City University
London

APPROXIMATE ALGEBRAIC COMPUTATIONS IN CONTROL THEORY

BY

STAVROS FATOUROS

THESIS SUBMITTED FOR THE
AWARD OF THE DEGREE OF
DOCTOR OF PHILOSOPHY

CONTROL ENGINEERING CENTRE
SCHOOL OF ENGINEERING AND MATHEMATICAL SCIENCE
CITY UNIVERSITY
LONDON EC1V 0HB

JULY 2003

DECLARATION

The University Librarian of the City University may allow this thesis to be copied in whole or in part without any reference to the author. This permission covers only single copies, made for study purposes, subject to normal conditions of acknowledgements.

ACKNOWLEDGEMENTS

This thesis is not the work of one person. The encouragement of my family and my friends during the years of my studies has been crucial for me to continue in all difficulties. But first I need to express my gratitude to my supervisor Professor Nicos Karcanias for the unmeasured assistance and guidance, for his generosity in sharing with me his scientific knowledge and vision, for the continuous encouragement and above all for his friendship.

I shall always be grateful to my professors in the Mathematical Department for supporting me in my difficult times in 1998 and especially to Professor Stavros Giotopoulos, Professor Grigorios Kalogeropoulos and Professor Charalambos Papageorgiou. My friends from ΑΕΠ do not need to be mentioned, they know my devotion to them.

I had the privilege to work with Dr George Halikias who unconditionally shared with me his knowledge and ideas in the field of my research. I was also honoured with the kindness and the advice of Dr Stathis Milonidis.

Special thanks to Joan Rivellini and Linda Carr in the departmental office for offering me their unlimited assistance.

I would like also to thank all my friends in the CM322 office and especially Costas, Daniel, Mufid, Ntia, Ramnath, Panagiotis and Dr Ziad Abu el Zeet. I need here to mention the late friend Lackson Chishimba for introducing me to the “culture” of the office; his spirit supervises eternally the CM322 room.

Special thanks to Jimmy Makris not only for his friendship but for his technical support as well. I have also to thank for their continuous encouragement my friends Ermina, Mano, Tania and Dimitris. At last but not least to thank my partner Elena for her understanding and for her emotional support.

My gratitude to my parents Dr Haritini Siouna Fatourou and Dr. Michael Fatouros is the least reward for their moral and financial support throughout the years of my studies. Their inspiration was fundamental in the setting of my personal goals. My sister Evangelia has been a big support, a tender and honestly advisor.

This thesis is finally dedicated to all my teachers and especially to Maria Anestou, Dimitris Pyrgakis, Orestis Katsanos and George Valtas.

ABSTRACT

The present thesis deals with some significant algebraic computations of Control Theory. The main problem examined in the Thesis concerns the properties of the Greatest Common Divisor (GCD) of a set of polynomials; these properties may be investigated using the Sylvester Resultant. New properties of the Sylvester Resultant linked to GCD are established and these lead to canonical factorisations of resultants expressing the extraction of common divisors from the elements of the original set. These results lead to a new representation of the GCD in terms of a canonical factorisation of the Sylvester Resultant obtained by a reduced Sylvester Resultant and a Toeplitz matrix representing the GCD.

The Sylvester resultant factorisation establishes the framework for the characterisation of the "approximate" GCD. The evaluation of the "optimal" approximate GCD and its "strength" comes as a result of the above framework. The problem of approximate factorisation of polynomials, a problem related to root clustering, is also considered and solved using the new techniques. The approximate GCD framework is applied to the case of Linear System properties and metrics measuring distances from fundamental properties are introduced.

Finally, an additional contribution consists of a detailed account of the parameterisation of the family of proper controllers as a solution of a scalar polynomial Diophantine equation.

Table of Contents

Chapter ***1***:

INTRODUCTION

The theory of algebraic and geometric invariants in Linear Systems is instrumental in describing system properties and it is linked to solvability of fundamental Control Theory problems [Rosenbrock, 1979], [Kailath, 1980], [Kucera, 1979]. These invariants are defined on rational, polynomial matrices and matrix pencils under different transformation groups (coordinate, compensation, feedback type) and their computation relies on algebraic algorithms. The use of symbolic tools may thus be considered as natural in developing algorithms for their computation. This introduces some considerable problems with the framework based on exact symbolic tools and motivates the need for approximate algebraic computations, which is the topic considered in this thesis. The underlying assumption behind the use of symbolic computations is that mathematical models always have numerical inaccuracies and this has a significant effect on the selection and development of the computational tools. The existence of certain types and/or values of invariants and system properties may be either generic or nongeneric on a family of linear models. Computing or evaluating nongeneric type or values of invariant and thus associated system properties on models with numerical inaccuracies is crucial for applications. For such cases, symbolic tools fail, since "almost always" leads to a generic solution, which does not represent the "approximate presence" of the value property on the set of models under considerations. The subject of the thesis is the discussion of the fundamentals of this important area of algebraic computations.

The development of a methodology for robust computation of nongeneric algebraic invariants, or nongeneric values of generic ones, has as prerequisites: (a) The development of a numerical linear algebra characterisation of the invariants, which may allow the measurement of degree of presence of the property on every point of the parameter set. (b) The development of special numerical tools, which avoid the introduction of additional errors. (c) The formulation of appropriate criteria which allow the termination of algorithms at certain steps and the definition of meaningful approximate solutions to the algebraic computation problem. It is clear that the formulation of the algebraic problem as an equivalent numerical linear algebra problem, is essential in transforming concepts of algebraic nature to equivalent concepts of analytic character; this property is referred to as numerical reducibility (NR) of the algebraic computation and it depends on the nature of the particular invariant. The last two prerequisites are referred to in short as numerical tools for nongeneric computations (NGC).

That effort goes back to the attempt to introduce the notion of almost zero of a set of polynomials [Karcanias et al., 1983], [Mitrouli et al., 1997] and study the properties of such zeros from the feedback viewpoint. This work was subsequently developed to a methodology for computing the approximate gcd of polynomials using numerical linear algebra methods, such as the ERES and matrix pencils [Karcanias, 1987]. The numerical methods in [Mitrouli et al., 1997], [Karcanias, 1987] have used a variety of procedures for NGC and a richer set of tools is given in [Mitrouli et al., 1995]. An overview of issues and a summary of results on some crucial problems are provided in [Karcanias et al., 1994]. The classification of types of computational problems on numerically uncertain linear models is first considered and then a set of fundamental tools for NGC are examined. Some recent results on the approximate factorization of polynomials and evaluation of the approximate least common multiple (LCM) of a set of polynomials were considered. The study of GCD and LCM of a set of polynomials is central in many algebraic synthesis problems and thus the derivation of methodologies for their approximate definition is crucial for the development of the field of approximate algebraic computations.

The development of robust algebraic computations procedures for engineering type models always has to take into account that the models have certain accuracy and that it is meaningless to continue computations beyond the accuracy of the original data set. In fact, engineering computations are defined not on a single model or a system S, but on a ball of system models $\Sigma\left(S_0, r(\varepsilon)\right)$, where S_0 is a nominal system and $r(\varepsilon)$ is some radius defined by the data error order ε. The result of computations has thus to be representative for the family $\Sigma\left(S_0, r(\varepsilon)\right)$ and not just the particular element of the family. From this viewpoint, symbolic computations carried out on an element of the $\Sigma\left(S_0, r(\varepsilon)\right)$ family may lead to results, which do not reveal the desired properties of the family. Numerical computations have to stop, when we reach the original data accuracy and an approximate solution to the computational task has be given. We consider algebraic computation problems which are numerically reducible (i.e. have a numerical linear algebra equivalent formulation). The classification of such computational tasks according to their behavior under numerical errors is important, since it reveals those requiring special attention; this classification is undertaken here and reveals an important class, that of nongeneric computations which form the main subject of this thesis.

The study of computations of nongeneric invariants uses the notion of genericity. To make the idea of genericity precise, we borrow some terminology from algebraic geometry. Consider polynomials $\varphi(\lambda_1,\dots,\lambda_n)$ with coefficients in $\mathbb{R}$. A variety $V\subset\mathbb{R}^N$ is defined to be the locus of common zeros of a finite number of polynomials $\varphi_1,\dots,\varphi_k$:

$$V=\left\{P\in\mathbb{R}^N:\varphi_i\left(P_1,\dots,P_N\right)=0,\ i\in\underset{\sim}{k}\right\}$$

For example one can prove that the set of all parameters describing a state-space model (A,B,C,D) of fixed dimensions modulo coordinate state transformations in a variety. A property Π on V is merely a function $\Pi:V\to\{0,1\}$, where $\Pi(P)=1$ (or 0) means Π holds (or fails) at P. Let V be a proper variety, we shall say that Π is *generic relative* to V provided $\Pi(P)=0$ only for points $P\in V'\subset V$ where V' is a proper subvariety of V; and that Π is *generic* provided such a V' exists. As V' is a locus of zeros of polynomials in V, the subset of V such that the property is not true is a negligible set (measure zero). On the basis of the above we are led to the following classification of algebraic computations:

Definition (1.1): Numerical computations dealing with the derivation of an approximate value of a property, function, which is nongeneric on a given model set, will be called *nongeneric computations* (NGC). If the value of a function always exists on every element of the model set and depends continuously on the model parameters, then the computations leading to the determination of such values will be called *normal numerical* (NNC). Computational procedures aiming at defining the generic value of a property, function on a given model set (if such values exists), will be called *generic* (GC).

■

On a set of polynomials with coefficients taking values from a certain parameter set, the existence of GCD is nongeneric; numerical procedures that aim to produce an approximate nontrivial value by exploring the numerical properties of the parameter set are typical examples of NG computations and approximate GCD procedures will be considered subsequently. NG computations refer to both continuous and discrete type system invariants. On the other hand, the eigenvalues of a square matrix, or the zeros of a square polynomial matrix are always defined on any model set and their numerical values continuously depend on the numerical values of the parameters set; such cases are examples of NNC computations and their study is covered well in numerical analysis books [Gantmacher, 1988] and are not considered here. For unstructured model sets, the generic value of discrete invariants is an issue that is usually simple and follows by the dimensionality of the matrices involved and genericity arguments. The various techniques which have been developed for the computation of approximate solutions of GCD and LCM are based on methodologies where exact properties of these notions are relaxed and appropriate solutions are sought using a variety of numerical tests. The basis of such approaches is the reduction of the algebraic problems to equivalent linear algebra, which are suitable for the study of approximation problems. A fundamental problem is the difficulty in characterising the accuracy of effectiveness of such methods, as well as, determining whether such solutions are “optimal” in some sense with respect with respect to all other techniques that may offer approximate solutions.

The main objective of this thesis is to introduce a new framework within which approximate solutions may be evaluated in terms of their error and optimal solutions may be determined as the outcome of some appropriate optimisation. A secondary objective of the thesis is the further development of some properties of the underlining algebraic framework which underpin the development of the main objective

Amongst the large number of the algebraic computation problems we focus here on the evaluation of strength of different gcd solutions, the development of the optimal approximate gcd and the application of such results to gcd related problems such as root clustering, approximate factorisation of polynomials and use of the results developed for sets of polynomials to the case of linear systems. The fundamental problems which relate to the main objectives are the representation of gcd of many polynomials in terms of the factorisation of Sylvester resultants and the parameterisation of proper solutions of scalar Diophantine equations as the natural tool for the study of Dynamic assignment and stabilisation problems in Linear Systems.

Chapter 2 provides an introduction to the fundamental of the algebraic problem linked to the Diophantine equation and serves as an overview and motivator for the algebraic computational problems involving gcd and which are considered in the thesis

Chapter 3 provides a detailed account of the parameterisation of the family of the proper controllers of a scalar polynomial Diophantine equation. This involves development of parameterisations, characterisation of McMillan degree of resulting solutions and some other related results linked to the family. These results although developed here for the scalar case may be extended to scalar Diophantine equation in many variables using a similar approach.

Chapter 4 starts as a review of the classical Sylvester results on resultants of many polynomials and their link to the gcd computation. The attempt to provide a new improved proof to the generalised Sylvester resultant result [Barnett, 1983], [Vardoulakis et al., 1978] has led to the characterisation of gcd in terms of a canonical factorisation of original resultant into a product of a reduced resultant and a canonical Toeplitz matrix defined by the coefficients of the gcd. This result does not provide a new procedure for gcd computation but it establishes the algebraic procedure for the extraction of gcd; it is of crucial importance in the development of an analytic framework within which approximate gcd evaluations may be assessed and the optimal approximate gcd can be defined.

Chapter 5 introduces the fundamentals of the approximate gcd evaluation framework. For sets of polynomials for a given number of elements and with fixed the two maximal degrees, a point in the projective space is defined, based on the coefficients of the polynomials in the set. The family of all sets which have a gcd with a given degree d is defined by the properties of the generalised resultant and it is shown to be a special variety of the projective space referred to as the d-gcd variety. The factorisation of the resultant allows us to define for any d-degree approximate gcd a subvariety of the d-gcd variety and thus the "strength" of the approximation , provided by the result of a given numerical method, may be completed as the evaluation of the distance of the given point (set of polynomials) from its subvariety of the d-gcd variety. This distance is worked out as the solution of a simple optimisation problem.

Chapter 6 deals with the characterisation of the optimal approximate gcd of a set of polynomials and although it follows the philosophy of the previous chapter, now provides also the means for computing the best approximate solution of a given degree for the gcd. For a given set of polynomials, the family of all perturbation sets (sets of polynomials which may disturb the original set) produce points (sets of polynomials) which belong to the d-gcd variety. The definition of the best d-degree approximate solution is equivalent to a computation of the distance of the given set from the d-gcd variety. This distance is computed by minimising the Frobenius norm of the resultant characterising the dynamic perturbations. Use of the resultant properties allows the reduction of the complex distance problem into two problems based on independent set of variables which are of considerable simpler nature than the original ones. The first of two optimisation sub problems is based on a standard minimisation of the norm of polynomials in many variables which are defined from the original set; this defines also the "optimal" approximate gcd and the "strength" (distance). The second problem always has a trivial solution when the solution of the first has been obtained. The results of this chapter provide a complete answer to the original objectives, that is the definition of optimal approximate gcd and its strength.

Chapter 7 uses the framework established in the last two chapters for the study of root clustering of polynomials, expressed as an approximate normal factorisation of a given polynomial. The results on the normal factorisation of a given polynomial [Karcanias et al., 2002] are based on the gcd algorithms and this allows the combinations of normal factorisation results to the approximate gcd to give solutions to the challenging root clustering problem.

Chapter 8 is a first attempt to extend the approximate gcd framework for a set of polynomials to the study of approximate matrix divisors of polynomial matrices and the study of properties inferred by such factorisations when applied to linear system problems. We use exterior algebra to associate with any polynomial matrix, a polynomial multivector [Marcus, 1973]. This expresses the classical Plücker embedding of an affine space into a projective space and yields the set of polynomials which is used for our study. These sets of polynomials are not random but they are characterised by the property of decomposability of multivectors which implies that the corresponding point of the projective space belongs to the Grassmann variety. The framework of the approximate gcd previously established is applicable in a straightforward manner to two partial cases which guarantee decomposability. In all other cases the optimal gcd results are estimates of the solution sort. In the general case the distance problem formulated as the distance between the given decomposable polynomial set and subvariety of the d-gcd variety defined as an intersection with the corresponding Grassmann variety. The results obtained here provide a characterisation of approximate zero polynomials, approximate of (input/output) zero polynomials and the errors or strength of approximation provide estimates of important system properties such as distance of a system from the set of uncontrollable systems and/or unobservable systems.

Finally Chapter 9 summarises the achievement and describes issues related to future research.

Chapter 2:

POLE ASSINGNMENT FOR SISO SYSTEMS: THE DIOPHANTINE EQUATION APPROACH

2.1. INTRODUCTION

The aim of this chapter is to introduce the fundamentals of the algebraic approach to the design of *Single Input*, *Single Output*, (SISO) linear control systems and provide some motivation for the algebraic computations problems which are considered in the thesis. Furthermore, the problem of characterising the family of proper solutions of polynomial Diophantine equations is considered in some detail. More specifically Chapter 2 considers the classical problem of pole assignment by dynamic compensation that introduces the Diophantine equation approach. Issues related to parameterisation of solutions and study of existence of proper solution are considered here. An alternative, Linear Algebra formulation of the problem is subsequently given and this opens the way for an algorithmic investigation of special types of solutions.

2.2. THE GENERAL FEEDBACK CONFIGURATION

We consider the general feedback configuration shown below

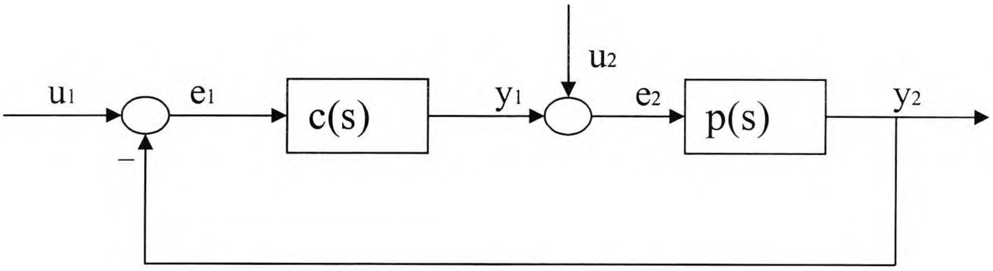


Figure (2.1): General Feedback Configuration

where $c(s)$ denotes the controller and $p(s)$ the plant transfer function respectively. The above configuration is described by the following equations:

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} - \begin{bmatrix} 0 & p \\ -c & 0 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}, \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} c & 0 \\ 0 & p \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{2.1}$$

where e_i, u_i, y_i denote Laplace transforms of the corresponding signals and c, p denote in short the corresponding transfer functions. If we define

$$\underline{e}=\begin{bmatrix} e_1 \\ e_2 \end{bmatrix}, \quad \underline{u}=\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad \underline{y}=\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad F=\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad G=\begin{bmatrix} c & 0 \\ 0 & p \end{bmatrix} \tag{2.2}$$

then (2.1) yield

$$\underline{e}=\underline{u}-FG\underline{e} \ , \quad \underline{y}=G\underline{e} \tag{2.3}$$

Given that for a SISO system, if $1+c(s)p(s)\neq 0$, we may always solve (2.3) and thus obtain

$$\underline{e}=(I_2+FG)^{-1}\cdot\underline{u}\triangleq H(p,c)\underline{u} \tag{2.4}$$

where $H(p,c)$ is the transfer function inputs to errors and is expressed as

$$H(p,c)=\begin{bmatrix} (1+pc)^{-1} & -p(1+cp)^{-1} \\ c(1+pc)^{-1} & (1+cp)^{-1} \end{bmatrix} \tag{2.5}$$

and the transfer function from inputs to the outputs is defined by

$$\underline{y}=W(p,c)\underline{u}=G(I_2+FG)^{-1}\underline{u} \tag{2.6}$$

and it is expressed as

$$W(p,c)=\begin{bmatrix} c(1+cp)^{-1} & -cp(1+cp)^{-1} \\ cp(1+cp)^{-1} & p(1+cp)^{-1} \end{bmatrix} \tag{2.7}$$

The above configuration is quite general and it has the following interpretations:

a) If u_1 is the signal to track and u_2 is a plant input disturbance, then we have the classical control scheme, where c is a precompensation.

b) If u_2 is the reference signal and u_1 is the measurement sensor noise disturbance, then $c(s)$ is interpreted as a feedback compensator.

