag{5.71}$$

$$\begin{cases}C_2-C_1=2\mathrm{Re}(a_n)\left(|a_n|^2-1\right) & \text{(a)}\\ C_2+C_1=-\left(|a_n|^4-1\right) & \text{(b)}\end{cases} \tag{5.72}$$

(a) and (b):

$$C_1=-\left(1-|a_n|^2\right)\times-\frac{1}{2}\left[|a_n|^2+2\mathrm{Re}(a_n)+1\right]$$

$$\boxed{C_1=-\left(1-|a_n|^2\right)\times A_{21}} \tag{5.73}$$

(*i) and (*ii):

$$C_2=\left(1-|a_n|^2\right)\times\frac{1}{2}\left[|a_n|^2-2\mathrm{Re}(a_n)+1\right]$$

$$\boxed{C_2=\left(1-|a_n|^2\right)A_{12}} \tag{5.74}$$

5.4.5 Summary

Based on the corroborations provided thus far, the state-space realization for any arbitrary sub-blocks in realization with ZD-OBF is summarized in below.

5.4.5.1 First-Order Blocks

For the two possible forms of 1st–order blocks associated with real stable pole as a_n it is:

SISO all-pass Section: $H_n(z) = \dfrac{1-a_n^* z}{z-a_n}$

$$\begin{aligned} A_n &= a_n \\ B_n &= 1 \\ C_n &= 1-|a_n|^2 \\ D_n &= a_n^* \end{aligned} \qquad (5.75)$$

SISO low-pass section: $H_n(z) = \dfrac{1}{z-a_n}$

$$\begin{aligned} A_n &= a_n \\ B_n &= 1 \\ C_n &= 1 \\ D_n &= 0 \end{aligned} \qquad (5.76)$$

5.4.5.2 Second-Order Blocks

For the second–order blocks associated with complex conjugate stable pole as a_n and $a_{n+1} = a_n^*$ it is:

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

$$\begin{aligned} A_{11} &= \frac{1}{2}\left[|a_n|^2 + 2\,\mathrm{Re}(a_n) - 1\right] \\ A_{12} &= \frac{1}{2}\left[|a_n|^2 - 2\,\mathrm{Re}(a_n) + 1\right] \\ A_{21} &= -\frac{1}{2}\left[|a_n|^2 + 2\,\mathrm{Re}(a_n) + 1\right] \\ A_{22} &= \frac{1}{2}\left[-|a_n|^2 + 2\,\mathrm{Re}(a_n) + 1\right] \end{aligned} \qquad (5.77)$$

$$\mathbf{B}=\begin{bmatrix}B_1\\B_2\end{bmatrix};\quad B_1=-1,\quad B_2=1 \tag{5.78}$$

$$\begin{aligned}&\mathbf{C}=[C_1,\,C_2]\\&C_1=-\left(1-|a_n|^2\right)A_{21}\\&C_2=\left(1-|a_n|^2\right)A_{12}\end{aligned} \tag{5.79}$$

$$D=|a_n|^2 \tag{5.80}$$

5.4.6 On Transfer Functions Realization

In light of the presented explanations, it is clear that for transfer function of the form:

$$H(z)=\sum_{n=1}^{P}\omega_n\varphi_n(z)+d=\sum_{n=1}^{P}\omega_n\overbrace{\left(\gamma_n X_n(z)\right)}^{\varphi_n(z)}+d \tag{5.81}$$

, where:

P: Number of required poles
$X_n(z)$: Refers to un-normalized orthogonal bases, (TF from State#i-to-input)
γ_n: Represents the normalization factor to form the orthonormal basis as $\|\phi_n(z)\|_2=1$
ω_n: Real coefficients for the bases

The minimal state-space realization can easily be attained by substituting realizations for each block in the general form for the $\mathbf{A}_{P\times P}$ and $\mathbf{B}_{P\times 1}$. The vector $\mathbf{C}_{1\times P}$ is formed as shown in (5.43) and scalar D is equal to d.

According to the importance, it is emphasized that presenting a real-valued impulse response is the main characteristic for transfer functions of physical systems. Hence, ensuring the property of $H^*(z^*) = H(z)$ for the resulted $H(z)$ requires all ω_n to appear as real-values.

CHAPTER 6. Error Estimation and LLS solution methods for VF Algorithms

The result from vector fitting (VF) process is the identification of an adequately accurate model for a system from experimental (noisy)/simulated observations. The error in the model or the deviation between model and actual system is measured analytically. Optimizing this error equation with respect to the parameters of the intended model is the important task undertaken through the vector fitting process. This optimization leads to a linear least square problem in each iteration, the solution vector of which contains the required parameters for system identification. Therefore, the numerical quality (solvability) of the resulting linear least square (LLS) equation is highly important. In addition, utilizing an accurate solution method is essential to attain an accurate solution.

This chapter first, reviews the 'Error Estimator and Optimization' concepts and techniques, utilized in vector fitting process. The formation of equations will be in 'frequency–domain'; however, the concepts are kept adequately general and valid for both time as well as frequency-domains. Afterward, linear least square problems in a detailed level will be considered. Existing efficient methods to solve LLS problem, formed in VF, will be reviewed. Furthermore, the 'rank-revealing QR' (RRQR) as an accurate solution method, will be discussed. In addition, a feasible formulation for

RRQR will be presented with mathematical details. The proposed formulation can be utilized in least square vector fitting algorithms regardless of the-domain and type of the basis functions.

6.1 Preliminary

Consider the dynamic LTI continuous-time system, which is originally known by a consistent set of experimental data. For the frequency-domain identification, the data can be a discrete set of measured input(X)-output(Y) data (spectra) at the ports of the network as shown in (6.1) or in the form of the frequency response shown in (6.2) .

$$\left\{X_i(f_k),\, Y_j(f_k)\right\}_{k=1,2,\ldots,K} \tag{6.1}$$

$$\left\{H_{ij}(f_k)\right\}_{k=1,2,\ldots,K} \tag{6.2}$$

where K is the length of the observed data.

A set of s-parameters from full-wave simulations can be named an example for the latter case.

In the vector fitting process, the goal is estimating real-valued coefficients $P_N = \{a_k\}$ and $P_D = \{b_k\}$ in the rational transfer function model H(s) of order n/d shown below.

$$\mathrm{H(s)} = \frac{N(s)}{D(s)} = \frac{\sum_{k=0}^{N} a_k s^k}{\sum_{k=0}^{D} b_k s^k} \tag{6.3}$$

It involves the real and complex conjugate poles.

To use the frequency-domain data with Laplace-domain functions, one should consider that, mathematically, there is an equivalency between the bilateral Laplace and continuous Fourier transforms by considering $s = j\omega = j2\pi f$.

$$H(\omega) = \mathcal{F}\{h(t)\} = \mathcal{L}\{h(t)\}\big|_{s=j\omega} = H(s)\big|_{s=j\omega} \tag{6.4}$$

Accordingly, with $s = j2\pi f$ the s-domain transfer function for dynamic system can be fed by the frequency spectrum data over some frequency range of interest. Several methods have been devised in the literature to fit (approximate) the experimental data denoted as $\tilde{H}(j\omega)$ with this rational function at the observed frequency points.

6.2 Error Estimator and Optimization

The absolute error between the observed data, $\tilde{H}(j\omega)$, and the one from approximation by rational fraction at any arbitrary frequency ω_i is defined below.

$$\varepsilon_k = \tilde{H}(j\omega_k) - H(j\omega_k) = \tilde{H}(j\omega_k) - \frac{N(j\omega_k)}{D(j\omega_k)} \tag{6.5}$$

ε_k is a function with respect to the ω and the parameters based on which $H(j\omega_k)$ has been defined. The goal is finding the parameters for which the magnitude of the error is the minimum at every frequency points. This goal is sensibly obtained by an attempt to minimize the mean square value of the error function. This optimization problem is solved by minimizing the following 'quadratic' cost function, by using one of the existing estimation methods[1].

[1] There are several estimation methods available in the literature. Ref. [17] can be referred for a recent (1994) survey.

$$\sum_{k=1}^{K} |\varepsilon_k|^2 = \sum_{k=1}^{K} \left| \tilde{H}(j\omega_k) - \frac{N(j\omega_k)}{D(j\omega_k)} \right|^2 \tag{6.6}$$

The problem of evaluating the model coefficients $\{a_k\}_{k=0,1,2,\ldots,N}$ and $\{b_k\}_{k=1,2,\ldots,D}$ by minimizing the sum of $|\varepsilon_k|^2$ at several experimental points leads to an overdetermined system of rational equations. Hence, minimizing the cost function (6.6) with respect to the model coefficients due to resulting in a non-linear least square problem is a cumbersome task[2]. In general, it is not possible in a fast and accurate way [15].

6.3 Linear Least Square Estimator

A first approximation to minimize error terms in (6.6) was given by Levi in [14], where he reduced the nonlinear least squares problem to a linear one by multiplying the denominator $D(j\omega_k)$ of the transfer function to the both side of the error equation.

$$\varepsilon'_k = D(j\omega_k) \times \varepsilon_k = D(j\omega_k) \times \left(\tilde{H}(j\omega_k) - \frac{N(j\omega_k)}{D(j\omega_k)} \right) \tag{6.7}$$

As a drawback, it should be considered that, when the roots of the denominator $D(j\omega_k)$ or poles of transfer function happen to have the angular frequency among the observed ω_k points, following the Levi's approach might cause a large error in approximation. Since multiplying by the real part of poles, which are normally very

[2] For the "Nonlinear Least Square Estimator" there are also optimization methods available in literature to minimize the nonlinear cost function. The proposed methods in [52] and [53] can be referred to for more details wherein the Newton-Gauss iteration scheme is utilized for s and z –domains. As it is always the drawback for Newton-Gauss, in this case the algorithm may converge to (be trapped at) a local extremum point.

small, can hide the error in fitting at those frequency points even if the error level is remarkably high.

By taking Levi's approach, the problem is simplified to minimizing the summation of the squared weighted error as below.

$$\sum_{k=1}^{K} \left|\varepsilon'_k\right|^2 = \sum_{k=1}^{K} \left|D(j\omega_k)\varepsilon_k\right|^2 = \sum_{k=1}^{K} \left|\tilde{H}(j\omega_k)D(j\omega_k) - N(j\omega_k)\right|^2 \qquad (6.8)$$

$$W_\varepsilon(\omega_k, P_D, P_N) = \sum_{k=1}^{K} \left|\tilde{H}(j\omega_k)D(j\omega_k) - N(j\omega_k)\right|^2 \qquad (6.9)$$

If $N(j\omega)$ and $D(j\omega)$ are given in the form of polynomials, the unknown parameters, P_N and P_D, will be the real coefficients in power series, $\{a_k\}$ and $\{b_k\}$, respectively. In this case, the resulting LLS problem is badly scaled and numerically ill-conditioned; seeing that "the columns in matrix **A** in equation (6.12) are multiplied with different powers of s. This fact limits the method to approximations of very low order, particularly if the fitting is over a wide frequency range" [13].

Alternatively, for curve-fitting formulation, both polynomials in numerator and denominator are approximated by a linear span of some possible bases[3]. Therefore, in resulting formulation, P_N and P_D represent the real coefficients for the bases (instead of coefficients for different powers of s). To form a linearized cost function with respect to the coefficients "a priori knowledge of poles [15]" is required to form and evaluate the intended bases functions at each frequency point.

Definition: For a real-valued function as $W_\varepsilon(\omega_k, P_D, P_N)$ in (6.9) with $\mathbb{R}$ as the domain (from which the valid values are selected) for parameters,

[3] Of the form of the Partial fractions or rational orthonormal basis functions

"$\arg \min_{P_D,P_N} W_\varepsilon(\omega_k,P_D,P_N)$ [†]" presents the set of parameters in $\mathbb{R}$ for which the function achieves the global minimum[4].

$$\arg \min_{P_D,P_N} \sum_{k=1}^{K} \left|\tilde{H}(j\omega_k)D(j\omega_k)-N(j\omega_k)\right|^2 =$$

$$=\{\, \mathcal{P}_D,\mathcal{P}_N \in \mathbb{R} \mid \quad W_\varepsilon(\omega_k,\mathcal{P}_D,\mathcal{P}_N)= \min_{\substack{\text{for any possible} \\ P_D,P_N \in \mathbb{R}}} \left(W_\varepsilon(\omega_k,P_D,P_N)\right) \,\} \qquad (6.10)$$

For a function of the form in (6.10) which is structured by a summation of absolute (positive) values, the expected global minimum sensibly is zero which is the minimum possible positive number. It understandably occurs only when all absolute value terms in summation are enforced to be zero.

Therefore:

$$\tilde{H}(j\omega_k)D(j\omega_k)-N(j\omega_k)=0, \quad k=1,2,\ldots,K \qquad (6.11)$$

Eq. (6.11) holds at all the experimental/simulated data points.

Writing (6.11) for several frequency points gives a set of "reasonably many" linear simultaneous algebraic equations. The resulting system in matrix notation resembles the general form in below.

$$[\mathbf{A}][\mathbf{X}]=[\mathbf{B}] \qquad (6.12)$$

6.4 Sanathanan and Koerner Interactive Weighted LLS Estimator

As it has been stated by Sanathanan and Koerner (SK) in [54]:

[†] "**arg min**"is a commonly used notation in the context of the optimization in least square sense.

[4] Compare to the concern in footnote 2

1) If the transfer function has to be determined for frequencies extending several decades, the elements of the matrix [A] are such that the lower frequency values have very little influence. Hence, a good fit cannot be obtained at lower frequencies.

2) If H(s) has poles in the complex s-plane and the angular frequencies for one or some of which happens to be quite close to the frequency points in the data set, then "the $|D(j\omega)|^2$ could vary widely throughout the experimental points and large error would be introduced [54]".

The SK[5] method in [54] by an iterative procedure overcame the above deficiency including the lack of sensitivity to low frequency errors of the linear least squares estimator.

Hence, the problem in the iteration number (i) of process consists of minimizing:

$$\sum_{k=1}^{K}\left(\left|\frac{D^{(i)}(j\omega_k)}{D^{(i-1)}(j\omega_k)}\right|^2|\varepsilon_k|^2\right)=\sum_{k=1}^{K}\left|\frac{1}{D^{(i-1)}(j\omega_k)}\right|^2\left|\tilde{H}(j\omega_k)D^{(i)}(j\omega_k)-N^{(i)}(j\omega_k)\right|^2 \tag{6.13}$$

where $\varepsilon''^{(i)}_k=\left(\frac{1}{D^{(i-1)}(j\omega_k)}\right)\times\varepsilon'^{(i)}_k=\left(\frac{D^{(i)}(j\omega_k)}{D^{(i-1)}(j\omega_k)}\right)\times\varepsilon^{(i)}_k$ is the weighted error associated with the iteration $\#i$ at the frequency point $\#k$. In the first iteration $(i=1)$ as $D^{(0)}(j\omega_k)$ is not known initially, it is assumed equal to one. This infers that for the very first iteration an estimate of the parameters (poles in curve-fitting) is obtained by minimizing the linearized cost function in (6.10). While for the rest of the iterations,

[5] Sanathanan and Koerner

until convergence occurs, the global minimum of weighted error is obtained from the following iteratively weighted-linearized cost function.

$$\arg \min_{P_D,P_N} \sum_{k=1}^{K} \left| \frac{1}{D^{(i-1)}(j\omega_k)} \right|^2 \left| \tilde{H}(j\omega_k) D^{(i)}(j\omega_k) - N^{(i)}(j\omega_k) \right|^2 \qquad (6.14)$$

In practice, this approach often gives favorable results for sufficiently high signal-to-noise ratios and sufficiently small modeling errors [15].

As the noteworthy remarks on the case, it is declared that:

i) By analyzing the gradients of the error criterion, it is straightforward to show that this method generates solutions that do not converge asymptotically to the solution of the linearized cost function in (6.10) even though the error criterion itself tends asymptotically to the fundamental least squares criterion [15], [59][6].

ii) In the original paper from Sanathanan and Koerner [54], it is stated, "the subsequent iterations tend to converge rapidly and the coefficients evaluated become effectively those obtained by minimizing the sum of $|\varepsilon_k|^2$ (non-linear estimator) at all the experimental points."

To avoid any possible confusion, it is highlighted that the first remark is about linearized estimator, while the second is about the nonlinear one. An authentic mathematical proof is available in the literature corroborating the latter property as "the solution of SK algorithm converges asymptotically to the solution of the nonlinear least square problem

[6] What has been observed within examining a convincing number of practical cases is experimentally verifying this conclusion.

as the iteration step $i \to \infty$ ". The convergence section in the next chapter presents more details in this regard.

6.5 On Linear Least Square Problems

Consider the problem of finding a vector $\mathbf{X} \in \mathbb{R}^n$ in the equation $\mathbf{AX} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$ are given. $\mathbf{A}$ and $\mathbf{B}$ are referred as "the data or coefficient matrix" and "the observation vector" respectively.

In each iteration of vector fitting process, a linear system of equations is required to be solved as error estimator. Having more equations than unknowns $(m > n)$, this estimator falls in a category known as overdetermined. Usually an overdetermined system has no exact solution [55]. Therefore solving the system explicitly means minimizing the suitable norm of the residual error. When a linearized form of error estimator is solved by minimizing the 2-norm of residual error, it is referred as linear least square (LLS) problem.

These 'error optimization' and 'linear least square' problems arise in many areas of science and engineering. In vector fitting process, utilizing an efficient algorithm to solve the large system of linear equations in least square sense is highly important. It ensures the final model to be sufficiently accurate and obtainable within a fast convergence.

6.6 Declaration in a More Mathematical Manner

Solving the full rank LLS problem consists in finding the vector $\mathbf{X}$ that satisfies:

$$\min_{\mathbf{X}\in\mathbb{R}^n} \|\mathbf{AX}-\mathbf{b}\|_2 \tag{6.15}$$

The least square problem for two following reasons is more tractable in contrast to the general p-norm minimization.

First, $\|\mathbf{AX}-\mathbf{b}\|_2$ is a differentiable function of $\mathbf{X}$ (and so it is $\phi(X)=\frac{1}{2}\|\mathbf{AX}-\mathbf{b}\|_2^2$), thus minimizing $\phi(X)$ satisfies the gradient equation as $\nabla\phi(X)=0$ [55]. Satisfying this gradient equation mathematically assures of obtaining a minimal point. Second, the 2-norm is invariant (preserved) under unitary (e.g. orthonormal) transformations. This means that we can seek an orthogonal $\mathbf{Q}$ such that the equivalent problem of minimizing $\frac{1}{2}\left\|\left(\mathbf{Q}^T\mathbf{A}\right)\mathbf{X}-\left(\mathbf{Q}^T\mathbf{b}\right)\right\|_2^2$ is easy to solve [55].

Here, orthonormal bases directly have been obtained by orthogonalizing the original bases utilizing any known orthogonalization algorithm (e.g. Gram-Schmidt). The attempt of utilizing these bases to approximate the observed data can improve the ill-conditioning in the original least squares formulation. Constructing a new system of equations as shown in above logically is an attempt of finding X from a better numerically conditioned LLS equation while minimizing the original estimator, $\|\mathbf{AX}-\mathbf{b}\|_2^2$ as well.

Similarly, solving a curve-fitting problem by using the orthogonal bases can be judged as forming the equivalent error estimator with better numerical condition.

It is a fact that, one can always increase the precision to overcome the ill-conditioning [56]. A better approach is to use orthogonal bases instead of using the standard partial fractional basis functions.

6.7 Noteworthy Considerations in Solving LLS Problems

- Consider overdetermined linear system when there are more equations than unknowns $(m \geq n)$; If $\mathbf{A}$ has full rank, $rank(\mathbf{A}) = n$, then there exists a unique solution to the LLS Problem. Otherwise, if $\mathbf{A}$ is rank-deficient, $rank(\mathbf{A}) < n$, there exists an infinite number of solutions.
- The set of all the "minimum two-norm solution $\aleph$ " can be defined as [55]:

$$\aleph = \left\{ \mathbf{X} \in \mathbb{R}^n : \|\mathbf{AX} - \mathbf{b}\|_2 = \min \right\} \tag{6.16}$$

In the above set of solutions, there is a unique element that has minimum two-norm, $\left(\min \|\mathbf{X}\|_2\right)$. It is considered as a unique solution to the LLS Problem.

Practically, the solution for the rank-deficient linear overdetermined system of equation is obtained by minimizing both: $\begin{cases} \|\mathbf{AX} - \mathbf{b}\|_2 = \min \\ \|\mathbf{X}\|_2 = \min \end{cases}$

6.8 A Review of Existing Methods for Solving LLS Problems

There are three standard solution methods: the "Normal Equation" (NE), the "QR decomposition", and the "singular decomposition" (SVD).

When the coefficient matrix has full rank, the solution can be obtained in a fast way by the first two methods. In contrast, when the matrix is rank-deficient or the rank is not known, The third category of solver methods such as SVD and the "Complete Orthogonal Decomposition" should be utilized.

6.8.1 Normal Equation Method

According to the above explanation, minimizing vector $\mathbf{X}$ in equation $\mathbf{AX} = \mathbf{b}$ is a solution of the 'normal equation' in the following form.

$$\left(\mathbf{A}^T\mathbf{A}\right)\mathbf{X} = \mathbf{A}^T\mathbf{b} \tag{6.17}$$

The matrix $\left(\mathbf{A}^T\mathbf{A}\right)$ on the left-hand side is an n-by-n square matrix, which is invertible if A has full column rank (that is, if the rank of $\mathbf{A}$ is n). In that case, the solution of the system of linear equations is unique and given by

$$\mathbf{X} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b} \tag{6.18}$$

It is worth noting that:

- since matrix $\mathbf{A}$ in above is not square, its direct inverse matrix does not exist. Considering the fact that Matrix $\left(\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\right)$ superficially works in place of the inverse of the matrix $\mathbf{A}$, it is called 'psedoinverse' of $\mathbf{A}$.
- The normal equations method squares the conditioning number of problem[7] so suffers the most round off error, but it is the fastest method [56].

6.8.2 On Efficient and Accurate Solution Methods

The SVD method offers a high numerical accuracy, but it is computationally expensive [57]. Its high computational cost sometimes makes it impractical for very large systems. The "complete orthogonal decomposition" is an alternative approach for

[7] By forming $\left(\mathbf{A}^T\mathbf{A}\right)$ as the square coefficient matrix, shown in equation (6.17).

solving the ill-condition cases. It is a faster method in comparison to SVD and performs accurately in practice.

To continue, the newly introduced algorithm cited as 'Complete Orthogonal Decomposition with RRQR', will be mathematically clarified. It is an attempt of lowering the computational cost while preserving numerical accuracy for ill-condition cases.

6.9 Rank-Revealing QR (RRQR) Factorization

Let the matrix **A** be an m-by-n (rectangular) matrix where $m \geq n$ (this assumption will not degrade the level of generality). Also, δ_i represents the "singular values" of the matrix **A**, where $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r \geq \delta_{r+1} \geq \cdots \geq \delta_n \geq 0$.

Definition 1: In the most general form, the **QR factorization** of matrix A is given by:

$$\mathbf{AP} = \mathbf{QR} = \mathbf{Q}\begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \bar{\mathbf{0}} & \mathbf{R}_{22} \\ \bar{\mathbf{0}} & \bar{\mathbf{0}} \end{pmatrix} \begin{matrix} r \\ n-r \\ m-n \end{matrix} \Big\} n \qquad (6.19)$$
$$\qquad\qquad r \quad n-r$$

in which $\mathbf{P}_{n\times n}$ is a permutation matrix, $\mathbf{Q} \in \mathbb{R}^{m\times m}$ is orthonormal $\left(\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_{m\times m}\right)$ and $\mathbf{R} \in \mathbb{R}^{m\times n}$ is upper triangular [55]. $\mathbf{R}_{11}$ is a $r\times r$ upper triangular matrix, $\mathbf{R}_{12}$ is a $r\times(n-r)$ matrix and $\mathbf{R}_{22}$ is $(n-r)\times(n-r)$ upper triangular matrix.

Definition 2: The numerical rank of A with respect to the threshold τ denoted as r is defined as shown below.

$$\mathbf{AP} = \mathbf{QR} = \mathbf{Q}\begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \overline{\mathbf{0}} & \overline{\mathbf{0}} \end{pmatrix}\begin{matrix} r \\ m-r \end{matrix} \qquad (6.20)$$
$$\begin{matrix} r & n-r \end{matrix}$$

Definition 3: If $r = n$, matrix **A** has **full column rank**, otherwise when $r < n$ **A** is known as rank-deficient.

Definition 4: The QR factorization of A as in (6.19) is an **RRQR factorization** if it happens to hold a property as:

$cond(\mathbf{R}_{11}) \approx \delta_1 / \delta_r$ *or* $cond(\mathbf{R}_{11}) \le \tau$, it infers that $\mathbf{R}_{11}$ is full rank

$\|\mathbf{R}_{22}\|_2 = \delta_{\max}(\mathbf{R}_{22}) \approx \delta_{r+1}$, it infers that $\mathbf{R}_{22}$ has small norm.

Whenever there is a well-determined gap in the singular-value spectrum between δ_r and δ_{r+1}, that means the numerical rank r is well defined, the RRQR factorization as shown in (6.19) reveals the numerical rank of **A** by having a well-conditioned (non-singular) upper triangular leading sub-matrix $\mathbf{R}_{11}$ and a trailing sub-matrix $\mathbf{R}_{22}$ of small norm [58]. By considering $\mathbf{R}_{22}$ being negligible, the truncated form of QR for any arbitrary rank-deficient matrix can be generalized as below.

Theorem: If $\mathbf{A}_{m\times n}$ is a rank deficient matrix, there exist **P**, **Q**, $\mathbf{R}_{11}$, and $\mathbf{R}_{12}$ such that

$$\mathbf{AP} = \mathbf{QR} = \mathbf{Q}\begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \overline{\mathbf{0}} & \overline{\mathbf{0}} \end{pmatrix}\begin{matrix} r \\ m-r \end{matrix} \qquad (6.21)$$
$$\begin{matrix} r & n-r \end{matrix}$$

In (6.21) the columns in the resulting **Q** are bases for the range of **A** spanning the same space as the column of **A**, while preserving the main property of **Q** as a unitary matrix. In the case of rank deficient QR with "Column Pivoting", the Householder QR factorization procedure can be modified in a simple way to produce (6.21) [55].

6.10 Adapting RRQR Factorization to Solve LLS Problems

Originally the "complete orthogonal decomposition" method called URV was suggested [by Stewart 1990] as an alternative to singular-value decomposition. The computational cost of this method is much cheaper than the SVD, but still provides acceptable results. By applying the rank-revealing QR, one would take advantage of the lower rank of matrix **A** to reduce the computational complexity. Also, the resulting method efficiently works for the very ill-condition (tends to be singular) cases.

This URV factorization decomposes a matrix as (6.22) where **U** and **V** are orthogonal and both $\mathbf{T}_{12}$ and $\mathbf{T}_{22}$ are of small norm of the order δ_{r+1} [58].

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_{22} \end{pmatrix}\mathbf{V}^T, \qquad \text{where, } \|\mathbf{T}_{12}\|_2 \approx \delta_{r+1} \text{ and } \|\mathbf{T}_{22}\|_2 \approx \delta_{r+1} \tag{6.22}$$

For a rank-deficient or nearly rank-deficient $m \times n$ matrix **A**, when there is a well-determined gap in the singular-value spectrum between δ_r and δ_{r+1}, then δ_{r+1} is adequately small.

Small value of δ_{r+1} directly means that all the entries in the matrices $\mathbf{T}_{12}$ and $\mathbf{T}_{22}$ are negligible and they can be approximated by zero matrices of the same size. Accordingly, the equation (6.22) would resemble the form as shown below [55]:

$$\mathbf{A} = \mathbf{UTV} = \mathbf{U}\begin{bmatrix} \mathbf{T}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{V} \tag{6.23}$$

Compared with QR (in which $\mathbf{A} = \mathbf{QRE}^T$) URV employs a general orthonormal matrix **V** instead of the permutation matrix **P**. This fact implies that, RRQR factorization algorithm (with slight modification) can be employed to compute an initial URV decomposition.

Considering RRQR as in (6.21) by applying a matrix transpose to both sides, it is:

$$(\mathbf{AP})^T = \mathbf{R}^T\mathbf{Q}^T = \begin{pmatrix} \mathbf{R}_{11}^T & \bar{\mathbf{0}} \\ \mathbf{R}_{12}^T & \bar{\mathbf{0}} \end{pmatrix}_{n\times m} \mathbf{Q}_{m\times m}^T \tag{6.24}$$

by using the block matrix arithmetic and by excluding the zero blocks in the matrix calculation, (6.24) can be re-written in the shrunken form:

$$(\mathbf{AP})^T = \begin{pmatrix} \mathbf{R}_{11}^T \\ \mathbf{R}_{12}^T \end{pmatrix}_{n\times r} \times \left(\mathbf{Q}^T\right)_{(1:r,\,1:m)} \tag{6.25}$$

The right hand side matrix can be further reduced if it is post-multiplied by an appropriate sequence of householder matrices [55] as shown below.

$$\mathbf{Z}_r \cdots \mathbf{Z}_1 \begin{bmatrix} \mathbf{R}_{11}^T \\ \mathbf{R}_{12}^T \end{bmatrix} = \begin{bmatrix} \mathbf{T}_{11}^T \\ \bar{\mathbf{0}} \end{bmatrix} \begin{matrix} r \\ n-r \end{matrix} \tag{6.26}$$

In (6.26), $\mathbf{Z}_i$ are Householder transformations and $\mathbf{T}_{11}^T$ is upper triangular. One of the possible (happens to be the most accurate) algorithm to perform QR decomposition is by utilizing the Householder transformations. Hence, the relationship between $\mathbf{Q}$ and householder matrices is attainable as:

$$\mathbf{Q} = \mathbf{Z}_1 \cdots \mathbf{Z}_r \quad \text{or} \quad \mathbf{Q}^T = \mathbf{Z}_r \cdots \mathbf{Z}_1 \tag{6.27}$$

Consequently with applying QR decomposition for the matrix denoted as $\mathfrak{R}$ in (6.28):

$$\underbrace{\begin{bmatrix} \mathbf{R}_{11}^T \\ \mathbf{R}_{12}^T \end{bmatrix}}_{\mathfrak{R}} = \underbrace{\mathbf{Z}_1 \cdots \mathbf{Z}_r}_{\mathbf{Q}_{\mathfrak{R}}} \overbrace{\begin{bmatrix} \mathbf{T}_{11}^T \\ \bar{\mathbf{0}} \end{bmatrix}}^{\mathbf{R}_{\mathfrak{R}}} \begin{matrix} r \\ n-r \end{matrix} \tag{6.28}$$

Next, by multiplying both sides of (6.25) by $\left(\mathbf{Q}_{\Re\,n\times n}\right)^T$, we will have:

$$\left(\mathbf{Q}_{\Re\,n\times n}\right)^T(\mathbf{AP})^T=\left(\mathbf{Q}_{\Re\,n\times n}\right)^T\times\begin{pmatrix}\mathbf{R}_{11}^T\\ \mathbf{R}_{12}^T\end{pmatrix}_{n\times r}\times\left(\mathbf{Q}^T\right)_{(1{:}r,\,1{:}m)} \tag{6.29}$$

by substituting (6.26) in (6.29):

$$\left(\mathbf{APQ}_{\Re\,n\times n}\right)^T=\begin{bmatrix}\mathbf{T}_{11}^T\\ \overline{\mathbf{0}}\end{bmatrix}_{n\times r}\times\left(\mathbf{Q}^T\right)_{(1{:}r,\,1{:}m)} \tag{6.30}$$

Returning the equations to the original size by including the zero blocks leads to the following expanded form:

$$\left(\mathbf{APQ}_{\Re\,n\times n}\right)^T=\begin{bmatrix}\mathbf{T}_{11}^T & \overline{\mathbf{0}}\\ \overline{\mathbf{0}} & \overline{\mathbf{0}}\end{bmatrix}_{n\times m}\times\mathbf{Q}_{m\times m}^T \tag{6.31}$$

Hence,

$$\mathbf{APQ}_{\Re\,n\times n}=\mathbf{Q}_{m\times m}\begin{bmatrix}\mathbf{T}_{11} & \overline{\mathbf{0}}\\ \overline{\mathbf{0}} & \overline{\mathbf{0}}\end{bmatrix}\begin{matrix}r\\ m-r\end{matrix}\qquad\begin{matrix}r & n-r\end{matrix} \tag{6.32}$$

$$\mathbf{A}=\mathbf{Q}_{m\times m}\begin{bmatrix}\mathbf{T}_{11} & \overline{\mathbf{0}}\\ \overline{\mathbf{0}} & \overline{\mathbf{0}}\end{bmatrix}_{m\times n}\left(\mathbf{Q}_{\Re\,n\times n}\right)^T\mathbf{P}_{n\times n}^T \tag{6.33}$$

$$\mathbf{A}=\mathbf{QT}\left(\mathbf{Q}_{\Re}\right)^T\mathbf{P}^T \tag{6.34}$$

Next, utilizing pseudoinverse as it is shown in continue would be an efficient approach.

Utilizing Pseudoinverse:

Considering the singular value decomposition of a rectangular matrix, $\mathbf{A}_{m\times n}$ with rank r. Then there exist a matrix $\Sigma_{m\times n}=diag(\delta_1,\cdots,\delta_r,0,\cdots,0)$, wherein $\delta_1\geq\delta_2\geq\cdots\geq\delta_r>0$, also, orthogonal matrices $U_{m\times m}$ and $V_{n\times n}$ such that:

$$\mathbf{A}=\mathbf{U}\Sigma\mathbf{V}^T \tag{6.35}$$

Based on matrix factors in (6.35), we define the matrix $\mathbf{A}^+\in\mathbb{R}^{n\times m}$ by

$$\mathbf{A}^+=\mathbf{V}\Sigma^+\mathbf{U}^T \tag{6.36}$$

where

$$\Sigma^+_{n\times m}=diag\left(\frac{1}{\delta_1},\cdots,\frac{1}{\delta_r},0,\cdots,0\right)\quad r=rank(\mathbf{A}) \tag{6.37}$$

Then $\mathbf{X}_{LS}$ solution for a least square problem as $\mathbf{AX}=\mathbf{b}$ would be $\mathbf{X}_{LS}=\mathbf{A}^+\mathbf{b}$ [55].

Typically, $\mathbf{A}^+$ is defined to be the unique *n*-by-*m* matrix that satisfies the four Moore-Penrose conditions [55]:

a) $\mathbf{AA}^+\mathbf{A}=\mathbf{A}$ b) $\mathbf{A}^+\mathbf{AA}^+=\mathbf{A}^+$

c) $\left(\mathbf{AA}^+\right)^T=\mathbf{AA}^+$ d) $\left(\mathbf{A}^+\mathbf{A}\right)^T=\mathbf{A}^+\mathbf{A}$

According to the above description, the pseudoinverse for $\mathbf{A}$ defined as in (6.33) would be:

$$\mathbf{A}^+=\mathbf{P}\mathbf{Q}_{\Re}\begin{bmatrix}\mathbf{T}_{11}^{-1} & \bar{\mathbf{0}}\\ \bar{\mathbf{0}} & \bar{\mathbf{0}}\end{bmatrix}\mathbf{Q}^T \tag{6.38}$$

This satisfies all above four Moore-Penrose conditions.

By neglecting the zero sub-matrices, an economy form of (6.38) would be:

$$\mathbf{A}^{+} = \mathbf{P}\mathbf{Q}_{\Re}\begin{bmatrix}\mathbf{T}_{11}^{-1}\\ \overline{0}\end{bmatrix}_{n\times r}\hat{\mathbf{Q}}^{T} \tag{6.39}$$

in which, $\hat{\mathbf{Q}}$ consists of first r columns of $\mathbf{Q}$, as shown to the right: $\hat{\mathbf{Q}} = \mathbf{Q}(:, 1:r)$.

Consequently the $\mathbf{X}_{LS}$ solution for a least square problem of interest is:

$$\mathbf{X}_{LS} = \mathbf{A}^{+}\mathbf{b} \Rightarrow \boxed{\mathbf{X}_{LS} = \mathbf{P}\mathbf{Q}_{\Re}\begin{bmatrix}\mathbf{T}_{11}^{-1}\hat{\mathbf{Q}}^{T}\mathbf{b}\\ \overline{\mathbf{0}}\end{bmatrix}} \tag{6.40}$$

6.11 The Computational Complexity

The original URV decompositions are more expensive to compute while being well suited for null-space updating [58]. It needs $\mathcal{O}\left(mn^2\right)$ floating point operations (flops). The RRQR factorization based method (described above), on the other hand, are more suited for least squares setting, since one need not store the orthogonal matrix $\mathbf{V}$. The complexity level of an efficient implementation of explained method may be of the order $\mathcal{O}(mnr)$. In the cases that the effective numerical rank 'r' is very small, the efficiency of the method would be superior.

CHAPTER 7. Formulation of z-Domain Orthonormal Vector Fitting

In chapter 2, the transformation methods to convert 'frequency-domain' and 'time-domain' data to z-domain were discussed. In addition, the feasible forms for the z-domain transfer function of physical systems were reviewed. Chapters 4 and 5 presented the new z-domain orthonormal functions and related state-space realization in discrete-domain. Chapter 6 provided insights regarding the error estimation methods and efficient solutions for the resulted LLS[1] problems.

Relying the outlined conclusions, this chapter will present the formulation for the proposed 'identification algorithm for rational transfer functions' (macromodeling) in z-domain by using novel orthonormal bases. This technique is capable of handling the 'measured / simulated' data in 'time-domain' as well as 'frequency-domain'. Moreover, it can carry out the macromodeling task for both 'continuous-time' and 'discreet-time' stable LTI systems. This broad validity scope for the proposed algorithm is one of its advantages over the existing continuous-domain methods. The fact that, the proposed method is capable of performing the macromodeling task with higher level of accuracy in the final model even when starting poles are not optimal is considered as another remarkable merit.

[1] Linear Least Squared

7.1 Problem Formulation

Conventional vector fitting algorithm was originally outlined in [13]. A formulation process utilizing the orthonormal basis recently has been presented in [15]. Both are in the frequency-domain.

Seemingly, in a parallel fashion with the above works, this thesis presents a complete account of formulation for z-domain vector fitting process. Basis functions in discrete-time z-domain will be denoted as $\phi(z)$. All other parameters and notations will also adopt the discrete-domain with involving "z" as the independent variable.

To continue, consider the linear dynamic time-invariant system, which is originally known by a consistent set of experimental data. The major goal of employing the proposed technique is to identify the mapping between the inputs and outputs of this (complex) system by an analytic model in the following form.

$$\mathrm{H}(z)=\frac{N(z)}{D(z)}=\frac{\sum_{k=0}^{N} a_k z^k}{\sum_{k=0}^{D} b_k z^k}, \quad \text{where } a_k \text{ and } b_k \in \mathbb{R} \tag{7.1}$$

7.1.1 Approximation of Rational TF Using a Linear Span of Basis Functions

The proper rational transfer function shown in (7.2) is utilized to fit the experimental data.

$$\mathrm{H}(z)=\frac{N(z)}{D(z)}=\frac{\sum_{k=0}^{N} a_k z^k}{\sum_{k=0}^{D} b_k z^k}=\frac{\sum_{k=0}^{N} a_k z^k}{\left(\sum_{k=0}^{D} b_k z^k\right)-\hat{d}+\hat{d}} \quad ,N \le D \text{ and } \hat{d} \in \mathbb{R} \qquad (7.2)$$

without loss of generality, the transfer function can be shown as in (7.3) where D defines the order of the model.

$$\mathrm{H}(z)=\frac{\sum_{k=0}^{N} a_k z^k}{\sum_{k=0}^{D} \hat{b}_k z^k+\hat{d}}, \quad N \le D \qquad (7.3)$$

If the model presents a physical process, the roots of the denominator will be a set of distinct real and complex conjugate poles. Let the length of the pole set be denoted as P that for a denominator polynomial of order D, it is $P = D$.

The polynomials in numerator and denominator can be well approximated by the linear combinations of intended basis functions as shown below.

$$H(z)=\frac{N(z)}{D(z)}=\frac{\sum_{n=1}^{P} c_n \varphi_n(z)}{\hat{d}+\sum_{n=1}^{P} \hat{c}_n \varphi_n(z)} \qquad (7.4)$$

It is given as the constraints that,

i) The same set of P poles is shared between the rational functions approximating both the numerator and denominator. Thus, the same bases are exploited to approximate them.

ii) $\hat{d}=1$

Recently (2006) an extension of standard vector fitting procedure has been proposed. It considers $\hat{d}$ in the formulation without any pre-assumption for its value. An improvement in the ability of VF to relocate poles to the better position is achieved [60], [61].

Considering the 'Sanathanan and Koerner iteratively weighted-linearized cost function' in z-domain, stated in (6.13) and (6.14) in section 4 of previous chapter:

$$\left(\frac{1}{D^{(i-1)}(z_k)}\right)\left(\tilde{H}(z_k)D^{(i)}(z_k)-N^{(i)}(z_k)\right)=0, \quad k=1,2,\ldots,K \tag{7.5}$$

where, K is the number of z-points and $\tilde{H}(\cdot)$ represents the observed data after being converted to the z-domain.

Combining (7.4) and (7.5) results in the following SK weighted cost function for all z points in data set at i-th recursion.

$$\left(\frac{1}{D^{(i-1)}(z_k)}\right)\left(\sum_{n=1}^{P} c_n^{(i)}\varphi_n^{(i)}(z_k) - \tilde{H}(z_k)\left(\hat{d}+\sum_{n=1}^{P}\hat{c}_n^{(i)}\varphi_n^{(i)}(z_k)\right)\right)=0, \quad k=1,2,\ldots,K \tag{7.6}$$

$$w_k^{(i-1)} \triangleq \frac{1}{D^{(i-1)}(z_k)} \tag{7.7}$$

by combining (7.6) and (7.7), we have:

$$\sum_{n=1}^{P} c_n^{(i)}\left(w_k^{(i-1)}\varphi_n^{(i)}(z_k)\right) - \sum_{n=1}^{P}\hat{c}_n^{(i)}\left\{\tilde{H}(z_k)\left(w_k^{(i-1)}\varphi_n^{(i)}(z_k)\right)\right\} = w_k^{(i-1)}\,\tilde{H}(z_k)\hat{d} \tag{7.8}$$

The resulting system in i-th iteration would consist of K linear simultaneous equation with $2P$ unknown $\left(\hat{d}=1\right)$.

In matrix notation, it resembles:

$$\mathbf{A}_{\text{complex}} =$$

$$\begin{bmatrix} \left(w_1^{(-1)}\varphi_1(z_1)\right) & \cdots & \left(w_1^{(-1)}\varphi_P(z_1)\right) & -\tilde{H}(z_1)\left(w_1^{(-1)}\varphi_1(z_1)\right) & \cdots & -\tilde{H}(z_1)\left(w_1^{(-1)}\varphi_P(z_1)\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \left(w_K^{(-1)}\varphi_1(z_K)\right) & \cdots & \left(w_k^{(-1)}\varphi_P(z_K)\right) & -\tilde{H}(z_K)\left(w_K^{(-1)}\varphi_1(z_K)\right) & \cdots & -\tilde{H}(z_K)\left(w_K^{(-1)}\varphi_P(z_K)\right) \end{bmatrix}_{K\times P} \tag{7.9}$$

Note: To simplify the form of equations, the superscript $(i-1)$ which is indicating that the parameters associated to the previous iteration has been shortened to (-1). Superscript (-1) identically shows that the term has been evaluated in the previous iteration.

$$\mathbf{B}_{\text{complex}\,K\times 1} = \left[w_1^{(-1)}\,\tilde{H}(z_1)\hat{d} \quad \cdots \quad w_K^{(-1)}\,\tilde{H}(z_K)\hat{d} \right]^T \tag{7.10}$$

To continue, it is shown that how the **A** and **B** matrices are constructed by defining their sub-blocks. [2]

i) Diagonal Weighting Matrix:

$$\mathbf{W}^{(i-1)} = \text{diag}\left(\left[w_k^{(i-1)}\right]_{1\times(1:K)}\right) = \begin{bmatrix} w_1^{(i-1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_K^{(i-1)} \end{bmatrix}_{K\times K} \tag{7.11}$$

[2] As an implementation concern, it should be considered that the scalar (or element-to-element) product between two matrices and using Kronecker tensor product are much more CPU time efficient in comparison to the multiplying two large matrices. Accordingly, some intelligent modification in the following mathematical formulation can result in dramatic improvement in execution time.

ii) ***Bases Matrix:***

$$\mathbf{\Phi} = \begin{bmatrix} \varphi_1(z_1) & \cdots & \varphi_P(z_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(z_K) & \cdots & \varphi_P(z_K) \end{bmatrix}_{K\times P} \tag{7.12}$$

iii) ***Weighted Bases Matrix:***

$$\mathbf{\Lambda}^{(i)} = \mathbf{W}^{(i-1)} \times \mathbf{\Phi}^{(i)} = \begin{bmatrix} w_1^{(-1)}\varphi_1(z_1) & \cdots & w_1^{(-1)}\varphi_P(z_1) \\ \vdots & \ddots & \vdots \\ w_K^{(-1)}\varphi_1(z_K) & \cdots & w_K^{(-1)}\varphi_P(z_K) \end{bmatrix}_{K\times P} \tag{7.13}$$

iv) ***Diagonal Data Matrix:***

$$\mathbf{H} = \mathrm{diag}\left(\left[\tilde{H}(z_k)\right]_{1\times(1:K)}\right) = \begin{bmatrix} \tilde{H}(z_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tilde{H}(z_K) \end{bmatrix}_{K\times K} \tag{7.14}$$

$$-\mathbf{H\Lambda}^{(i)} = -\mathbf{H} \times \mathbf{\Lambda}^{(i)} = \begin{bmatrix} -H(z_1)w_1^{(-1)}\varphi_1(z_1) & \cdots & -H(z_1)w_1^{(-1)}\varphi_P(z_1) \\ \vdots & \ddots & \vdots \\ -H(z_K)w_K^{(-1)}\varphi_1(z_K) & \cdots & -H(z_K)w_K^{(-1)}\varphi_P(z_K) \end{bmatrix}_{K\times P} \tag{7.15}$$

Hence, (7.13) and (7.15):

$$\mathbf{A}^{(i)}{}_{\text{complex}} = \begin{bmatrix} \mathbf{\Lambda}^{(i)} & -\mathbf{H\Lambda}^{(i)} \end{bmatrix}_{K\times 2P} \tag{7.16}$$

Then the least square equation $(\mathbf{Ax} = \mathbf{B})$ with real matrices is formed as below. It enforces the entries of the vector of unknown, which are the coefficients of the bases to appear as real values.

$$\mathbf{A}=\begin{bmatrix}\mathcal{R}e\left(\begin{bmatrix}\boldsymbol{\Lambda} & -\mathbf{H}\boldsymbol{\Lambda}\end{bmatrix}\right)\\ \mathcal{I}m\left(\begin{bmatrix}\boldsymbol{\Lambda} & -\mathbf{H}\boldsymbol{\Lambda}\end{bmatrix}\right)\end{bmatrix}_{2K\times 2P} \tag{7.17}$$

$$\text{Vector of unknowns: } \mathbf{X}_{(2P)\times 1}=\begin{bmatrix}c_1 & \cdots & c_P & \hat{c}_1 & \cdots & \hat{c}_P\end{bmatrix}^T \tag{7.18}$$

And for $\hat{d}=1$,

$$\mathbf{B}_{2K\times 1}=\left[\mathcal{R}e\left(\begin{bmatrix}w_1^{(-1)}\tilde{H}(z_1) & \cdots & w_K^{(-1)}\tilde{H}(z_K)\end{bmatrix}\right)\quad \mathcal{I}m\left(\begin{bmatrix}w_1^{(-1)}\tilde{H}(z_1) & \cdots & w_K^{(-1)}\tilde{H}(z_K)\end{bmatrix}\right)\right]^T \tag{7.19}$$

Now, the obtained least square problem is solved to attain the coefficients resulted from *i*-th iteration. Providing an accurate solution in this step is essential to accomplish a sufficient level of accuracy in the final model within a rapid convergence. For this reason, the feasible solution methods for linear system of equations in least square sense including RRQR method were presented in the previous chapter.

7.1.2 When Modeling with Improper Transfer Functions

In this section, the improper form of transfer functions is considered for the modeling. Considering the similarity, extracting the accurate formulation in both freq. and z-domains will be discussed.

7.1.2.1 For Frequency-Domain Vector-Fitting Techniques

Consider the Laplace-domain improper rational function as:

$$\mathrm{H}(s)=\frac{\sum_{k=0}^{N} a_k s^k}{\sum_{k=0}^{D} b_k s^k}+rs+q \quad \text{, where } D<N \text{ and } r,q \text{ are real} \tag{7.20}$$

It results in:

$$\mathrm{H}(s)=\frac{N(s)}{D(s)}=\frac{\sum_{k=0}^{D+1} \hat{a}_k s^k}{\sum_{k=0}^{D} b_k s^k} \tag{7.21}$$

without any loss of generality, (7.21) can be written in the form:

$$\mathrm{H}(s)=\frac{N(s)}{D(s)}=\frac{\sum_{n=1}^{P} c_n \varphi_n(s)+c''s+c'}{\hat{d}+\sum_{n=1}^{P} \hat{c}_n \varphi_n(s)} \tag{7.22}$$

It is still assumed that the numerator and denominator share the same poles. Mathematically, it is seen that the numerator in (7.22) is a rational function of a polynomial with order $(P+1=D+1)$ over a polynomial of the form $\prod_{k=1}^{P}(s-p_k)$. The latter is repeated in the denominator. The rest of the formulation is performed in a similar manner and the resulting matrices would resemble the ones in below.

Other than the matrix $\Lambda^{(i)}$ associated with the denominator, matrix $\Lambda_N^{(i)}$ is defined as:

$$\boldsymbol{\Lambda}_{\mathrm{N}}^{(i)} = \mathbf{W}^{(i-1)} \times \boldsymbol{\Phi}_{1:\mathrm{N}}^{(i)} = \begin{bmatrix} w_1^{(-1)}\varphi_1(s_1) & \cdots & w_1^{(-1)}\varphi_P(s_1) & s_1 & 1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ w_K^{(-1)}\varphi_1(s_K) & \cdots & w_K^{(-1)}\varphi_P(s_K) & s_K & 1 \end{bmatrix}_{K\times(P+2)} \tag{7.23}$$

The form of $\boldsymbol{\Lambda}_{\mathrm{N}}^{(i)}$ is decided by the transfer function according to the type (strictly proper, proper or improper). Associated with the intended type for transfer function there is:

$$\mathbf{X}_{(2P)\times 1} = [c_1 \quad \cdots \quad c_P \quad c'' \quad c' \quad \hat{c}_1 \quad \cdots \quad \hat{c}_P]^T \tag{7.24}$$

7.1.2.2 For z-Domain Vector-Fitting Techniques

The details regarding the possible formulation of z-domain transfer functions for physical systems have been explained in the chapter titled as 'Signals and Systems in Discrete Time-domain'. Referring to the outlined results in the mentioned chapter, H(z) for engineering applications will be in the following form:

$$\mathrm{H}(z) = \frac{a_0 z^N + b_1 z^{N-1} + \cdots + a_N}{b_0 z^D + b_1 z^{D-1} \cdots + b_D}, \quad (N \le D) \tag{7.25}$$

It also has been shown that (7.25) as a z-domain transfer function for LTI passive systems, the most general form can only occur in the following two possible forms:

One appears as $H(z) = z\hat{H}(z)$ and the other may have the form of $H(z) = \hat{H}(z) + d$.

Here, $\hat{H}(z)$ is in strictly proper format. Thus, it can be approximated by an expansion of either partial fractions or orthonormal bases.

It is importantly remarked that H(z) as a transfer function of LTI stable systems can not include a linear term because it causes a causality violation.

For the first case, the standard formulation can be utilized and after obtaining a sufficient model for $\hat{H}(z)$, the final model is analytically worked out.

For the second cases, in a similar fashion with frequency-domain, matrix $\mathbf{\Lambda}_N^{(i)}$ is defined as:

$$\mathbf{\Lambda}_{\mathrm{N}}^{(i)} = \mathbf{W}^{(i-1)} \times \mathbf{\Phi}_{1:\mathrm{N}}^{(i)} = \begin{bmatrix} w_1^{(-1)}\varphi_1(z_1) & \cdots & w_1^{(-1)}\varphi_P(z_1) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ w_K^{(-1)}\varphi_1(z_K) & \cdots & w_K^{(-1)}\varphi_P(z_K) & 1 \end{bmatrix}_{K\times(P+1)} \tag{7.26}$$

7.1.3 Multi–Input and Multi–Output Systems

There is no significant difficulty in extending the above formulation to multi-variable systems. The core idea in the formulation for passive LTI multiport (MIMO) systems is enforcing all the entries of the complex-valued transfer function matrix, $H_{ij}(z)$, to share the same dynamic modes.

As a soft approach, the method is demonstrated for a 2-port network. The less involved form of equations may help to develop the overall idea for larger networks with *n* ports. The four sub-transfer functions defining the port-to-port relationship for MIMO network as shown in (7.27) are treated one by one.

$$\mathbf{H}(z) = \begin{bmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \end{bmatrix} \tag{7.27}$$

Using the formulation for each single port cases induces four independent subsystems of equations as shown in (7.28).

$$\begin{aligned} \mathbf{A}_{11}^{(i)}\begin{bmatrix}\mathbf{C}_{11}^{(i)}\\ \widehat{\mathbf{C}}_{11}^{(i)}\end{bmatrix}=\mathbf{B}_{11}, \quad \mathbf{A}_{12}^{(i)}\begin{bmatrix}\mathbf{C}_{12}^{(i)}\\ \widehat{\mathbf{C}}_{12}^{(i)}\end{bmatrix}=\mathbf{B}_{12} \\ \mathbf{A}_{21}^{(i)}\begin{bmatrix}\mathbf{C}_{21}^{(i)}\\ \widehat{\mathbf{C}}_{21}^{(i)}\end{bmatrix}=\mathbf{B}_{21}, \quad \mathbf{A}_{22}^{(i)}\begin{bmatrix}\mathbf{C}_{22}^{(i)}\\ \widehat{\mathbf{C}}_{22}^{(i)}\end{bmatrix}=\mathbf{B}_{22} \end{aligned} \tag{7.28}$$

In the resulting LLS problems, all the transfer functions should be enforced to share the same coefficients for their denominators,

$$\left[\widehat{\mathbf{C}}_{11}^{(i)}\right]_{P\times 1}=\left[\widehat{\mathbf{C}}_{12}^{(i)}\right]_{P\times 1}=\left[\widehat{\mathbf{C}}_{21}^{(i)}\right]_{P\times 1}=\left[\widehat{\mathbf{C}}_{22}^{(i)}\right]_{P\times 1} \tag{7.29}$$

To ensure the criterion as (7.29), a compact global matrix equation including the 4 simultaneous linear sub-system of equations in (7.28) would be in the form of:

$$\mathbf{A}_{\text{complex}}=\begin{bmatrix}\boldsymbol{\Lambda}_N & 0 & 0 & 0 & -\mathbf{H}_{11}\boldsymbol{\Lambda}\\ 0 & \boldsymbol{\Lambda}_N & 0 & 0 & -\mathbf{H}_{12}\boldsymbol{\Lambda}\\ 0 & 0 & \boldsymbol{\Lambda}_N & 0 & -\mathbf{H}_{21}\boldsymbol{\Lambda}\\ 0 & 0 & 0 & \boldsymbol{\Lambda}_N & -\mathbf{H}_{22}\boldsymbol{\Lambda}\end{bmatrix} \tag{7.30}$$

$$\mathbf{B}_{ij}=\begin{bmatrix} w_1^{(-1)}\,\tilde{H}_{ij}(z_1) & \cdots & w_K^{(-1)}\,\tilde{H}_{ij}(z_K)\end{bmatrix} \tag{7.31}$$

$$\mathbf{B}_{\text{Complex}}=\begin{bmatrix}\mathbf{B}_{11} & \mathbf{B}_{12} & \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix}^T \tag{7.32}$$

$$\mathbf{A}=\begin{bmatrix}\mathcal{R}e\left(\mathbf{A}_{\text{Complex}}\right)\\ \mathcal{I}m\left(\mathbf{A}_{\text{Complex}}\right)\end{bmatrix} \tag{7.33}$$

$$\mathbf{B}=\begin{bmatrix}\mathcal{R}e\left(\mathbf{B}_{\text{Complex}}\right)\\ \mathcal{I}m\left(\mathbf{B}_{\text{Complex}}\right)\end{bmatrix} \tag{7.34}$$

$$\mathbf{X}=\begin{bmatrix}\mathbf{C}_{11}\\ \mathbf{C}_{12}\\ \mathbf{C}_{21}\\ \mathbf{C}_{22}\\ \widehat{\mathbf{C}}\end{bmatrix} \tag{7.35}$$

Defining **A** and **B** matrices in LLS equation (**AX=B**) guarantees the fact that all sub-transfer functions for multi port network would have the same set of poles.