The function $f(s)=1+c(s)p(s)$ is known as the *return difference* for the feedback configuration. If we denote

$$c(s)=\frac{n_c(s)}{d_c(s)}\ ,\quad p(s)=\frac{n_p(s)}{d_p(s)} \tag{2.8}$$

then clearly

$$f(s)=1+c(s)p(s)=\frac{n_c(s)n_p(s)+d_c(s)d_p(s)}{d_c(s)d_p(s)}=\frac{\varphi_c(s)}{\varphi_0(s)} \tag{2.9}$$

where $\varphi_0(s)=d_c(s)d_p(s)$ is referred to as the *open-loop pole polynomial* and

$$\varphi_c(s)=n_c(s)n_p(s)+d_c(s)d_p(s) \tag{2.10}$$

is defined as the *closed-loop polynomial* of the feedback configuration. The reason for the latter terminology is that $\varphi_c(s)$ is the denominator of all transfer function elements of $H(p,c)$ and $W(p,c)$ as it can be readily verified. In fact, for the SISO case we have

$$W(p,c)=G\cdot H(p,c)=\frac{1}{\varphi_c(s)}\begin{bmatrix} n_c(s)d_p(s) & -n_c(s)n_p(s) \\ n_c(s)n_p(s) & d_c(s)n_p(s) \end{bmatrix} \tag{2.11}$$

<u>**Remark (2.1):**</u> If the plant and the controller are represented by their transfer functions, then the feedback configuration of Figure (2.1) has $\varphi_c(s)$ as its pole polynomial.

■

2.3. ALGEBRAIC SYNTHESIS PROBLEMS AND THE DIOPHANTINE EQUATION APPROACH

Given the system represented by the transfer function $p(s)$, or the coprime pair $(n_p(s), d_p(s))$, the fundamental *Pole Assignment Problem*, (PAP) [Kucera 1979], is to define whether there exists a controller $c(s)$. Represented by the coprime pair $(n_c(s), d_c(s))$ such that

$$n_p(s)n_c(s) + d_p(s)d_c(s) = \varphi(s) \ , \quad \varphi(s) \in \mathbb{R}[s] \tag{2.12}$$

where $\varphi(s)$ is a given polynomial. If in addition we require that

$$\deg\{n_c(s)\} \leq \deg\{d_c(s)\} \tag{2.13}$$

then we have the Pole Assignment Problem with Proper controllers, or *Proper-PAP* (P-PAP) [Kucera 1979],.

In addition to the PAP or the P-PAP we have the special case of the *Stabilisation Problem* (SP) if we require that $\varphi(s)$ is stable rather than having arbitrarily assignable roots.

Equation (2.12) is known as *Diophantine Equation*. A special form of the latter is obtained when $\varphi(s) = 1$, i.e.

$$n_p(s)n_c(s) + d_p(s)d_c(s) = 1 \tag{2.14}$$

The solvability of the general equation (2.12) is facilitated by the study of (2.14) first.

Theorem (2.1): Necessary and sufficient condition for equation (2.14) to have a solution is that the pair $(n_p(s), d_p(s))$ is coprime

Proof:

(Necessity): Let us assume that $t(s)$ is the greatest common divisor of $(n_p(s), d_p(s))$. Then (2.14) may be written as

$$t(s)\left(n_c(s)n_p'(s)+d_c(s)d_p'(s)\right)=1$$

Since $t(s)\neq c$, then for all roots of $t(s)$, say s_i, we get that

$$t(s_i)\left(n_c(s_i)n_p'(s_i)+d_c(s_i)d_p'(s_i)\right)=0\neq 1$$

(Sufficiency): Assume that $(n_p(s), d_p(s))$ is a coprime pair. Equation (2.14) may be written as

$$\begin{bmatrix} n_p(s) & d_p(s) \end{bmatrix}\begin{bmatrix} n_c(s) \\ d_c(s) \end{bmatrix}=1 \tag{2.14a}$$

Let $Q(s)\in\mathbb{R}^{2\times 2}[s]$, $Q(s)=c\in\mathbb{R}$, $c\neq 0$ be a unimodular matrix which reduces $\left[n_p(s), d_p(s)\right]$ to its Smith form. Then

$$\begin{bmatrix} n_p(s) & d_p(s) \end{bmatrix}\begin{bmatrix} q_{11}(s) & q_{12}(s) \\ q_{21}(s) & q_{22}(s) \end{bmatrix}=\begin{bmatrix} 1 & 0 \end{bmatrix} \tag{2.15}$$

from which is clear that

$$\begin{bmatrix} n_p(s) & d_p(s) \end{bmatrix}\begin{bmatrix} q_{11}(s) \\ q_{21}(s) \end{bmatrix}=1$$

and thus equation (2.14) has a solution for $n_c(s)=q_{11}(s)$, $d_c(s)=q_{21}(s)$

■

Corollary (2.1): Let $Q(s)\in\mathbb{R}^{2\times 2}[s]$, $|Q(s)|=c\in\mathbb{R}$, $c\neq 0$ be a right unimodular transformation matrix which reduces $\left[n_p(s), d_p(s)\right]$ to its Smith form

$$\begin{bmatrix} n_p(s) & d_p(s) \end{bmatrix} \begin{bmatrix} q_{11}(s) & q_{12}(s) \\ q_{21}(s) & q_{22}(s) \end{bmatrix} = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

The general solution of (2.14) is then given by

$$\begin{cases} n_c(s) = q_{11}(s) + t(s) q_{12}(s) \\ d_c(s) = q_{21}(s) + t(s) q_{22}(s) \end{cases} \text{where, } t(s) \in \mathbb{R}[s] \text{ arbitrary} \tag{2.16}$$

Proof:

By (2.15) it follows that the general unimodular matrix which reduces $\left[n_p(s), d_p(s)\right]$ to its Smith form is

$$Q'(s) = \begin{bmatrix} q_{11}(s) & q_{12}(s) \\ q_{21}(s) & q_{22}(s) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t(s) & k \end{bmatrix} = \begin{bmatrix} q'_{11}(s) & q'_{12}(s) \\ q'_{21}(s) & q'_{22}(s) \end{bmatrix}$$

and thus the first column of $Q'(s)$ gives a family of solutions. In order to show that (2.16) gives the whole family, let us assume that $\left(n_c(s), d_c(s)\right)$, $\left(n'_c(s), d'_c(s)\right)$ are two solutions, then

$$\begin{cases} n_p(s) n_c(s) + d_p(s) d_c(s) = 1 \\ n_p(s) n'_c(s) + d_p(s) d'_c(s) = 1 \end{cases}$$

and by subtracting we have

$$n_p(s)\left(n_c(s) - n'_c(s)\right) + d_p(s)\left(d_c(s) - d'_c(s)\right) = 0$$

or

$$\left[n_p(s), d_p(s)\right] \begin{bmatrix} n_c(s) - n'_c(s) \\ d_c(s) - d'_c(s) \end{bmatrix} = 0$$

By (2.15) it is clear that the second column of $Q(s)$ defines a least degree basis [Marcus et al., 1969] for the right null space of $\left[n_p(s), d_p(s)\right]$. Thus

$$\begin{bmatrix} n_c(s) - n'_c(s) \\ d_c(s) - d'_c(s) \end{bmatrix} = \begin{bmatrix} q_{12}(s) \\ q_{22}(s) \end{bmatrix} t(s)$$

By assuming that one solution is defined by the first column of $Q(s)$, the result follows.

■

Remark (2.2): Note that $q_{12}(s)=d_p(s)$, $q_{22}(s)=-n_p(s)$ and thus the general solution of (2.14) may be written as

$$\begin{cases} n_c(s)=q_{11}(s)+t(s)d_p(s) \\ d_c(s)=q_{21}(s)-t(s)n_p(s) \end{cases}, \text{ where, } t(s)\in\mathbb{R}[s] \text{ arbitrary} \tag{2.17}$$

■

We may now consider the solution of (2.12). Note that if $(n_c(s),d_c(s))$ is a solution of (2.14) then by multiplying both sides of (2.14) by $\varphi(s)$ we have

$$\{\varphi(s)n_c(s)\}n_p(s)+\{\varphi(s)d_c(s)\}d_p(s)=\varphi(s) \tag{2.18a}$$

and thus, $n_c'(s)=\varphi(s)n_c(s)$, $d_c'(s)=\varphi(s)d_c(s)$ is a solution of (2.12). However such a solution is not acceptable since $q(s)\triangleq y_2(s)/u_1(s)$ is

$$q(s)=\frac{n_c(s)\varphi(s)n_p(s)}{\varphi(s)}=n_c(s)n_p(s)\in\mathbb{R}[s] \tag{2.18b}$$

Other types of solution of (2.12) are examined next. We first note that (2.12) may be written as

$$\left[n_p(s),d_p(s),-\varphi(s)\right]\begin{bmatrix} n_c(s) \\ d_c(s) \\ 1 \end{bmatrix}=0 \tag{2.19}$$

Thus, the solutions are those vectors in $\mathcal{N}_r\left(\left[n_p(s),d_p(s),-\varphi(s)\right]\right)$ which are of the type $\left[a(s),b(s),1\right]$. If $Q(s)$ is the unimodular matrix that reduces $\left[n_p(s),d_p(s)\right]$ to its Smith form $[1,0]$, then

$$\left[n_p(s),d_p(s),-\varphi(s)\right]\left[\begin{array}{cc|c} q_{11}(s) & q_{12}(s) & q_{11}(s)\varphi(s) \\ q_{21}(s) & q_{22}(s) & q_{21}(s)\varphi(s) \\ \hdashline 0 & 0 & 1 \end{array}\right]=[1,0,0] \tag{2.19a}$$

and thus a least degree basis for $\mathcal{N}_r\left(\left[n_p(s),d_p(s),-\varphi(s)\right]\right)$ (that is a matrix which has no finite zeros) is defined by

$$T(s)=\begin{bmatrix} d_p(s) & q_{11}(s)\varphi(s) \\ -n_p(s) & q_{21}(s)\varphi(s) \\ 0 & 1 \end{bmatrix} \tag{2.20}$$

The general vector in $\mathcal{N}_r\left(\left[n_p(s),d_p(s),-\varphi(s)\right]\right)$ may be expressed as

$$\underline{x}(s)=t(s)\begin{bmatrix} d_p(s) \\ -n_p(s) \\ 0 \end{bmatrix}+z(s)\begin{bmatrix} q_{11}(s)\varphi(s) \\ q_{21}(s)\varphi(s) \\ 1 \end{bmatrix} \tag{2.20a}$$

and thus the general vector of (2.12) ($z(s)=1$) is

$$\underline{x}(s)=t(s)\begin{bmatrix} d_p(s) \\ -n_p(s) \\ 0 \end{bmatrix}+\begin{bmatrix} q_{11}(s)\varphi(s) \\ q_{21}(s)\varphi(s) \\ 1 \end{bmatrix} \tag{2.21}$$

<u>**Corollary (2.2):**</u> The general solution of the Diophantine equation

$$n_c(s)n_p(s)+d_c(s)d_p(s)=\varphi(s) \tag{2.12}$$

is given by

$$\begin{cases} n_c(s)=q_{11}(s)\varphi(s)+t(s)d_p(s) \\ d_c(s)=q_{21}(s)\varphi(s)-t(s)n_p(s) \end{cases} \tag{2.21a}$$

■

Note that now the closed-loop transfer function $q(s)=y_2(s)/u_1(s)$ is

$$q(s)=\frac{\left\{q_{11}(s)\varphi(s)+t(s)d_p(s)\right\}n_p(s)}{\varphi(s)} \tag{2.21a}$$

The question that still remains to be answered is how we choose $t(s)$ such that $c(s)$ is proper and $q(s)$ is also proper. This problem will be considered in detail on Chapter 3.

<u>**Example (2.1):**</u> Let $p(s)=\dfrac{(s+1)}{s(s-1)}$, i.e. $n_p(s)=s+1$, $d_p(s)=s^2-s$. Then $\left[s+1,s^2-s\right]$ may be transformed to its Smith form as :

$$\begin{tikzcd}
\left[s+1,s^2-s\right] \arrow[d] & \begin{bmatrix}1 & -s\\ 0 & 1\end{bmatrix}=Q_1(s) \arrow[l] \\
\left[s+1,-2s\right] \arrow[d] & \begin{bmatrix}1 & 0\\ 0 & 1/2\end{bmatrix}=Q_2(s) \arrow[l] \\
\left[s+1,-s\right] \arrow[d] & \begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix}=Q_3(s) \arrow[l] \\
\left[1,-s\right] & \begin{bmatrix}1 & s\\ 0 & 1\end{bmatrix}=Q_4(s) \arrow[l]
\end{tikzcd}$$

Thus

$$Q(s)=Q_1(s)Q_2(s)Q_3(s)Q_4(s)=\frac{1}{2}\begin{bmatrix}2-s & -s^2+s\\ 1 & s+1\end{bmatrix}$$

verify that

$$\left[s+1, s^2-s\right]\begin{bmatrix}2-s & -s^2+s\\ 1 & s+1\end{bmatrix}\frac{1}{2}=[1,0]$$

and

$$q_{11}(s)=\frac{1}{2}(2-s),\ q_{21}(s)=\frac{1}{2}$$

From which

$$q(s)=\frac{\left\{\frac{1}{2}(2-s)\varphi(s)+t(s)\left(s^2-s\right)\right\}(s+1)}{\varphi(s)}$$

$$c(s)=\frac{\frac{1}{2}(2-s)\varphi(s)+t(s)\left(s^2-s\right)}{\frac{1}{2}\varphi(s)-t(s)(s+1)}=\frac{(2-s)\varphi(s)+2t(s)\left(s^2-s\right)}{\varphi(s)-2t(s)(s+1)}$$

Let us chose a $t(s)$ such that $\varphi(s)=s^2+as+b$. One way of doing that is to carry the division

$$(2-s)\varphi(s)=-2t(s)\left(s^2-s\right)+r(s)$$

from which $t(s)=s$ and $r(s)=(1-a)s^2+(2a-b)s+2b$. Then

$$c(s)=\frac{r(s)}{s^2+as+b-s(s+1)}=\frac{(1-a)s^2+(2a-b)s+2b}{(a-1)s+b}$$, non-proper

Similarly, let us choose $t(s)$ by the following division

$$\varphi(s)=s^2+as+b=s(s+1)+(a-1)s+b$$

i.e. $t(s)=s$, then this is clearly the same non-proper solution. If $t(s)=c\in\mathbb{R}$ then again is a non-proper compensator. Similarly, it may be verified that any other value of $t(s)=\gamma s+\delta$ or $t(s)$ with $\deg\{t(s)\}>1$ gives rise to a non-proper compensator. Thus, there is no proper solution for $\varphi(s)$ of second order. The procedure should be repeated for $\varphi(s) = s^3 + as^2 + bs + \gamma$.

■

2.4. DISCUSSION

The problem of the polynomial Diophantine equation related to the pole assignment of SISO system has been introduced and the family of solutions have been classified. It has been obvious from Example (2.1) that is difficult to find a proper compensator. The problem may be avoided by working in the ring of proper rational functions, or proper and stable rational functions. An alternative approach to the study of properness, that reduces the overall study of solvability of Diophantine equation to a linear algebra problem, is considered on Chapter 3.

Chapter 3:

POLYNOMIAL DIOPHANTINE EQUATIONS, TOEPLITZ MATRICES AND PROPERTIES OF SOLUTION

3.1. INTRODUCTION

The use of the Smith form for the solution of polynomial Diophantine equations implies resorting to symbolic computations. An alternative approach is formulated in this chapter that reduces the problem to standard linear algebra and thus allows the use of numerical methods. This alternative method reduces the problem to the investigation of a system expressed with Toeplitz matrices and apart from its computational advantages, also permits the study of a number of theoretical tasks, which are harder to define in the initial algebraic setup.

Firstly we will reduce the problem of the polynomial Diophantine equation to its Toeplitz matrix form and we will investigate the solvability of the system. Then the main task will be to examine if the solutions are proper. A number of results will be given for this purpose. Finally, some further properties of the family of solutions related to McMillan degree are considered

3.2. TOEPLITZ MATRIX FORMULATION OF THE DIOPHANTINE EQUATION

Let us consider the system and controller represented by the transfer functions

$$p(s)=\frac{n_p(s)}{d_p(s)}=\frac{b_0 s+b_1 s+\cdots+b_m s^m}{a_0+a_1 s+\cdots+a_{n-1}s^{n-1}+s^n}\equiv\frac{b_p(s)}{a_p(s)} \tag{3.1}$$

$$c(s)=\frac{n_c(s)}{d_c(s)}=\frac{c_0 s+c_1 s+\cdots+c_\mu s^\mu}{d_0 s+d_1 s+\cdots+d_\nu s^\nu} \tag{3.2}$$

We assume that $\left(n_c(s),d_c(s)\right)$ is a coprime pair. Thus the pair $\left(n_c(s),d_c(s)\right)$, which satisfies the Diophantine equation (2.14) is also coprime. We consider now the Diophantine equation

$$n_p(s)n_c(s)+d_p(s)d_c(s)=\varphi(s) \tag{3.3}$$

where it is assumed that $\varphi(s)$ is an arbitrary polynomial with $\deg\varphi(s)=k=\max\{m+\mu,n+\nu\}$ and expressed as

$$\varphi(s)=\varphi_0+\varphi_1 s+\ldots+\varphi_{k-1}s^{k-1}+s^k \tag{3.4}$$

Substituting (3.1), (3.2) and (3.4) into (3.3) and using the notation of the Toeplitz matrices, it may be shown that polynomials $n_c(s),d_c(s)$ with degrees μ,ν respectively, which solve (3.3), exist if and only if the following equation has a solution

$$\underline{e}^{t}_{m+\mu}(s)T_\mu(\underline{b}_p)\underline{c}_\mu+\underline{e}^{t}_{n+\nu}(s)T_\nu(\underline{a}_p)\underline{d}_\nu=\underline{e}^{t}_{k}(s)\underline{\varphi}_k\ , \tag{3.5a}$$

where

$$T_\mu(\underline{b}_p)=\begin{array}{cc} \longleftarrow\ \mu+1\ \longrightarrow & \\ \begin{bmatrix} b_0 & 0 & \cdots & 0\\ b_1 & b_0 & & \vdots\\ \vdots & b_1 & \ddots & \\ b_m & \vdots & \ddots & b_0\\ 0 & b_m & & b_1\\ \vdots & 0 & \ddots & \vdots\\ & & & \\ 0 & \cdots & 0 & b_m \end{bmatrix} & \updownarrow\ m+\mu+1 \end{array},\quad T_\nu(\underline{a}_p)=\begin{array}{cc} \longleftarrow\ \nu+1\ \longrightarrow & \\ \begin{bmatrix} a_0 & 0 & \cdots & 0\\ a_1 & a_0 & & \vdots\\ \vdots & a_1 & \ddots & \\ a_m & \vdots & \ddots & a_0\\ 0 & a_m & & a_1\\ \vdots & 0 & \ddots & \vdots\\ & & & \\ 0 & \cdots & 0 & a_m \end{bmatrix} & \updownarrow\ n+\nu+1 \end{array}$$

$$e^{t}_{i}(s)=[1,s,\ldots,s^{i}] \tag{3.5b}$$

The above equation is readily reduced to an equivalent matrix formulation introduced below:

Proposition 3.1: The solution of the Diophantine equation (3.3) is reduced to the study of solvability of the following matrix equation:

i) If $n+\nu>m+\mu$ and $\rho=n+\nu-m-\mu$, then (3.5) yields:

$$\begin{bmatrix} T_{\mu}\left(\underline{b}_p\right) \\ \hdashline \mathbf{0}_{\rho,\mu+1} \end{bmatrix}\underline{c}_{\mu}+T_{\nu}\left(\underline{a}_p\right)\underline{d}_{\nu}=\underline{\varphi}_k \tag{3.6a}$$

ii) If $n+\nu<m+\mu$ and $\rho'=m+\mu-n-\nu$, then (3.5) yields:

$$T_{\nu}\left(\underline{b}_p\right)\underline{c}_{\nu}+\begin{bmatrix} T_{\mu}\left(\underline{a}_p\right) \\ \hdashline \mathbf{0}_{\rho',\nu+1} \end{bmatrix}\underline{d}_{\mu}=\underline{\varphi}_k \tag{3.6b}$$

iii) If $n+\nu=m+\mu$ then (3.5) is equivalent to:

$$T_{\mu}\left(\underline{b}_p\right)\underline{c}_{\mu}+T_{\nu}\left(\underline{a}_p\right)\underline{d}_{\nu}=\underline{\varphi}_k \tag{3.6c}$$

■

Note that $T_{\mu}\left(\underline{b}_p\right)\in\mathbb{R}^{(m+\mu+1)\times(\mu+1)}$, $T_{\nu}\left(\underline{a}_p\right)\in\mathbb{R}^{(n+\nu+1)\times(\nu+1)}$, $\underline{c}_{\mu}\in\mathbb{R}^{\mu+1}$, $\underline{d}_{\nu}\in\mathbb{R}^{\nu+1}$ and $\underline{\varphi}_k\in\mathbb{R}^{k+1}$. The above conditions may be expressed in a more compact form as shown below:

$$T_{\mu,\nu}^{m,n}\left(\underline{b}_p,\underline{a}_p\right)\underline{t}_{\mu,\nu}=\underline{\varphi}_k \tag{3.7}$$

where $\underline{t}_{\mu,\nu}=\left[\underline{c}_{\mu}^{t}\underline{d}_{\nu}^{t}\right]^{t}$ and

$$T_{\mu,\nu}^{m,n}\left(\underline{b}_p,\underline{a}_p\right)\underline{t}_{\mu,\nu}\triangleq\left[\begin{array}{c|c} \begin{matrix} T_{\mu}\left(\underline{b}_p\right) \\ \hdashline \mathbf{0}_{\rho,\mu+1} \end{matrix} & T_{\nu}\left(\underline{a}_p\right) \end{array}\right],\ \text{if}\ \ \rho=n+\nu-m-\mu>0 \tag{3.8a}$$

$$T_{\mu,\nu}^{m,n}\left(\underline{b}_p,\underline{a}_p\right)\underline{t}_{\mu,\nu}\triangleq\left[\begin{array}{c|c} T_{\mu}\left(\underline{b}_p\right) & \begin{matrix} T_{\nu}\left(\underline{a}_p\right) \\ \hdashline \mathbf{0}_{\rho',\nu+1} \end{matrix} \end{array}\right],\ \text{if}\ \ \rho'=m+\mu-n-\nu>0 \tag{3.8b}$$

$$T_{\mu,\nu}^{m,n}\left(\underline{b}_p,\underline{a}_p\right)\underline{t}_{\mu,\nu}\triangleq\left[T_{\mu}\left(\underline{b}_p\right)\ \middle|\ T_{\nu}\left(\underline{a}_p\right)\right],\ \text{if}\ \ \rho=n+\nu-m-\mu=0 \tag{3.8c}$$

The study of equations (3.7) for different values of μ,ν indices is the subject of the following investigation. The pair of indices (μ,ν) characterises the different types of solutions, since the different types of solutions, since $\mu\triangleq\deg\left[n_c(s)\right]$, $\nu\triangleq\deg\left[d_c(s)\right]$

and will be referred to as the *order* of the solution. The order is significant for the parameterisation and it is linked to the notion of McMillan degree introduced below:

Definition (3.1): Let $c(s)=n_c(s)/d_c(s)\in\mathbf{R}(s)$, where $n_c(s)$, $d_c(s)$ are coprime and $\deg\left[n_c(s)\right]=\mu$, $\deg\left[d_c(s)\right]=\nu$. Them number defined as

$$\delta_M\left(c(s)\right)\triangleq\max(\mu,\nu) \tag{3.9}$$

is defined as the *extended McMillan degree* of $c(s)$. ∎

It is well known that $\delta_M\left(c(s)\right)$ denotes the total number of finite and infinite poles of the rational function $c(s)$. A parameterisation of solutions may be done in terms of δ_M number. Before we embark to the study of solutions, we give a useful alternative interpretation of Proposition (3.1).

Theorem (3.1): The Diophantine equation (3.3), or equivalently the linear system (3.7) has a solution of order (μ,ν) for any arbitrary vector $\underline{\varphi}_k$, $k=\max\{\mu+m,\nu+n\}$ if and only if

$$\operatorname{rank}\left\{T_{\mu,\nu}^{m,n}\left(\underline{b}_p,\underline{a}_p\right)\right\}=\max\{\mu+m,\nu+n\} \tag{3.10}$$

∎

The proof readily follows by inspections of dimensions of the blocks (3.7). This alternative form of the result allows the formulation of algorithmic procedures and be used subsequently. We first notice the following important result.

Theorem (3.2): Let $p(s)=n_p(s)/d_p(s)$ be a plant with $m=\deg\left[n_p(s)\right]$, $n=\deg\left[d_p(s)\right]$, $\left(n_p(s),d_p(s)\right)$ coprime. There always exists a solution $c(s)=n_c(s)/d_c(s)$ with order $(n-1,m-1)$ for arbitrary polynomial $\varphi(s)$,

$\deg\left[\varphi(s)\right]=\max\{m+\mu,n+\nu\}$. Furthermore, the coefficient vectors $\underline{c}_{n-1}$, $\underline{d}_{m-1}$ are respectively defined by

$$\left[\begin{array}{c}\underline{c}_{n-1}\\ \hdashline \underline{d}_{m-1}\end{array}\right]=\left\{T^{m,n}_{n-1,m-1}\left(\underline{b}_p,\underline{a}_p\right)\right\}^{-1}\underline{\varphi}_k \tag{3.11}$$

Proof:

For $\mu=n-1$, $\nu=m-1$, $T^{m,n}_{n-1,m-1}\left(\underline{b}_p,\underline{a}_p\right)\in\mathbb{R}^{(n+m)\times(n+m)}$ and thus

$$T^{m,n}_{n-1,m-1}\left(\underline{b}_p,\underline{a}_p\right)=\left[T_{n-1}\left(\underline{b}_p\right)\mid T_{m-1}\left(\underline{a}_p\right)\right] \tag{3.11a}$$

To prove the result, it is sufficient to show that $T^{m,n}_{n-1,m-1}\left(\underline{b}_p,\underline{a}_p\right)$ has full rank; in fact, if it has full rank, then (3.7) has a solution and it is given as in (3.11).

We consider $T^{m,n}_{n-1,m-1}\left(\underline{b}_p,\underline{a}_p\right)=\left[T_{n-1}\left(\underline{b}_p\right)\mid T_{m-1}\left(\underline{a}_p\right)\right]$. By reordering the two blocks, we produce $\left[T_{m-1}\left(\underline{a}_p\right)\mid T_{n-1}\left(\underline{b}_p\right)\right]$ which by transposition leads to

$$R\left(\underline{b}_p(s)\underline{a}_p(s)\right)=\left[\begin{array}{ccccccccc}
1 & a_{n-1} & \cdots & & a_0 & 0 & \cdots & & 0\\
0 & 1 & a_{n-1} & \cdots & a_1 & a_0 & 0 & \cdots & 0\\
\vdots & \ddots & \ddots & & & & \ddots & & \\
0 & \cdots & 0 & 1 & a_{n-1} & \cdots & & & a_0\\
\hdashline
b_m & b_{m-1} & \cdots & & b_0 & 0 & & & 0\\
0 & b_m & \cdots & & b_1 & b_0 & \cdots & & 0\\
\vdots & \ddots & \ddots & & & \ddots & \ddots & & \vdots\\
0 & \cdots & 0 & b_m & b_{m-1} & \cdots & b_1 & & b_0
\end{array}\right]\begin{array}{c} \\ m \\ \\ \\ \\ n \\ \\ \\ \end{array}$$

$$R\left(\underline{b}_p(s)\underline{a}_p(s)\right)\in\mathbb{R}^{(n+m)\times(n+m)} \tag{3.11b}$$

However $R\left(\underline{b}_p(s)\underline{a}_p(s)\right)$ is the Sylvester resultant of $n_p(s),d_p(s)$ polynomials, which are coprime and thus $\left|R\left(\underline{b}_p(s)\underline{a}_p(s)\right)\right|\neq 0$ [Barnett, 1983]. However, $R\left(\underline{b}_p(s)\underline{a}_p(s)\right)$ and $T^{m,n}_{n-1,m-1}\left(\underline{b}_p,\underline{a}_p\right)$ are equivalent and thus they have the same rank. ■

The above result provides the means for the computation of a particular solution of the Diophantine equation without resorting to the use of algebraic means. (Smith form and Symbolic computations) and thus allows the parameterisation of the family of solutions. The particular solution is given by

$$c(s)=\frac{n_c^*(s)}{d_c^*(s)}=\frac{c_0+c_1s+\ldots+c_{n-2}s^{n-2}+c_{n-1}s^{n-1}}{d_0+d_1s+\ldots+d_{m-2}s^{m-2}+d_{m-1}s^{m-1}} \tag{3.12a}$$

where the c_i, d_i are defined by solving the matrix equation

$$\left[T_{n-1}\left(\underline{b}_p\right) \mid T_{m-1}\left(\underline{a}_p\right)\right]\left[\begin{array}{c}\underline{c}_{n-1}^* \\ \hdashline \underline{d}_{m-1}^*\end{array}\right]=\underline{\varphi}_{m+n} \tag{3.12b}$$

The matrix $T_{m,n}\left(\underline{b}_p,\underline{a}_p\right)=\left[T_{n-1}\left(\underline{b}_p\right) \mid T_{m-1}\left(\underline{a}_p\right)\right]\in\mathbb{R}^{(m+n)\times(m+n)}$ has always full rank and can be used in the study of properties of the solution of the Diophantine equation. The solution $\left(n_c^*(s),d_c^*(s)\right)$ will be called the *Sylvester Solution* of the Diophantine equation. If $\mathcal{W}\left(\underline{b}_p,\underline{a}_p,\underline{\varphi}\right)$ denotes the whole family of solutions of (3.3), then according to corollary (2.2), this family may be expressed as:

Corollary (3.1): If $\left(n_c^*(s),d_c^*(s)\right)$ is the Sylvester solution pair of the Diophantine equation (3.3) for the $m+n-1$ degree polynomial $\varphi(s)$, $\mathcal{W}\left(\underline{b}_p,\underline{a}_p,\underline{\varphi}\right)$ is defined by

$$\left\{\begin{array}{l} n_c(s)=n_c^*(s)+t(s)d_p(s) \\ d_c(s)=d_c^*(s)-t(s)n_p(s)\end{array}\right., \quad t(s)\in\mathbb{R}[s] \text{ arbitrary} \tag{3.13}$$

■

Remark (3.1): For a strictly proper system $(m < n)$ the Sylvester solution is clearly non-proper. However, for the of bi-proper case $(m = n)$, or for the case of non-proper systems $(m > n)$ the Sylvester Solution is proper. ■

3.3. PROPER SOLUTIONS OF THE DIOPHANTINE EQUATION AND PARAMETRISATION ISSUES

The above analysis provides alternative tools for studying the properties of solutions of Diophantine equations. Some issues related to properness and McMillan degree parameterisation are considered next. We first note;

Corollary (3.2): Let $\left(n_c^*(s), d_c^*(s)\right)$ be the Sylvester solution of the Diophantine equation defined by (3.12b) and let $m < n$. There always exists a proper solution $\left(\tilde{n}_c(s), \tilde{d}_c(s)\right)$with McMillan degree $n-1$ which arbitrarily assigns any polynomial of degree $2n-1$. This solution is defined by

$$\left[\begin{array}{c:c} \begin{array}{c} T_{n-1}\left(\underline{b}_p\right) \\ \hdashline \mathbf{0} \end{array} & T_{n-1}\left(\underline{a}_p\right) \end{array}\right]\left[\begin{array}{c} \tilde{c}_{n-1} \\ \hdashline \tilde{d}_{n-1} \end{array}\right] = \underline{\tilde{\varphi}}_{2n-1} \tag{3.14}$$

$$\left[\begin{array}{c:c} \begin{array}{c} T_{n-1}\left(\underline{b}_p\right) \\ \hdashline \mathbf{0} \end{array} & T_{n-1}\left(\underline{a}_p\right) \end{array}\right] \triangleq T_n\left(\underline{b}_p, \underline{a}_p\right)$$

Proof:

We consider first the Sylvester solution $\left(n_c^*(s), d_c^*(s)\right)$ which corresponds to polynomials with maximal degree $m+n-1$. The corresponding equation may be extended to that of (3.14). Equation (3.14) may be written explicitly as

$$\left[\begin{array}{c|c|c} T_{n-1}\left(\underline{b}_p\right) & T_{m-1}\left(\underline{a}_p\right) & Y \\ \hline \mathbf{0}_{n-m,n} & \mathbf{0}_{n-m,m} & X \end{array}\right]\left[\begin{array}{c} \underline{\tilde{c}}_{n-1} \\ \hline \underline{\tilde{d}}_{m-1} \\ \hline \underline{\tilde{d}}_{n-m} \end{array}\right]=\underline{\tilde{\varphi}}_{2n} \tag{3.14a}$$

where the matrix X has the form

$$X=\begin{bmatrix} 1 & a_{n-1} & \cdots & \\ 0 & 1 & \ddots & \\ \vdots & \ddots & \ddots & a_{n-1} \\ 0 & \cdots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{(n-m)\times(n-m)} \tag{3.14b}$$

and $\left[T_{n-1}\left(\underline{b}_p\right) \,\vdots\, T_{m-1}\left(\underline{a}_p\right)\right]$ is the $T_{m,n}\left(\underline{b}_p,\underline{a}_p\right)$ $(m+n)\times(m+n)$ full rank matrix. Thus, the matrix in (3.14b) is upper block triangular with full rank diagonal blocks and this proves that $T_n\left(\underline{b}_p,\underline{a}_p\right)$ has full column rank, which implies (3.14). ■

The above result implies the following:

<u>Remark (3.2)</u>: If a non-proper solution of order $(k-1,l-1)$, $k>l$ may be found, which assigns a polynomial $\varphi(s)$, then we can always find a proper solution of order $(k-1,k-1)$ that assigns a polynomial $\varphi'(s)$ which contains $\varphi(s)$ as a factor, but it is otherwise arbitrary. ■

<u>Remark (3.3)</u>: Every polynomial with degree $k \geq 2n-1$ may be always assigned by a proper controller of some appropriate McMillan degree.

<u>Example (3.1)</u>: For the case where $m=\deg\left[n_p(s)\right]=2$, $n=\deg\left[d_p(s)\right]=5$, the Sylvester solution of the Diophantine equation is non-proper and it is defined by

$$\begin{array}{cc}
\begin{array}{c} \\ \\ \\ \uparrow \\ n\text{-}1=4 \\ \\ \downarrow \end{array} &
\left[\begin{array}{ccccc|cc}
b_0 & 0 & 0 & 0 & 0 & a_0 & 0 \\
b_1 & b_0 & 0 & 0 & 0 & a_1 & a_0 \\
b_2 & b_1 & b_0 & 0 & 0 & a_2 & a_1 \\
0 & b_2 & b_1 & b_0 & 0 & a_3 & a_2 \\
0 & 0 & b_2 & b_1 & b_0 & a_4 & a_3 \\
0 & 0 & 0 & b_2 & b_1 & 1 & a_4 \\
0 & 0 & 0 & 0 & b_2 & 0 & 1
\end{array}\right]
\left[\begin{array}{c} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ \hline d_0 \\ d_1 \end{array}\right]
=
\left[\begin{array}{c} \varphi_0 \\ \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \end{array}\right] \\
 & \longleftarrow n\text{-}1=4 \longrightarrow
\end{array}$$

Such a solution may be expanded to a proper solution by solving the following equation

$$\begin{array}{cc}
\begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \uparrow \\ n\text{-}m=3 \\ \downarrow \end{array} &
\left[\begin{array}{ccccc|cc|ccc}
b_0 & 0 & 0 & 0 & 0 & a_0 & 0 & 0 & 0 & 0 \\
b_1 & b_0 & 0 & 0 & 0 & a_1 & a_0 & 0 & 0 & 0 \\
b_2 & b_1 & b_0 & 0 & 0 & a_2 & a_1 & a_0 & 0 & 0 \\
0 & b_2 & b_1 & b_0 & 0 & a_3 & a_2 & a_1 & a_0 & 0 \\
0 & 0 & b_2 & b_1 & b_0 & a_4 & a_3 & a_2 & a_1 & a_0 \\
0 & 0 & 0 & b_2 & b_1 & 1 & a_4 & a_3 & a_2 & a_1 \\
0 & 0 & 0 & 0 & b_2 & 0 & 1 & a_4 & a_3 & a_2 \\ \hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & a_4 & a_3 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & a_4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right]
\left[\begin{array}{c} c_0' \\ c_1' \\ c_2' \\ c_3' \\ c_4' \\ \hline d_0' \\ d_1' \\ d_2' \\ d_3' \\ d_4' \end{array}\right]
=
\left[\begin{array}{c} \varphi_0' \\ \varphi_1' \\ \varphi_2' \\ \varphi_3' \\ \varphi_4' \\ \varphi_5' \\ \varphi_6' \\ \varphi_7' \\ \varphi_8' \\ \varphi_9' \end{array}\right] \\
 & \longleftarrow n=5 \longrightarrow \;\; \longleftarrow m=2 \longrightarrow
\end{array}$$

As long as $\varphi_9' \neq 0$, the above solution is proper. ■

The parameterisation of the family $\mathcal{W}\left(\underline{b}_p, \underline{a}_p, \underline{\varphi}\right)$ is considered next. We assume that $\deg\left[\varphi(s)\right] = k$ and shall examine the properties of the entire family.