An attempt to generalize the idea for any arbitrary n-ports MIMO system with following the same template would lead to the following concise results.

$$\mathbf{A}_{\text{complex}}=\begin{bmatrix}\boldsymbol{\Lambda}_{\mathrm{N}} & \cdots & 0 & -\mathbf{H}_{11}\boldsymbol{\Lambda}\\ \vdots & \ddots & \vdots & \vdots\\ 0 & \cdots & \boldsymbol{\Lambda}_{\mathrm{N}} & -\mathbf{H}_{nn}\boldsymbol{\Lambda}\end{bmatrix}_{\left(Kn^2\right)\times\left(Nn^2+P\right)} \tag{7.36}$$

$$\mathbf{B}_{\text{Complex}}=\begin{bmatrix}\mathbf{B}_{11}\\ \vdots\\ \mathbf{B}_{nn}\end{bmatrix}_{Kn^2\times 1} \tag{7.37}$$

$$\mathbf{X}=\begin{bmatrix}\mathbf{C}_{11}\\ \vdots\\ \mathbf{C}_{nn}\\ \widehat{\mathbf{C}}\end{bmatrix}_{\left(Nn^2+P\right)\times 1} \tag{7.38}$$

K: The number of the z points in the tabulated data
P: The number of poles (order of the intended model)
n: The number of ports
N: Depends on the type of the model $P\le N\le P+2$

Similarly, to enforce all the coefficients to be real:

$$\mathbf{A}=\begin{bmatrix}\mathcal{R}e\left(\mathbf{A}_{\text{Complex}}\right)\\ \mathcal{I}m\left(\mathbf{A}_{\text{Complex}}\right)\end{bmatrix}\ \text{and}\ \mathbf{B}=\begin{bmatrix}\mathcal{R}e\left(\mathbf{B}_{\text{Complex}}\right)\\ \mathcal{I}m\left(\mathbf{B}_{\text{Complex}}\right)\end{bmatrix}\rightarrow\ \mathbf{AX}=\mathbf{B}$$

7.1.4 Relocated Poles Resulting in each Iteration

After parameterization of **x** by solving the LLS equation for the unknown vector, the function in (7.4)† will be known. The obtained function can be further simplified by cancelling out the common poles from numerator and denominator. In the final form the polynomial numerator of the rational function pointed as $D^{(i)}(z)$ remains as denominator of transfer function. This fact is convincing enough to accept the zeros of $D^{(i)}(z)$ as the poles for the next round of iteration (or final poles).

Thus, the task of finding the relocated poles is simplified to the calculation of the zeros for a dynamic system presented by $D^{(i)}(z)$ shown below.

$$D(z) = \hat{d} + \sum_{n=1}^{P} \hat{c}_n \varphi_n(z) \tag{7.39}$$

Equation (7.39) as a proper transfer function with distinct poles presents a dynamic for which the minimal LTI state-space realization of the form is possible, as shown below.

$$z\mathbf{X}(z) = \mathbf{A}\mathbf{X}(z) + \mathbf{B}U(z) \tag{7.40}$$

$$Y(z) = \mathbf{C}\mathbf{X}(z) + dU(z) \tag{7.41}$$

Chapter 5 can be referred to for more details, wherein the procedure of forming **A**, **B**, **C**, and D matrices when using the z-domain orthonormal bases is precisely explained.

from (7.41):

$$U(z) = \left(-d^{-1}\mathbf{C}\right)\mathbf{X}(z) + \left(d^{-1}\right)Y(z) \tag{7.42}$$

by substituting (7.42) in (7.40):

$$z\mathbf{X}(z) = \mathbf{A}\mathbf{X}(z) + \mathbf{B}\left(-d^{-1}\mathbf{C}\right)\mathbf{X}(z) + \mathbf{B}\left(d^{-1}\right)Y(z)$$

† or equivalently (7.22)

$$z\mathbf{X}(z)=\left(\mathbf{A}-\mathbf{B}d^{-1}\mathbf{C}\right)\mathbf{X}(z)+\left(\mathbf{B}d^{-1}\right)Y(z) \tag{7.43}$$

For the linear system identified by (7.43) and (7.42) as in below:

$$\begin{cases} z\mathbf{X}(z)=\left(\mathbf{A}-\mathbf{B}d^{-1}\mathbf{C}\right)\mathbf{X}(z)+\left(\mathbf{B}d^{-1}\right)Y(z) \\ U(z)=\left(-d^{-1}\mathbf{C}\right)\mathbf{X}(z)+\left(d^{-1}\right)Y(z) \end{cases} \tag{7.44}$$

The transfer function for the interchanged input and output dynamic model is:

$$H_D^{-1}(z)=\frac{U(z)}{Y(z)} \tag{7.45}$$

The eigenvalues of system's dynamic matrix in the state space equation (7.44) as $\{a_i\}=eig\left(\mathbf{A}-d^{-1}\mathbf{BC}\right)$ would be poles of the system. They also are roots of the characteristic equation or the denominator of the transfer function in (7.45).

By considering the relationship as $H_D(z)=\frac{Y(z)}{U(z)}=\frac{1}{H_D^{-1}(z)}$, it is proven that the poles of $H_D^{-1}(z)$ would be the zeros for $H_D(z)$.

Reminding the fact that d has been initially set to unity, then

$$\{zeros\}=eig(\mathbf{A}-\mathbf{BC})^{\dagger} \tag{7.46}$$

The vector-fitting process should insure that, the $\mathbf{H}=\mathbf{A}-\mathbf{BC}$ as a real matrix, for which the eigenvalues are in either real or complex conjugate pairs. It is a mandatory property for the poles of a physical LTI system.

---

† when the (7.41) is written of the form $Y(z)=\mathbf{C}^T\mathbf{X}(z)+dU(z)$, the equation (7.46) will appear as: $\{zeros\}=eig(\mathbf{A}-\mathbf{BC}^T)$, which is just a cosmetic change. The state space equation for a SISO system was of interest in this proof. A conceptually similar attempt to prove has been reported in ref [62]

The resulting poles at this step will be the starting poles for the next iteration until the poles converge to the final locations. The resulted final poles can be used to determine the residues of the partial fractions or coefficients for the orthonormal bases.

7.2 The Choice of Initial Pole Locations

The choice of initial pole positions plays an important role in the quality of the approximated model. As a mater of fact, taking the first guess for the orders and starting poles is in the most part heuristic. An educated guess results in faster convergence and better (more accurate) model with optimum degree.

7.2.1 Selection of Initial Poles in Frequency-Domain

The recommended procedure and criteria for the selection of the s-domain starting poles outlined in [13] and [16] is still considered as a reliable guideline for that-domain. A quick review of the standard methods for selecting the number and locations of initial poles in s-domain can be instrumental. This is useful to obtain an initial idea regarding the order of the intended model too.

- **Functions with distinct resonance peaks [13]:**

The starting poles should be complex conjugate $p_n = -\alpha - j\beta$ and $p_{n+1} = -\alpha + j\beta$. This is not a firm rule; however, one can obtain an idea regarding the number of required poles by associating each local peak or minimum of the frequency-domain spectrum curves with one pair of complex conjugate poles. The imaginary part of poles should be linearly distributed over the frequency range of interest, where $\alpha = \beta/100$. The sufficiently small real parts for the poles help to avoid the ill-conditioning problem.

- **Smooth functions [13]:**

To fit this type of the data one can use real poles that are spaced either linearly or logarithmically.

7.2.2 Selection of Initial Poles in z-Domain

Choosing initial poles in the ZD-VF is performed in one of the following ways.

1) Using the same initial poles with SD-VF

The interchangeability between the initial poles in s and z-domains is explained in the chapter 2.

Table 7-1: Mapping initial poles from s-plane to z-plane

Initial Poles in S-plane	Mapped into	Initial Poles in z-plane
$s = -\alpha \mp j\beta$ Where: $0 < \beta < 2\pi F_{\max}$ $\alpha = \beta/100$	$\mapsto$	$z = \left(e^{-\alpha T_s}\right) e^{\mp j\left(\frac{\beta}{\omega_s}\right)2\pi} = \rho e^{\mp j\theta}$ Thus: $0 < \lvert\rho\rvert < 1$, $\theta \in \left[-\left(\frac{F_{\max}}{F_s}\right)\pi, 0\right] \cup \left(0, \left(\frac{F_{\max}}{F_s}\right)\pi\right]$

As it is in frequency-domain vector fitting algorithm, the initial poles can be chosen between logarithmically spaced real poles and linearly spaced complex poles within the "frequency range of interest". It is experienced that the frequency spectrum wherein the data has been measured is a proper choice for the "frequency range of interest".

This fact gives rise to the conclusion that the initial complex poles in z-plane will not be spread all over the unit disk along the 2π angle. It will be declared in continue.

2) **Directly in z-plane as** $\{P_{\text{initial}}\} = \rho_{\text{initial}} e^{\mp j\theta_0}$

The suggested value for radius can be $\rho_{initial} \approx 0.95$. However, the curve over which the initial poles are selected should be limited within proper angle such that the initial poles still lie along the range of the frequency of interest.

If the initial complex poles vary on a curve with radius of $\rho_{initial}$ limited to the $[-\theta_0, 0) \cup (0, \theta_0]$, θ_0 can be obtained as:

$$\theta_0 = \underset{\substack{\text{in the range}\\ \text{of interest}}}{\omega_{\text{max freqency}}} \times T_{\text{sampling}} = 2\pi \underset{\substack{\text{in the range}\\ \text{of interest}}}{F_{\text{max freqency}}} \times \left(\frac{1}{(2n) \underset{\substack{\text{in the range}\\ \text{of interest}}}{F_{\text{max freqency}}}} \right) = \frac{\pi}{n}$$

Consequently,

$$Z_{\text{Complex initial Poles}} = \rho_{\text{initial}} e^{\pm j\theta}$$
$$\theta \in \left(-\frac{\pi}{n}, 0\right) \cup \left(0, \frac{\pi}{n}\right)$$
$$\rho_{\text{initial}} = 0.95$$

The absolute value of the phase for the complex conjugate pairs is uniformly distributed within the range of $\left(0, \frac{\pi}{n}\right)$.

7.3 Convergence Issues

Based on the experimental results, the vector fitting process as an iterative algorithm has favorable convergence properties. However, there is no theoretical justification on this fact or even conditions to ensure the convergence.

This section presents some experimental measures to judge when the process is converged.

- As the iterative algorithm proceeds, the cost function (error estimator) is minimized. In the mean time, iteration by iteration, the relocated poles tend to their final positions for which the model fits the data with minimum error.
- When the value of the cost function consistently drops below a priori error bound, it indicates a sufficiently accurate fitting has been achieved then process can be stopped.
- The maximum difference between the real and imaginary parts of the poles in two consecutive iterations should fall below a priori threshold. This threshold level is decided according to the frequency-scaling factor explained in the previous chapter.
- Since the poles from *i*-th iteration happens to be very close to the poles from *(i-1)*-th iteration; the ratio between modulus of denominator in these two consecutive steps tends to one.

$$\frac{D^{(i)}(j\omega_k)}{D^{(i-1)}(j\omega_k)} = \frac{\prod_{n=1}^{N}\left(jw_k - p_n^{(i)}\right)}{\prod_{n=1}^{N}\left(jw_k - p_n^{(i-1)}\right)}, \quad \text{where N is the order of the model} \tag{7.47}$$

Then:

$$\left|\frac{D^{(i)}(j\omega_k)}{D^{(i-1)}(j\omega_k)}\right| \to 1 \qquad p_n^{(i-1)} \to p_n^{(i)} \tag{7.48}$$

As the iterations proceed, it asymptotically occurs $\varepsilon''_k = \left(\frac{D^{(i)}(j\omega_k)}{D^{(i-1)}(j\omega_k)} \right) \varepsilon_k \to \varepsilon_k$.

It is understood that when minimizing the weighted SK cost function equivalently the absolute error from nonlinear estimator is minimized. This proves that the parameters evaluated based on minimizing the SK cost function tends to those obtained by minimizing the original nonlinear problem.

The ratio $\sum_{k=1}^{K} \left(\left| \frac{D^{(i)}(j\omega_k)}{D^{(i-1)}(j\omega_k)} \right| \right)$ can be also watched as a measure for occurrence of convergence. Once its average tends to one and becomes close enough to one, this can be an indication for convergences; also a sign for accomplishing an accurate approximation.

7.4 Relocated Poles Refinement Strategy

Recognizing the unstable poles in ZD and outlining the pole refinement strategy to ensure the stability for the poles during iterations is outlined as below.

I. Push (flip) the poles at the outside of the unit circle to the inside:

$$z_{pole} = \rho e^{j\theta} \text{ for } |\rho| > 1 \quad \to \quad z_{\substack{refined \\ Pole}} = \frac{1}{\rho} e^{j\theta} = \frac{1}{z^*_{Pole}};$$

$$z^*_{Pole} = conj(z_{pole})$$

Note: This correction causes the possible unstable pure negative real poles still appear as pure negative real poles on the unit disk

II. Push the poles on the unit circle to the inside:

$$z_{pole} = \rho e^{j\theta};\ \text{where } \rho = 1 \ \rightarrow \ \underset{Pole}{z_{refined}} = \rho_{initial} \times e^{j\theta}$$

A special case is the pure positive real poles on the unit circle:

$$z_{pole} = 1 \ \rightarrow \ \underset{Pole}{z_{refined}} = \rho_{initial}$$

III. Omit the poles at the origin of the z-plane and replace them with proper one:

$$z_{pole} = 0 \rightarrow \underset{Pole}{z_{refined}} = \rho_{initial}$$

IV. Omit pure negative real poles and replaced them with proper one:

$$Z_{Pole} = -\rho \ \rightarrow \ \underset{Pole}{z_{refined}} = \rho$$

V. Cancel out the possible duplicate poles (Poles multiplicity) and replaced it with proper distinct poles.

$$if \left| z_{Pole} - z'_{pole} \right| < Threshold \ \rightarrow \ \underset{Pole}{z'_{refined}} = \rho' e^{j\theta'} \quad ,\theta' = \frac{\sum_{i=1}^{P} \theta_i}{P} \ and \ \rho' = \frac{\sum_{i=1}^{P} \rho_i}{P}$$

VI. Check if all the complex poles have occurred in complex conjugate pair.

7.5 Bounding the Angular Frequencies of the Poles

The final stable complex poles resulted from the ZD-VF algorithm of form in (7.49) are scattered inside the unit circle.

$$\underset{2\nu+1}{P_{2\nu}} = \rho_\nu e^{\mp j\theta}, \ |\rho_\nu| < 1 \text{ and } 0 \le \theta < +\pi \qquad (7.49)$$

This directly means by using ZD-VF, one enforces the equivalent s-domain complex conjugate poles $\left(\begin{matrix} P_{2\nu} \\ 2\nu+1 \end{matrix} = \alpha_\nu \mp j\omega_\nu \right)$ to vary along the $j\omega$ axel just within the primary region. In other words, the radian frequency of the poles would be bounded to the aliasing margin, as: $\omega_\nu < \omega_{\max} = \frac{sampling\ rate\ (\omega_S)}{2}$.

In contrary, the vector fitting process in freq.-domain results in the final poles scattered in the left half plane that can lie anywhere along the $j\omega$ axel from $\omega = -\infty$ to $+\infty$. Constraining the poles to the primary region (frequency bounded) is one of the remarkable characteristic of the z-domain vector fitting process.

7.6 Frequency Scaling

When identifying continuous-time systems in the Laplace-domain, "it is indispensable to scale the frequency axis (and hence also the parameters) to guarantee the numerical stability [17]" of the least square estimator equation. Without scaling, identification in the s-domain often faces the poor conditioning (even rank deficiency) in the equations particularly when the intended transfer function includes the direct coupling (constant) term. At that time, preserving computation precision within identification process is impossible even for modest orders of the transfer function.

At the end, the parameters for the final macromodel should be scaled back to support the original data. This fact causes the recently introduced frequency-domain OVF algorithm to face a rank deficient equation when identifying the final model by orthonormal bases. In contrary, for the proposed ZD-OVF, in which the continuous time frequency-domain

data is converted to the z-domain, scaling the frequency is not required. ZD-OVF can work with both scaled and un-scaled frequencies and results in the same numerical quality of system equations and the same resulting model.

CHAPTER 8. Computational Results

In this chapter, two example data are modeled utilizing the proposed technique and the results are compared with those from the previous methods (FD-VF, FD-OVF, and ZD-VF). Meanwhile, the performance of the employed techniques in inducing the accurate macromodel are demonstrated and judged. The numerical sensitivity of the system equations to the choice of the initial poles is also studied. This investigation is performed, by trying the same application examples with all the above techniques in the identical condition. Afterward, the results from ZD-OVF are compared with those from above three methods. Moreover, The effect of linear least square (LLS) solution methods on the resulting model from the proposed method is examined.

8.1 Preliminaries

First, frequency-domain data (e.g. tabulated scattering parameters) is converted to the z-domain as given in sections 2.5.2 and 2.5.3 in chapter 2. Resulting z-domain data is used as input to the algorithm. To continue, according to the explanation in section 2 in chapter 7, a set of the prescribed real and complex conjugate starting poles is selected. This selection of initial poles can be performed either in s or in z domain. For an educated guess of poles in s-domain, the angular frequencies of the initial poles are uniformly distributed along the frequency of interest and the real parts are properly small in comparison to the imaginary parts. Initial poles in z-domain can be obtained

either by converting this set of optimal initial poles to the z-domain as it is in example one, or it can be selected directly in z-domain as shown in example two.

The intended form for transfer functions (TF), given in section 2.2.4 of chapter 2 is decided. According to the form of TF and type of the basis functions, the Sanathanan-Koerner weighted cost function (SK), as given in (6.13) in section 6.4, is formed and solved. The coefficients c_p and $\hat{c}_p$ is estimated by optimizing the cost function in linear least square sense. To serve the evaluation purposes, the linearized cost function as explained in (6.11) in section 6.3 also, can be tested and compared. In each iteration, using the poles and the resulted coefficients for the denominator $\hat{c}_p$, the minimal state space realization (**A**, **B**, **C**, D) is constructed, as shown in chapter 5. The zeros of the denominator which are the eigenvalues of the real matrix **H=A-BC** will be the starting poles for the next iteration. This sequence of iterations continuous until the convergence is achieved. Once the final poles are obtained, the real coefficients (residues) for the bases are worked out. Final macromodel is constructed using either the orthonormal or the partial fraction bases by solving the following linear least squares problem, where φ_i is either orthonormal or partial fraction. The parameter d is the direct coupling term. If d was included in the initial form for TF and was used to obtain the poles in the vector fitting process, it should be considered nonzero in below.

$$\arg\min \sum_{k=0}^{K} \left| \hat{H}(z_k) - \left(\sum_{i=1}^{P} c_i \varphi_i(z_k) + d \right) \right|^2 \tag{8.1}$$

8.2 Example One

The S parameters[1] data from a 2-port radial stub (symmetric network) is examined. A strictly proper z-domain transfer functions are considered to approximate the model by utilizing the proposed ZD-OVF technique, then the resulting model will be evaluated in the s-domain. The order of the model is 27 when the set of starting poles consists of 1 real pole and 13 complex conjugate pairs. The optimal initial poles were selected in the frequency domain. Then, they were converted to the z-domain and used as starting poles for the first iteration of ZD-OVF, for the comparison purpose.

Figure 8-1 illustrates the location of initial poles inside the unit disk in z-plane.

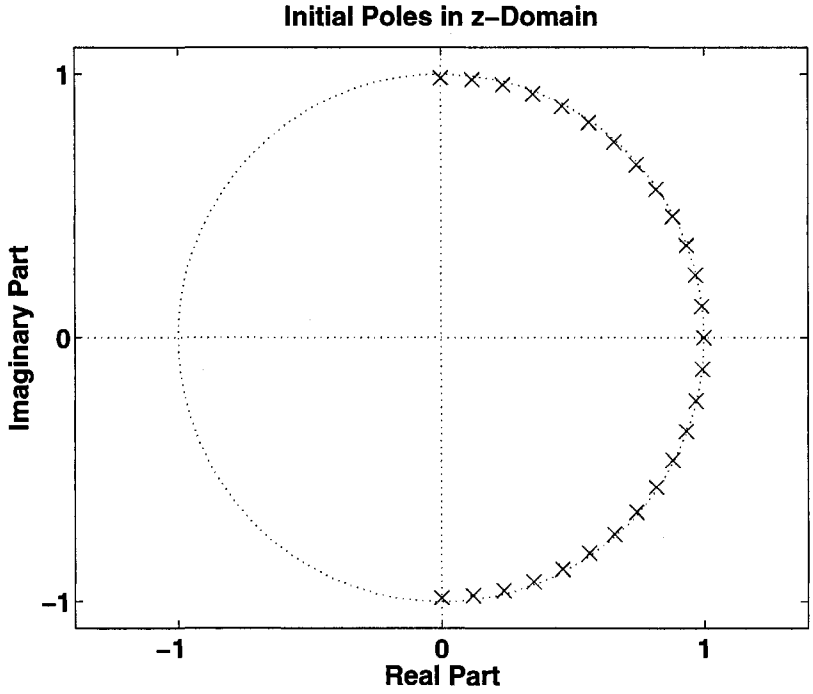


Figure 8-1: Initial pole placement in z-plane: Example one.

The location of the final poles, resulted from ZD-OVF are also, demonstrated in the next graph.

[1] Reflection Coefficients

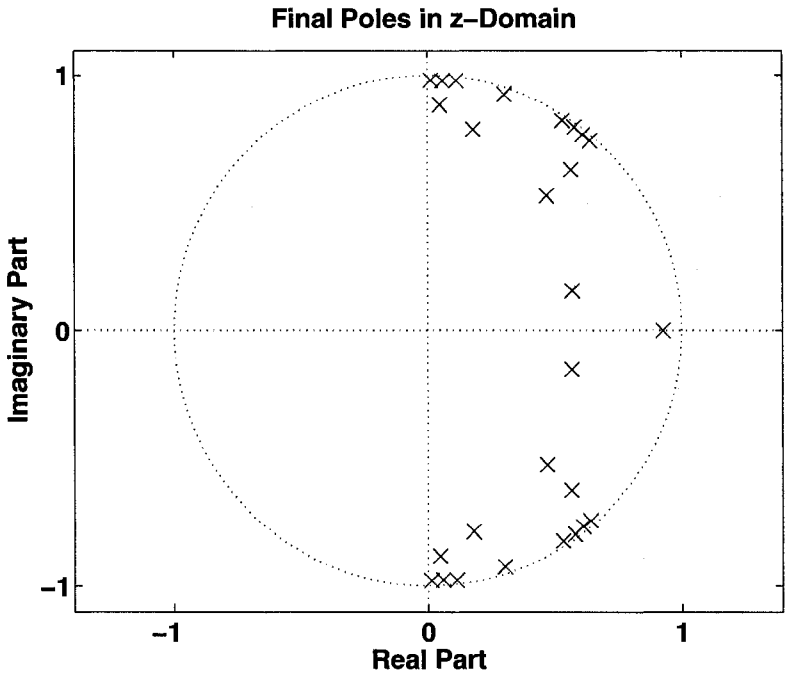


Figure 8-2: Final pole locations in z-plane: Example one.

To continue, the convergence pace in process, when ZD-OVF vector fitting algorithm is used for the data of example one, is presented in the following two plots.

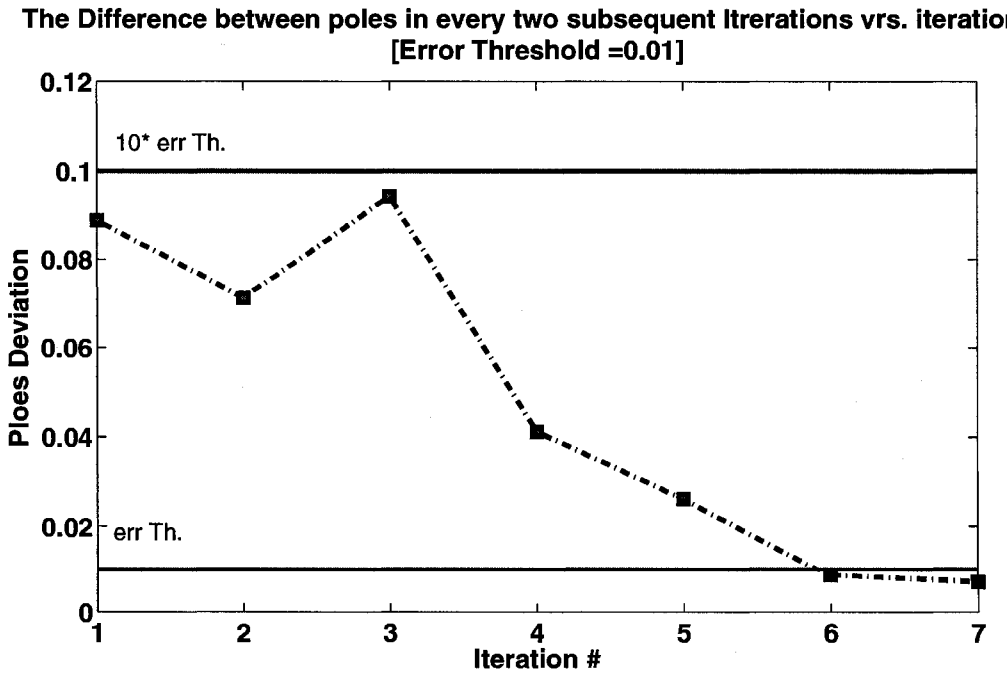


Figure 8-3: In 7 iterations, poles have been converged to their optimal locations: Example one

As shown in the Figure 8-3, at the higher iterations after convergence the poles are tending to their final optimal locations with smaller steps (accurately). This means that the difference between poles resulting from every two consecutive iterations become consistently smaller.

Figure 8-4 in below shows the convergence in SK weighting factor according to the explanation provided in section 7.4. Since in each iteration, the denominator of H(z) is evaluated at every frequency points, as SK weighting factor for the next iteration, thus, evaluation of the ration given in (7.48) is performed with negligible computational cost.