Remark (3.4): As long as $\left(n_p(s), d_p(s)\right)$ is a coprime pair, then for any $\varphi(s)$, the family $\mathcal{W}\left(\underline{b}_p, \underline{a}_p, \underline{\varphi}\right)$ is nonempty and thus the parameterisation of the family is a problem that makes sense. Every particular solution is characterised by a pair of (μ, ν)-

orders and thus parameterisation in terms of the McMillan degree is a problem that may be considered. However, the family $\mathcal{W}\left(\underline{b}_p, \underline{a}_p, \underline{\varphi}\right)$ may not necessarily contain proper solutions, if $\deg\left[\varphi(s)\right] = k$ is any arbitrary number. ■

Although any $\varphi(s)$ with arbitrary $\deg\left[\varphi(s)\right] = k$ may be assigned, it is interesting to investigate the minimal degree of $\varphi(s)$. We have to distinguish between two cases:

i) $\varphi(s)$, $\deg\left[\varphi(s)\right] = k$ is a generic polynomial, i.e. $\underline{\varphi}$ is a generic vector of $\mathbb{R}^{k+1}$

ii) $\varphi(s)$, $\deg\left[\varphi(s)\right] = k$ is a non-generic polynomial, i.e. $\underline{\varphi}$ is associated with a proper variety of $\mathbf{P}^k$ projective space.

The following result provides necessary conditions for the order (μ, ν) of solutions for the generic polynomial $\varphi(s)$, $\deg\left[\varphi(s)\right] = k$.

<u>Lemma (3.1)</u>: Let $\underline{b}_p \in \mathbb{R}^{m+1}$, $\underline{a}_p \in \mathbb{R}^{n+1}$ and consider the Toeplitz matrix $T_{\mu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right)$ defined for some (μ, ν) pair as in (3.8). Necessary conditions for $T_{\mu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right)$ to be epic, i.e. $\operatorname{Im}\left\{T_{\mu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right)\right\} = \mathbb{R}^\tau$, $\tau = \max\{m+\mu+1, n+\nu+1\}$ are that

$$\mu \geq n-1, \ \nu \geq m-1 \tag{3.15}$$

Proof:

Note that the matrix $T_{\mu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right)$ has dimensions $\left[\max\{m+\mu+1, n+\nu+1\}\right] \times \left[(\mu+1)+(\nu+1)\right]$ and thus necessary condition for the matrix to be epic is that $\left[\max\{m+\mu+1, n+\nu+1\}\right] \times \left[(\mu+1)+(\nu+1)\right]$ and thus necessary condition for the matrix to be epic is that

$$(\mu+1)+(\nu+1) \geq \max(m+\mu+1, n+\nu+1) \tag{3.16}$$

The above is equivalent to

$$\mu+\nu+2\geq 1+\max\left(m+\mu,n+\nu\right),\text{ or, }\mu+\nu+1\geq \max\left(m+\mu,n+\nu\right) \tag{3.16a}$$

The problem we now have is to parameterise the family of all (μ,ν) pairs which satisfy (3.16a). We distinguish the following cases:

I) $\underline{m=n}$: (3.16a) then becomes

$$\mu+\nu+1\geq \max\left(m+\mu,n+\nu\right)=m+\max\left(\mu,\nu\right) \tag{3.17a}$$

i) If $\mu\geq\nu$, then (3.17a) leads to

$\mu+\nu+1\geq m+\mu \quad \rightarrow \quad \underline{\nu\geq m-1}$ and $\underline{\mu\geq\nu\geq m-1}$

ii) If $\mu\leq\nu$, then (3.17a) leads to

$\mu+\nu+1\geq m+\nu \quad \rightarrow \quad \underline{\mu\geq m-1}$ and $\underline{\nu\geq\mu\geq m-1}$

which establishes the result for $m=n$

II) $\underline{m\geq n}$: Let us then write $m=n+\delta$, $\delta>0$. In this case (3.15a) becomes

$$\mu+\nu+1\geq \max\left(n+\delta+\mu,n+\nu\right)=n+\max\left(\delta+\mu,\nu\right) \tag{3.17b}$$

i) If $\delta+\mu\geq\nu$, then condition (3.17b) yields

$\mu+\nu+1\geq n+\delta+\mu \quad \rightarrow \quad \nu\geq n-1+\delta=n-1+m-n \quad \rightarrow \quad \underline{\nu\geq m-1}$

and

$\delta+\mu\geq\nu\geq m-1 \quad \rightarrow \quad m-n+\mu\geq m-1 \quad \rightarrow \quad \underline{\mu\geq n-1}$

ii) If $\delta+\mu\leq\nu$, then condition (3.17b) yields

$\mu+\nu+1\geq n+\nu \quad \rightarrow \quad \underline{\mu\geq n-1}$

and

$$\nu \geq \delta + \mu = m - n + \mu \quad \rightarrow \quad \nu - m + n \geq \mu \geq n - 1$$

or

$$\nu - m + n \geq n - 1 \quad \rightarrow \quad \underline{\nu \geq m - 1}$$

which establishes the result for the $m > n$ case. The case $m < n$ is established in similar lines. ■

The above Lemma together with Theorem (3.2) leads to the following result.

<u>Theorem (3.2)</u>: Let $(n_p(s), d_p(s))$ be a coprime pair of polynomials with $\deg(n_p(s)) = m$, $\deg(d_p(s)) = n$ and let $(n_c(s), d_c(s))$, $\deg(n_c(s)) = \mu$, $\deg(d_c(s)) = \nu$, be a solution of (3.3):

$$n_p(s)n_c(s) + d_p(s)d_c(s) = \varphi(s)$$

Necessary and sufficient condition for the orders (μ, ν) of the solution to correspond to a generic $\varphi(s)$ is that

$$\mu \geq n - 1, \ \nu \geq m - 1 \qquad (3.15)$$

Furthermore, the solution for $\mu^* = n - 1$, $\nu^* = m - 1$ is the Sylvester solution, which is uniquely defined and has the minimal McMillan degree $\delta^* = \max(m, n) - 1$ amongst all solutions associated with the generic $\varphi(s)$ polynomial.

Proof:

If $\varphi(s)$ is generic, then equation (3.7) must have a solution for all vectors of $\mathbf{R}^\tau$, $\tau = 1 + \max\{m + \mu, n + \nu\}$, i.e. $T_{\mu,\nu}^{m,n}(\underline{b}_p, \underline{a}_p)$ must be epic. By Lemma (3.1) it then follows that necessary condition is that

$$\mu \geq n - 1, \ \nu \geq m - 1$$

which establishes the necessity. By theorem (3.2) we have that for $\mu^* = n - 1$, $\nu^* = m - 1$, there always exists the Sylvester solution which has the minimal McMillan degree

$\delta_M^* = \max\{m,n\}-1$. Using the $(n_c^*(s), d_c^*(s))$ Sylvester solution, then corollary (3.1) shows clearly that for every $\mu \geq n-1$, $\nu \geq m-1$ we can select the $t(s)$ parameter to get a corresponding order solution. This establishes the sufficiency. ∎

The above result leads to the following remark:

Remark (3.5): If $\varphi(s)$ is an arbitrary polynomial, it can always be assigned by a controller $(n_c(s), d_c(s))$, but there is no controller with order (μ,ν) that violates conditions (3.15). For non-generic polynomials $\varphi(s)$, there exist solutions with order (μ,ν) for which $\mu < n-1$ and $\nu < m-1$. ∎

The family of non-generic polynomials which are assigned by controllers with order (μ,ν) violating (3.15) is defined below.

Corollary (3.3): If (μ,ν) order of solution is fixed, $\mu < n-1$, $\nu < m-1$, then the set of polynomials that can be assigned, is defined in the vectors in $\operatorname{Im}\left\{T_{\mu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right)\right\}$. ∎

The McMillan degree is important in the parameterisation of the $\mathcal{W}\left(\underline{b}_p, \underline{a}_p; \underline{\varphi}\right)$. We first state the following result:

Proposition (3.2): Let $\mathcal{W}\left(\underline{b}_p, \underline{a}_p; \underline{\varphi}\right)$ be the family of solutions of the Diophantine equation. The following properties hold true:

i) The relationship $\mathcal{R}_M$ on $\mathcal{W}\left(\underline{b}_p, \underline{a}_p; \underline{\varphi}\right)$ defined $\forall\ t_1, t_2 \in \mathcal{W}\left(\underline{b}_p, \underline{a}_p; \underline{\varphi}\right)$ by

$$t_1 \mathcal{R}_M t_2 \Leftrightarrow \delta_M(t_1) = \delta_M(t_2) \tag{3.18}$$

is an equivalence relation.

ii) Let $\ell_M(t)$ denote the equivalence class of $t \in \mathcal{W}(\underline{b}_p, \underline{a}_p; \underline{\varphi})$ under $\mathcal{R}_M$. The family $\langle \ell_M(t) \rangle$ of all equivalence classes $\ell_M(t)$ form a partition of $\mathcal{W}(\underline{b}_p, \underline{a}_p; \underline{\varphi})$.

■

Showing that $\mathcal{R}_M$ is an equivalence relation is trivial. Part (ii) follows immediately from part (i). The partitioning of $\mathcal{W}(\underline{b}_p, \underline{a}_p; \underline{\varphi})$ under $\mathcal{R}_M$ is indicated below:

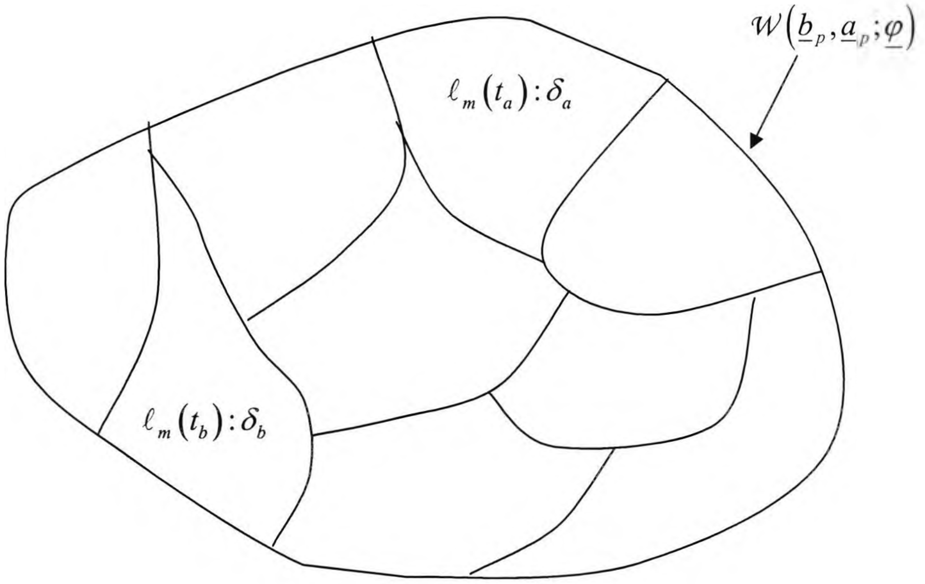

$\mathcal{W}(\underline{b}_p, \underline{a}_p; \underline{\varphi})$
$\ell_m(t_a) : \delta_a$
$\ell_m(t_b) : \delta_b$

Figure (3.1)

Note that each equivalence class is parameterised by a distinct number, which is the extended McMillan degree of the class. For the $\ell_M(t_i)$ class we shall simply denote $\delta_m(t_i) = \delta_i$. The value of δ_i characterises the $\ell_M(t_i)$ class and will be called the *McMillan index* of the class; in turn, the family $\ell_M(t_i)$ will be referred to as the δ_i *-family.* The set of indices

$$I_M(\underline{b}_p, \underline{a}_p; \underline{\varphi}) \triangleq I_M \triangleq \{\delta_i : \delta_i \text{ index of } \ell_M(t_i), \quad \forall \ell_M(t_i) \in \langle \ell_M \rangle\} \tag{3.19}$$

will be called the *McMillan index set* of $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$. Clearly I_M is a subset of the nonnegative integers $\mathbb{Z}_{\geq 0}$. Two of the important problems that emerge in relation to the parameterisation are defined below:

PARAMETERISATION PROBLEMS: Given a system defined by a pair $\left(n_p(s),d_p(s)\right)$, $\deg\left[n_p(s)\right]=m$, $\deg\left[d_p(s)\right]=n$ of coprime polynomials and a polynomial $\varphi(s)$, $\deg\left[\varphi(s)\right]=k$ with $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$ family of solutions, we consider the following problems:

i) Minimal Design Problem (MDP): Define the minimal McMillan degree amongst all elements of $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$.

ii) Parameterisation of Proper Solutions Problem (PPSP): Investigation of proper solutions and study of properties of the proper family $\widetilde{\mathcal{W}}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$ for any given polynomial $\varphi(s)$.

iii) Index Parameterisation Problem (IPP): Define the McMillan index set I_M associated with $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$ and $\widetilde{\mathcal{W}}$ subfamily.

iv) Class Parameterisation Problem (CPP): For every $\delta\in I_M$ define a parametric expression of the equivalent class $\ell_M(t)$ for which $\delta_M(t)=\delta$. Investigation of parameterisation aspects of the proper subclass with this degree δ.

■

3.4. THE MINIMAL McMILLAN DEGREE PROBLEM

The study of the above problems involves the polynomial $\varphi(s)$, $\deg\left[\varphi(s)\right]=k$ and its associated coefficient vector $\underline{\varphi}\in\mathbb{R}^{k+1}$, as well as the order (μ,ν) of the sought solutions. If we denote $\mathbb{Z}_0^+$ the set of natural numbers (positive integers including zero) and

$$Q_{m,n} \triangleq \{(\mu,\nu): \ \mu \geq n-1, \ \nu \geq m-1\} \tag{3.19a}$$

then by $\widehat{Q}_{m,n}$ we denote the complementary set of $Q_{m,n}$ with respect to $\mathbb{Z}_0^+ \times \mathbb{Z}_0^+$

$$\widehat{Q}_{m,n} \triangleq \{(\mu,\nu): \ \mu < n-1 \ \forall \nu \in \mathbb{N} \text{ or } \nu < m-1 \ \forall \mu \in \mathbb{N}\} \tag{3.19b}$$

The set $Q_{m,n}$ will be referred as *regular* and $\widehat{Q}_{m,n}$ as *irregular* set of orders. The solution to the minimal design problem is considered now for regular and irregular orders and for given degree polynomials $\varphi(s)$.

Proposition (3.3): Let $\varphi(s)$ be such that $k \leq m+n-1$. The minimal McMillan degree solution is defined by searching through all pairs (μ,ν) for which

$$\mu \leq n-1, \ \nu \leq m-1 \tag{3.20}$$

and the following condition is satisfied

$$\underline{\varphi} \in \operatorname{Im}\left(T_{\mu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right)\right) \tag{3.21}$$

and selecting the one with $\min\{\max\{\mu,\nu\}\}$

Proof:

Note that the dimension of $\underline{\varphi}$ vector is $\max\{m+\mu+1, n+\nu+1\}$ and thus if $k = \deg[\varphi(s)]$ then

$$k+1 = \max\{m+\mu+1, n+\nu+1\}$$

The above, together with $k+1 < m+n$ imlies that

$$\max\{m+\mu+1, n+\nu+1\} < m+n \tag{3.22}$$

we distinguish the cases

i) Assume $m+\mu \geq n+\nu$. Then (3.22) implies

$$1+m+\mu < m+n \;\rightarrow\; \mu < n-1$$

From $\quad m+\mu \geq n+\nu \;\rightarrow\; \mu \geq n-m+\nu \quad$ and thus

$$n-1 > \mu \geq n-m+\nu \;\rightarrow\; n-1 > n-m+\nu \;\rightarrow\; \nu < m-1$$

ii) Assume $m+\mu < n+\nu$. Then (3.22) implies

$$1+\nu+n < m+n \;\rightarrow\; \nu < m-1$$

From $\quad m+\mu < n+\nu \;\rightarrow\; \nu > m-n+\mu \;\rightarrow\; \mu < n-1$

The conditions $\mu < n-1$, $\nu < m-1$ imply that (μ,ν) is irregular and a solution exists only when $\underline{\varphi} \in \operatorname{Im}\left(T_{\mu,\nu}^{m;n}\left(\underline{b}_p,\underline{a}_p\right)\right)$.

If there is no (μ,ν) such that the above condition is satisfied, then for $\mu = n-1$, $\nu = m-1$ we have a solution, the Sylvester solution which is the minimal. ■

The above case clearly corresponds to irregular orders if a solution exists for some $\mu < n-1$, $\nu < m-1$ and a searching procedure for determing the minimum can be defined. Note that such search is restricted to the testing of a finite number of conditions, since if one of the μ,ν becomes larger than $n-1, m-1$, then it has to be constrained by the inequality

$$k = \max\left(m+\mu, n+\nu\right) < m+n-1 \tag{3.23}$$

because otherwise

$$k = \max\left(m+\mu, n+\nu\right) \geq m+n-1 \tag{3.24}$$

then for the polynomial $\varphi(s)$, we always have the Sylvester solution as the minimal one. This leads to the next result readily reduced from the above.

Corollary (3.4): Let $m \le n$, $k \le n+m-1$. The search for the minimum (μ^*, ν^*)-order solution of the Diophantine equation is reduced to testing condition (3.21) successively for matrices $T_{\mu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right)$ corresponding to the pairs:

$$\begin{aligned}
&\nu = 0, \quad \mu = 0,1,\ldots,n-m \\
&\nu = 1, \quad \mu = 0,1,\ldots,n-m+1 \\
&\qquad\qquad \vdots \\
&\nu = m-1, \qquad \mu = 0,1,\ldots,n-m+\nu = n-1
\end{aligned} \tag{3.25}$$

The first matrix for which (3.21) is satisfied reveals the optimal (μ^*, ν^*). ■

Example (3.2): Consider the system of Example (3.1) where $m = 2$, $n = 5$ and since $k+1 = m+n = 7$, we examine the assignment of polynomials with degree $k \le 6$. We thus distinguish the following cases:

Case (a): $k \le 5$: and thus $k+1 = n+1 = 6$. In this case $\nu = 0$ and μ may take the values $\mu = 0$, $\mu = 1$, $\mu = 2$, $\mu = n-m+\nu = 5-2+0 = 3$. This implies testing (3.21) for the matrix

$$T_1 = \left[\begin{array}{cccc:c}
b_0 & 0 & 0 & 0 & a_0 \\
b_1 & b_0 & 0 & 0 & a_1 \\
b_2 & b_1 & b_0 & 0 & a_2 \\
0 & b_2 & b_1 & b_0 & a_3 \\
0 & 0 & b_2 & b_1 & a_4 \\
0 & 0 & 0 & b_2 & 1
\end{array}\right]$$

If (3.20) is satisfied for the given $\varphi(s)$ with the vector $\underline{\varphi}$ the 6-dimentional coefficient vector and the matrix T_1 above, then $\nu = 0$ and the exact value of μ is revealed by testing successfully (3.21) for the following set of submatrices of T_1

$$T_1^0=\left[\begin{array}{c|c} b_0 & a_0 \\ b_1 & a_1 \\ b_2 & a_2 \\ 0 & a_3 \\ 0 & a_4 \\ 0 & 1 \end{array}\right],\quad T_1^1=\left[\begin{array}{cc|c} b_0 & 0 & a_0 \\ b_1 & b_0 & a_1 \\ b_2 & b_1 & a_2 \\ 0 & b_2 & a_3 \\ 0 & 0 & a_4 \\ 0 & 0 & 1 \end{array}\right],\quad T_1^2=\left[\begin{array}{ccc|c} b_0 & 0 & 0 & a_0 \\ b_1 & b_0 & 0 & a_1 \\ b_2 & b_1 & b_0 & a_2 \\ 0 & b_2 & b_1 & a_3 \\ 0 & 0 & b_2 & a_4 \\ 0 & 0 & 0 & 1 \end{array}\right],\quad T_1^3=T_1$$

$$(\mu=0)\qquad\qquad (\mu=1)\qquad\qquad (\mu=2)\qquad\qquad (\mu=3)$$

The above tests reveal that the minimal value of μ as the order of the first submatrix from the sequence $\left\{T_1^0,T_1^1,T_1^2,T_1^3\right\}$ for which (3.21) is satisfied. This specifies the minimal order $\left(\mu^*,\nu^*\right)$.

Case (b): $k\le 6$: If $k=6$ or $k\le 6$ and the testing of condition (3.21) has failed for T_1, then we consider $\nu=1$ and μ may take the values $\mu=0$, $\mu=1$, $\mu=2$, $\mu=3$, $\mu=4=n-m+1$. This implies testing (3.21) for the matrix

$$T_2=\left[\begin{array}{cccc|cc} b_0 & 0 & 0 & 0 & a_0 & 0 \\ b_1 & b_0 & 0 & 0 & a_1 & a_0 \\ b_2 & b_1 & b_0 & 0 & a_2 & a_1 \\ 0 & b_2 & b_1 & b_0 & a_3 & a_2 \\ 0 & 0 & b_2 & b_1 & a_4 & a_3 \\ 0 & 0 & 0 & b_2 & 1 & a_4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

Note that T_2 corresponds to $\mu=3$, since for $\mu=4$, $\nu=1$, the resulting matrix is the Sylvester and (3.21) is satisfied. If condition (3.21) is satisfied for a given $\varphi(s)$ with $\underline{\varphi}$ 7-dimensional coefficient vector and the matrix T_2 above, then $\nu=1$ and the exact value of μ is revealed by testing successively (3.21) for the following submatrices of T_2:

$$T_2^0\begin{bmatrix} b_0 & \vdots & a_0 & 0 \\ b_1 & \vdots & a_1 & a_0 \\ b_2 & \vdots & a_2 & a_1 \\ 0 & \vdots & a_3 & a_2 \\ 0 & \vdots & a_4 & a_3 \\ 0 & \vdots & 1 & a_4 \\ 0 & \vdots & 0 & 1 \end{bmatrix}, \quad T_2^1\begin{bmatrix} b_0 & 0 & \vdots & a_0 & 0 \\ b_1 & b_0 & \vdots & a_1 & a_0 \\ b_2 & b_1 & \vdots & a_2 & a_1 \\ 0 & b_2 & \vdots & a_3 & a_2 \\ 0 & 0 & \vdots & a_4 & a_3 \\ 0 & 0 & \vdots & 1 & a_4 \\ 0 & 0 & \vdots & 0 & 1 \end{bmatrix}, \quad T_2^2\begin{bmatrix} b_0 & 0 & 0 & \vdots & a_0 & 0 \\ b_1 & b_0 & 0 & \vdots & a_1 & a_0 \\ b_2 & b_1 & b_0 & \vdots & a_2 & a_1 \\ 0 & b_2 & b_1 & \vdots & a_3 & a_2 \\ 0 & 0 & b_2 & \vdots & a_4 & a_3 \\ 0 & 0 & 0 & \vdots & 1 & a_4 \\ 0 & 0 & 0 & \vdots & 0 & 1 \end{bmatrix}, \quad T_2^3 = T_1$$

$$(\mu=0) \qquad\qquad (\mu=1) \qquad\qquad (\mu=2) \qquad\qquad (\mu=3)$$

The above tests reveal the minimal value of μ as the order of the first submatrix from the sequence $\left\{T_2^0, T_2^1, T_2^2, T_2^3\right\}$ for which (3.21) is satisfied. This leads to the specification of $\left(\mu^*, \nu^*\right)$ minimal order. ■

For polynomials with degree $k \geq m+n-1$ the minimal McMillan degree solution is defined below.

Corollary (3.5): Let $m \leq n$ and $k = \deg\left[\varphi(s)\right] \geq n+m-1$. For every such $\varphi(s) \in \mathbf{R}[s]$ the Diophantine equation has a solution, which has the following properties:

i) If k: $m+n-1 \leq k \leq 2n-1$, then the minimal order is defined by $\left(\mu^*, \nu^*\right) = (n-1, k-n)$ and $\delta^* = n-1$.

ii) If k: $2n-1 \leq k$, then the minimal order is $\left(\mu^*, \nu^*\right) = (n-1, k-n)$ and $\delta^* = k-n$

Proof:

For the case $k = m+n-1$ we have the Sylvester solution and thus if $T_{n-1,m-1}^{m,n}\left(\underline{b}_p, \underline{a}_p\right) = T^*$, then for any $k > m+n-1$ we can consider the matrix $T_{\mu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right)$ where $\mu = n-1$ and $\nu = \tau + (m-1) = k-n$. This matrix is clearly of the form:

$$T_{\mu,\nu}^{m,n}\left(\underline{b}_p,\underline{a}_p\right)=\left[\begin{array}{c|cccc} & \times & & 0\cdots & 0\\ T^* & & & \ddots & \vdots\\ & & & & 0\\ & \times & & & \\ \hline & 1 & \times & & \times\\ & 0 & 1 & \times & \\ 0 & & \ddots & \ddots & \\ & & & & \times\\ & 0 & & 0 & 1\end{array}\right], \quad \tau=k-1-n-m$$

and has clearly a full rank. Thus $\underline{\varphi}$ can be assigned with $\mu^*=n-1$ and $\nu^*=k-n$. Clearly, if $k\leq 2n-1$, $\delta^*=\max\{n-1,k-n\}=n-1$, whereas, if $k\geq n-1$, then $\delta^*=k-n$.

■

3.5. PARAMETERISATION OF PROPER SOLUTIONS

The analysis so far deals with the minimal design problem and provides a complete solution. The existence of proper solutions of the Diophantine equation has been established by remark (3.3), which also defines an upper bound for proper controller. In the study of properness we will use the following Toeplitz matrices:

$$T_{\nu}^{m,n}\triangleq T_{\nu,\nu}^{m,n}\left(\underline{b}_p,\underline{a}_p\right)\in\mathbb{R}^{(n+\nu+1)\times(2\nu+2)} \qquad (3.26)$$

where it is assumed that $m\leq n$, $\nu=0,1,2,\ldots,n-1$ and we denote by $\mathcal{T}_{\nu}^{m,n}$ the space $\mathcal{T}_{\nu}^{m,n}=\operatorname{colsp}\left\{T_{\nu}^{m,n}\right\}$. A vector $\underline{x}\in\mathcal{T}_{\nu}^{m,n}$ will be called a *normal* vector of the subspace, if $\underline{x}\notin\operatorname{colsp}T_{\nu,\nu-1}^{m,n}\left(\underline{b}_p,\underline{a}_p\right)$. In terms of this notation we may state:

Theorem 3.4: Let $m \leq n$, $k = \deg\left[\varphi(s)\right]$ and $\underline{\varphi}$ be the coefficient vector of $\varphi(s)$. The family of proper solutions of the Diophantine equation has the following properties:

i) If $k < n$, then there exists no proper solution.

ii) If $n \leq k \leq 2n-2$, then proper solutions exist, if and only if $\underline{\varphi}$ is for some $\nu = 0,1,2,\ldots,n-1$, a normal vector of $\mathcal{T}_{\nu}^{m,n}$ space. Furthermore, if such solution exists, it is uniquely defined and the minimal ν for which the above property holds determines the proper minimal McMillan degree solution.

iii) If $k \geq 2n-1$, then for every polynomial $\varphi(s)$, there exists a proper solution with minimal McMillan degree $\delta^{*} = k-n$.

Proof:

i) By the structure of $T_{\mu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right)$ we see that if $\mu < n-m$ and $k < n$, then for any ν, equation (3.5) implies that $d_0 = d_1 = \cdots = d_\nu = 0$ and thus there is no solution. Thus we have to consider the case $\mu \geq n-m$ (i.e. $\mu = n-m+\tau$, $\tau \geq 0$ and $\nu \geq \mu$ for properties with $d_\nu \neq 0$. Assume now that $\nu = \mu + a$, $k < n$ and that in equation (3.5) we have the form

$$
\begin{array}{c}
\\ \\ \\ m+1+\mu \\ \\ \\ \\ \\ n \\ \\
\end{array}
\begin{array}{cccc|cccc}
b_0 & 0 & \cdots & 0 & a_0 & 0\cdots & & 0 \\
\vdots & \ddots & & \vdots & \vdots & \ddots & & \\
\vdots & & \ddots & 0 & & & & a_0 \\
b_m & & & b_0 & & & & \vdots \\
0 & \ddots & & & \vdots & & & \\
\vdots & \ddots & \ddots & & & & & \\
0 & \cdots & 0 & b_m & a_{n-1} & & & \\
\cline{1-4}
 & & & & 1 & \ddots & & \vdots \\
 & & & & & & & a_{n-1} \\
 & & & & & & & 1
\end{array}
\begin{array}{c}
\\ \\ \\ \\ n+1+\nu \\ \\ \\ \\ \\ \\
\end{array}
\tag{3.27a}
$$

and for properness $\nu = n-m+\tau+a$, $\left(\nu \geq \mu,\ a \geq 0\right)$. From the above, it follows that the condition $k<n$ implies that the last n of the coordinates of $\underline{d}$ vector have to be zero i.e. $d_{\nu}=d_{\nu-1}=\cdots=d_{\nu-n}=0$ where

$n=n+1+\nu-(m+1-\mu)=n-m-\nu-\mu=n-m+a$

The above implies that by increasing ν above $n-m$ the possibility of finding proper solutions does not improve since the additional a coordinates in $\underline{d}$ vector (from the end) have to be zero. Thus we consider $a=0$ and $\nu=\mu=n-m+\tau$. Given then that the number of zero coordinates in $\underline{d}$ will always be $n-m$, any solutions that may exist will have the property that

$d_{\nu}=d_{\nu-1}=\cdots=d_{\nu-(n-m)}=0$

The above implies that the structure of $T_{\mu,\nu}^{m,n}\left(\underline{b}_p,\underline{a}_p\right)$ for $\nu=\mu-(n-m)$ is of the type described below

$$T_{\mu,\nu}^{m,n}\triangleq\left[\begin{array}{cccc|ccc}
b_0 & 0 & \cdots & 0 & a_0 & 0\cdots & 0\\
\vdots & \ddots & \ddots & \vdots & & \ddots & \\
\vdots & & \ddots & 0 & & & a_0\\
\vdots & & & b_0 & & & \\
\vdots & & & \vdots & & & \\
b_m & & & \vdots & a_{n-1} & & \\
0 & \ddots & & \vdots & 1 & \ddots & \\
\vdots & \ddots & \ddots & \vdots & 0 & \ddots & a_{n-1}\\
0 & \cdots & 0 & b_m & 0\ldots & 0 & 1
\end{array}\right]\begin{bmatrix} c_0\\ c_1\\ \vdots\\ c_{\mu}\\ \hdashline d_0\\ \vdots\\ d_{\nu}\end{bmatrix}=\begin{bmatrix}\varphi_0\\ \varphi_1\\ \vdots\\ \varphi_k\\ 0\\ \vdots\\ 0\end{bmatrix} \qquad (3.27b)$$

and thus the condition $k<n$ implies that the above yields

$$m+\mu-k\left\updownarrow\vphantom{\begin{array}{c}1\\1\\1\\1\\1\\1\end{array}}\right.\underbrace{\left[\begin{array}{cccccc|cccc}
b_m & b_{m-1} & \cdots & & & & a_k & & & \\
0 & b_m & \ddots & & & & \vdots & \ddots & & \\
\vdots & 0 & \ddots & \ddots & & & 1 & & \ddots & \\
\vdots & & \ddots & \ddots & \ddots & \vdots & 0 & \ddots & & a_k\\
\vdots & & & \ddots & \ddots & b_{m-1} & \vdots & \ddots & \ddots & \vdots\\
0 & \cdots & & \cdots & 0 & b_m & 0 & \cdots & 0 & 1
\end{array}\right]}_{m+\mu-k}\begin{bmatrix} c_{k-\mu}\\ \vdots\\ c_{\mu}\\ \hdashline d_0\\ \vdots\\ d_{\nu}\end{bmatrix}=\mathbf{0} \qquad (3.27c)$$

The only way (3.27c) can imply properness is when some of the $c_\mu, c_{\mu-1}, \ldots$, become zero.

However, if $c_\mu = 0$, then (3.27) implies that $d_\nu = 0$ and so forth. Thus, whatever solution they exist, they have $\mu > \nu$ and they are non-proper.

ii) From the proof of part (i) it follows that it suffices to consider the case $\mu = \nu$ and for $k \geq n$. Note that the matrices $T_\nu^{m,n}$ have more rows than columns when

$$n+1+\nu > 2(\nu+1) \rightarrow \nu < n-1$$

and thus the maximal degree of k has to be such that $1+k \leq n+1+\nu = n+1+n-2 = 2n-1$ i.e. $k \leq 2n-2$. Solvability with proper controller implies testing $\underline{\varphi} \in \mathcal{T}_\nu^{m,n}$ with $\nu = 0,1,\ldots,n-2$ and with the additional properties that if for some ν this condition is satisfied, then we would like to guarantee normal membership. The latter is important, since otherwise the solution may have a reduced ν, which implies non-properness. Clearly, testing the condition $\underline{\varphi} \in \mathcal{T}_\nu^{m,n}$, starting $\nu = 0,1,\ldots$ reveals the proper minimal solution.

iii) By corollary (3.2) the result is established for $k = 2n-1$ and with a McMillan degree $\delta = n-1$. For any $k > 2n-1$, condition (3.14a) may be extended to

$$\begin{array}{c} \left[\begin{array}{c|c|c} T_{n-1}\left(\underline{b}_p\right) & T_{m-1}\left(\underline{a}_p\right) & Y \\ \hline \mathbf{0}_{k+1-(m+n),n} & \mathbf{0}_{k+1-(m+n),m} & X \end{array}\right] \\ \qquad\qquad\qquad\qquad\quad \xleftrightarrow{} \\ \qquad\qquad\qquad\qquad\quad n-m+\tau \end{array} \begin{array}{c} \left[\begin{array}{c} \underline{c}_{n-1} \\ \hline \underline{d}_{m-1} \\ \hline \underline{d}_{n-m+\tau} \end{array}\right] \\ \\ \\ \end{array} \begin{array}{c} = \underline{\varphi}_k \\ \\ \\ \end{array} \tag{3.28}$$

where we assume $\nu = k-n$ for $\forall\ k \geq 2n-1$ write such k as $k = 2n-1+\tau$, $\tau \geq 0$. Clearly, X has the form

$$X=\begin{bmatrix} 1 & a_{n-1} & \cdots & & \\ 0 & 1 & a_{n-1} & \cdots & \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & a_{n-1} \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}$$

and it is full rank. Thus, a proper solution exists and it is of minimal McMillan degree, which by construction is $\delta = m-1+n-m+\tau = n-1+\tau = k-n$ (since $k = 2n-1+\tau$). ■

Corollary (3.6): The minimal degree polynomial for which there always exists a proper controller has McMillan degree $\delta^* = n-1$. For polynomials with k: $n \le k \le 2n$-2 the existence of proper solutions is a non-generic property. If $k < n$ there is no proper solution for any $\varphi(s)$. ■

Example (3.3): The search for proper solutions is examined here for the case of polynomials with $m = 2$, $n = 5$ (example (3.1)). We first note that

$$T_{0,0}^{2,5}=\begin{bmatrix} b_0 & a_0 \\ b_1 & a_1 \\ b_2 & a_2 \\ 0 & a_3 \\ 0 & a_4 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} b_0 & a_0 \\ b_1 & a_1 \\ b_2 & a_2 \\ 0 & a_3 \\ 0 & a_4 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} c_0 \\ d_0 \end{bmatrix}=\begin{bmatrix} \varphi_0 \\ \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ 0 \end{bmatrix}$$

implies $d_0 = 0$ and thus there is no proper solution for $k = 4 < 5$. Solutions for $k < 5$ may exist if we consider the case for instance of the Sylvester system

$$\left[\begin{array}{ccccc:cc} b_0 & 0 & 0 & 0 & 0 & a_0 & 0 \\ b_1 & b_0 & 0 & 0 & 0 & a_1 & a_0 \\ b_2 & b_1 & b_0 & 0 & 0 & a_2 & a_1 \\ 0 & b_2 & b_1 & b_0 & 0 & a_3 & a_2 \\ 0 & 0 & b_2 & b_1 & b_0 & a_4 & a_3 \\ 0 & 0 & 0 & b_2 & b_1 & 1 & a_4 \\ 0 & 0 & 0 & 0 & b_2 & 0 & 1 \end{array}\right]\left[\begin{array}{c} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ \hdashline d_0 \\ d_1 \end{array}\right]=\begin{bmatrix} \varphi_0 \\ \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \end{bmatrix}=\underline{\varphi}, \qquad \forall \underline{\varphi} \in \mathbb{R}^7$$

which however for $\forall \underline{\varphi}$ (or a generic $\underline{\varphi}$) has a non-proper solution.