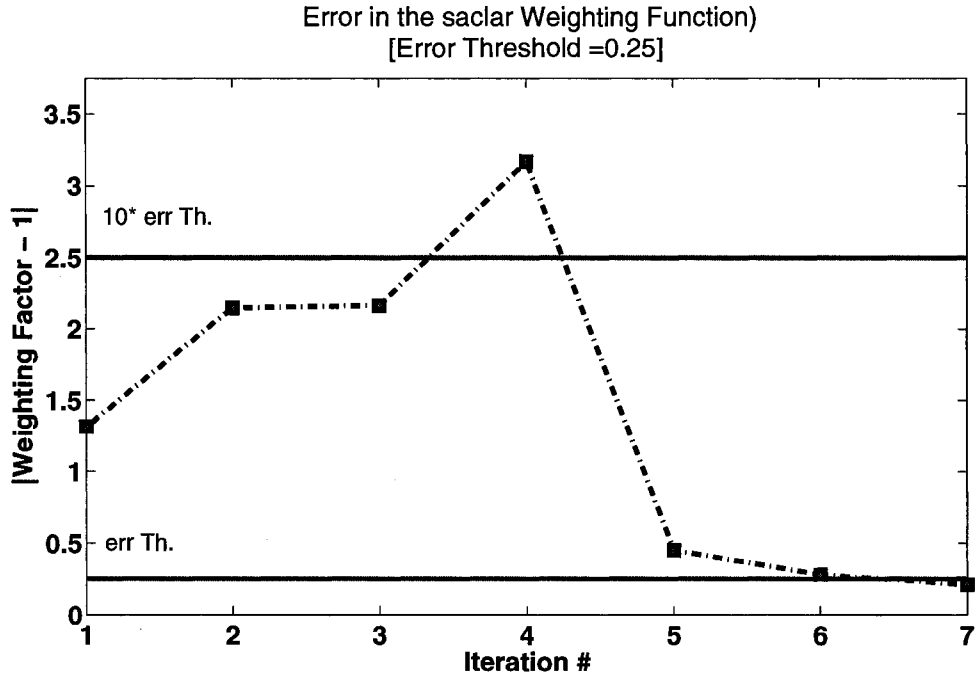


Figure 8-4: When convergence happens the SK weighting factor tends to one: Example one

In the graphs, shown in Figure 8-3 and Figure 8-4 it is seen that the convergence in poles and convergence in the SK weighting factor are confirming each other. Since checking the convergence in SK weighting factor is a CPU efficient task, considering it as a confirming sign for convergence can avoid possible misjudgments in deciding the convergence to stop the vector fitting process.

Figure 8-5 and Figure 8-6 in below compare the real and imaginary parts of the spectral response versus the original data over the frequency range of interest [0Hz -15GHz].

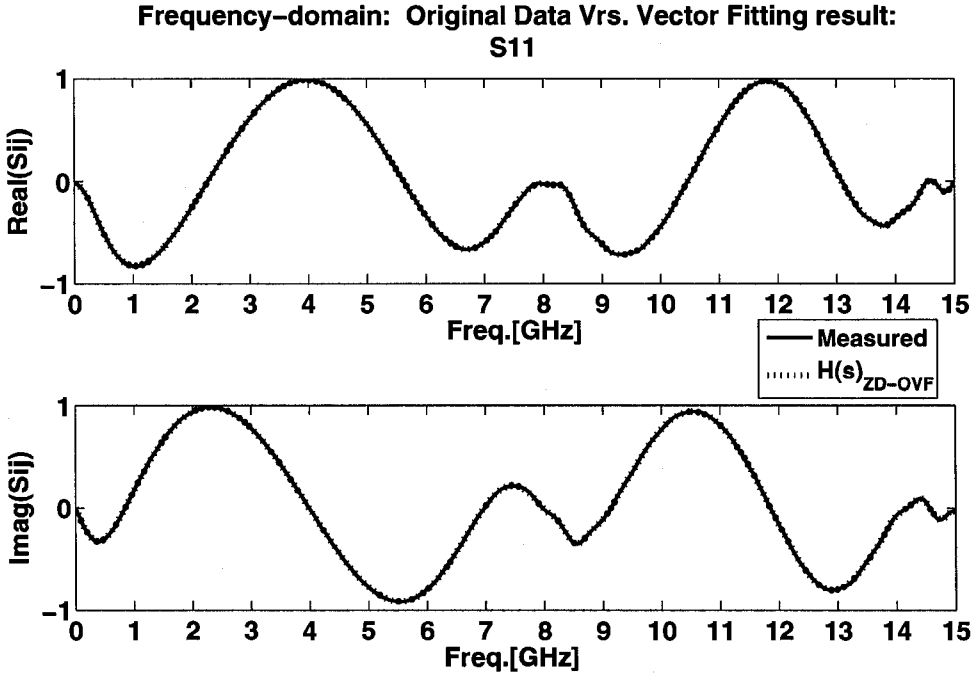


Figure 8-5: Reflection coefficients (S_{11}) of the Radial Stub: Example one

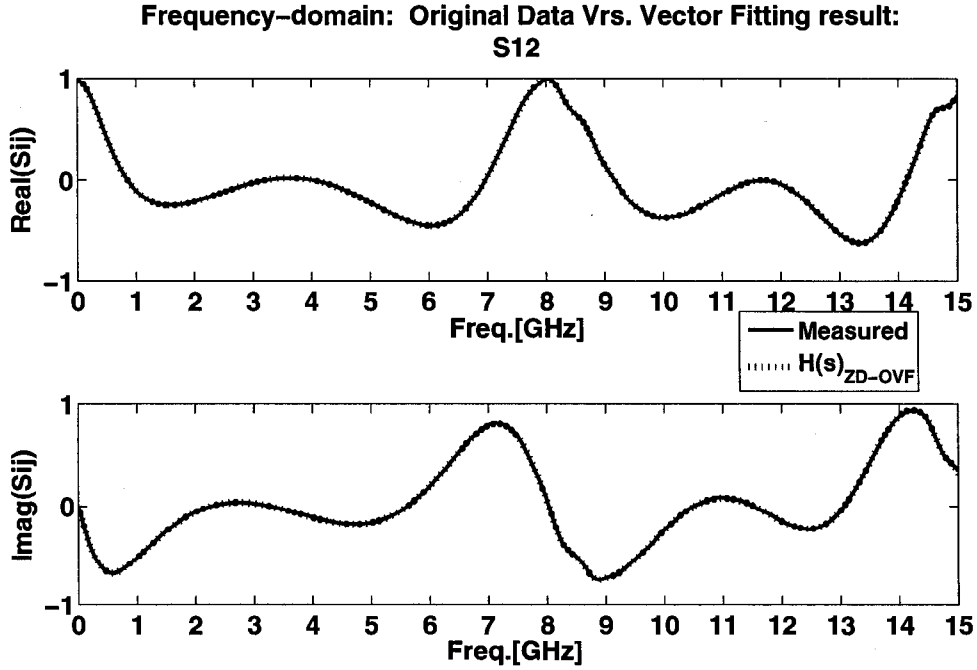


Figure 8-6: Reflection coefficients (S_{12}) of the Radial Stub: Example one

Figure 8-7 shows that a good fitting has been obtained in lower frequency range and the error is almost in the close to the numerical nose level.

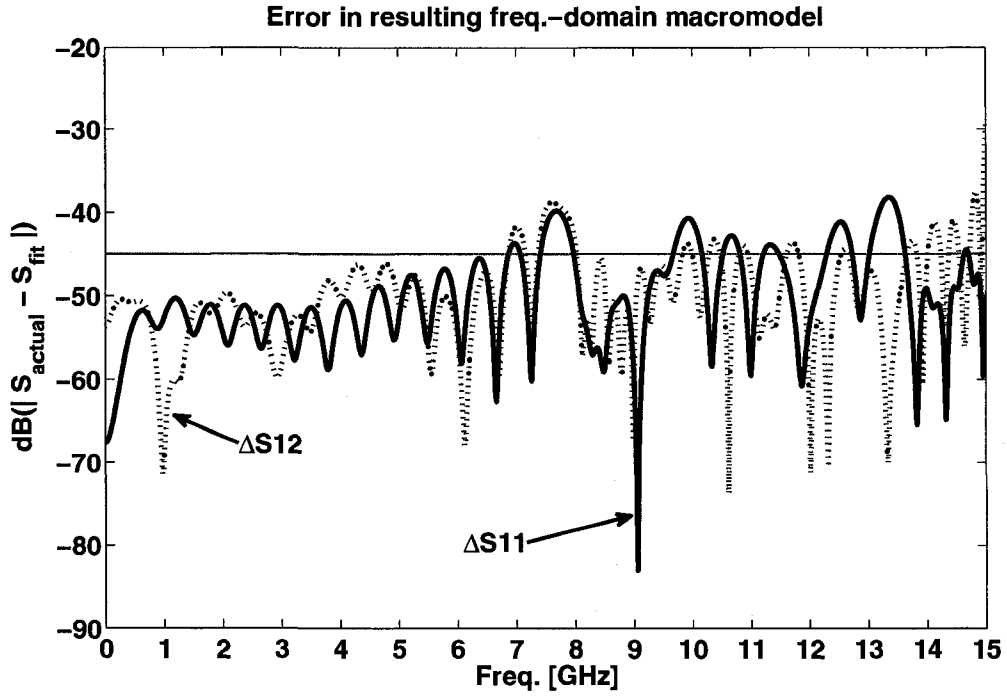


Figure 8-7: Error fitting model: Example one

8.2.1 Examining Four Vector Fitting Algorithms for Example One

To provide an accurate comparison, example one was solved by using four curve-fitting techniques FD-VF, FD-OVF, ZD-VF, and ZD-OVF. In all 4 cases, strictly proper transfer functions have been used to model the data. The same set of 31 optimal poles was used as the starting poles in all four methods.

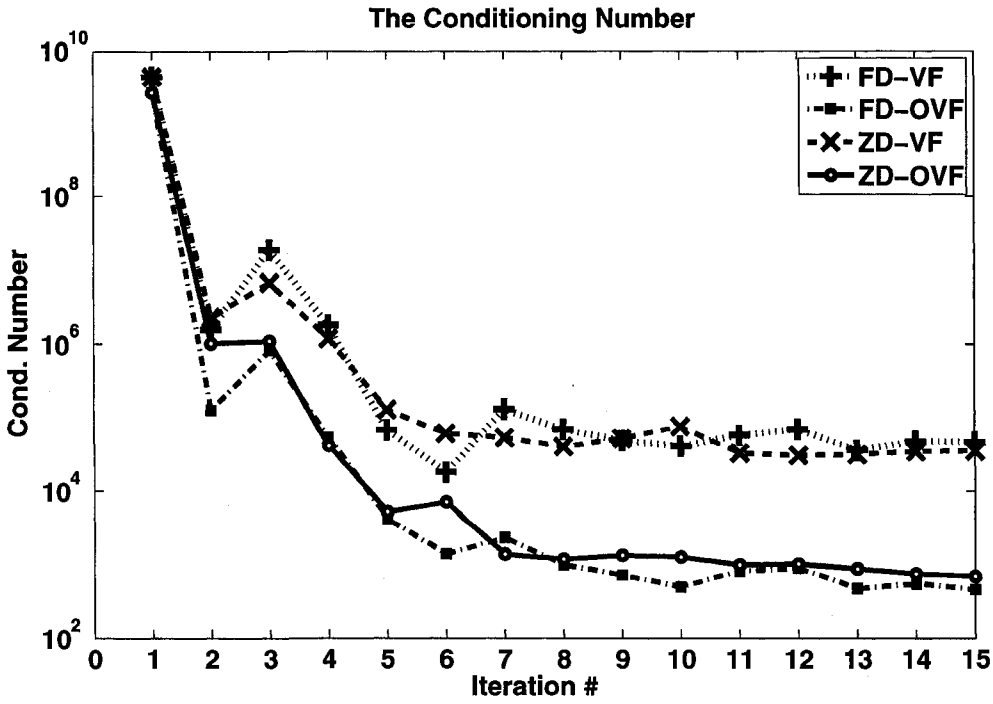


Figure 8-8: Comparison between the condition numbers per iteration for the four pole identification algorithms using the same set of 31 optimal starting poles: Example one

Above figure shows that, the numerical conditioning of the system equations are consistently improving as iterations proceed. When utilizing partial fractions, the equations are of poorer numerical quality in comparison to the orthonormal bases. The condition number for the ZD-OVF and FD-OVF are closely following each other, while theirs are better than the ones from the non-orthogonal VF methods. In other words, the orthonormal bases improve the numerical stability of the methods.

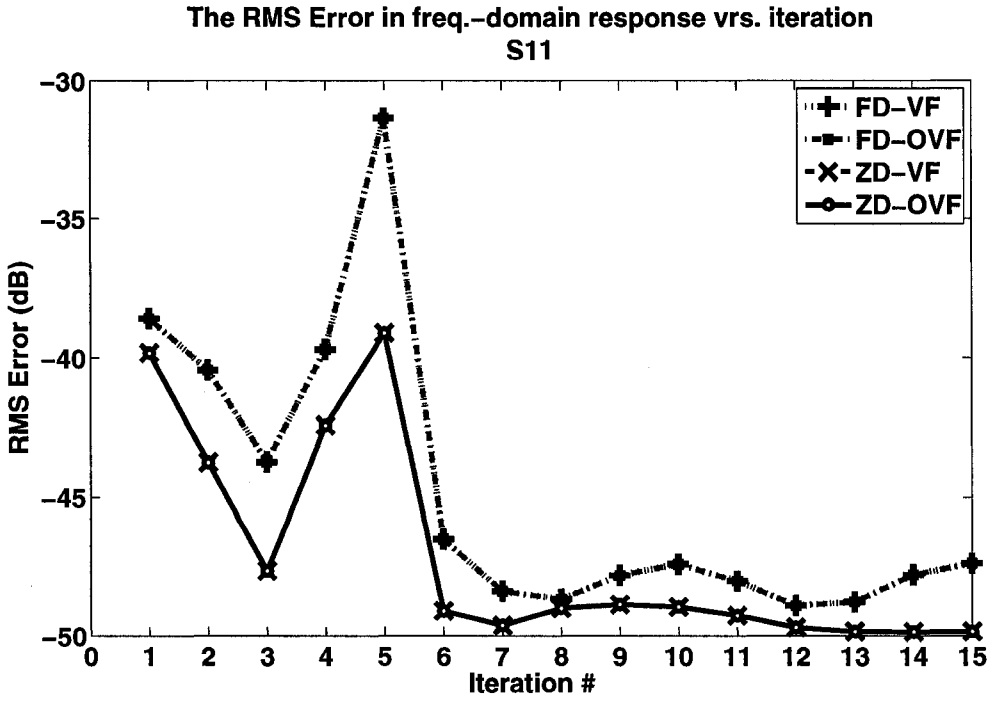


Figure 8-9: RMS error in (S_{11}) resulting from different VF algorithms with SK weighting factor per iteration: Example one

Figure 8-9 shows the descending pace of the RMS error in all 4 algorithms. Throughout the iterations, the z-domain vector fitting techniques provide results that are more accurate. In addition, z-domain methods in higher iterations show consistent behavior and stabilize the RMS error in resulting model in a decreasing pace.

Moreover, the results from the vector-fitting methods with orthonormal and partial fraction bases in each domain are closely following each other. Precisely writing, in each domain the both methods converge to the almost same final model, as it is shown in below:

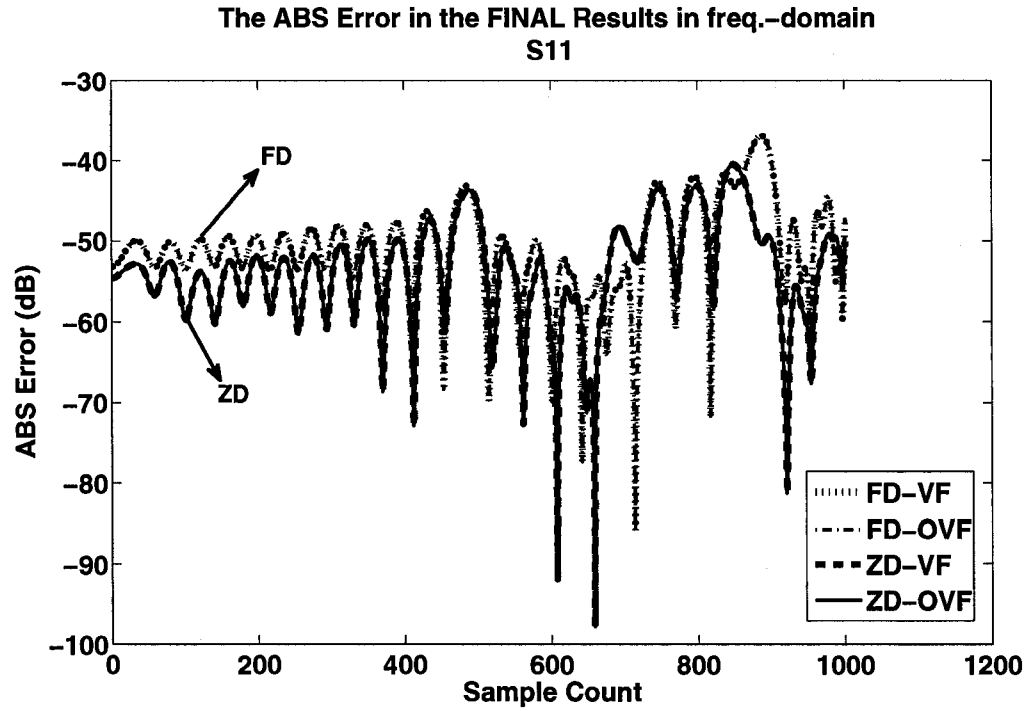


Figure 8-10: The absolute errors in the S_{11} responses induced from final models: Example one

It is also explicitly inferred by Figure 8-10 that:

- The final models obtained from ZD vector fitting techniques represents the behavior of the original network better with less absolute error along the spectrum (except for a few frequency samples in the mid-spectra) .
- The almost same results have been produced by two methods in each domain.

Based on many observations, it can be judged that:

When the pole-relocation algorithms start from the optimal initial poles, and it does not tackle a burden of solving the severely ill-conditioned system equations, almost the same poles and consequently the same level of accuracy are expected from VF and OVF.

For this specific example, the fact is investigated in the following plots.

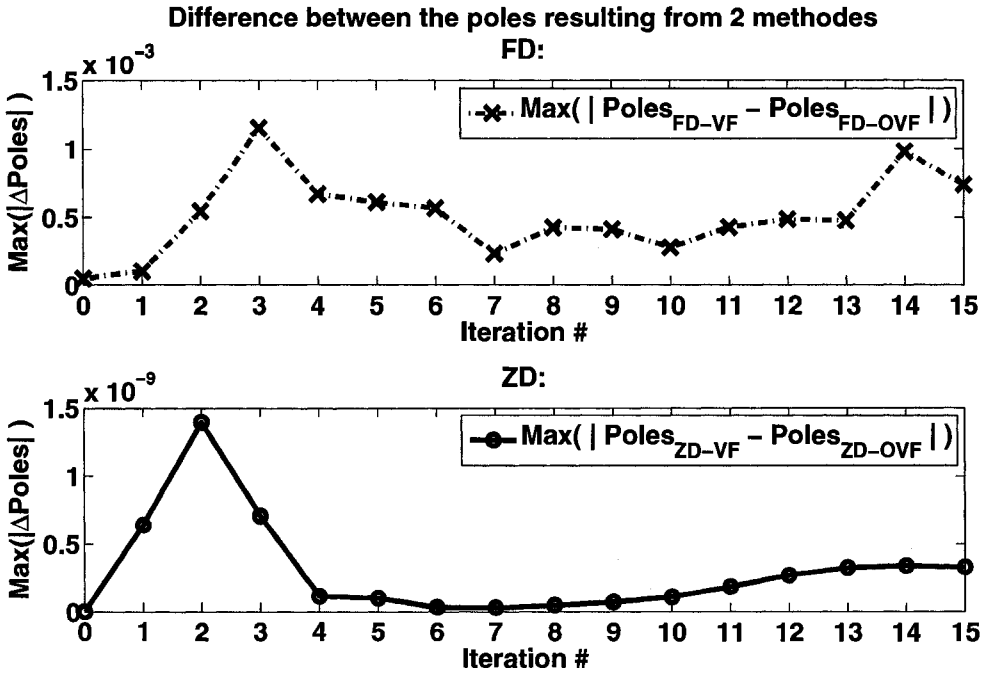


Figure 8-11: The maximum modulus of the displacement vectors between poles: Example one

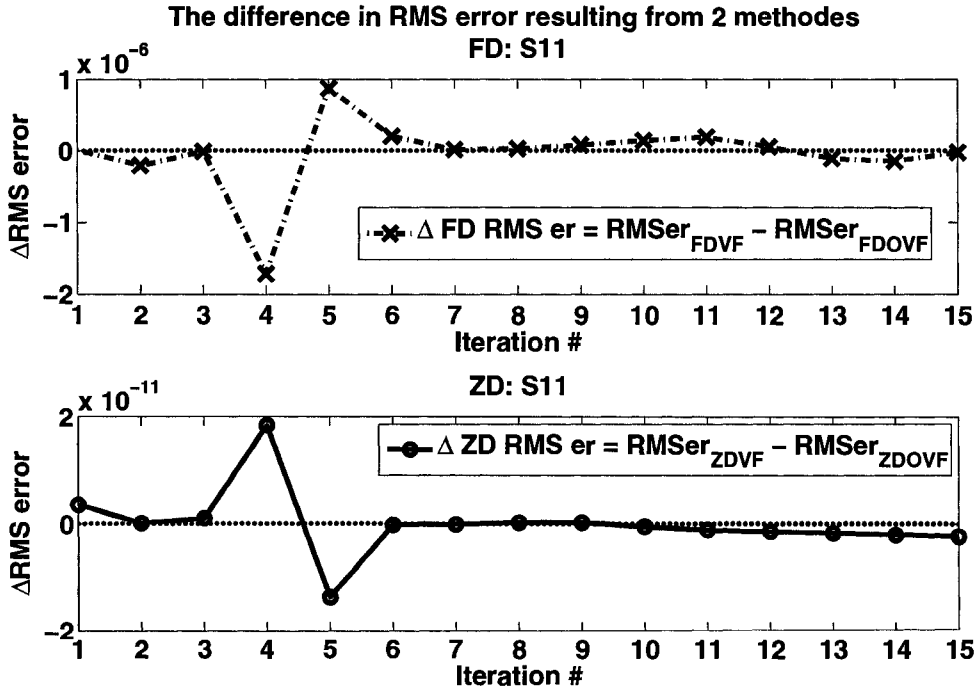


Figure 8-12: Shows that the models induced in each step are noticeably close and it is so for the final models: Example one

A sample of execution times for algorithms when handling the data from a 2-port network in example one are compared in the following table. The order of models is 31 and formulations are based on the SK method. It has examined in a machine with 1.6GHz Intel dual core processor without utilizing the parallel processing techniques.

Table 8-1: Comparison of the sample execution times: Example one

Algorithm →	FD-VF	FD-OVF	ZD-VF	ZD-OVF
Ave. elapsed time / itr. [sec.]	1.24	1.26	1.24	1.26

8.2.2 Effect of SK Weighting Function

Each one of the four processes has been repeated with and without the SK weighting factor. The effect of utilizing the SK cost function for the example one can be compared and judged within the following graphs.

In Figure 8-13 and Figure 8-14, it is seen:

- **For initial iterations:** using the SK weighting factor degrades the numerical quality of the equations.
- **Close to the convergence:** formulation based on SK weighting factor improved the conditioning number of the equations when process is converging. Then, it leads a better LLS problem to be solved.
- **Within the iterations after the convergence:** Using the SK formulation does not noticeably improve the condition number of LLS equations. This is the point that was theoretically expected.

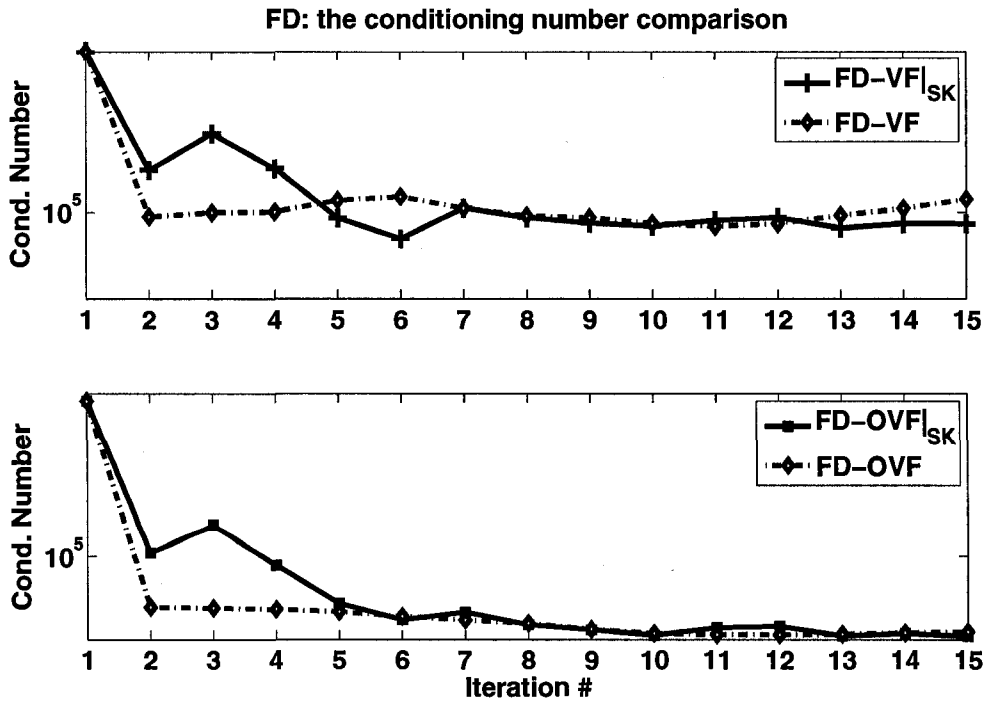


Figure 8-13: Comparison between numerical quality of the resulting system equations in frequency-domain techniques: Example one

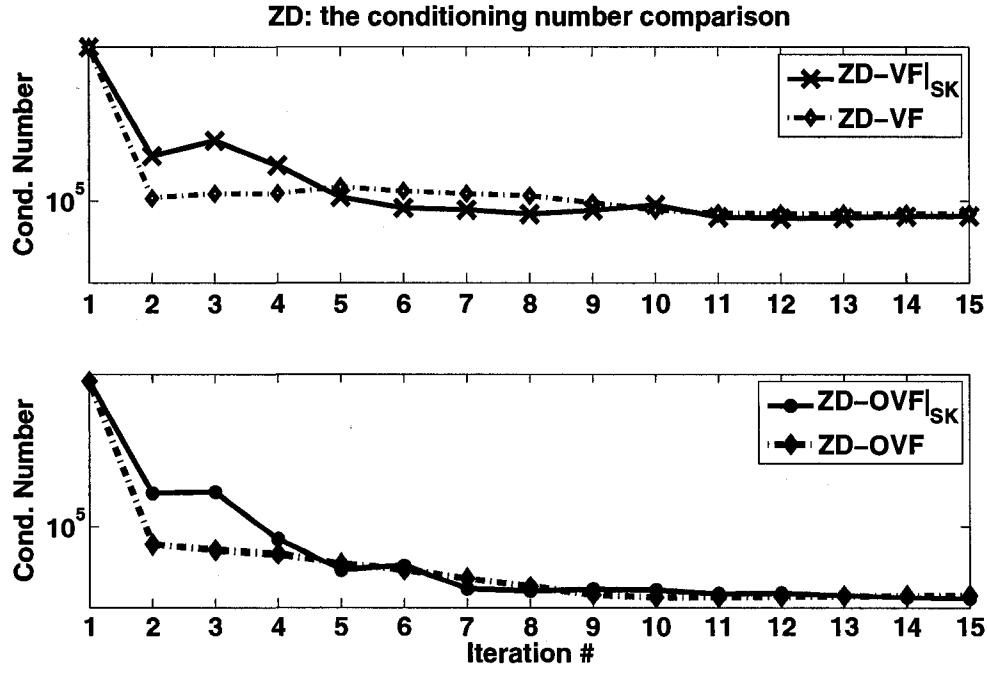


Figure 8-14: Comparison between numerical quality of the resulting system equations in discrete-domain techniques: Example one