Proper solution for families of polynomials may exist but such families are defined as sets with measure zero. In fact by considering the matrices

$$T_{0,0}^{2,5} = \left[\begin{array}{c|c} b_0 & a_0 \\ b_1 & a_1 \\ b_2 & a_2 \\ 0 & a_3 \\ 0 & a_4 \\ 0 & 1 \end{array}\right], \quad T_{1,1}^{2,5} = \left[\begin{array}{cc|cc} b_0 & 0 & a_0 & 0 \\ b_1 & b_0 & a_1 & a_0 \\ b_2 & b_1 & a_2 & a_1 \\ 0 & b_2 & a_3 & a_2 \\ 0 & 0 & a_4 & a_3 \\ 0 & 0 & 1 & a_4 \\ 0 & 0 & 0 & 1 \end{array}\right], \quad T_{2,2}^{2,5} = \left[\begin{array}{ccc|ccc} b_0 & 0 & 0 & a_0 & 0 & 0 \\ b_1 & b_0 & 0 & a_1 & a_0 & 0 \\ b_2 & b_1 & b_0 & a_2 & a_1 & a_0 \\ 0 & b_2 & b_1 & a_3 & a_2 & a_1 \\ 0 & 0 & b_0 & a_4 & a_3 & a_2 \\ 0 & 0 & 0 & 1 & a_4 & a_3 \\ 0 & 0 & 0 & 0 & 1 & a_4 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right],$$

$$T_{3,3}^{2,5} = \left[\begin{array}{cccc|cccc} b_0 & 0 & 0 & 0 & a_0 & 0 & 0 & 0 \\ b_1 & b_0 & 0 & 0 & a_1 & a_0 & 0 & 0 \\ b_2 & b_1 & b_0 & 0 & a_2 & a_1 & a_0 & 0 \\ 0 & b_2 & b_1 & b_0 & a_3 & a_2 & a_1 & a_0 \\ 0 & 0 & b_2 & b_1 & a_4 & a_3 & a_2 & a_1 \\ 0 & 0 & 0 & b_2 & 1 & a_4 & a_3 & a_2 \\ 0 & 0 & 0 & 0 & 0 & 1 & a_4 & a_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & a_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

$$T_{4,4}^{2,5} = \left[\begin{array}{ccccc|cc|ccc} b_0 & 0 & 0 & 0 & 0 & a_0 & 0 & 0 & 0 & 0 \\ b_1 & b_0 & 0 & 0 & 0 & a_1 & a_0 & 0 & 0 & 0 \\ b_2 & b_1 & b_0 & 0 & 0 & a_2 & a_1 & a_0 & 0 & 0 \\ 0 & b_2 & b_1 & b_0 & 0 & a_3 & a_2 & a_1 & a_0 & 0 \\ 0 & 0 & b_2 & b_1 & b_0 & a_4 & a_3 & a_2 & a_1 & a_0 \\ 0 & 0 & 0 & b_2 & b_1 & 1 & a_4 & a_3 & a_2 & a_1 \\ 0 & 0 & 0 & 0 & b_2 & 0 & 1 & a_4 & a_3 & a_2 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & a_4 & a_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & a_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

we see that the column spaces of $T_{0,0}^{2,5}$, $T_{1,1}^{2,5}$, $T_{2,2}^{2,5}$, $T_{3,3}^{2,5}$ define vector sets characterising the coefficients of polynomial vectors, which may be assigned by proper compensators. Clearly the column spans of such matrices, which have always a component along the lost column (needed to guarantee properness) define the assignable polynomials by proper controllers. The matrix $T_{4,4}^{2,5}$ is the smallest dimension matrix that guarantees the existence of a proper solution for any arbitrary polynomial of degree $2n-1=9$.

■

In the study of proper solutions we deal with matrices which are ordered (as far as column ordering due to Toeplitz structure). For such matrices, or ordered sets of vectors $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$ we define the notion of the proper span by

$$\widehat{\mathrm{sp}}\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\} \triangleq \{\underline{x}:\ \underline{x} = c_1\underline{x}_1 + c_2\underline{x}_2 + \ldots + c_k\underline{x}_k,\ \forall c_i:\ c_k \neq 0\} \tag{3.29}$$

The notion of the proper span is more restrictive than the usual span notion, since it excludes vectors which are in $\mathrm{sp}\{\underline{x}_1, \ldots, \underline{x}_{k-1}\}$. In terms of the notion of the proper span we may now define the following sets, which characterise special solutions of the Diophantine equation; these sets characterise the family of normal vectors in the spaces $\mathcal{T}_\nu^{m,n}$, $\nu = 0,1,\ldots,n-1$ and characterise the polynomials that may be assigned by certain McMillan degree compensators.

Definition (3.2): Let $\underline{b}_p \in \mathbb{R}^{m+1}$, $\underline{a}_p \in \mathbb{R}^{n+1}$, $m \leq n$, and consider the Toeplitz matrices defined by (3.25) for $\nu = 0,1,\ldots,n-1$ as

$$T_\nu^{m,n} \triangleq T_{\nu,\nu}^{m,n}\left(\underline{b}_p, \underline{a}_p\right) \in \mathbb{R}^{(n+\nu+1)\times 2(\nu+1)}$$

If $\mathcal{T}_\nu^{m,n} = \mathrm{colsp}\{T_\nu^{m,n}\}$, then we define the *proper subset* of $\mathcal{T}_\nu^{m,n}$, and shall denote it by $\widehat{\mathcal{T}}_\nu^{m,n}$, the set of all proper vectors of $\mathcal{T}_\nu^{m,n}$.

■

Corollarv (3.7): For any $\underline{b}_p \in \mathbb{R}^{m+1}$, $\underline{a}_p \in \mathbb{R}^{n+1}$, $m \le n$, the family of polynomials with degree k: $k = n + \nu$, $\nu = 0,1,\ldots,n-1$ which may be assigned by proper compensators of McMillan degree ν is given by the vectors in $\widehat{\mathcal{T}}_{\nu}^{m,n}$.

We consider next the parameterisation of the family of all solutions for a given $\varphi(s)$ and in particular the parameterisation of the proper family of solutions. By considering as a basis for the parameterisation the minimal solution (proper, or non-proper) that has been previously defined, we have from corollary (3.1):

Remark (3.6): If $\left(\tilde{n}_c(s), \tilde{d}_c(s)\right)$ is the minimal McMillan degree solution of the Diophantine equation (3.3) that corresponds to the k degree polynomial $\varphi(s)$, then the whole family of solutions for the same polynomial $\varphi(s)$, $\mathcal{W}\left(\underline{b}_p, \underline{a}_p; \underline{\varphi}\right)$ is defined by

$$\begin{cases} n_c(s) = \tilde{n}_c(s) + t(s) d_p(s) \\ d_c(s) = \tilde{d}_c(s) - t(s) n_p(s) \end{cases}, \quad \begin{array}{c} t(s) \in \mathbb{R}[s] \\ \text{arbitrary} \end{array} \tag{3.30}$$

■

The above leads to the following results dealing with the parameterisation issues:

Corollary(3.8):_ If δ^* is the minimal McMillan degree of the family that corresponds to a given $\varphi(s)$, then the McMillan index set I_M of $\mathcal{W}\left(\underline{b}_p, \underline{a}_p; \underline{\varphi}\right)$ is defined by:

$$I_M = \left\{\delta^*;\ \max\{m,n\} + k,\ k = 0,1,2,\ldots\right\} \tag{3.31}$$

Proof:

By (3.30), if $t(s) = 0$, then the minimal solution is δ^*. If $\deg\left[t(s)\right] = k$, then given that for the minimal solution $\mu^* \le n-1$, $\nu^* \le m-1$, we have that

$\delta_M\left(g_c(s)\right)=\max\{n+k,m+k\}$, since $a_n,b_m\neq 0$, or otherwise $\max\{m,n\}+k$ for $k=0,1,2,\ldots$ ■

Corollary(3.9): The element of $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$ family that corresponds to the minimal McMillan degree, where $\delta_j=\max\{m,n\}+j$ is parameterised by

$$\begin{cases} n_c(s)=\tilde{n}_c(s)+t(s)d_p(s) \\ d_c(s)=\tilde{d}_c(s)-t(s)n_p(s) \end{cases} \tag{3.32}$$

where $t(s)\in\mathbb{R}[s]$, $\deg\{t(s)\}=j$

Proof:

For any $t(s)\in\mathbb{R}[s]$ with $\deg\left[t(s)\right]=j$ it is clear that $\delta_M\left(g(s)\right)=\max\{m,n\}+j$ and thus by fixing the degree of $t(s)$ we remain within the family of the given $\delta_j=\max\{m,n\}+j$ index. For any other choice of degree for $t(s)$, i.e. $\deg\left[t(s)\right]=k\neq j$, then the corresponding controller will have McMillan degree $\max\{m,n\}+k\neq\max\{m,n\}+j$ and thus it will belong to a different subfamily.

The above results deal with the parameterisation issues of the general solution family. Next we consider the parameterisation issues for the proper subfamily of controllers. In the following we shall use the Toeplitz representation of the general family, which follows from Remark (3.6) and it is described below:

Remark (3.7): If $\left(\tilde{n}_c(s),\tilde{d}_c(s)\right)$ is the minimal McMillan degree solution of the Diophantine equation that corresponds to the k degree polynomial, then the Toeplitz representation of the general solution vector is given by:

$$\left[\begin{array}{c} c_0 \\ c_1 \\ \vdots \\ \\ c_p \\ \hdashline d_0 \\ d_1 \\ \vdots \\ \\ d_r \end{array}\right] = \left[\begin{array}{c} \tilde{c}_0 \\ \tilde{c}_1 \\ \vdots \\ \tilde{c}_{\tilde{\mu}} \\ 0 \\ \vdots \\ 0 \\ \hline \tilde{d}_0 \\ \tilde{d}_1 \\ \vdots \\ \tilde{d}_{\tilde{\nu}} \\ 0 \\ \vdots \\ 0 \end{array}\right] + \left[\begin{array}{cccc} a_0 & 0 & \cdots & 0 \\ \vdots & a_0 & \ddots & \vdots \\ & & \ddots & 0 \\ \vdots & & & a_0 \\ a_{n-1} & & & \vdots \\ 1 & \ddots & & \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1} \\ 0 & \cdots & 0 & 1 \\ \hdashline b_0 & & & \\ \vdots & b_0 & & \\ & & \ddots & \\ & & & b_0 \\ \vdots & & & \vdots \\ b_0 & & & \\ 0 & \ddots & & \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & b_0 \end{array}\right] \begin{bmatrix} \mathrm{t}_0 \\ \mathrm{t}_1 \\ \vdots \\ \\ \mathrm{t}_\theta \end{bmatrix} \tag{3.33}$$

where $\tilde{n}_c(s) = \tilde{c}_0 + \ldots + s^{\tilde{\mu}} \tilde{c}_{\tilde{\mu}}, \tilde{d}_c(s) = \tilde{d}_0 + \ldots + s^{\tilde{\nu}} \tilde{d}_{\tilde{\nu}}$ and the general solution is described by $n_c(s) = c_0 + \ldots + s^p c_p$, $d_c(s) = d_0 + \ldots + s^r d_r$, where $p \triangleq \deg[n_c(s)] = \max\{\tilde{\mu}, \theta + n\} = \theta + n$ $\quad r \triangleq \deg[d_c(s)] = \max\{\tilde{\nu}, \theta + m\} = \theta + m$, and $\theta \triangleq \deg[t(s)]$, $t(s) = t_0 + \ldots + s^\theta t_\theta$ ■

Note that a similar expression to the one given above may be given, if we use as fundamental solution any other solution apart from the minimal. The parameterisation of proper solutions of the Diophantine equation has the following properties:

Corollary (3.10): Let $m<n$, $k=\deg\left[\varphi(s)\right]$ and δ^* be the minimal McMillan degree solution for the given $\varphi(s)$. The family of proper solutions of the Diophantine equation for different polynomials $\varphi(s)$ has the following properties:

a) For all $\varphi(s)$ with $k\le 2n-2$ for which $\delta^*\le n-1$, then the following properties hold true:

i) If the minimal solution is non-proper, then all solutions in $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$ are non-proper

ii) If the minimal solution is proper, then it is the only proper element of the $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$ family

b) For all $\varphi(s)$ with $k\ge 2n-1$, there exists a proper minimal McMillan degree solution $\left(n_c^*(s),d_c^*(s)\right)$, with $\delta^*=k-n\ge n$, where

$$n_c^*(s)=c_0+\ldots+s^{\delta^*}c_{\delta^*},\ d_c^*(s)=d_0+\ldots+s^{\delta^*}d_{\delta^*},\ d_{\delta^*}\ne 0 \tag{3.34a}$$

Furthermore, the $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$ family contains a maximal subfamily of proper solutions $\widetilde{\mathcal{W}}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$, all of them having McMillan degree δ^*, which is defined in parametric form by

$$\begin{aligned} n_c(s)&=n_c^*(s)+d_p(s)t(s)\\ d_c(s)&=d_c^*(s)-n_p(s)t(s) \end{aligned} \tag{3.34b}$$

where $t(s)\in\mathbb{R}[s]$ arbitrary with $\deg\left[t(s)\right]=\sigma\le\delta^*-n=k-2n$.

Proof:

a) By using Theorem (3.4) and expression (3.33) it follows that since $\delta^*=\max\left(\tilde{\mu},\tilde{v}\right)\le n-1$, then the general solution will have

$$n_c(s)=n_c^*(s)+t(s)d_p(s),\quad d_c(s)=d_c^*(s)-t(s)n_p(s)$$

and thus

$$\deg\left[n_c(s)\right]=\max\left\{\deg\left[n_c^*(s)\right],\sigma+n\right\}=\max\left\{\tilde{\mu},\sigma+n\right\}=\sigma+n$$

$$\deg\left[d_c(s)\right]=\max\left\{\deg\left[d_c^*(s)\right],\sigma+m\right\}=\max\left\{\tilde{\nu},\sigma+m\right\}=\sigma+m$$

If $\left(n_c^*(s),d_c^*(s)\right)$ is non-proper then from the above it follows that for all $t(s)$ $\deg\left[n_c(s)\right]=\sigma+n>\deg\left[d_c(s)\right]=\sigma+m$ and thus all other solutions are non-proper. If $\left(n_c^*(s),d_c^*(s)\right)$ is proper, then the above proves that all other solutions are non-proper and thus the minimal solution is the unique proper one.

b) Since $\delta^*\geq n$, (3.33) indicates that if $\sigma=\deg\left[t(s)\right]\leq\delta^*-n=k-2n$, then

$$\deg\left[d_c^*(s)-n_p(s)t(s)\right]=\delta^*\geq\deg\left[n_c^*(s)-d_p(s)t(s)\right]$$

$$=\max\left\{\deg\left[n_c^*(s)\right],n+\sigma\right\}=\max\left\{\delta^*,n+\sigma\right\}$$

Note, that if $\deg\left[t(s)\right]>k-2n$, then

$\deg\left[d_c^*(s)-n_p(s)t(s)\right]<\deg\left[n_c^*(s)-d_p(s)t(s)\right]$ then all solutions are non-proper.

■

For the case bi-proper systems $(m=n)$ we have the following result describing the properties of $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$.

Corollary (3.11): Let $(m=n)$, $k=\deg\left[\varphi(s)\right]$ and δ^* be the minimal McMillan degree solution for the given order k. For the family $\mathcal{W}\left(\underline{b}_p,\underline{a}_p;\underline{\varphi}\right)$ we have the properties:

a) If $\varphi(s)$ is generic, then the minimal k for which there exists a proper solution is $k=2n-1$. Furthermore, this solution is uniquely defined and $\delta_{\min}^*=\tilde{\delta}=n-1$.

b) For every $\varphi(s)$ such that $k > 2n-1$, the family $\mathcal{W}(\underline{b}_p, \underline{a}_p; \underline{\varphi})$ has all its elements proper, the minimal solutions have $\delta^* = k-n$, they are not uniquely defined and generically bi-proper.

c) If $(n_c^*(s), d_c^*(s))$ is a $\delta^* = k-n$ minimal McMillan degree solution for $k > 2n-1$, the family of solutions

$$\begin{aligned} n_c(s) &= n_c^*(s) + d_p(s)t(s) \\ d_c(s) &= d_c^*(s) - n_p(s)t(s) \end{aligned} \tag{3.35}$$

has all its elements proper for every $t(s)$. Furthermore this family may be partitioned into subfamilies with fixed McMillan degree as shown below:

i) If $\deg[t(s)] = \sigma \le k-2n$, then the McMillan degree of the subfamily is $\delta = k-n$.

ii) If $\deg[t(s)] = \sigma > k-2n$, then for all such $t(s)$ the subfamily has $\delta = \sigma + n > k-n$ degree.

■

The proof of the result is rather obvious and follows along the same lines with the previous proofs.

Example (3.4): Consider the case $m = n = 3$ and shall investigate the parameterisation of the overall family of solutions $\mathcal{W}(\underline{b}_p, \underline{a}_p; \underline{\varphi})$ for all possible values of k.

Case (I): Clearly the minimal k for is for $k = 2n-1 = 6-1 = 5$ and this corresponds to the solution of

$$\left[\begin{array}{ccc:ccc} b_0 & 0 & 0 & a_0 & 0 & 0 \\ b_1 & b_0 & 0 & a_1 & a_0 & 0 \\ b_2 & b_1 & b_0 & a_2 & a_1 & a_0 \\ b_3 & b_2 & b_1 & 1 & a_2 & a_1 \\ 0 & b_3 & b_2 & 0 & 1 & a_2 \\ 0 & 0 & b_3 & 0 & 0 & 1 \end{array}\right] \left[\begin{array}{c} \tilde{c}_0 \\ \tilde{c}_1 \\ \tilde{c}_2 \\ \hdashline \tilde{d}_0 \\ \tilde{d}_1 \\ \tilde{d}_2 \end{array}\right] = \left[\begin{array}{c} \varphi_0 \\ \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \end{array}\right]$$

\which is uniquely defined. For $t(s) = t_0$ we have

$$\left[\begin{array}{c} c_0 \\ c_1 \\ c_2 \\ c_3 \\ \hdashline d_0 \\ d_1 \\ d_2 \\ d_3 \end{array}\right] = \left[\begin{array}{c} \tilde{c}_0 \\ \tilde{c}_1 \\ \tilde{c}_2 \\ 0 \\ \hdashline \tilde{d}_0 \\ \tilde{d}_1 \\ \tilde{d}_2 \\ 0 \end{array}\right] + \left[\begin{array}{c} a_0 \\ a_1 \\ a_2 \\ 1 \\ \hdashline -b_0 \\ --b_1 \\ -b_2 \\ -b_3 \end{array}\right] t_0 \ , \quad \text{with } \delta = \sigma + n = 0 + 3 > k - n = 5 - 3 = 2$$

For $t(s) = t_1 s + t_0$ we have

$$\left[\begin{array}{c} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ \hdashline d_0 \\ d_1 \\ d_2 \\ d_3 \\ d_3 \end{array}\right] = \left[\begin{array}{c} \tilde{c}_0 \\ \tilde{c}_1 \\ \tilde{c}_2 \\ 0 \\ 0 \\ \hdashline \tilde{d}_0 \\ \tilde{d}_1 \\ \tilde{d}_2 \\ 0 \\ 0 \end{array}\right] + \left[\begin{array}{cc} a_0 & 0 \\ a_1 & a_0 \\ a_2 & a_1 \\ 1 & a_2 \\ 0 & 1 \\ \hdashline -b_0 & 0 \\ -b_1 & -b_0 \\ -b_2 & -b_1 \\ -b_3 & -b_2 \\ 0 & -b_3 \end{array}\right] \left[\begin{array}{c} t_0 \\ t_1 \end{array}\right], \quad \text{with } \delta = \sigma + n = 1 + 3 > k - n = 2$$

and similarly for $t(s)=t_0+t_1s+\ldots+t_rs^r$, $t_r\neq 0$ we obtain the corresponding family with $\delta=\sigma+n=r+n$.

Case (II): Consider now the case where we could like to assign a polynomial with degree $k=7$ (not the absolute minimum). A minimal McMillan degree that corresponds to $k=7$ is defined with $\delta^*=k-n=7-3=4$ and computed from the equation

$$\left[\begin{array}{ccc:ccc:cc} b_0 & 0 & 0 & a_0 & 0 & 0 & 0 & 0 \\ b_1 & b_0 & 0 & a_1 & a_0 & 0 & 0 & 0 \\ b_2 & b_1 & b_0 & a_2 & a_1 & a_0 & 0 & 0 \\ b_3 & b_2 & b_1 & 1 & a_2 & a_1 & 0 & 0 \\ 0 & b_3 & b_2 & 0 & 1 & a_2 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 & 1 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]\left[\begin{array}{c} c_0' \\ c_1' \\ c_2' \\ \hdashline d_0' \\ d_1' \\ d_2' \\ d_3' \\ d_4' \end{array}\right]=\left[\begin{array}{c} \varphi_0 \\ \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \\ \varphi_7 \end{array}\right], \quad d_4'=\varphi_7\neq 0$$

We consider now the families of solutions defined by

$$\left[\begin{array}{c} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ \hdashline d_0 \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array}\right]=\left[\begin{array}{c} c_0' \\ c_1' \\ c_2' \\ 0 \\ 0 \\ 0 \\ 0 \\ \hdashline d_0' \\ d_1' \\ d_2' \\ d_3' \\ d_4' \\ 0 \\ 0 \end{array}\right]+\left[\begin{array}{cccc} a_0 & 0 & 0 & 0 \\ a_1 & a_0 & 0 & 0 \\ a_2 & a_1 & a_0 & 0 \\ 0 & a_2 & a_1 & a_0 \\ 0 & 0 & a_2 & a_1 \\ 0 & 0 & 0 & a_2 \\ 0 & 0 & 0 & 0 \\ \hdashline -b_0 & 0 & 0 & 0 \\ -b_1 & -b_0 & 0 & 0 \\ -b_2 & -b_1 & -b_0 & 0 \\ -b_3 & -b_2 & -b_1 & -b_0 \\ 0 & -b_3 & -b_2 & -b_1 \\ 0 & 0 & -b_3 & -b_2 \\ 0 & 0 & 0 & -b_3 \end{array}\right]\begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{bmatrix}$$

Then we distinguish the cases:

i) $\deg\left[t(s)\right]=\sigma \leq k-2n=7-6=1$, i.e. $t_3=t_2=0$. All solutions are proper with $\delta=k-n=7-3=4$, which is equal to the minimal.

ii) $\deg\left[t(s)\right]=\sigma > k-2n=1$. All solutions are proper and if $t_3=0$, then $\delta=5$, whereas if $t_3\neq 0$, i.e. $\sigma=3$, then all resulting solutions are proper with $\delta=6$. Clearly, if $\sigma=r$, then the corresponding family has McMillan degree $\delta=n+\sigma$

■

3.6. DISCUSSION

In this chapter we have developed a complete analysis of the properties of the solutions of scalar Diophantine equations. We focus on the investigation of the proper solution, a problem that is related to the pole assignment for SISO systems. The minimal McMillan degree solution was characterised and parameterisations of the family of solutions according to McMillan degree was given. The results presented here for the scalar Diophantine equation may extend to the case of the vector Diophantine equation, or the generalised scalar Diophantine equation

$$x_1(s)a_1(s)+x_2(s)a_2(s)+\ldots+x_n(s)a_n(s)=\varphi(s)$$

where $a_i(s)\in\mathbb{R}[s]$ given and $x_i(s)\in\mathbb{R}[s]$ are the unknowns.

Chapter 4:

STRUCTURE AND PROPERTIES OF RESULTANTS AND THEIR USE IN THE CHARACTERISATION OF THE GCD OF MANY POLYNOMIALS

4.1. INTRODUCTION

The computation of the greatest common divisor (gcd) of a set of polynomials has attracted a lot of attention in the recent years ([Barnett, 1983], [Vardoulakis et al., 1978], [Pace et al., 1973], [Karcanias, 1987], [Mitrouli et al., 1993], [Karcanias et al., 1994] and references therein) and has widespread applications in linear system, network theory and control [Rosenbrock, 1970], [Kailath, 1980]. The methodologies dealing with gcd computation may be distinguished into those based on Euclid's algorithm and the matrix based methodologies. The class of matrix based methodologies is based on the properties of the Basis matrix [Karcanias, 1987], or the Generalised Sylvester Resultant [Barnett, 1972] of the given set of polynomials and have the advantage that can deal simultaneously with many polynomials and reduce computation of gcd to standard linear algebra problems. In particular, the matrix pencil approach [Karcanias et al., 1994] provides an efficient procedure and establishes links with standard problems of system theory. The ERES method [Mitrouli et al., 1993] exploits the invariance properties of the gcd under row transformations and shifting [Karcanias, 1987] and allows the derivation of "approximate gcd", when the system data are not accurate. The Generalised Sylvester Resultant [Barnett, 1990], [Vardoulakis et al., 1978] provides a simple characterisation of coprimeness and a procedure for gcd computation, but increases the number of polynomials used considerably. As far as improving existing computational methods, there are clear advantages in linking ERES, Matrix pencil and Generalised Resultant approaches.

This chapter deals with the development of some basic properties of resultants and with two matrix based algorithms for the computation of the *greatest common divisor* (gcd) of a set of polynomials. A new proof of the resultant theorem is shown in this chapter based on a new property of the Sylvester matrix. This proof also establishes the isomorphism expressing the factorisation out of factors or gcd as an equivalent Toeplitz matrix in the resultant representation set up. This property expresses also the invariance of the remaining coprime factors of the polynomials under appropriate column operations. An implementation of one of these algorithms using MATLAB 5.3, is also suggested. All basic propositions, in this paragraph, are proved by using simple algebraic properties.

The new results in combination with the ERES and Matrix pencil methods will then provide simplified procedures for the gcd computation and a new representation of the gcd. Central to these developments are results on the factorisation of resultants in terms of reduced order resultants and square Toeplitz matrices. Such factorisations are equivalent to the extraction of common factors from the set of polynomials; the gcd is then represented by a Toeplitz matrix with an irreducible reduced order generalised resultant expressing the remaining polynomials in the factorisation. This new representation of the gcd allows the unification of a number of known results and the development of a simplified version of ERES method that avoids the operation of shifting. The new representation of the gcd based on the canonical factorisation of the Generalised Resultant into a reduced resultant and a Toeplitz matrix defining the gcd opens new ways for study of approximate solutions to gcd evaluation given that it provides a minimal parametric description for possible gcds.

The matrix-pencil algorithm, introduced by [Karcanias, 1987], combines the mathematical theory on the matrix-pencils developed by [Gantmacher, 1988], with properties of Control Theory. In this Chapter a variation of the matrix pencil algorithm is described that uses the resultant set of polynomials associated with the given set. The advantage of this augmentation of the original set is that the right kernel of the resultant set completely characterises the GCD and this leads to a simpler formulation of the matrix pencil algorithm. The factorisation of resultants in terms of reduced resultant and Toeplitz representations of the GCD provide the means for discussing approximate GCD ideas in subsequent chapters.

4.2. DEFINITIONS AND PRELIMINARY RESULTS

Construction of Sylvester Resultant

The construction of the Sylvester matrix [Barnett, 1990] is described here initially for the case of two polynomials. In this case the Sylvester matrix is square. The Sylvester matrix for a set of three or more polynomials is based on the case of two as the dimensions are defined by the two polynomials of the highest degree.

Consider two polynomials $a(s) = s^n + a_{n-1}s^{n-1} + \ldots a_1 s + a_0$ and $b(s) = b_p s^p + b_{p-1}s^{p-1} + \ldots b_1 s + b_0$ where $a(s)$ is monic, and suppose that $\deg\{b(s)\} \leq \deg\{a(s)\}$. If the polynomials have a common factor, this means that there exists a value of $s = s_0$ for which the equations $a(s_0) = 0$, $b(s_0) = 0$ are simultaneously satisfied. If we multiply the first equation by, $s^{p-1}, s^{p-2}, \ldots, s, 1$ respectively, we obtain the n equations

$$\left.\begin{aligned}
s^{n+p-1} + a_{n-1}s^{n+p-2} + \ldots + a_1 s^n + a_0 s^{n-1} &= 0 \\
s^{n+p-2} + a_{n-1}s^{n+p-3} + \ldots + a_1 s^{p-1} + s^{p-2} &= 0 \\
\vdots \qquad\qquad\qquad & \vdots \\
s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0 &= 0
\end{aligned}\right\} \tag{4.1}$$

Similarly, if we multiply the second equation by $s^{n-1}, s^{n-2}, \ldots, s, 1$, respectively we obtain the m equations

$$\left.\begin{aligned}
b_p s^{n+p-1} + b_{p-1}s^{n+p-2} + \ldots + b_1 s^n + b_0 s^{n-1} &= 0 \\
b_p s^{n+p-2} + b_{p-1}s^{n+p-3} + \ldots + b_1 s^{n-1} + s^{n-2} &= 0 \\
\vdots \qquad\qquad\qquad & \vdots \\
b_p s^p + b_{p-1}s^{p-1} + \cdots + b_1 s + b_0 &= 0
\end{aligned}\right\} \tag{4.2}$$

It is readily established by using factorisation of the above polynomials, that such operations do not affect the g.c.d. of the original two polynomials. In fact the set $\mathcal{P} = \{a(s), b(s)\}$ and the resulting set

$S[\mathcal{P}] = \{a(s), sa(s), \ldots, s^{p-1}a(s), b(s), sb(s), \ldots, s^{n-1}b(s)\}$ has the same g.c.d.. The set $S[\mathcal{P}]$ is called the *Sylvester Resultant set* of the original set $\mathcal{P} = \{a(s), b(s)\}$ of polynomials. Such a construction leads to the following definition:

<u>Definition 4.1</u>: [Barnett, 1990] Given a set of polynomials $\mathcal{P} = \{a(s), b(s)\}$ the matrix of coefficients $S \in \mathbb{R}^{(n+p)\times(n+p)}$ of the $(n+p)$ Sylvester Resultant set of $\mathcal{P}$, $S[\mathcal{P}]$, defined as:

$$S=\left[\begin{array}{ccccccccccc}
1 & a_{n-1} & a_{n-2} & & \cdots & a_0 & 0 & & \cdots & 0 & 0 \\
0 & 1 & a_{n-1} & & \cdots & a_1 & a_0 & 0 & \cdots & 0 & 0 \\
\cdot & \cdot & \cdot & & \cdots & \cdot & \cdot & & \cdots & \cdot & \cdot \\
0 & 0 & 0 & \cdots & 1 & \cdots & & & & a_1 & a_0 \\
b_p & b_{p-1} & b_{p-2} & \cdots & & b_0 & 0 & 0 & \cdots & 0 & 0 \\
0 & b_p & b_{p-1} & \cdots & & b_1 & b_0 & 0 & \cdots & 0 & 0 \\
\cdot & \cdot & \cdot & \cdots & & \cdot & \cdot & \cdots & & \cdot & \cdot \\
0 & 0 & 0 & \cdots & b_p & b_{p-1} & & \cdots & & b_1 & b_0
\end{array}\right]
\begin{array}{c}
\uparrow \\ \\ p\text{ rows} \\ \downarrow \\ \uparrow \\ \\ n\text{ rows} \\ \downarrow
\end{array}
\tag{4.3}$$

is called the ***Sylvester's Resultant Matrix*** associated with the polynomials $a(s)$ and $b(s)$.

■

A first observation is that S is a basis matrix of the polynomials set $S[\mathcal{P}]=\{a(s),sa(s),\dots,s^p a(s),b(s),sb(s),\dots s^n b(s)\}$. The following result is obvious from the factorization property of the resulting polynomials.

Consider a set of polynomials $\mathcal{P}=\{a(s),b_i(s)\in\mathbb{R}[s],\ i\in\underset{\sim}{h}\}$ which has $h+1$ elements and with the two largest values of degrees (n,p). Without loss of generality we may assume $a(s)$ monic and represent the polynomials with respect to the n degree as $a(s)=s^n+a_{n-1}s^{n-1}+\dots+a_1 s+a_0$ and $b_i(s)=b_{in}s^n+\dots+b_{i1}s+b_{i0}$, $i=1,2,,\dots,h$. Whenever we want denote the number of elements and the maximal degree we shall use the notation $\mathcal{P}_{h+1,n}$. The greatest common divisor (gcd) of $\mathcal{P}$ will be denoted by $\varphi(s)\triangleq\gcd\{\mathcal{P}\}$. With the set $\mathcal{P}$ we may associate the polynomial vector

$$\begin{aligned}
\underline{p}_{h+1}(s)&=\begin{bmatrix} a(s) \\ b_1(s) \\ \vdots \\ b_h(s) \end{bmatrix}=\left[\underline{p}_n,\underline{p}_{n-1},\dots,\underline{p}_1,\underline{p}_0\right]\underline{e}_n(s) \\
&=P_{h+1}\underline{e}_n(s),\quad \underline{e}_n(s)=\left[s^n,s^{n-1},\dots,s,1\right]^t
\end{aligned}\tag{4.4}$$

where $P_{h+1} \in \mathbb{R}^{(h+1)\times(n+1)}$ is referred to as a *basis matrix* (bm) of $\mathcal{P}$ and $\underline{p}_{h+1}(s)$ as a *vector representative* (vr) of $\mathcal{P}$. If c is the integer for which $\underline{p}_n = \underline{p}_{n-1} = \cdots = \underline{p}_{c-1} = \underline{0}$, $\underline{p}_c \neq \underline{0}$, then c will be referred to as the *order* of $\mathcal{P}$ and it is denoted by $c \triangleq \omega(\mathcal{P})$. The set is called *proper* if $c = 0$ and *nonproper*, if $c \geq 1$.

<u>Remark (4.2):</u> If $c \geq 1$,then the set $\mathcal{P}$, then s^c is an elementary divisor (ed) of $\varphi(s)$ and if $c = 0$, then the set $\mathcal{P}$ is coprime at $s = 0$ ($\varphi(0) \neq 0$).

■

On sets of polynomials we may introduce a notion of equivalence that preserves the properties of gcd and which provide a usefull framework for its computation (Karcanias, 1987).

<u>Definition (4.3):</u> Let $\mathcal{P}_{m,n}$ be a set of polynomials and P_m be its bm. On P_m and thus also on $\mathcal{P}_{m,n}$ we may define the following operations:

(i) Elementary row operations with scalars from $\mathbb{R}$ on P_m.

(ii) Addition, or elimination of zero rows on P_m.

(iii) If $\underline{a}^t = [a_n, \ldots, a_\varepsilon, 0, \ldots, 0] \in \mathbb{R}^{1\times(n+1)}$, $a_\varepsilon \neq 0$ is a row of P_m, then we define as the shifting operation shf : $\text{shf}\left(\underline{a}^t\right) = \underline{a}^{*t} = [0, \ldots 0, a_n, \ldots, a_\varepsilon] \in \mathbb{R}^{1\times(n+1)}$, $a_\varepsilon \neq 0$

■

Type (i), (ii) and (iii) operations are referred to as *extended row equivalence and shifting* (ERES) operations, while *extended row equivalence* (ERE) will be simply used for types (i) and (ii). These operations on P_m have an interpretation on the set of polynomials. In fact, type (i) operations imply that we can rearrange the order, scale the coefficients by non-zero constants and substitute a polynomial by a linear combination of polynomials of the set. Type (ii) operation imply that we may eliminate all zero polynomials, or add any number of zero polynomials, or linear combinations of polynomials of the set. Type (iii) operations, imply that if we have a polynomial $p(s) = s^c p'(s)$ in the set, then we may substitute it by the polynomial

$p'(s)$, which clearly has less degree. By $\mathrm{shf}\left(\mathcal{P}_{m,n}\right)=\mathcal{P}_{m,n}^{*}$ we shall denote the set obtained from $\mathcal{P}_{m,n}$ by applying shift on every polynomial. A number of properties of the gcd under the above transformations are summarised below [Karcanias, 1987]:

Theorem (4.1): For any set $\mathcal{P}_{m,n}$ with a bm P_m, $\rho(P_m)=r$ and $\gcd\{\mathcal{P}_{m,n}\}=\varphi(s)$ we have the following properties:

(i) If $\mathcal{R}$ is the row space of P_m, then $\varphi(s)$ is invariant of $\mathcal{R}$. Furthermore, if $r=\dim\mathcal{R}=n+1$, then $\varphi(s)=1$.

(ii) If $\omega\left(\mathcal{P}_{m,n}\right)=c\geq 1$ and $\mathrm{shf}\left(\mathcal{P}_{m,n}\right)=\mathcal{P}_{m,n}^{*}$, then $\varphi(s)=\gcd\left\{\mathcal{P}_{m,n}\right\}=s^{c}\gcd\left\{\mathcal{P}_{m,n}^{*}\right\}$

(iii) If $\mathcal{P}_{m,n}$ is proper, then $\varphi(s)$ is invariant under ERES operations.

■

The above result forms the basis for the ERES methodology where Gaussian transformations with partial pivoting are used together with shifting to produce in a finite number of steps the gcd, or approximations of the gcd [Mitrouli et al., 1993]. Some important properties emerging from this result are:

Remark (4.2): In developing methodologies for computing the gcd the following points should be taken into account:

(a) Not all polynomials in $\mathcal{P}_{m,n}$ are needed for the computation of the gcd; in fact, only a subset with the property that it provides a basis for the row space $\mathcal{P}$ of P_m is required.

(b) The computation of gcd may always be reduced to computing the gcd of a proper set, if shifting is applied on the original set.

■

The above suggests that studies of gcd for a set $\mathcal{P}_{m,n}$ may be restricted to any subset $\tilde{\mathcal{P}}_{r,n}$ of $\mathcal{P}_{m,n}$ where $r=\rho(P_m)$ and with the property that the rows of $\tilde{P}_r$ define a basis for the row space $\mathcal{R}$ of P_m. Any such subset will be referred to as a *normal* subset of $\mathcal{P}_{m,n}$ and it has at most $n+1$ elements ($\rho(P_M)\leq n+1$). For numerical computations the selection of the "best" normal subset of $\mathcal{P}_{m,n}$ becomes an important issue [Mitrouli et al., 1993].

The classical approaches for the study of coprimeness and determination of the gcd makes use of the Sylvester Resultant which in the case of many polynomials is defined as shown below [Barnett, 1983].

Definition (4.4): Consider the set $\mathcal{P}_{h+1,n}=\left\{a(s),b_i(s),i\in \underset{\sim}{h},\ n=\deg\left\{a(s)\right\}\right.$, $n\geq\deg\left\{b_i(s)\right\}\forall i\in \underset{\sim}{h},\ p\text{=}\max\left\{\deg\left\{b_i(s)\right\},i\in\underset{\sim}{h}\right\}\}$, where $a(s)$, $b(s)$ are described as:

$$a(s)=s^n+a_{n-1}s^{n-1}+\ldots+a_1s+a_0,\ b_i(s)=b_{i,p}s^p+\ldots+b_{i,1}s+b_{i,0},\ i=1,2,\ldots h \quad (4.5)$$

(i) We can define a $p\times(n+p)$ matrix associated with $a(s)$:

$$S_0=\begin{bmatrix}1 & a_{n-1} & a_{n-2} & \cdots & a_1 & a_0 & 0 & \cdots & 0\\ 0 & 1 & a_{n-1} & \cdots & a_2 & a_1 & a_0 & \cdots & 0\\ \vdots & & \ddots & & & & & & \vdots\\ 0 & 0 & \cdots & 1 & a_{n-1} & \cdots & & a_1 & a_0\end{bmatrix} \quad (4.6a)$$

and an $n\times(n+p)$ matrix associated with $b_i(s)$:

$$S_i=\begin{bmatrix}b_{i,p} & b_{i,p-1} & b_{i,p-2} & \cdots & b_{i,1} & b_{i,0} & 0 & \cdots & 0\\ 0 & b_{i,p} & b_{i,p-1} & \cdots & b_{i,2} & b_{i,1} & b_{i,0} & \cdots & 0\\ \vdots & & \ddots & & & & & & \vdots\\ 0 & \cdots & 0 & b_{i,p} & b_{i,p-1} & \cdots & & b_{i,1} & b_{i.0}\end{bmatrix} \quad (4.6b)$$

for each $i=1,2,\cdots,h$. An ***extended Sylvester matrix*** for the set $\mathcal{P}$ is then defined by:

$$S_{\mathcal{P}}=\begin{bmatrix}S_0\\ S_1\\ \vdots\\ S_h\end{bmatrix}\in\mathbb{R}^{(p+hn)\times(n+p)} \quad (4.6c)$$

(ii) The matrix S_P is the basis matrix of the set of polynomials

$$S[\mathcal{P}]=\left\{a(s),sa(s),\ldots,s^{p-1}a(s);b_1(s),\ldots,b_h(s),sb_h(s),\ldots,s^{n-1}b_h(s)\right\} \quad (4.7)$$

which is also referred to as the *Sylvester Resultant set* of the given set $\mathcal{P}$.