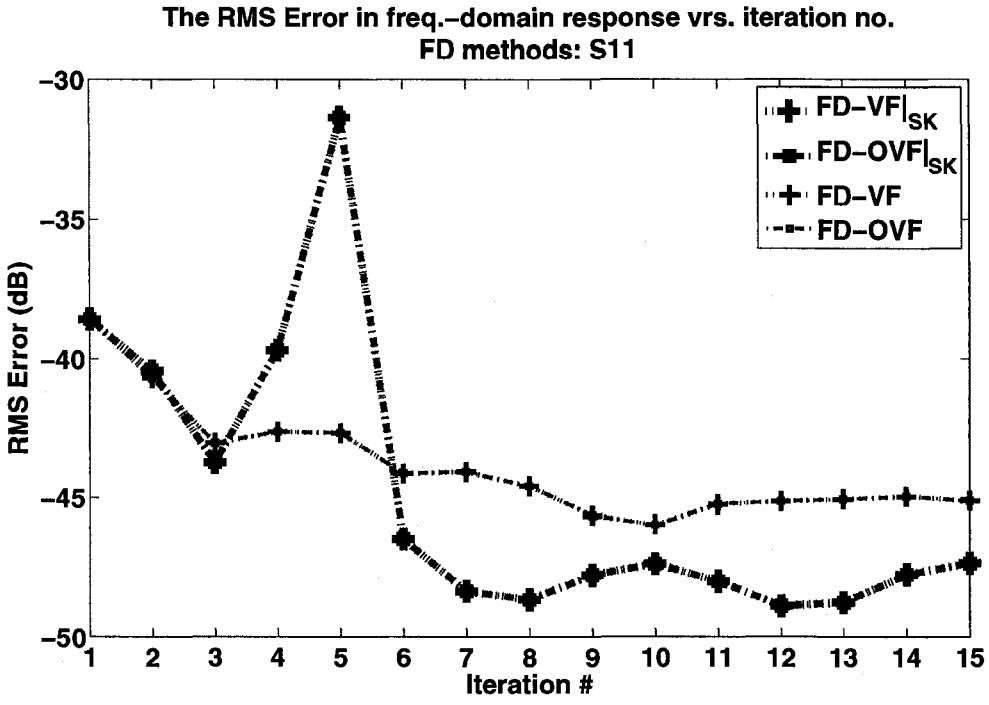


Figure 8-15: Comparison of the RMS errors per iteration for S_{11} resulting from FD algorithms: Example one

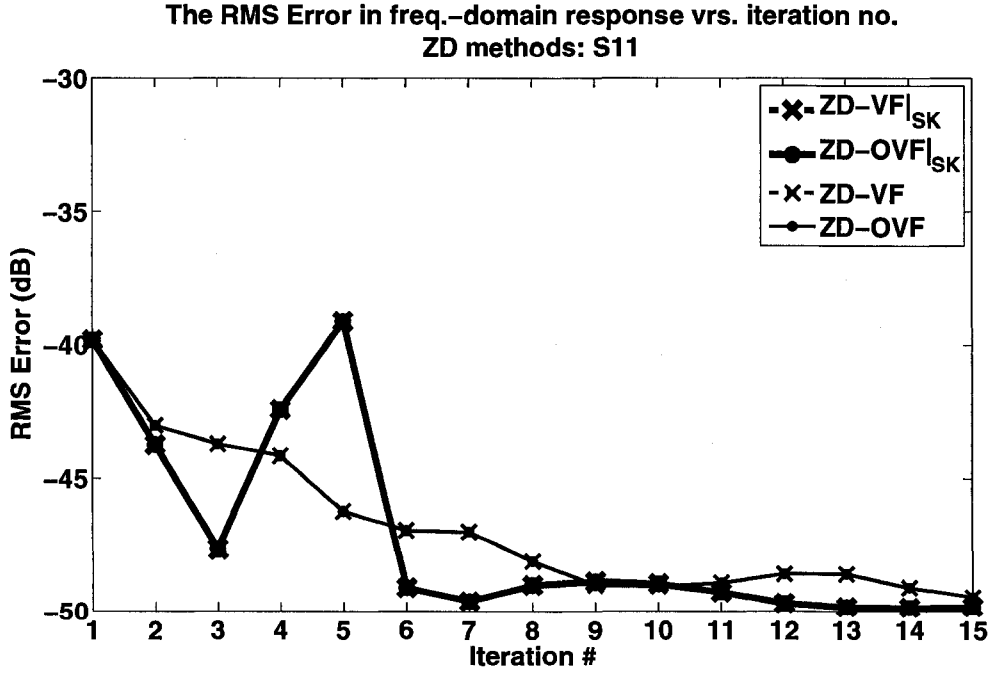


Figure 8-16: Comparison between RMS errors per iteration for S_{11} resulting from ZD algorithms: Example one

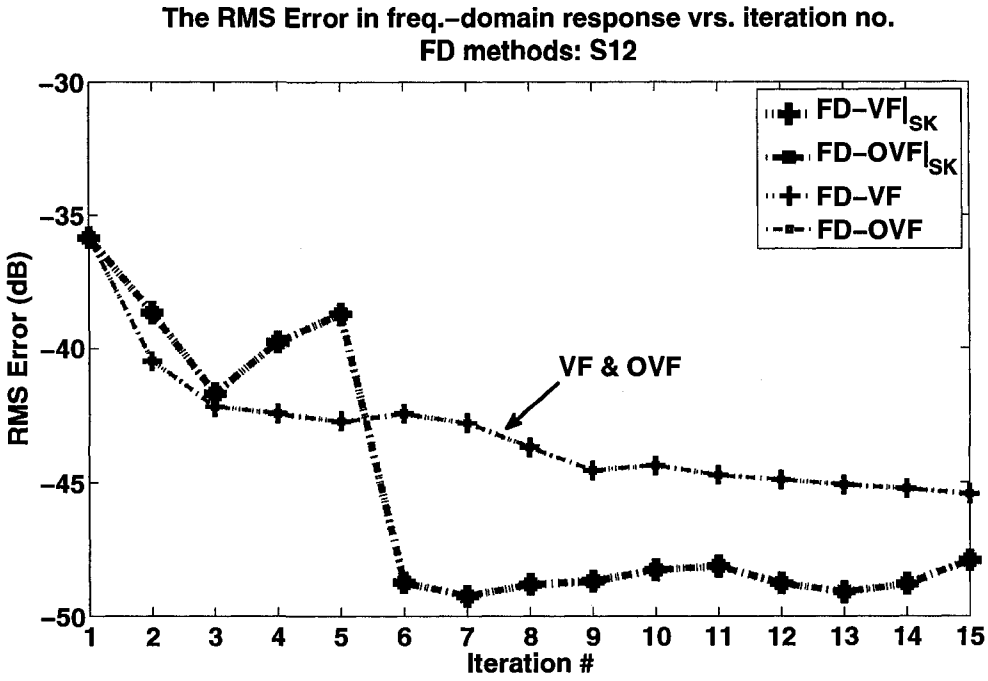


Figure 8-17: Comparison of the RMS errors per iteration for S_{12} resulting from FD algorithms: Example one

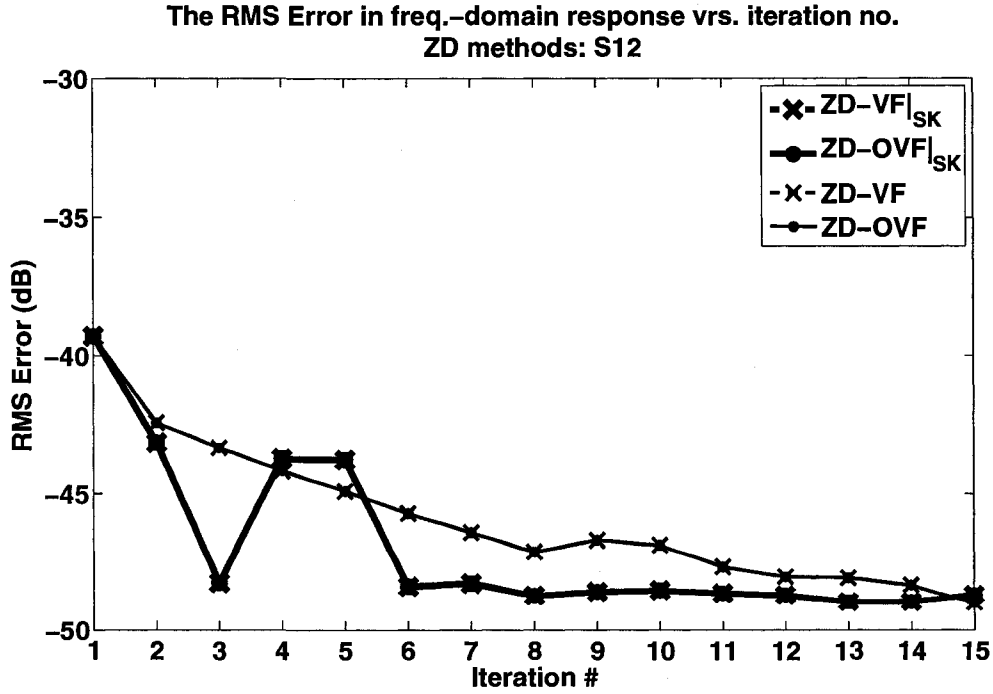


Figure 8-18: Comparison between RMS errors per iteration for S_{12} resulting from ZD algorithms: Example one

Table 8-2: Summarizing the effect of SK method on the RMS error for example one

Comparison	Initial Iterations	close to convergence	Iterations after convergence	Far after the convergence
FD-VF$\|_{SK}$ vs. FD-VF	–	+	+	+
FD-OVF$\|_{SK}$ vs. FD-OVF	–	+	+	+
ZD-VF$\|_{SK}$ vs. ZD-VF	$\cong$	+	+	+
ZD-OVF$\|_{SK}$ vs. ZD-OVF	$\cong$	+	$\approx$+	+
Effect on in the numerical condition of the matrix in system equations: + : Improvement (better) – : Degradation (worse) $\cong$: Almost equal				

According to the above table, which is comparing RMS errors within the processes, it is concluded:

- ZD techniques provide better and more accurate results in comparison to their FD counterpart methods, with and without the SK factor. Only for the S_{12}, **FD-OVF$|_{SK}$** and **ZD-OVF$|_{SK}$** follow each other closely at the iterations after convergence.
- In general, using the SK factor improves the accuracy of the results.
- SK improves the quality of the FD methods noticeably such that for S_{12} by using SK factor FD technique can approach to the accuracy level of ZD methods. For the ZD it meliorates too but with smaller effect.

8.3 Example Two

The scattering-parameters of a 2-port circular resonator structure, measured in the frequency range of 1Hz – 12GHz, is used as the second example to demonstrate the accuracy and fast convergence of the ZD-OVF and comparing it with previous methods. Proper form of z-domain transfer functions including constant terms are intended to fit the data, converted from s-domain. The number of 2 real and 80 complex conjugate starting poles have been selected directly in z-domain as:

$$\{P_{\text{initial}}\} = \rho_{\text{initial}} e^{\mp j\theta_0}$$

wherein $\rho_{initial} = 0.95$ and θ_0 is the angle in the equally spaced scale within $0 \le \theta_0 \le \frac{\pi}{2}$.

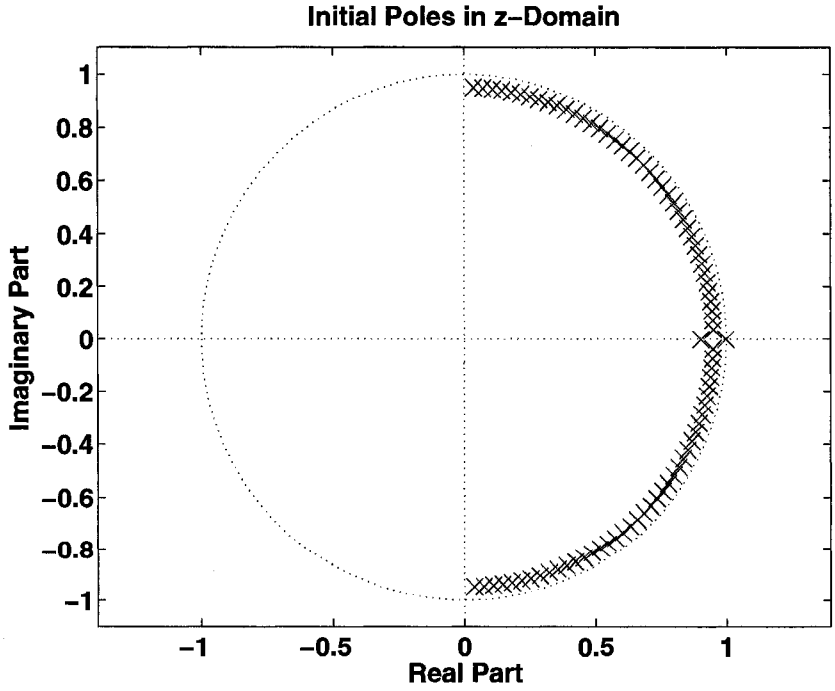


Figure 8-19: Initial pole placement in z-plane: Example two

By utilizing the proposed ZD-OVF technique, fitting process has been managed in z-domain to obtain an accurate model, based on which s-domain results will be evaluated. When the iterative pole-relocation process converged, the final refined poles in z-plane would resemble:

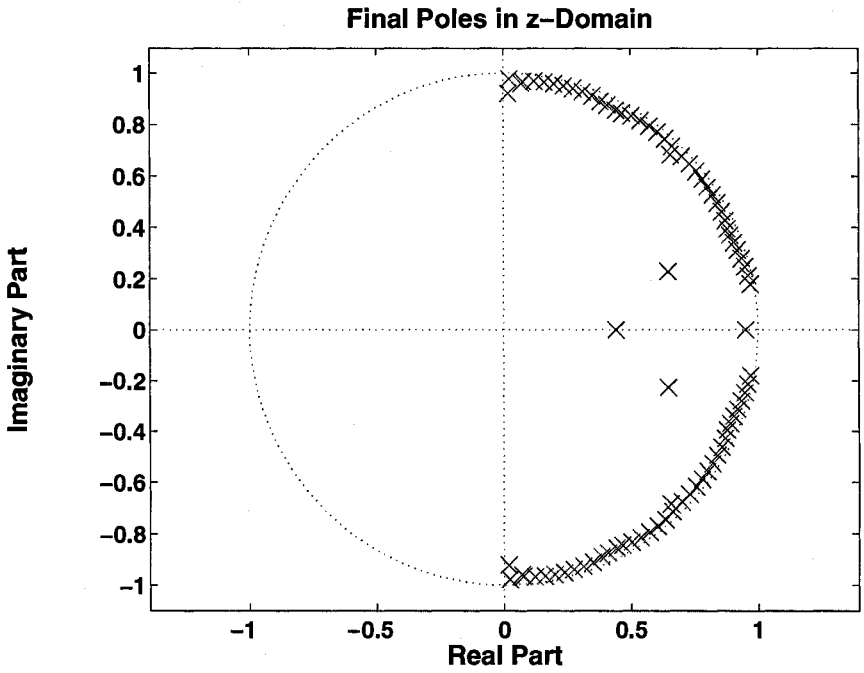


Figure 8-20: Plot of the final pole location in z-plane for example two.

The plot showing the poles convergence pace is illustrated in below:

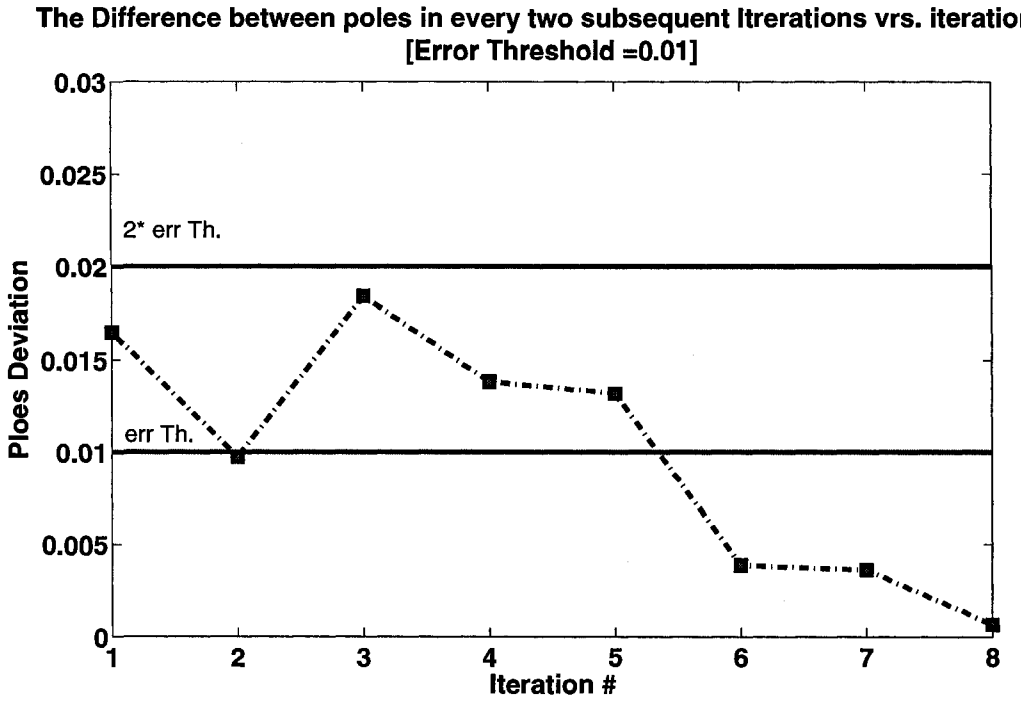


Figure 8-21: After 5 iterations poles have been consistently converged to their final optimal locations: Example two

Although at the second iteration the maximum modulus of 'poles replacement vectors' drops below the intended threshold l, algorithm does not stop. This is because of the fact that, merely when the poles consistently tend to the final optimal location the weighting factor criteria will be satisfied. This fact conforms to the idea that, examining the modulus of 'poles replacement vectors' is not enough (criterion) to decide the convergence. Thus, considering the SK scalar-weighting factor, as shown below, can be instrumental in recognizing the occurrence of convergence. Convergence in this factor is a confirming sign for the convergence of vector fitting process, which can be checked with a negligible CPU cost. It can prevent possible misjudgments of convergence to ensure the accuracy of the resulting model. The following graph presents the weighting factors convergence pace within steps of the process.

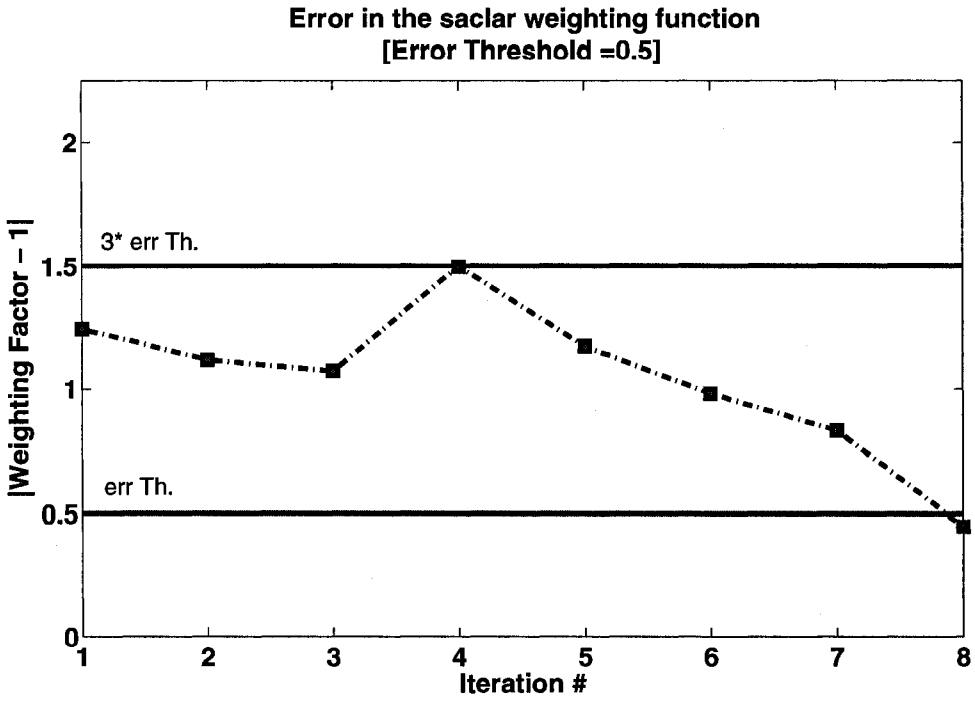


Figure 8-22: Convergence pace of the weighting factor in SK cost function: Example two

Following figures show the real and imaginary part of resulting s_{11} and s_{12} and compare them with experimental data on the same graph.

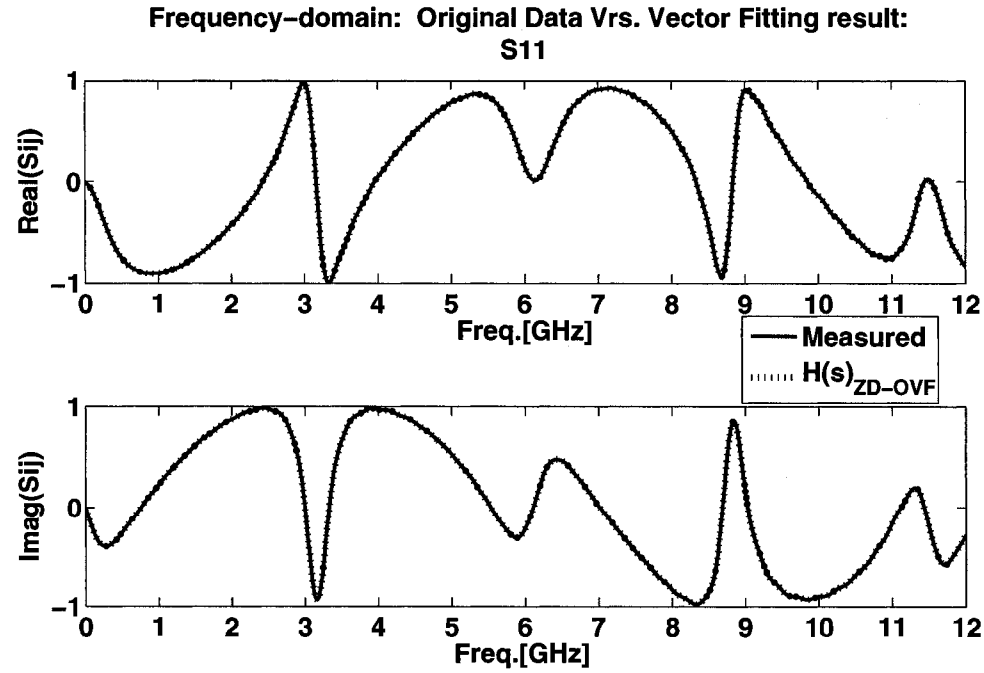


Figure 8-23: Reflection coefficients (S_{11}) of the circular resonator: Example two

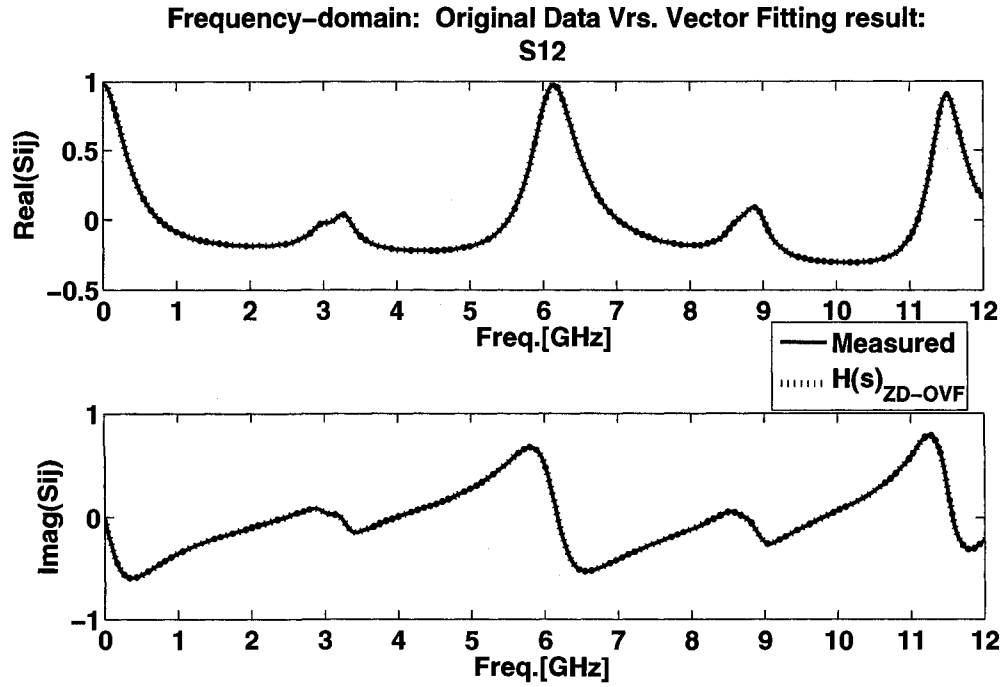


Figure 8-24: Reflection coefficients (S_{12}) of the circular resonator: Example two

The following figures compare the magnitude and phase of the spectral response with the original data over the frequency range if interest [1Hz -12GHz].

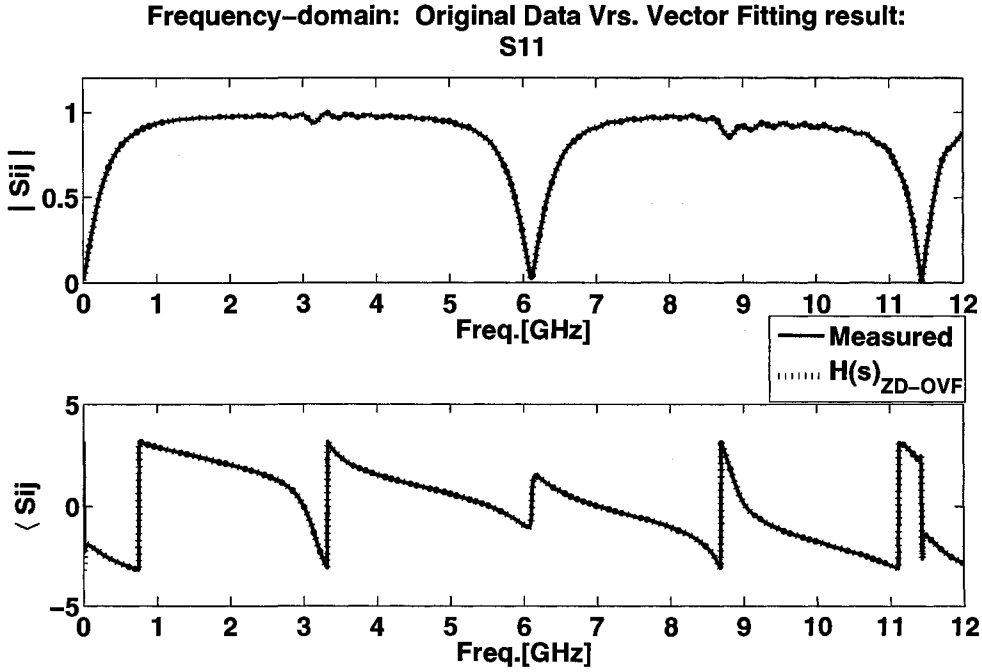


Figure 8-25: The magnitude and phase of the spectral response for S_{11}: Example two

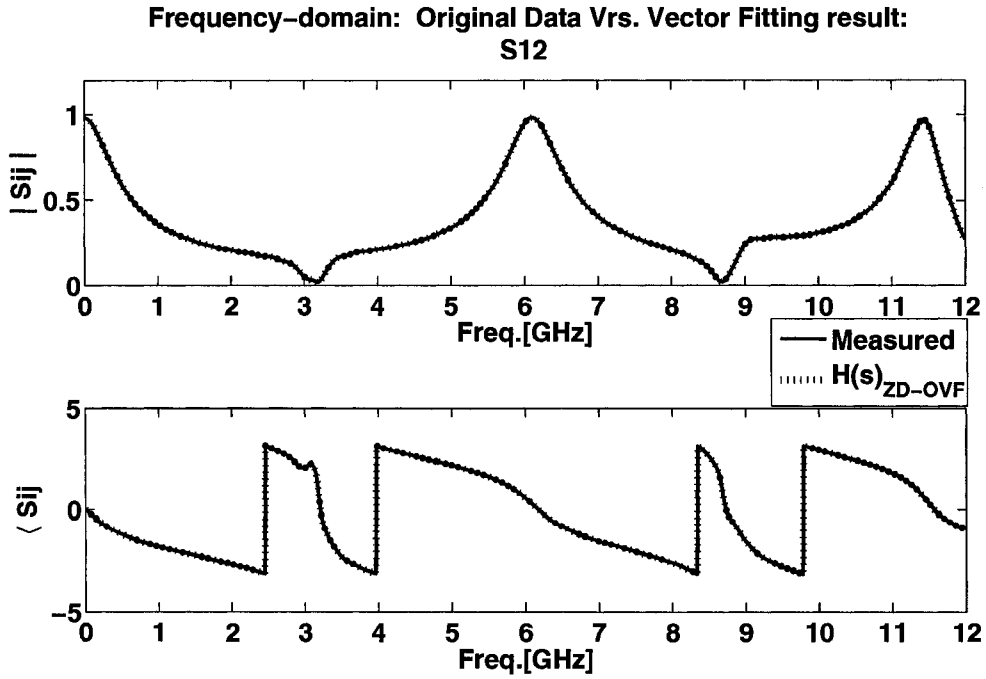


Figure 8-26: The magnitude and phase of the spectral response for S_{12}: Example two

Figure 8-27 (below) shows that, an accurate fitting has been achieved. The absolute error is bounded to the -60dB and for S11 (and S22) it is in the range of the numerical noise within the most part of the spectrum.

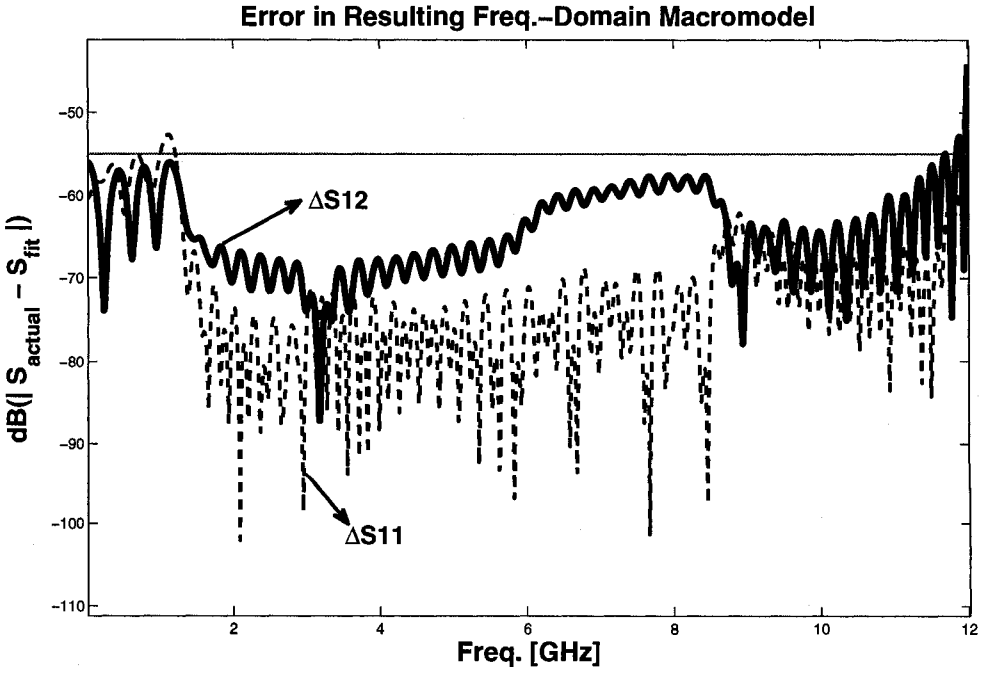


Figure 8-27: Error fitting model: Example two

8.3.1 Examining Four Vector Fitting Algorithms for Example Two

To serve the comparison purpose, the four vector fitting methods, FD-VF, FD-OVF, ZD-VF, and ZD-OVF are applied on the data in example two. In all cases, strictly proper transfer functions have been used to model the data. The set of 82 optimal poles has been chosen in s-domain primarily and the same set of poles after conversion is used as starting poles for all four methods. The following graphs present the comparison between the performances.

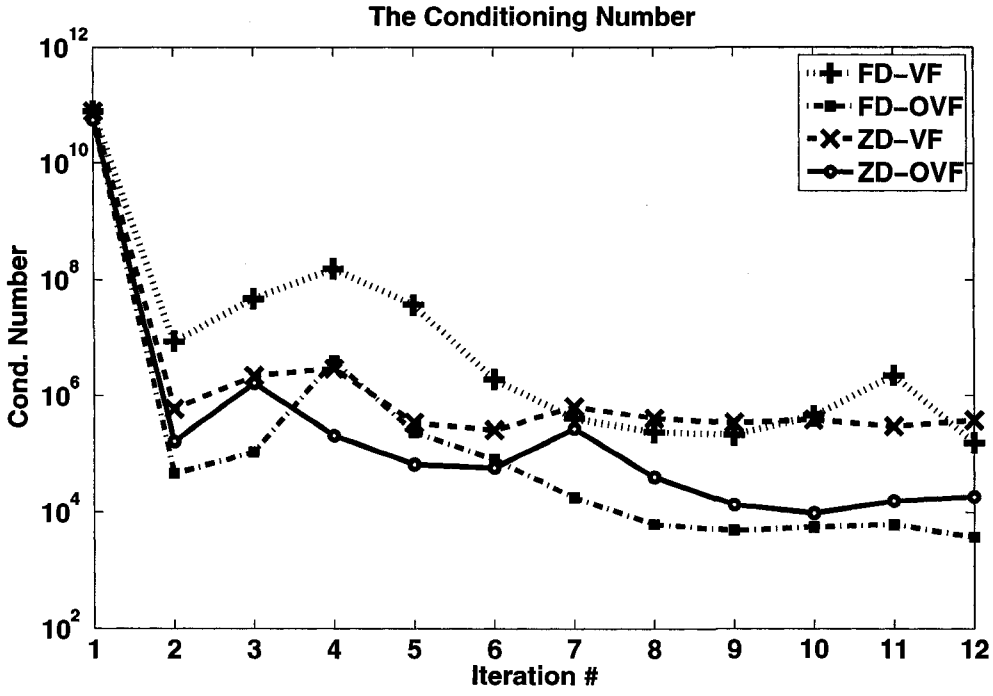


Figure 8-28: Comparison between the condition numbers per iteration using the optimal starting poles: Example two

In above figure, the fact that, utilizing the orthonormal bases enhances the numerical stability of vector fitting is verified. The condition number for the ZD-OVF and FD-OVF are following each other closely. After initial iterations, FD-OVF provides numerically better-conditioned system equations. However, they are close, in the sense

that, the conditioning number for ZD-OVF is in the range of 10^4 in comparison to the 10^3 for FD-OVF.

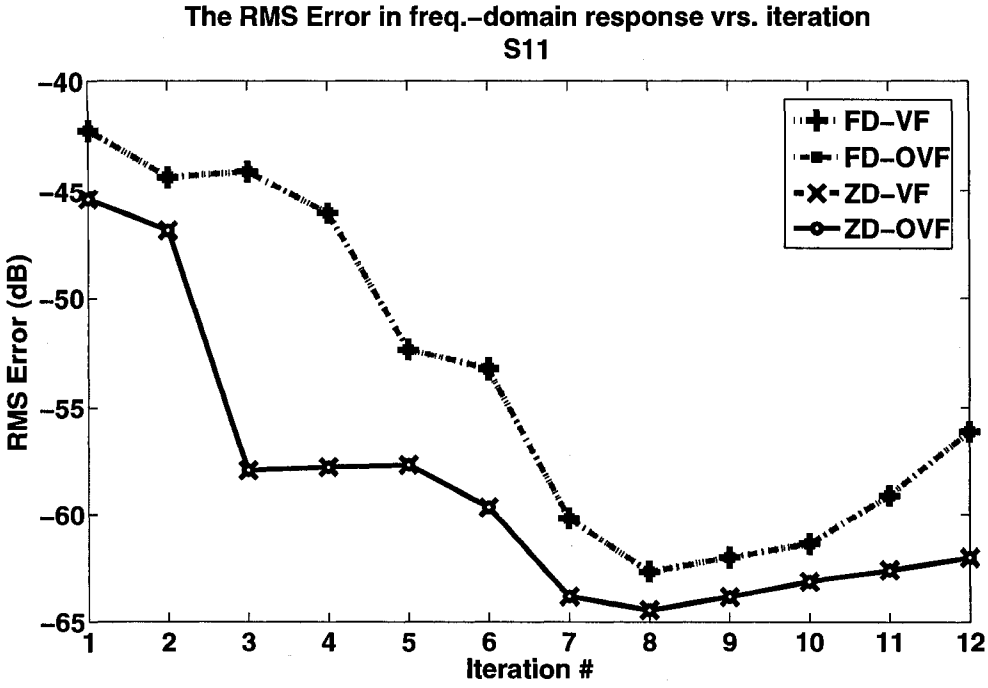


Figure 8-29: Per iteration RMS error in resulting S_{11} from 4 algorithms: Example two

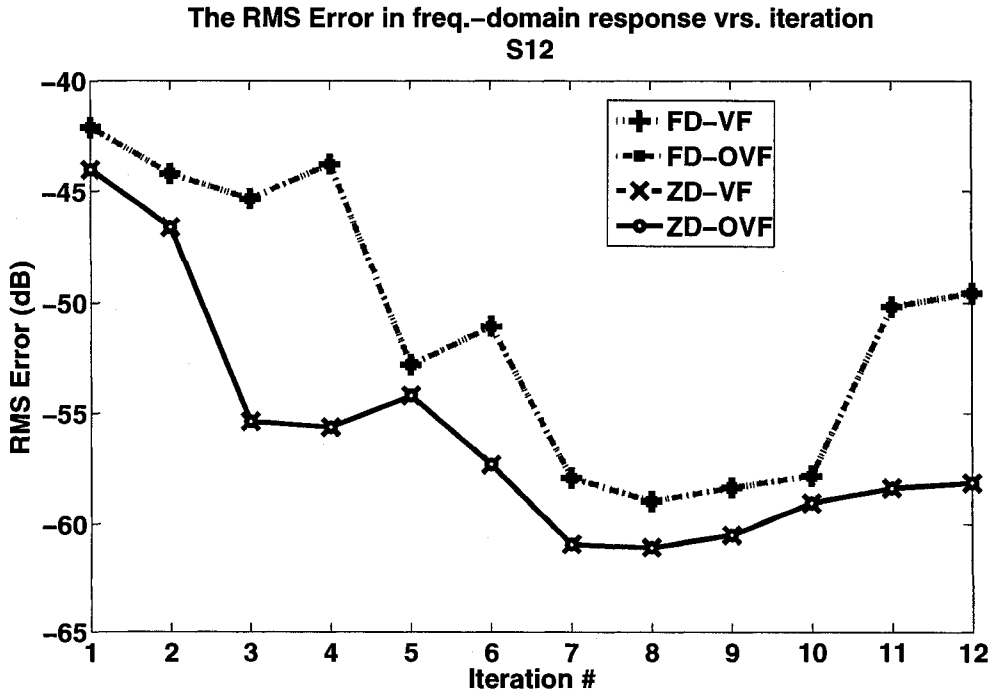


Figure 8-30: Per iteration RMS error in resulting S_{12} from 4 algorithms: Example two

Figure 8-29 and Figure 8-30 show that the z-domain pole-identification techniques provide results that are more accurate. Moreover, when the optimal starting poles are used, results from the vector-fitting methods with orthonormal and partial fraction bases in each domain are closely following each other. Consequently, both methods converge to almost same result, as shown in below.

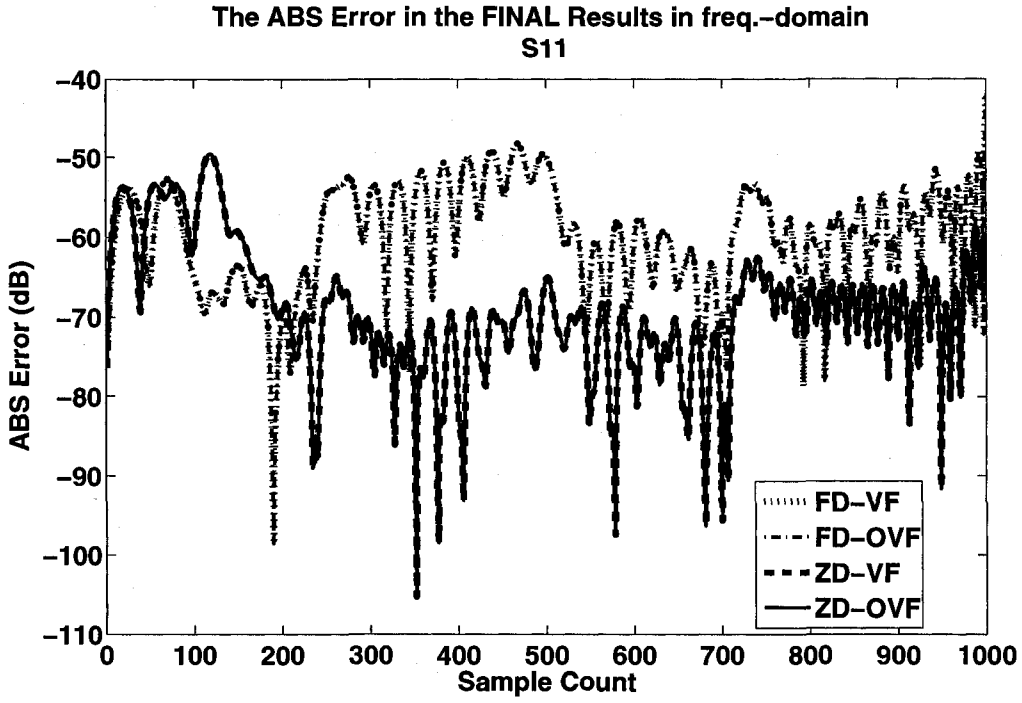


Figure 8-31: The absolute errors in the S_{11}response induced from final models: Example two

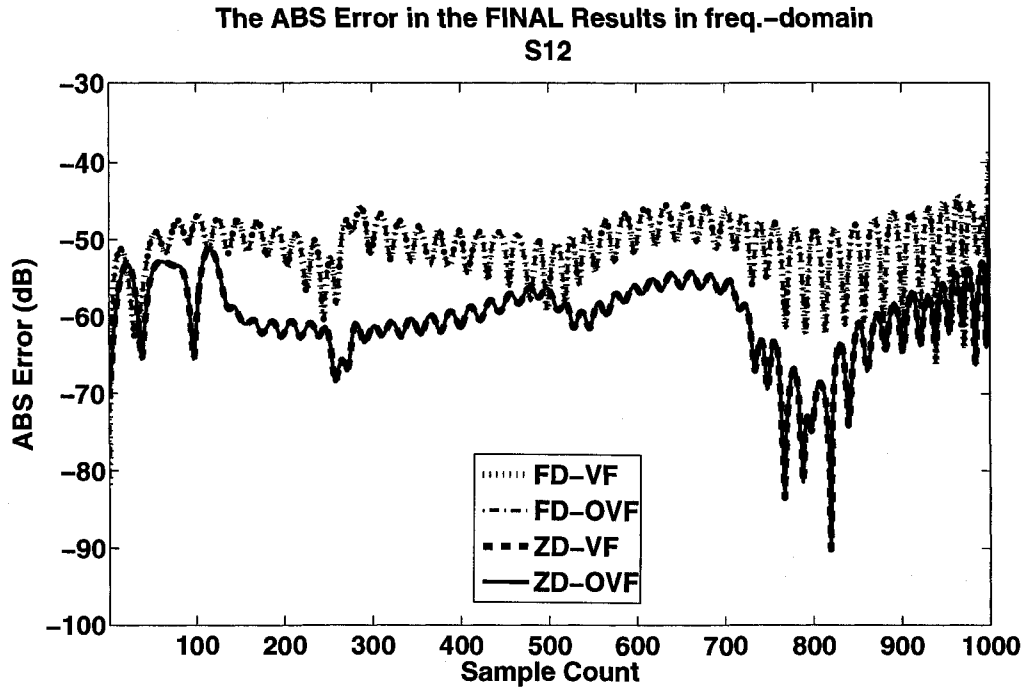


Figure 8-32: The absolute errors in the S_{12} response induced from final models: Example two

According to the Figure 8-31 and Figure 8-32, it can be judged that:

- The ZD vector fitting techniques result in the more accurate final models.
- As it was stated for the first example, when the pole identification algorithms start from optimal initial poles, and proceed without facing the numerically ill-conditioned system of equations, then both methods (of each domain) result in almost same model.

8.3.2 The Effect of SK Weighting Function for Example Two

To broaden the scope of observation in this report, the effect of utilizing the SK cost function is tested and illustrated in the following graphs.

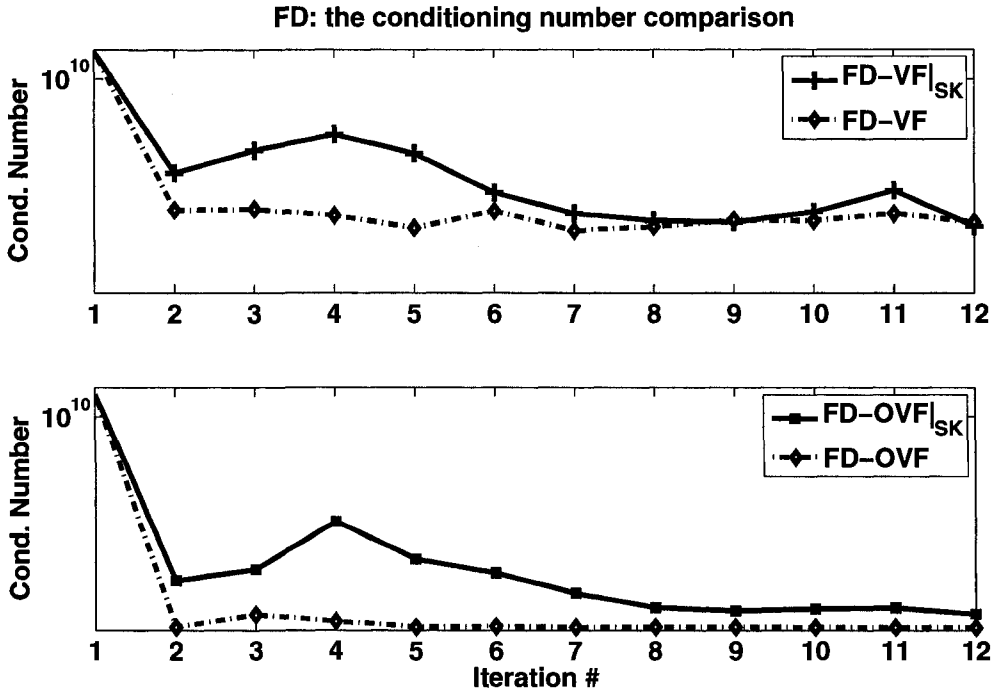


Figure 8-33: Comparison between numerical quality of the resulting system equations in frequency-domain techniques: Example two

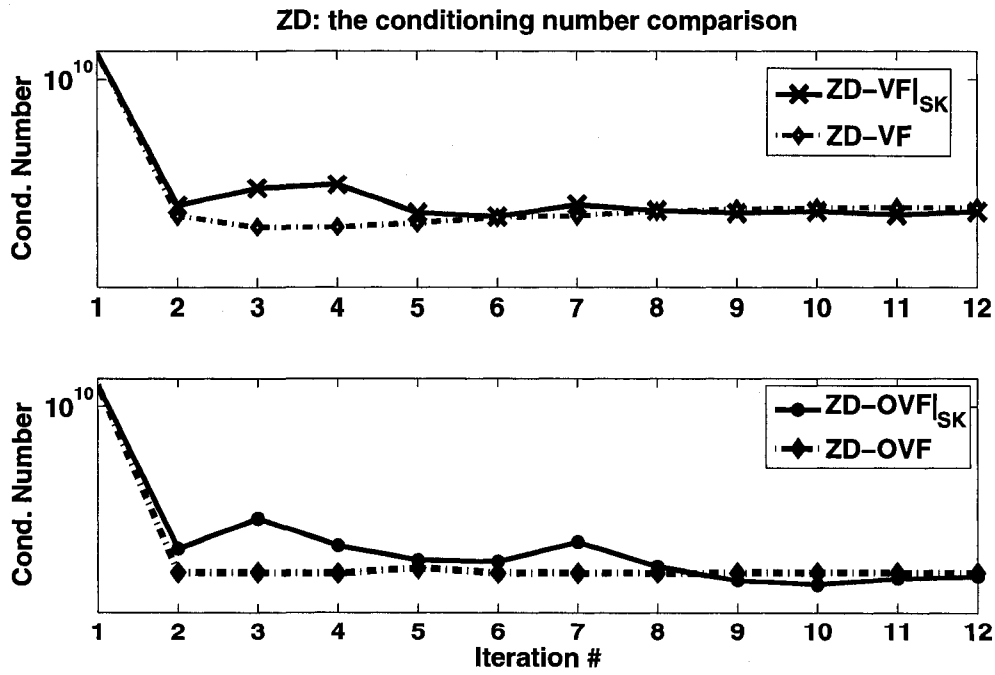


Figure 8-34: Comparison between numerical quality of the resulting system equations in discrete-domain techniques: Example two

Table 8-3: Summarizing the effect of SK method on conditioning numbers for example two

Comparison	Initial Iterations	close to convergence	Iterations after convergence	Far after the convergence
FD-VF$\|_{SK}$ vs. FD-VF	–	$\cong$	$\cong$	–
FD-OVF$\|_{SK}$ vs. FD-OVF	–	–	–	–
ZD-VF$\|_{SK}$ vs. ZD-VF	–	$\cong$	$\cong$	$\cong$
ZD-OVF$\|_{SK}$ vs. ZD-OVF	–	$\cong$	+	$\approx$+

Effect on in the numerical condition of the matrix in system equations:
+ : Improvement (better)
– : Degradation (worse)
$\cong$: Almost equal

For this set of data, one may conclude that using the SK generally leads to numerical quality degradation in the equations. As it is shown below, the solutions of the resulting equations however, approximate the original data better, especially, for z-domain techniques.

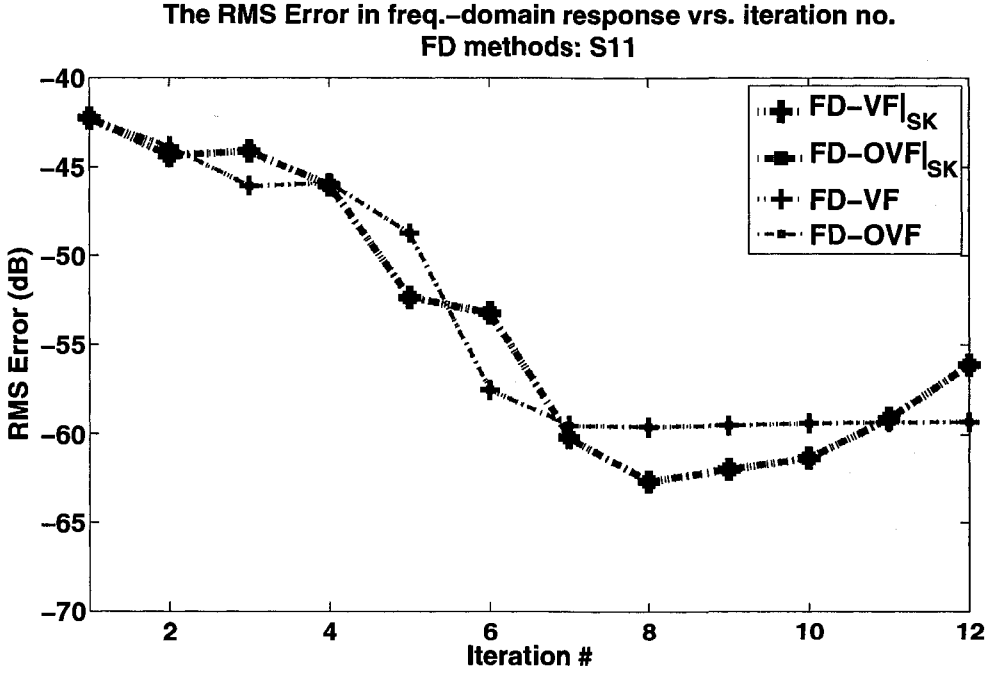


Figure 8-35: Comparison of RMS errors per iteration for S_{11} resulting from FD algorithms: Example two

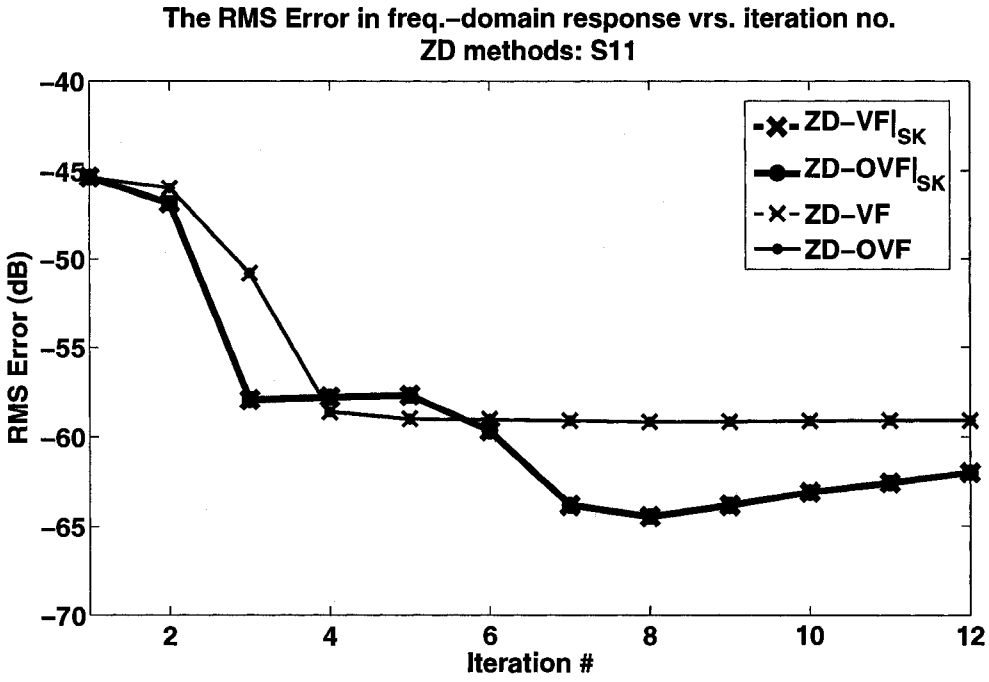


Figure 8-36: Figure 8-37: Comparison of RMS errors per iteration for S_{11} resulting from ZD algorithms: Example two

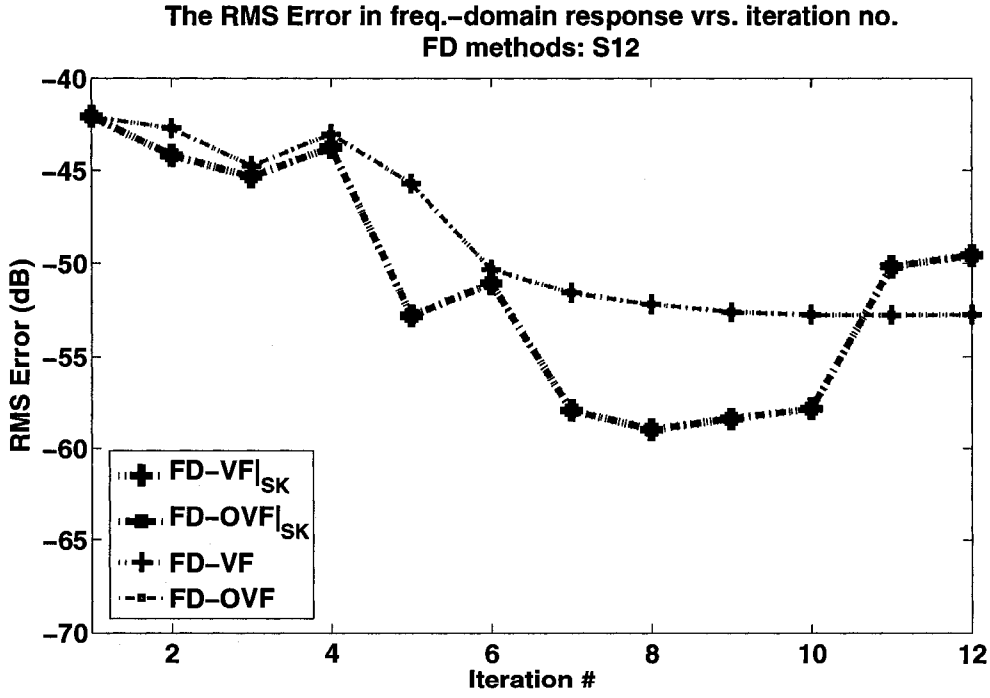


Figure 8-38: Comparison of RMS errors per iteration for S_{12} resulting from FD algorithms: Example two

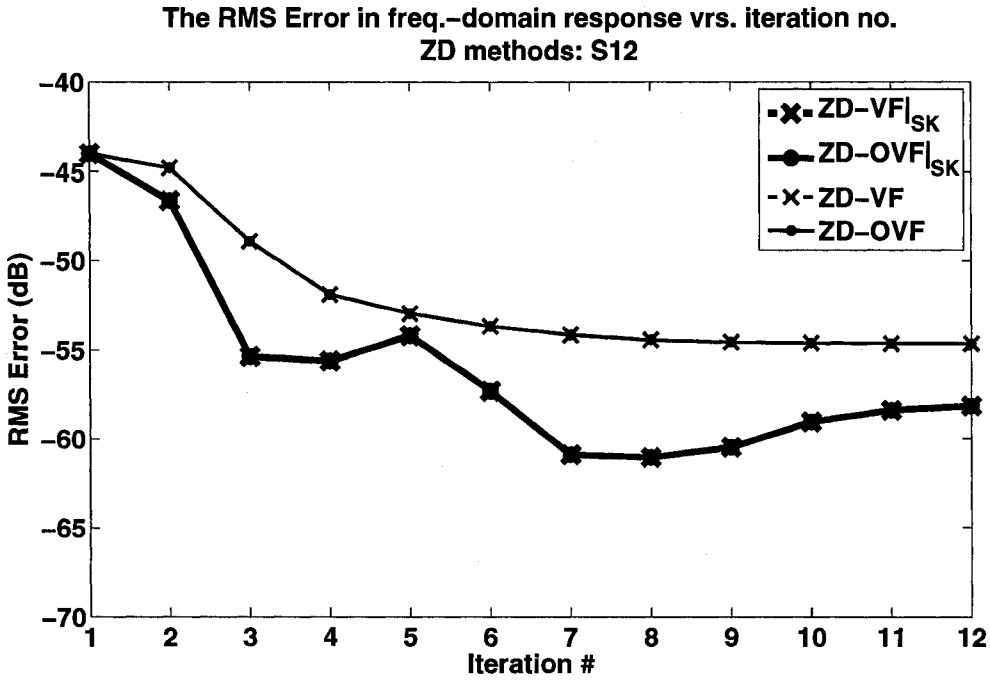


Figure 8-39: Comparison of RMS errors per iteration for S_{12} resulting from ZD algorithms: Example two

Table 8-4: Summarizing the effect of SK method on the RMS error for example two

Comparison	Initial Iterations	close to convergence	Iterations after convergence	Far after the convergence
FD-VF$\|_{SK}$ vs. FD-VF	$\cong$	+	+	–
FD-OVF$\|_{SK}$ vs. FD-OVF	$\cong$	+	+	–
ZD-VF$\|_{SK}$ vs. ZD-VF	+[†]	+	+	+
ZD-OVF$\|_{SK}$ vs. ZD-OVF	+[†]	+	+	+

Effect on in the numerical condition of the matrix in system equations:
+ : Improvement (better)
– : Degradation (worse)
$\cong$: Almost equal

[†]: Except for iterations number 4 and 5 in S11

By comparing the RMS errors and considering the results in above table, it is concluded that, ZD techniques provide more accurate results in comparison to their counterpart methods. It is also seen that, using the SK factor improves the accuracy in results. The resulting improvement on the z-domain techniques is more consistent.

8.3.3 Ill-Conditioned System Equations

It may happen that the overdetermined linear system of equations in the pole-identification process becomes ill-conditioned. One of the known circumstances in which a poor numerically conditioned linear problem in vector fitting may happen, is when the starting poles are real. Complex starting poles, whose real parts are not sufficiently small, may also lead to ill-conditioning. In particular, a severe condition is expected when there are dominant resonance peaks in the considered frequency interval of the response to be fitted. This issue has been studied in [13] and to overcome the problem it has been suggested that, optimal starting poles to be complex conjugate with sufficiently small real parts.

In the z-domain vector fitting techniques, similar concerns were experienced when the moduli of the starting poles are not sufficiently close to one; in other words, when starting poles are not close enough to the unit circle.

For the cases in which algorithm faces numerical instability in system equations, to overcome the problem, there is always a possibility of going through an iterative procedure. It includes the selection of sensible location of starting poles and then running a round of vector fitting to optimize this set of poles. Afterward, resulting poles from the first round can be used as starting poles for the next round of trial. Following this procedure is explicitly considered as 'trading time with accuracy'. Thus, the method that can handle the ill-conditioning problem with better accuracy is superior and leads to a reduction in the overall computation time.

To continue, a set of 82 real initial poles uniformly spaced along the frequency range of interest is chosen. A comparison of the conditioning numbers and RMS errors per iteration is illustrated in the following figures.

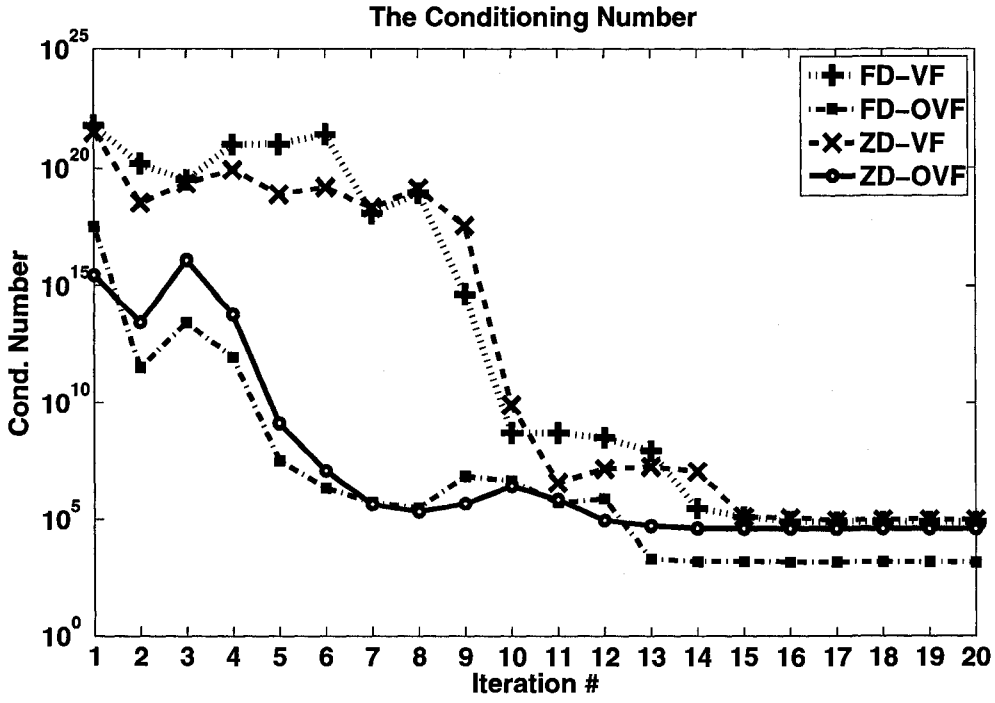


Figure 8-40: Condition number per iteration for proposed ZD-OVF versus 3 other methods: Real initial poles

It is seen that, in the convergence region after iteration #13, the conditioning number of matrix **A** in frequency-domain orthonormal vector fitting formulation drops under the one from proposed method; however, within the first iterations they closely follow each other.

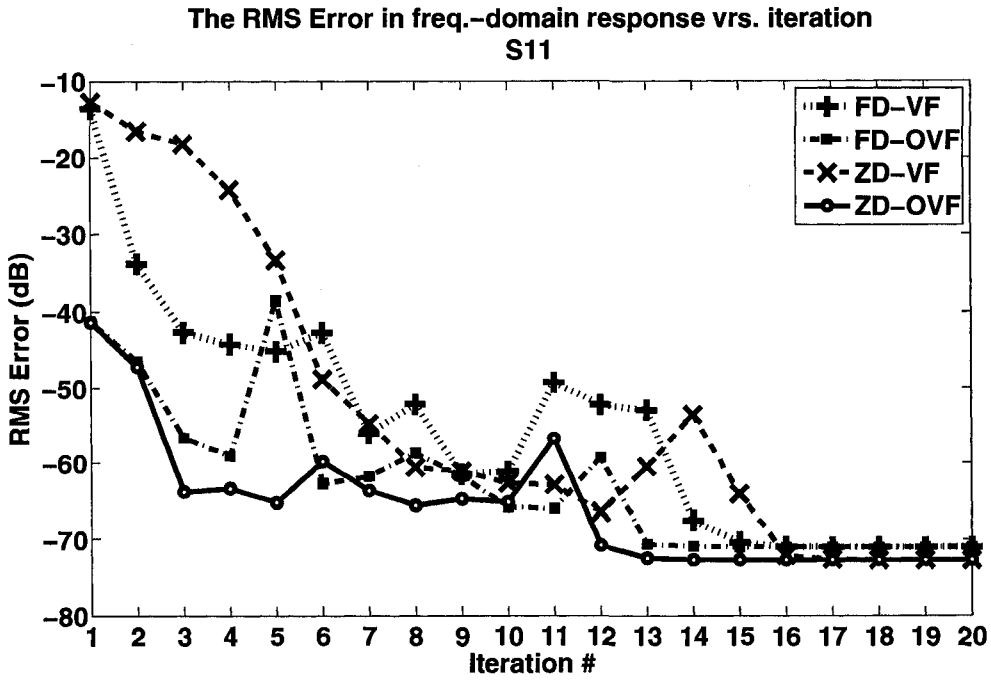


Figure 8-41: RMS error per iteration from proposed ZD-OVF versus 3 other methods: Real initial poles

In Figure 8-41, it is shown that,

- The proposed ZD-OVF has lower RMS error in almost all iterations.
- It converges faster before any other methods at iteration#13, while to get the same accuracy ZD-VF needs 3 more iteration.
- When we let all the pole-identification processes continue until they overcome the ill-conditioning and converge, in each domain VF and OVF methods result in the same model.
- The z-domain techniques result in more accurate final models. This fact is verified in Figure 8-42 too.

In Figure 8-41 it is seen that, among the previous methods FD-OVF is showing better performance in handling the unoptimal starting poles. Figure 8-42 shows absolute error for the resulting S_{11} from proposed ZD-OVF in the same graph with the one from FD-

OVF. It is apparently seen that the absolute error $\left(\left|S_{\text{original}} - S_{\text{model}}\right|_{dB}\right)$ in the result from VZ-OVF is lower that the one from FD-OVZ throughout the frequency spectrum (except for few frequency points).

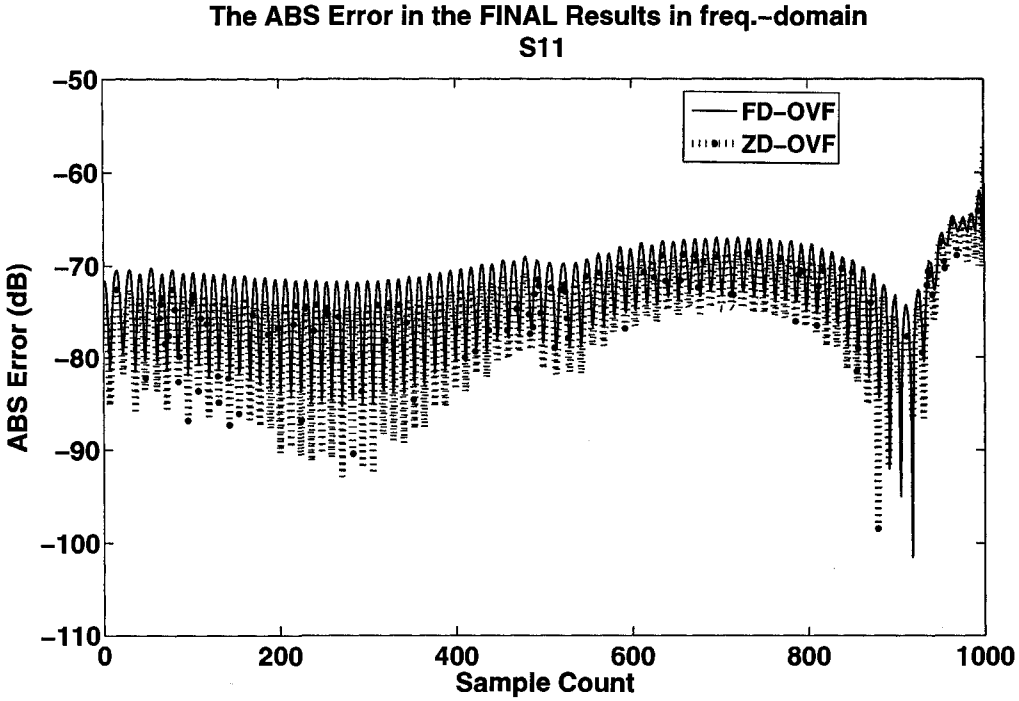


Figure 8-42: Absolute error per iteration from proposed ZD-OVF versus FD-OVF method – Real initial poles

8.4 Accurate LLS Equation Solvers to Enhance the Accuracy

A successful application of vector fitting algorithm requires that the linear least squares problem, formed in the process, can be solved with sufficient accuracy [13]. When these system equations become ill-conditioned, using the 'normal equations method' (NE) would not be an efficient approach. Since it squares the conditioning number of problem, it suffers the most round off error. To shed light on this aspect of VF, for the same example in the section 8.3.3, the results of following two experiments are considered.

i) Using Optimal Starting Poles

A set of optimal complex conjugate poles was used to fit the data by exploiting the proposed ZD-OVF algorithm. Using these poles leads to properly conditioned system equations. The LLS equations resulting from ZD-OVF were recorded and then examined by different LLS solving methods.

A comparison between the performances and the results from different solution methods when solving the two LLS equations resulted in the first two iterations is presented in the following tables.

Table 8-5: Comparison between solutions for the LLS eqn. resulted from ZD-OVF in first iteration, with optimal starting poles

Method	norm(residual)= sqrt(sum(|b-Ax|^2))	No of zero entry in x (answer vector)	norm(x)
Normal Eqn.	+3.68689593e-002	0	+5.37209724e+000
Using \	+2.16896622e-002	0	+6.31623250e+003
QR (X = R\Q'*b)	+2.16896622e-002	0	+6.31623250e+003
QR (PseudoInverse)	+2.16896622e-002	0	+6.31623250e+003
SVD	+2.16896622e-002	0	+6.31623250e+003
RRQR	+2.16896622e-002	0	+6.31623250e+003

- In Table 8-5 it is seen that due to the full rank, almost properly conditioned matrix all the examined methods except 'normal equation' are resulted in the acceptable solutions. It shows that, because 'NE' squares the conditioned number of the matrix **A,** the resulting solution is less accurate.

After initial poles were relocated and optimized in the first iteration, they were used as the starting poles for the second iteration.

Table 8-6: Comparison between solutions for the LLS eqn. resulting from ZD-OVF in the second iteration, using optimal poles resulted from the first iterations

Method	norm(residual)= sqrt(sum(|b-Ax|^2))	No of zero entry in x (answer vector)	norm(x)
Normal Eqn.	+1.68049711e-003	0	+8.39554025e-001
Using \	+1.68049711e-003	0	+8.39554025e-001
QR (X = R\Q'*b)	+1.68049711e-003	0	+8.39554025e-001
QR (PseudoInverse)	+1.68049711e-003	0	+8.39554025e-001
SVD	+1.68049711e-003	0	+8.39554025e-001
RRQR	+1.68049711e-003	0	+8.39554025e-001

- Table 8-6 shows that due to using better poles, in the second iteration, the numerical condition of the LLS equation has been sufficiently improved, such that, all the methods were able to capture the same solution for the system of linear equations.[2]

ii) Using Real Starting Poles

To worsen the numerical quality of equations even in comparison to the explained condition in section 8.3.3, the starting real poles are distributed logarithmically over the frequency range from f_{min} to $2f_{max}$. After converting to z-domain, these poles were used as the initial poles for the proposed ZD-OVFD method to create a proper transfer function model (including the constant term). Due to these poor starting poles the ZD vector fitting process has been faced a severely ill-conditioned matrix especially within initial iterations. The resulting system equations were examined with different solution methods and the comparisons are presented in the following tables.

[2] For this unique solution both the 'mean square error' in column 2 and the norm of the parameters vector in column 4 of the table were uniquely minimized.

Table 8-7: Comparison between solutions for the LLS eqn. resulted from ZD-OVF in the first iteration, by using real initial poles

Matrix A is rank deficient!

Method	norm(residual)= sqrt(sum(|b-Ax|^2))	No of zero entry in X (answer vector)	norm(x)
Normal Eqn.	+2.10515150e+004	0	+2.31117745e+006
Using \	+7.23600989e-001	113	+5.35134525e+003
QR (X = R\Q'*b)	+5.66895220e-001	105	+1.01721487e+006
QR (PseudoInverse)	+3.05596310e+007	0	+6.59507597e+016
SVD	+4.54784373e+008	0	+1.32856099e+017
RRQR	+7.58186222e-001	0	+5.48813613e+005

Table 8-8: Comparison between the solutions for the LLS eqn. resulted from ZD-OVF in the third iteration, using real initial poles

Matrix A is rank deficient!

Method	norm(residual)= sqrt(sum(|b-Ax|^2))	No of zero entry in X (answer vector)	norm(x)
Normal Eqn.	+3.14150357e+010	0	+1.48493710e+019
Using \	+3.17780425e+000	110	+6.27257054e+006
QR (X = R\Q'*b)	+3.07250666e+000	108	+1.24228389e+008
QR (PseudoInverse)	+1.18656307e+003	0	+5.10380656e+024
SVD	+1.24078194e+006	0	+5.16107016e+024
RRQR	+3.07176349e+000	0	+8.34511788e+007

According to the above observations, it is judged:

- The system equations in these iterations are severely ill-conditioned.
- Applying 'normal equation' (NE) method in the 1-st and 2-nd iterations results in 2.1e+4 and 3.14e+10 error in systems parameters respectively. It shows that NE is not a trustworthy choice to obtain the required accuracy.
- The Matlab function 'mldivide[3] or' \' 'and its close counterpart shown as QR in the table provide lower absolute error and minimum norm of the solution vector. This

[3] The specific algorithm used for solving the simultaneous linear equations denoted by **X=A\B** depends upon the structure of the coefficient matrix **A**. E.g. when **A** is not square in MATLAB, Householder reflections are used to compute an orthogonal-triangular factorization. $AP = QR$, where **P** is a

minimum norm of **X**, however, obtained by enforcing many of the system parameters to zero[4]. This is not a desirable situation, because, the effect (existence) of many poles is practically neglected.

- It is seen that RRQR method gives better solution with non-zero parameters. Even it is more accurate in comparison to the "economy size" SVD, which is conventionally known as a very accurate method.

8.5 Proposed ZD-OVF with Different LLS Equation Solvers

Next, the behavior of the ZD-OVF by utilizing the following three methods is examined.

- MATLAB 'mldivide' function denoted by ' \ '
- Rank Revealing QR factorization (RRQR)
- "Economy size" Singular Value Decomposition (SVD)

i) **Experiment 1:**

When the same severely ill-conditioned problem introduced in section 8.4 is solved; according to the graphs in Figure 8-43 and Figure 8-44:

- The result from RRQR is more accurate.
- It is seen that utilizing '\' function also leads to the acceptably accurate results, even when the vector fitting process started from the poor starting poles. Practically, it scatters initial poles over the unit disk within initial iterations, the new complex

permutation, **Q** is orthogonal and **R** is upper triangular. The least squares solution **X** is computed with $X = P(R\backslash(Q'B)$ [46]. For more details, [65] and standard MATLAB documentation can be the referred to.

[4] Any value that drops under the round off error level (1e-12) is counted as zero in this experiment.

poles can be a better starting choice for the subsequent iterations. Therefore, the numerical quality of the system equations can be improved in consequent iterations.

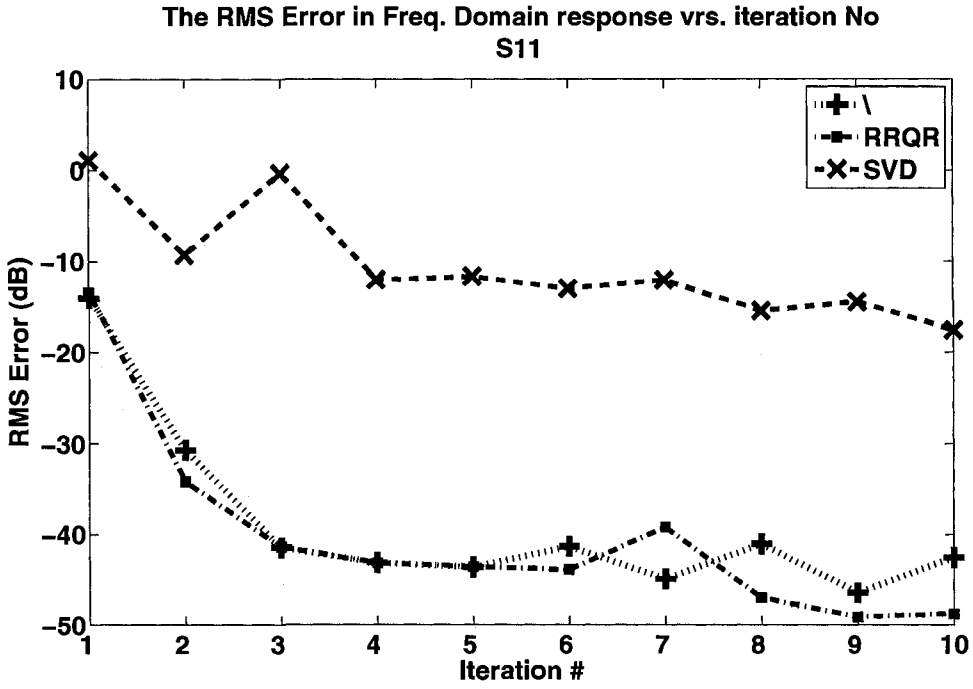


Figure 8-43: RMS error in the final model per iteration, using real initial poles

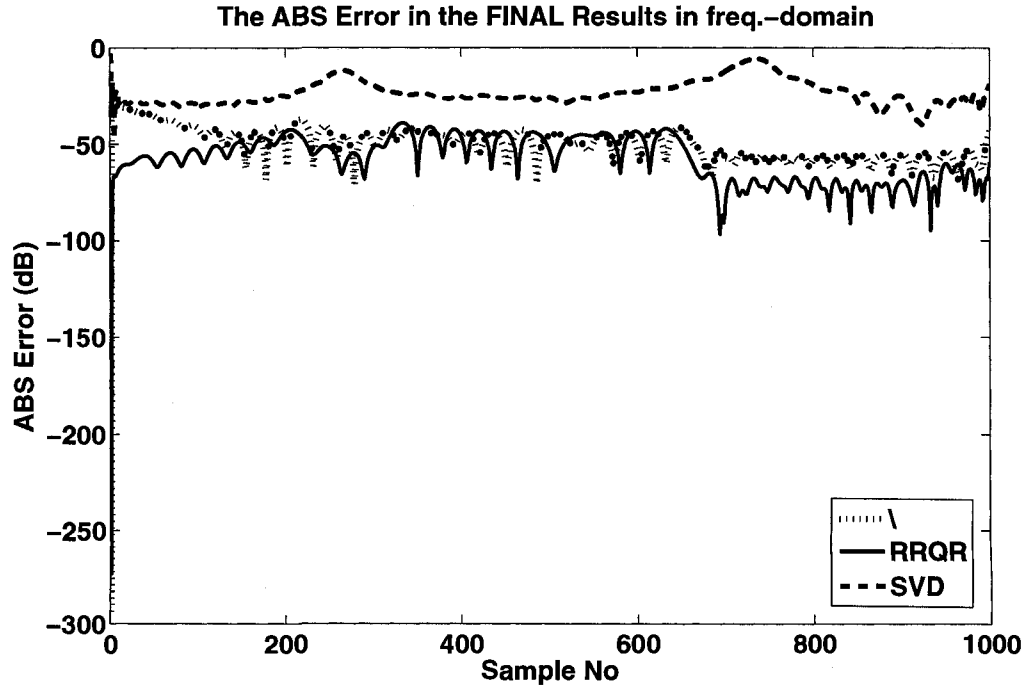


Figure 8-44: Absolute error in the final model per iteration, using real initial poles

ii) Experiment 2:

The above condition may seem as a rare instance of ill-conditioning issue in practical vector fitting. To examine a more realistic situation, the problem in 8.3.3 was tested; results are compared in following figure.

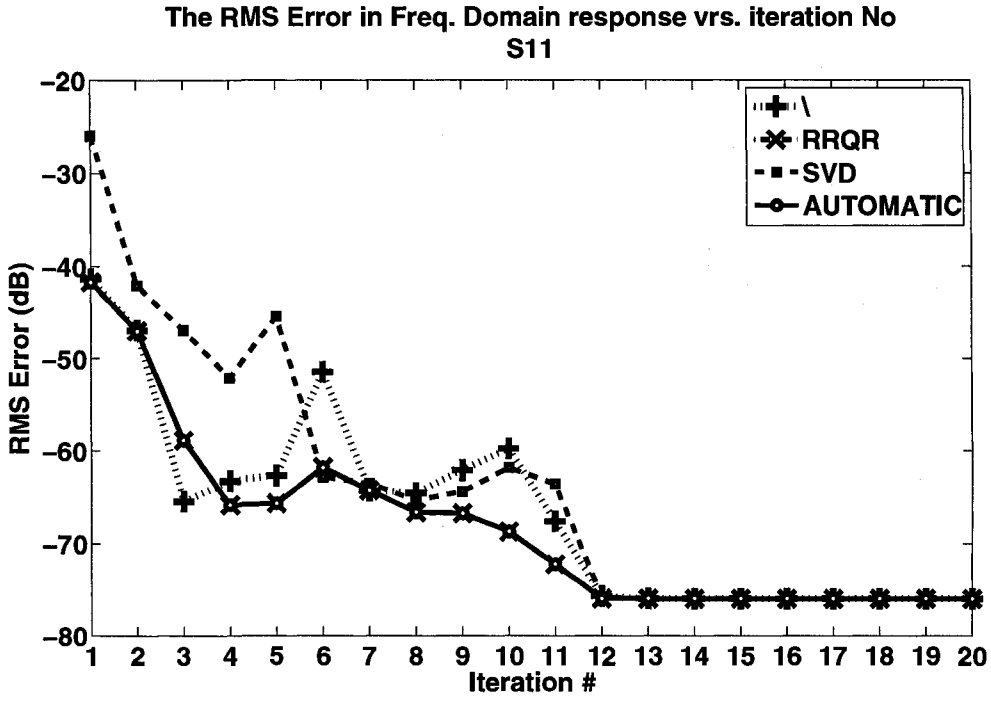


Figure 8-45: RMS error in the final model per iteration: for same circumstance outlined in section 8.3.3

- As it is expected, after initial iterations when ZD-OVF by iteratively optimizing the poles location to be recovered from ill-conditioning, the solutions for full rank system equations by all the above methods converge to almost same results.
- However, within the first 12 iterations RRQR provides better results.
- The 'automatic' algorithm, examined above, dynamically changes the solution method to RRQR when the linear equation is severely ill conditioned. As shown in

the graph, it provides almost same accuracy with RRQR and shows the importance of solving ill-conditioned system equations to ensure accurate resulting models.

CHAPTER 9. Conclusions and Future Work

This chapter contains concise conclusions of the work that was presented in this thesis. In addition, the directions for future work are outlined.

9.1 Conclusions

In this thesis, the 'z-Domain orthonormal vector fitting technique' (ZDOVF), as a robust and efficient multiport algorithm for the fast linear macromodeling of port-to-port response, has been introduced and examined. This work also comparatively studies the four vector fitting algorithms in two different domains from the methodological perspective.

Following is a summarized account of the major characteristics and advantages of the method presented in this thesis.

1) The method provides a high numerical stability and does not suffer from the problem of ill-conditioning while computing the response over a broadband frequency spectrum.

2) In the presented technique in this work, the numerical nature of the system equations formed by using the proposed z-domain basis functions is more compatible with the Sanathanan-Koerner formulation in comparison to the FD techniques. Thus, the

algorithm can compatibly enjoy the advantages of the SK method in increasing the numerical stability and enhancing the accuracy.

3) The proposed multiport z-domain VF technique allows the computation of common multiport poles in an efficient manner, while the angular frequency of the resulted complex s-domain poles can be bounded to a maximum frequency of interest.

4) The consistently convergent nature of the iterative procedure is confirmed by introducing a new method to measure the convergence within the process, while in similar to the other VF processes the resulting poles is tending to their final optimal locations.

5) The algorithm intrinsically has less cumulative numerical error. This can be judged from the consistent behavior throughout the higher iterations far after convergence, where it stabilizes the RMS errors in the resulting model within an acceptable margin.

6) It is capable of evaluating the parameters of the system from the 'frequency-domain' responses or directly from 'time-domain' (usually truncated) responses. Therefore, it can even be considered as an alternative for the conventional time-domain vector fitting techniques.

7) This technique can carry out the macromodeling task for both 'continuous-time' and 'discreet-time' stable LTI systems.

8) Exploiting the z-transformation in formulation makes this method capable of handling the data for the delayed systems / signals.

9) As it is stated in chapter 7, when vector fitting in the frequency-domain, it is essential to scale the frequency axis and hence the system's parameters, to guarantee the numerical stability of the system equation. Afterward, the practical concerns become involved in obtaining the final model to support the original data. However, for the proposed ZD-OVF, It is dispensable to scale the frequency.

9.2 Future Research

1) Improving the numerical quality for LLS equations

The formulation of the proposed method has been structured based on a set of orthonormal basis that was proposed to adapt the Takenaka-Malmquist functions to ensure the real time-domain impulse response. Mathematically, the proposed set is not unique. Hence, it may be instrumental that future work to be invested to extract other formation for the functions to further improves the numerical conditioning of the matrix **A** in ZD-OVF formulation. This can enhance the numerical stability of the method and guarantee the lowest conditioning number in formulation in comparison to its FD counterpart method.

2) Approximation with higher-order poles multiplicity

The classical vector fitting macromodeling algorithm may deliver a poor fitting model if some poles of the basis functions occur with a higher order multiplicity [63], especially if the 'normal equations' (NE) are solved [19]. Prototyping RRQR as an efficient linear least square equation solver and examining a number of practical cases in which matrix **A** was suffering from rank deficiency, provides a sufficient background to tackle this established problem by means of the proposed formulation.

3) Passivity study and compensation

It will be desirable to enforce the passivity criteria while generating the macromodel. This may increase the vector fitting time dramatically while reducing the efficiency as well attainable accuracy of the method. The effect of enforcing the passivity constrains in the convergence of the pole-relocation process and the attainable level of accuracy in the final model within a reasonable number of iterations seems an interesting subject for further investigation.

4) Delay extraction applications

An efficient macromodeling using sampled time-domain or frequency-domain data from lossy-coupled lines with long delay is still an interesting subject for further research. Utilizing the z-transforms in the formulation of ZD-OVF provides an easy means to handle the delay extraction task prior to approximating the time-domain data of the line. It is expected to guarantee the efficiency and accuracy in macromodeling process.

5) Negative real z-domain poles

There are still open research subjects on the equivalence of z-domain and s-domain models in system identifications. Handling the negative real poles is cited among the concerns in the step-invariant or ZOH-transformation of discrete-time transfer function models to continuous time. Practically it is an unsolved problem in the 'computer-aided control systems design' context. In this work, It is shown that, to obtain a reasonable s-domain equivalent of the discrete-time model, the negative real poles, which are commonly considered as not transformable, can be mapped into a pair of complex conjugate poles on the aliasing margin. This causes the order of the s-domain

model to be higher than the order of its counterpart in z-domain by the number of the negative real poles. However, It is suggested that to escape the controversy of incompatibility, the frequency boundary for the primary region to be chosen such that the negative poles can be avoided. Handling the negative poles, when inverse mapping from z-domain to the s-domain while preserving the isomorphism property, is an interesting subject for further investigation.

6) Vector relaxation technique in the ZD-OVF formulation

Recently (2006) in [60] and [61], an extension of the standard vector fitting procedure by replacing the high-frequency asymptotic constraint of scaling function with a milder summation requirement has been suggested. For the conventional frequency-domain VF, it has been shown that this replacement can improve the ability of algorithm to relocate poles to better positions and reduces the significance of the choice of initial poles. The effect of a similar extension on the behavior of ZD-OVF and evaluating the attainable improvement in its performance is worth being investigated.

7) Studying the convergence of time-domain sequence

The inverse z-transform of the z-domain macromodel, structured by using proposed ZD-OBF, yields the corresponding time sequence. According to the stability criterion for stable systems, it is expected that every bounded excitation would lead to a bounded response. Consequently, studying the conditions and providing a mathematical justification on the convergence of the resulting time-domain sequence would be a noteworthy subject for a future theoretical attempt. [20, p153 and Appendix C] can be referred to for a useful theoretical background.

References

[1] H. Johnson, M. Graham, *High-Speed Digital Design.* Upper Saddle River, NJ: Prentice-Hall, 1993

[2] J. E. Schutt-Aine and R. Mittra, "Scattering Parameter Transient analysis of transmissions lines loaded with nonlinear terminations," *IEEE Transactions on Microwave Theory and Techniques*, vol. 36, pp. 529-536, 1988.

[3] J. E. Bracken, V. Raghavan, and R. A. Rohrer," Interconnect Simulation with Asymptotic Waveform Evaluation," *IEEE Trans. Circuits Sys.*, vol. 39, pp. 869-878, Nov. 1992.

[4] R. Achar and M. Nakhla, "Simulation of high-speed interconnects," *Proceedings of the IEEE*, vol. 49, pp. 693-728, May 2001.

[5] A. Deutsch, "Electrical characteristics of interconnections for high-performance systems," *Proceedings of the IEEE*, vol. 86, pp. 315-355, Feb. 1998.

[6] C. R. Paul, *Analysis of multiconductor transmission lines.* New York, NY: John Wiley and sons, 1994.

[7] C. Yen, Z. Fazarinc and R. L. Wheeler, "Time-domain skin-effect model for transient analysis of lossy transmission lines," *Proceedings of the IEEE*, vol. 70, pp. 759-757, July 1982.

[8] T. Vu Dinh, B. Cabon and J. Chilo,"Time domain analysis of skin effect on lossy interconnects," *Electronics letters*, vol. 26, pp. 2057-2058, Sep. 1990.

[9] D. Ioan, G. Ciuprina, M. Radulescu, and E. Seebacher, "Compact modeling and fast simulation of on-chip interconnect lines," *IEEE Trans. Magn.*, vol. 42, no. 4, pp. 547–550, Apr. 2006.

[10] W. T. Beyene and J. E. Schutt-Aine, "Efficient transient simulation of high-speed interconnects characterized by sample data," *IEEE Trans. Component.*, vol. 21, pp. 105-114, Feb. 1998.

[11] Y. S. Mekonnen and J. E. Schutt-Aine, "Broadband macromodeling of sampled frequency data using z-domain vector-fitting method," *SPI Conference*, pp.45-48, 2007.

[12] E. Chiprout and M. S. Nakhla, *Asymptotic waveform evaluation and moment matching of interconnect analysis.* Boston, MA: Kluwer, 1994.

[13] B. Gustavsen and A. Semlyen, "Rational approximation of frequency domain responses by vector fitting," *IEEE Trans. Power Delivery*, vol. 14, pp. 1052-1061, July 1999.

[14] E. C. Levi, "Complex curve fitting," *IEEE Trans. Automatic Control*, vol. AC-4, no. 1, pp. 37–43, Jan. 1959.

[15] D. Deschrijver, B. Haegeman, and T. Dhaene, "Orthogonal vector fitting: A robust macromodeling tool for rational approximation of frequency domain responses," *IEEE Trans. Advanced Packaging*, vol. 30, no. 2, pp. 216-225, May 2007.

[16] B. Gustavsen, "Computer code for rational approximation of frequency dependent admittance matrices," *IEEE Trans. Power Delivery*, Vol. 17, pp. 1093-1098, Oct. 2002.

[17] R. Pintelon, P. Guillaume, Y. Rolain, J. Schoukensand H. Van hamme, " Parametric identification of transfer functions in the frequency domain-A survey," *IEEE Transactions on Automatic Control*, vol. 39, no. 11, Nov. 1994.

[18] A. Taflove, Computational Electrodynamics: *The Finite-Difference Time-Domain Method.* Norwood, MA: Artech House, 1995.

[19] S. Grivet-Talocia, "Package macromodeling via time-domain vector fitting," *IEEE Microw. Wireless Compon. Lett.*, vol. 13, no. 11, pp.472–474, Nov. 2003.

[20] P. W. Broome, "Discrete orthonormal sequence," *Journal of the association for computing Machinery*, Vol.12, no. 2, pp.151-168, 1965

[21] A. V. Oppenheim, A. S. Willsky, and I. T. Young, *Signals and Systems*. 1983.

[22] P. S. C. Heuberger, P. M. J. Van Den Hof, and O. H. Bosgra, "A generalized orthonormal basis for linear dynamical system," *IEEE Trans. On Automatic control,* vol. 40 no. 3, pp.451-465, March 1995.

[23] E. Kreyszig, *Advanced Engineering Mathematics*. 6th ed., New York, John Wiley

and sons, 1988.

[24] J. W. Brown and R. V. Churchill, *Complex Variables and Applications.* 7th ed. McGrawHill, 2004.

[25] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing.* 2nd ed., Prentice Hall Inc., 1998.

[26] L.R. Rabiner and B. Gold, *Theory and application of digital signal processing,* Englewood Cliffs, NJ: Prentice-Hall, 1975. pp.393-399, 1975.

[27] K. Ogata, *Discrete-Time Control Systems.* 2nd ed., Prentice-Hall, 1995.

[28] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* Prentice Hall Inc., 1975.

[29] J. Schoukens, R. Pintelon, and H. Van Hamme, "Identification of linear dynamic systems using piecewise constant excitations: Use, misuse, and alternatives," *Automatica,* vol. 30, no. 7, pp. 1153-1 169, 1994.

[30] C. Evans, D. Rees, L. Jones, and D. Hill, "Time and frequency domain identification of jet engine dynamics: Problems and solutions," in Sysid'94, 10th IFAC Symposium on System Identification, Copenhagen, July 4-6, 1994. Vol.2, pp. 2.243-48, 1994.

[31] N. K. Sinha and G. J. Lastman, "Transformation of discrete-time models," in *Identification of Continuous-Time Systems (N. K. Sinha and G. P. Rao ed.),* ch. 4, pp. 123-137, Dor-drecht: Kluwer Academic Publishers, 1991

[32] G. F. Franklin, J . D. Powell, and M. L. Workman, *Digital control of dynamic systems.* 3rd ed., Reading, MA: Addison-Wesley, 1998.

[33] I. Kollar, G. Franklin, and R. Pintelon, "On the equivalence of z-domain and s-domain models in system identification," *IEEE instrumentation and Measurement Technology Conf.*, Brussels, Belgium, pp.14-19, June 4-6, 1996.

[34] The MathWorks™, *Signal Processing Toolbox 6 User's Guide, MATLAB Version 6.8 (R2007b).*

[35] P. Quast, M. K. Sain, B. F. Spencer Jr. ASCE, and S. J. Dykeb, "Microcomputer implementation of digital control strategies for structural response reduction," [online] available at cee.uiuc.edu/sstl/papers/micro9.pdf

[36] P. S. C. Heuberger, P. M. J. Van Den HOF and B. Wahlberg, *Modeling and Identification with Rational Orthogonal Basis Functions*, London, Springer-Verlag, 2005.

[37] T. Oliveira e Silva, *Rational orthonormal functions on the unit circle and on the imaginary axis, with application in system identification. A*vailable at:

URL:ftp://inesca.inesca.pt/pub/tos/English/rof.ps.gz, October 1995.

[38] R. H. Middleton and G. C. Goodwin, *Digital control and estimation.* Englewood Cliffs, NJ: Prentice-Hall, 1990.

[39] H. Akcay, S. Islam, and B. Ninness, "Orthonormal basis functions for continuous-time systems and $\mathcal{L}_p$ convergence," *Mathematics of Control, Signal and Systems*, vol. 12, pp.295-305, 1999.

[40] W. H. Kautz, "Transient synthesis in the time domain," *IRE Trans. Circuit Theory*, vol. 1, pp. 29-39, 1954.

[41] T. Paatero, M. Karjalainen, and A. Harma, "Modeling and equalization of audio systems using Kautz filters," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, pp.3313-3316, 2001.

[42] B. Wahlberg,"System identification using Laguerre models," *IEEE Trans. Automatic Control*, vol. 36, no. 5, pp. 551–562, 1991.

[43] B. Wahlberg, "System identification using Kautz models," *IEEE Trans. Automatic Control*, vol. 39, no. 6, pp.1276–1282, 1994.

[44] Y. W. Lee, *Statistical theory of communication.* New York, John Wiley and sons, 1960.

[45] T. O. Silva, "Laguerre filters- An introduction," *REVISTA DO DETUA*, Vol.1,

No. 3, Janerio 1995.

[46] The Wolfram website, [Online] available at:, *http://mathworld.wolfram.com*

[47] C.T. Chen, *Linear System Theory and Design.* 3rd ed., New York: Oxford University Press, 1999.

[48] B. Friedland, *Control System Design- An Introduction to State Space Methods.* McGraw-Hill, 1986.

[49] Maciejowski, J.M., *Multivariable Feedback Design*, Addison-Wesley Publishing Company, 1989.

[50] H. Akcay and B. Ninness, "Rational basis functions for robust identification from frequency and time domain measurements," *Proceedings of the American Control Conference,* vol. 6, pp.3559 - 3563 vol.6, June1998.

[51] B. Ninness, S. Gibson, and S. R. Weller, "Practical aspects of using orthonormal system parameterizations in estimation problems," *12th IFAC Symposium on System Identification*, Sant Barbara USA, 2000.

[52] A. W. M. Van den Enden, G. C. Groenendaal, and E. Van de Zee, "An improved complex curve-fitting method," *Proc. Conf. Computer Aided Design of Electronic, Microwave Circuits and Systems,* Hull, United Kingdom, pp. 53-58, 1977.

[53] A. W. M. Van den Enden and G. A. L. Leenknegt, "Design of optimal filters with arbitrary amplitude and phase requirements," *Signal Processing III: Theories and Applications, (L. T. Young et al. ed.),* pp. 183-186, North Holland: Elsevier Science, 1986.

[54] C. K. Sanathanan and J. Koerner, "Transfer function synthesis as a ratio of two complex polynomials, " *IEEE Trans. Automatic Control,* vol. AC-8, no. 1, pp. 56-58, Jan.1963.

[55] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., London, UK: The Jon Hopkins University Press, 1996.

[56] G. J. Fee, "Linear Least-square data fitting with orthogonal polynomials,"

Department of Mathematics, Simon Fraser University, Burnaby, Canada.

Available at: www.cecm.sfu.ca/CAG/papers/gregMSW02.ps

[57] G. Q. Orti and E. S. Q. Orti, "Efficient algorithm for solving the linear least square problem," *Informe Te'cnico DI 1-5/96*, Department de informatica, University Jaime I.

[58] C. H. Bischof and G. Quintana-Qrti, "Computing Rank-Revealing QR factorization of dense matrices," *ACM Transaction on Mathematical Software*, Vol. 24, NO.2, Pages 226-253, June 1998.

[59] H. Whitfield, "Asymptotic behavior of transfer function synthesis methods," *Int. J. Control*, vol. 45, pp. 1083–1092, 1987.

[60] B. Gustavsen, "Improving the pole relocating properties of vector fitting," *IEEE Trans. Power Delivery*, vol. 21, no. 3, pp. 1587–1592, Jul. 2006.

[61] B. Gustavsen, "Relaxed vector fitting algorithm for rational approximation of frequency domain responses," *10th IEEE Workshop on Signal Propagation On Interconnects*, Berlin, Germany, May 2006

[62] B. Gustavsen and A. Semlyen, "Simulation of transmission line transients using vector fitting and modal decomposition," *IEEE Trans. on Power Delivery*, vol. 13, no 2, pp. 605 – 614, Apr. 1998.

[63] The Vector Fitting website [Online] Available at:

http://www.energy.sintef.no/produkt/VECTFIT/index.asp.

[64] D. Deschrijver and T. Dhaene, "A note on the multiplicity of poles in the vector fitting macromodeling method", *IEEE Transactions on Microwave Theory and Techniques*, vol. 55, no. 4, pp.736–741, April 2007.

[65] E. Anderson, Z. Bai, C. Bishof, S. Blackford , J. Demmel, J. Dongrra, J. Du Croz, A. Greenbaum, S. Hammastry, A. McKenney, and D. Sorensen, *LAPACK User's Guide,* 3rd ed., Philadelphia, PA: SIAM Press, 1999.

[66] D. R. Brillinger, *Time Series: Data Analysis and Theory.* New York: McGraw-

Hill, 1981.

[67] E. W. Weisstein, "Sampling Theorem," [Online] from MathWorld, a Wolfram Web Resource: http://mathworld.wolfram.com/SamplingTheorem.html .

[68] B.Wahlberg and P. Makila, "On approximation of stable linear dynamical systems using laguerre and Kautz functions," *Automatica*, vol. 32, pp. 693–708, 1996.

[69] S. Lin and E. S. Kuh, "Transient simulation of lossy interconnects based on the recursive convolution formulation" *IEEE Trans. Circuits Sys.*, vol. 39, pp. 879-892, Nov. 1992.

[70] C. Gordon, T. Blazeck, and R. Mittra, "Time-domain simulation of multi-conductor transmission lines with frequency-dependent losses," *IEEE Trans. CAD*, vol.11, pp. 1372-1387, Nov. 1992.

[71] L. M. Silveria, I. M. Elfadel, and J. K. White, "An efficient approach to transmission line simulation using measured or tabulated s-parameter data,"*31th ACM/IEEE Design Automation Conference*, 1994.

[72] Online document available at:

http://ccrma.stanford.edu/~jos/StateSpace/Markov_Parameters.html.

Appendix A. Review of Markov Parameters

Assume a discrete time process is realized by state-space equation as:

$$\begin{cases} \mathbf{x}(n+1) = \mathbf{A}\mathbf{x}(n) + \mathbf{B}u(n) \\ y(n) = \mathbf{C}\mathbf{x}(n) + Du(n) \end{cases}$$

The impulse response for this model at different "time moments" can be obtained as:

$t \le 0$, $\mathbf{x}(\bullet) = \overline{0}$, system is at its zero or rest state

$$t = 0, \quad \begin{cases} h_0 = \mathbf{C}\mathbf{x}(0) + D\delta(0) \\ \mathbf{x}(1) = \mathbf{A}\mathbf{x}(0) + \mathbf{B}\delta(0) \end{cases} \Rightarrow \begin{cases} \boxed{h_0 = D} \\ \mathbf{x}(1) = \mathbf{B} \end{cases}$$

$$t = T_S, \quad \begin{cases} h_1 = \mathbf{C}\mathbf{x}(1) + D\delta(1) \\ \mathbf{x}(2) = \mathbf{A}\mathbf{x}(1) + \mathbf{B}\delta(1) \end{cases} \Rightarrow \begin{cases} \boxed{h_1 = \mathbf{CB}} \\ \mathbf{x}(2) = \mathbf{AB} \end{cases}$$

$$t = 2T_S, \quad \begin{cases} h_2 = \mathbf{C}\mathbf{x}(2) + D\delta(2) \\ \mathbf{x}(3) = \mathbf{A}\mathbf{x}(2) + \mathbf{B}\delta(2) \end{cases} \Rightarrow \begin{cases} \boxed{h_2 = \mathbf{C}\mathbf{A}^1\mathbf{B}} \\ \mathbf{x}(3) = \mathbf{A}^2\mathbf{B} \end{cases}$$

...

$$t = nT_S, \quad h_n = \mathbf{C}\mathbf{x}(n-1) + D\delta(n-1) \Rightarrow \boxed{h_n = \mathbf{C}\mathbf{A}^{n-1}\mathbf{B}}$$

Thus, the impulse response of the state-space model can be summarized as

$$h_n = \begin{cases} D, & n = 0 \\ \mathbf{C}\mathbf{A}^{n-1}\mathbf{B}, & n > 0 \end{cases}$$

These impulse response terms $h_n = \mathbf{C}\mathbf{A}^{n-1}\mathbf{B}$ for $n \ge 0$ are known as the *Markov Parameters* [72].

Appendix B. Review of the Kautz Series

The Kautz functions have been introduced in [40], [68]. In the original form, they were associated with the continuous time domain based on the Laplace transforms. In [40], Kautz adapted the mathematical form of Takenaka-Malmquist orthogonal functions such that they can be exploited for the "transient synthesis" problem in electrical networks. The nature of the problem of network synthesis for prescribed transient response would necessary require that Kautz filter to have a real-valued time domain realization. Then the value of the Kautz's work can be understood from this perspective. The original mathematical form of an orthogonal set has been modified to make sure the inverse Laplace transform of every system transfer function approximated with a linear combination of Kautz functions (with real coefficients) appears as a real-valued time domain function.

The original Kautz functions outlined in [40] have been covering the two specific conditions when all poles are either real or in complex conjugate pairs as follows:

I. The transforms of the functions when all poles are real and lied at $-\alpha_1,-\alpha_2,\ldots,-\alpha_n$:

$$\Phi_n(s)=\sqrt{2\alpha_n}\times\left(\prod_{i=1}^{n-1}\frac{s-\alpha_i}{s+\alpha_i}\right)\times\frac{1}{s+\alpha_n} \qquad \text{(B.1)}$$

The zeros of each function are located at the negative of the poles (all-pass structure) except for the new poles not present in the previous function ($\Phi_{n-1}(s)$) [40].

II. If poles are in complex pairs and lie at: $p_{2\nu-1} = -\alpha_\nu - j\beta_\nu$ and $p_{2\nu} = p^*_{2\nu-1} = -\alpha_\nu + j\beta_\nu$. The Kautz transforms resemble the following:

$$\left.\begin{matrix}\varphi_{2\nu-1}(s)\\ \varphi_{2\nu}(s)\end{matrix}\right\} = \sqrt{2\alpha_\nu} \times \frac{\left[(s-\alpha_1)^2+\beta_1^2\right]\times\cdots\times\left[(s-\alpha_{\nu-1})^2+\beta_{\nu-1}^2\right]\times\left(s\pm|p_{2\nu-1}|\right)}{\left[(s+\alpha_1)^2+\beta_1^2\right]\times\cdots\times\left[(s+\alpha_{\nu-1})^2+\beta_{\nu-1}^2\right]\times\left[(s+\alpha_\nu)^2+\beta_\nu^2\right]}$$

$$; \ (\nu = 1,2,\ldots,\frac{N}{2}) \qquad \text{(B.2)}$$

'*N*' referrers to the number of the poles (order of the model) also $\varphi_{2\nu-1}(s)$ and $\varphi_{2\nu}(s)$ are the basis functions corresponding to the $p_{2\nu-1} = -\alpha_\nu - j\beta_\nu$ and $p^*_{2\nu-1} = p_{2\nu} = -\alpha_\nu + j\beta_\nu$ respectively.

Within a few steps of trivial algebraic manipulation, a shorthand writing form for (B.2) can be obtained like:

$$\left.\begin{matrix}\varphi_{2\nu-1}(s)\\ \varphi_{2\nu}(s)\end{matrix}\right\} = \sqrt{2\alpha_\nu} \times \left(\prod_{k=1}^{2\nu-2}\frac{s+p_k^*}{s-p_k}\right)\times\frac{s\pm|p_{2\nu-1}|}{(s-p_{2\nu-1})(s-p^*_{2\nu-1})} \quad (\nu = 1,2,\ldots,\frac{N}{2}) \qquad \text{(B.3)}$$

The upper (+) sign pertains to $\varphi_{2\nu-1}(s)$ and the lower (-) sign pertains to $\varphi_{2\nu}(s)$.

Appendix C. Adapted Partial Fraction Bases in Continuous Domain

Assume that the transfer function is approximated by linear combination of $\Phi_i(s)$ with rigorously real-valued coefficients and poles strictly occurring either in real or in a complex conjugate pairs. To make sure that the resulting function can absolutely make real-valued impulse response in time domain, the adapted form of the partial fraction bases schemed in below should be utilized for fitting problem formulation.

a) When complex stable[1] poles lie at $p_i = -\alpha_i - j\beta_i$, $p_{i+1} = p_i^* = -\alpha_i + j\beta_i$ for $\alpha_i > 0$ also assuming $\beta_i \geq 0$ does not hurt the generality.

The combination of corresponding partial fraction transforms would resemble to

$$F(s) = C_i \times \left(\frac{1}{s - p_i}\right) + C_{i+1} \times \left(\frac{1}{s - p_i^*}\right),$$

in which residues come in complex conjugate pair $C_i = C_i' + jC_i''$, and $C_{i+1} = C_i^* = C_i' - jC_i''$, when C_i' & $C_i'' \in \mathbb{R}$.

Hence: $F(s) = \left(C_i' + jC_i''\right) \times \left(\frac{1}{s - \left(-\alpha_i - j\beta_i\right)}\right) + \left(C_i' - jC_i''\right) \times \left(\frac{1}{s - \left(-\alpha_i + j\beta_i\right)}\right) =$

[1] The case of stable physical systems is of interest. However, where unstable poles are allowed, still one can resort to the partial fraction bases.

$$C_i^{'} \times \underbrace{\left(\frac{1}{s-p_i} + \frac{1}{s-p_i^*} \right)}_{\phi_i(s)} + C_i^{''} \times \underbrace{\left(\frac{j}{s-p_i} - \frac{j}{s-p_i^*} \right)}_{\phi_{i+1}(s)}$$

♣- Adapted bases for complex conjugate poles would resemble:

$$\Phi_i(s) = \frac{1}{s-p_i} + \frac{1}{s-p_i^*} \tag{C.1}$$

$$\Phi_{i+1}(s) = \frac{j}{s-p_i} - \frac{j}{s-p_i^*} \tag{C.2}$$

b) when p_i is real stable pole lied on negative real axis in S-plane, $p_i = -\alpha_i \quad , \alpha_i > 0$,:

♣- The adapted basis would be as:

$$\Phi_i(s) = \frac{1}{s-p_i} \tag{C.3}$$