■

From the factorisation property of the polynomials in $\mathcal{S}[\mathcal{P}]$ we have the following obvious results.

Proposition (2.1): The gcd of $\mathcal{P}$ is the same as the gcd of $\mathcal{S}[\mathcal{P}]$, that is

$$\gcd\{\mathcal{P}\} = \gcd\{\mathcal{S}[\mathcal{P}]\} \tag{4.8}$$

■

The above suggests that the resultant set may be used for the evaluation of gcd. Note that $\mathcal{S}[\mathcal{P}]$ has more elements than $\mathcal{P}$, but it is advantageous to use it due to certain special properties. The Sylvester Resultant characterises the gcd in a simple way [Barnett, 1972]. The classical result is usually given for two polynomials and this is also extended to the case of many polynomials [Vardoulakis et al., 1978]. Here we examine certain properties of extraction of divisors from the set $\mathcal{P}$, which are equivalently expressed as factorisation of resultant matrices. This leads to establishing a link between factorisation of resultants and a matrix representation of the gcd. The new representation of the gcd provides the means for deriving an alternative proof to the classical Resultant Theorem for the cases of many polynomials; Furthermore, they enable the simplification of ERES numerical method [Mitrouli, et al., 1991] and the Matrix Pencil Method [Karcanias et al., 1994] for the computation of the gcd.

4.3. FACTORISATION OF SYLVESTER RESULTANTS AND THE MATRIX REPRESENTATION OF GCD

Some known results [Barnett, 1990] related with the Sylvester matrix of two polynomials will be described first. New proofs for these classical theorems and new properties will be introduced then for the case of many polynomials

4.3.1 Properties of Resultant of two polynomials

The following theorem is a classical result for the study of gcd of two polynomials based on the Sylvester resultant matrix.

Theorem 4.2: (Resultant theorem) [Barnett, 1990] Given two polynomials $a(s)$ and $b(s)$ then:

i) A necessary and sufficient condition for two polynomials $a(s)$ and $b(s)$ to have nontrivial g.c.d. is that the Sylvester's Resultant associated with $a(s)$ and $b(s)$ is singular.

ii) The degree of the greatest common divisor of two polynomials $a(s)$ and $b(s)$, of degree n and p respectively, is equal to:

$$\deg \varphi(s) = n + p - \operatorname{rank} S \tag{4.9}$$

where S is the Sylvester's Resultant associated with $a(s)$ and $b(s)$.

iii) If we reduce the Sylvester's Resultant matrix to its echelon form, the last non-vanishing row defines the coefficients of the greatest common divisor

■

Example 4.1: Let $a(s) = 3s^3 - 2s^2 + 3s - 2$, $b(s) = 3s^2 + s - 2$ then the associated Sylvester's matrix is:

$$S = \left[\begin{array}{ccccc} 3 & -2 & 3 & -2 & 0 \\ 0 & 3 & -2 & 3 & -2 \\ \hline 3 & 1 & -2 & 0 & 0 \\ 0 & 3 & 1 & -2 & 0 \\ 0 & 0 & 3 & 1 & -2 \end{array}\right]$$

Note that $\operatorname{rank}(S) = 4$ and thus the degree of the g.c.d $d(s)$ equals to 1. If we reduce S to his echelon row form, using only row operations we find

$$S_e = \begin{bmatrix} 1 & \frac{1}{3} & -\frac{2}{3} & 0 & 0 \\ 0 & 1 & \frac{1}{3} & -\frac{2}{3} & 0 \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{2}{3} \\ 0 & 0 & 0 & 1 & -\frac{2}{3} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

From the last non-vanishing row and Theorem 4.3 it is implied that

$$d(s) = s - \frac{2}{3}$$

■

In some applications it is more convenient to use a slightly different form of the Sylvester matrix in which the last n rows are in reversed order. In this case we can construct centrally situated sub-matrices by successively deleting a row and a column all the way round. Then we can define the sub-resultants as follows:

Definition 4.5: We denote the i-th sub-resultant S_i of two polynomials the determinant obtained by striking out the first i and the last i rows and also the first i columns and the last i columns from the resultant of these polynomials.

■

Example 4.2: If we have the polynomials $a(s) = a_3 s^3 + a_2 s^2 + a_1 s + a_0$ and $b(s) = b_2 s^2 + b_1 s + b_0$ of degree 3 and 2 respectively then their resultant R is a determinant of order 5 R_1 is the sub-resultant of order 3 and R_2 the sub-resultant of order 1 as it is shown below.

$$R = \begin{vmatrix} a_3 & a_2 & a_1 & a_0 & 0 \\ 0 & a_3 & a_2 & a_1 & a_0 \\ 0 & R_1 = 0 & R_2 = b_2 & b_1 & b_0 \\ 0 & b_2 & b_1 & b_0 & 0 \\ b_2 & b_1 & b_0 & 0 & 0 \end{vmatrix}$$

■

The propositions below use the semi-reversed form of the resultant in order to find the degree and the coefficients of the g.c.d. polynomial.

Proposition 4.2: [Barnett, 1990]: The degree of the greatest common divisor of $a(s)$ and $b(s)$ where $\deg a(s) = n$ and $\deg b(s) = m$, is equal to the subscript of the first of the sub-resultants which does not vanish.

■

The next theorem describes an alternative procedure for the evaluation of the g.c.d. of two polynomials:

Theorem 4.3 [Barnett, 1990]: If i is the degree of the greatest common divisor of two polynomials $a(s)$ and $b(s)$, then this greatest common divisor may be obtained from the i-th sub-resultant of $a(s)$ and $b(s)$ by replacing the last element in the last row of coefficients of $a(s)$ by $a(s)$, the element just above this by $s \cdot a(s)$, the one just above this by $s^2 a(s)$, etc.; and replacing the last element in the first row of elements of $b(s)$ by $b(s)$, the element just below this by $s \cdot b(s)$ etc.

■

Example 4.3: Referring to example 4.1, we may use the semi-reduced formula of S for the polynomial then the resultant would become:

$$R = \begin{vmatrix} 3 & -2 & 3 & -2 & 0 \\ 0 & 3 & -2 & 3 & -2 \\ 0 & 0 & 3 & 1 & -2 \\ 0 & 3 & 1 & -2 & 0 \\ 3 & 1 & -2 & 0 & 0 \end{vmatrix} = 0$$

and the first sub-resultant is:

$$R_1 = \begin{vmatrix} 3 & -2 & 3 \\ 0 & 3 & 1 \\ 3 & 1 & -2 \end{vmatrix} = -54$$

By applying the procedure described in Theorem 4.3 we conclude that $\deg d(s) = 1$ and $d(s)$ is given from:

$$d(s) = \begin{vmatrix} 3 & -2 & 3s^3 - 2s^2 + 3s - 2 \\ 0 & 3 & 3s^2 + s - 2 \\ 3 & 1 & 3s^3 + s^2 - 2s \end{vmatrix} = -54\left(s - \frac{2}{3}\right)$$

■

Properties of Resultant of many polynomials and its factorisation

The aim of the work here is to consider the fundamentals behind the properties of the resultants and in doing so derive new properties of the Sylvester Resultant linked to GCD and its evaluation. We start by considering the classical approaches and results and then derive some important new results. The properties of the

Sylvester resultant matrix we examine below apply to any polynomial set. The significance of these resultant results is that they lead as to conclusions about the degree of the g.c.d. of the polynomial set and, under certain type of operations, methods for its evaluation. For the establishment of the properties of the Sylvester resultant matrix we need to examine some general properties of real polynomials.

The factorisation of common divisors from the set of polynomials $\mathcal{P}_{h+1,n}$ leads to factorised or reduced sets and has certain implications on the resultant of the set. In fact, such a factorization of polynomials leads to a factorization of the corresponding resultant, which in turn provides the basis for the matrix representation of gcd. Establishing a representation of gcd is the subject of this section and it is equivalent to a factorization of the original resultant into a reduced resultant, (corresponding to the remaining factors after extracting the gcd) and a square Toeplitz type matrix representing the gcd.

We consider first the case of non-proper sets and then examine the factorization of proper sets.

Let $\mathcal{P}_{h+1,n} = \left\{a(s), b_i(s), i \in \underset{\sim}{h},\ n = \deg\{a(s)\}, n \geq \deg\{b_i(s)\}\ \forall i \in \underset{\sim}{h},\ p\text{=}\max\left\{\deg\{b_i(s)\}\right\}\right\}$.

If $c = \omega\left(\mathcal{P}_{h+1,n}\right)$, then we have the obvious factorisation

$$a(s) = s^c \tilde{a}(s),\ b_i(s) = s^c \tilde{b}_i(s),\ \forall i \in \underset{\sim}{h} \tag{4.10}$$

where $\deg\{\tilde{a}(s)\} = n - c$, $\tilde{p} = \max\left\{\deg\{b_i(s)\}\right\} = p - c$ and the set $\tilde{\mathcal{P}}_{h+1,n-c} = \left\{\tilde{a}(s), \tilde{b}_i(s), i \in \underset{\sim}{h}\right\}$ is the reduced set of factorisation of the s^c factor. The presence of the s^c common factor in $\mathcal{P}_{h+1,n}$ implies that the corresponding generalised resultant $S_{\mathcal{P}}$ has the following form

$$S_{\mathcal{P}} = \left[\bar{S}_{\tilde{\mathcal{P}}} \mid \mathbf{0}_c\right] \in \mathbb{R}^{(p+hm)\times(n+p)} \tag{4.11a}$$

where $\mathbf{0}_c$ is c column block of zeros and $\bar{S}_{\tilde{\mathcal{P}}}$ is the nonzero part of the $S_{\mathcal{P}}$ Sylvester resultant. Note that $\bar{S}_{\tilde{\mathcal{P}}}$ has the form of the resultant and in fact is the expansion of the resultant of the $\tilde{\mathcal{P}}_{h+1,n-c}$ set of polynomials from the $(n-c, p-c)$ degrees to (n, p); that is we assume that the two maximal degrees are n, p respectively and then we drop the first c zero columns of the matrix. Because of the links of $\bar{S}_{\tilde{\mathcal{P}}}$ to the resultant of $\tilde{\mathcal{P}}_{h+1,n-c}$ when we assume that the two maximal degrees are (n,p), $\bar{S}_{\tilde{\mathcal{P}}}$ will be called an (n,p)*-expanded resultant* of the $\tilde{\mathcal{P}}_{h+1,n-c}$ set. Expression (4.11a) readily implies the following result.

Proposition (4.3): The Sylvester resultant of the c-order set $\mathcal{P}_{h+1,n}$ may be expressed as

$$S_{\mathcal{P}} = \left[\mathbf{0}_c \mid \bar{S}_{\tilde{\mathcal{P}}} \right] \begin{bmatrix} \mathbf{0} & I_c \\ I_{n+p-c} & \mathbf{0} \end{bmatrix} = \bar{S}^c_{\tilde{\mathcal{P}}} \, \hat{Q}_c \tag{4.11b}$$

■

This obvious result (block permutation) implies a general property that is examined next, that is the extraction of common divisor (in this case s^c) is equivalent to a factorization of the Sylvester resultant. This factorization is expressed in terms of two matrices $\bar{S}^c_{\tilde{\mathcal{P}}}$ and $\hat{Q}_c$, where $\bar{S}^c_{\tilde{\mathcal{P}}}$ is referred to as a *reduced resultant* of $S_{\mathcal{P}}$ and $\hat{Q}_c$ is a Toeplitz matrix of $(n+p)\times(n+p)$ dimensions characterising the divisor s^c. Thus, (4.11b) may be interpreted as the representation of the (4.10) factorization.

The representation of the s^c common divisor is now extended to the case of general, not necessarily divisors at $s=0$. Thus let $\mathcal{P}_h = \{a(s), b_1(s), \ldots, b_h(s)\}$ be a set of polynomials such that $a(s) = a_n s^n + \ldots + a_1 s + a_0$ $b_i(s) = b_{i,p} s^p + \ldots + b_{i,1} s + b_{i,0}$, $p \le n$ and let $(s-r)$, $r \ne 0$ be a common factor of $a(s)$ and $b_i(s)$ $i=1,\ldots,h$. Then we can write:

$$a(s) = (s-r)\left(a'_{n-1}s^{n-1} + a'_{n-2}s^{n-2} + \ldots + a'_1 s + a'_0\right) \tag{4.12}$$

$$b_i(s) = (s-r)\left(b'_{i,n-1}s^{n-1} + b'_{i,n-2}s^{n-2} + \ldots + b'_{i,1}s + b'_{i,0}\right) \quad i=1,\ldots,h \tag{4.13}$$

which express algebraically the factorisation of $(s-r)$ factor. The matrix representation of this factorisation is expressed as shown below:

Proposition(4.4): Let $\mathcal{P}_h = \{a(s), b_1(s), \ldots, b_h(s)\}$ be a set of polynomials such that $\deg a(s) = n$, $\deg b_i(s) \le p$, $i=1,\ldots,n$, $p \le n$ and let $(s-r)$ be a common devisor of the elements of $\mathcal{P}_h$, with $r \neq 0$. Then, there always exists a transformation Q_r defined as in (4.14) and a reduced resultant $\overline{S}_{\mathcal{P}'}^{(1)}$ defined as shown in (4.15) where $\mathcal{P}'$ is the reduced polynomial set after extraction of $(s-r)$ factor such that if $S_{\mathcal{P}}$ is the resultant of $\mathcal{P}$, there is a reduced resultant $\overline{S}_{\mathcal{P}'}^{(1)}$ such that

$$Q_r^{n+p} = -\begin{bmatrix} \frac{1}{r} & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ \frac{1}{r^2} & \frac{1}{r} & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & & \vdots \\ \frac{1}{r^{k-1}} & \frac{1}{r^{k-1}} & \cdots & \frac{1}{r} & 0 & \cdots & 0 & 0 & 0 \\ \frac{1}{r^k} & \frac{1}{r^k} & \cdots & \frac{1}{r^2} & \frac{1}{r} & \cdots & 0 & 0 & 0 \\ \vdots & & & & & \ddots & & & \vdots \\ \frac{1}{r^{n+p-1}} & \frac{1}{r^{n+p-2}} & \cdots & \frac{1}{r^{l-2}} & \frac{1}{r^{l-3}} & \cdots & \frac{1}{r^2} & \frac{1}{r} & 0 \\ \frac{1}{r^{n+p}} & \frac{1}{r^{n+p-1}} & \cdots & \frac{1}{r^{l-1}} & \frac{1}{r^{l-2}} & \cdots & \frac{1}{r^3} & \frac{1}{r^2} & \frac{1}{r} \end{bmatrix} \tag{4.14}$$

$$\overline{S}_{\mathcal{P}'}^{(1)} = S_{\mathcal{P}} Q_r^{n+p} \tag{4.15}$$

$$\overline{S}_{\mathcal{P}'}^{(1)} = \left[\begin{array}{cccccccc} 0 & a'_{n-1} & \cdots & & a'_0 & 0 & \cdots & 0 \\ 0 & 0 & a'_{n-1} & \cdots & & a'_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & a'_{n-1} & & & a'_0 \\ \hdashline 0 & b'_{1,p-1} & \cdots & & b'_{1,0} & 0 & \cdots & 0 \\ 0 & 0 & b'_{1,p-1} & \cdots & & b'_{1,0} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & b'_{1,p-1} & \cdots & & b'_{1,0} \\ \hdashline \vdots & & & & & & & \\ & & \vdots & & \vdots & & & \\ \hdashline 0 & b'_{h,p-1} & \cdots & & b'_{h,0} & 0 & \cdots & 0 \\ 0 & 0 & b'_{h,p-1} & \cdots & & b'_{h,0} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & b'_{h,p-1} & \cdots & & b'_{h,0} \end{array}\right] \tag{4.16}$$

Proof:

By (4.12) (4.13) and the definitions of the polynomials it is implied that

$$a_n = a'_{n-1},\quad a_i = a'_{i-1} - ra'_i\ ,\ i = 1,\dots,n-1,\quad a_0 = ra'_0 \tag{4.17}$$

$$b_{i,p} = b'_{i,p-1},\quad b_{i,j} = b'_{i,j-1} - rb'_{i,j}\ ,j = 1,\dots,p-1,\quad b_{i,0} = rb'_{i,0},\ i = 1,\dots,h \tag{4.18}$$

and thus the respective Sylvester matrix will have the form indicated in (4.19). Supposing $r \neq 0$, we can perform the transformations:

i) multiply the last column by $-\frac{1}{r}$ and

ii) subtract it from the $n+p-1$ column of S_φ.

Then the resulting matrix that is derived from (4.19) has the form indicated in (4.20)

$$S'_\varphi = \left[\begin{array}{cccccccc}
a'_{n-1} & a'_{n-2} - ra'_{n-1} & \cdots & & -ra'_0 & 0 & \cdots & 0 \\
\vdots & \ddots & & & & \ddots & & \vdots \\
0 & \cdots & a'_{n-1} & \cdots & \cdots & a'_0 - ra'_1 & -ra'_0 & 0 \\
0 & \cdots & 0 & a'_{n-1} & \cdots & a'_1 - ra'_2 & a'_0 - ra'_1 & -ra'_0 \\
\hline
b'_{1,p-1} & b'_{1,p-2} - rb'_{1,p-1} & & -rb'_{1,0} & 0 & \cdots & & 0 \\
\vdots & & & & & & & \vdots \\
0 & \cdots & 0 & b'_{1,p-1} & b'_{1,p-2} - rb'_{1,p-1} & \cdots & -rb'_{1,0} & 0 \\
0 & & \cdots & 0 & b'_{1,p-1} & \cdots & b'_{1,0} - rb'_{1,1} & -rb'_{1,0} \\
\hline
 & & & & & & & \\
 & & & \vdots & & & & \\
 & & & & & & & \\
\hline
b'_{h,p-1} & b'_{h,p-2} - rb'_{h,p-1} & & -rb'_{h,0} & 0 & \cdots & & 0 \\
\vdots & & & & & & & \vdots \\
0 & \cdots & 0 & b'_{h,p-1} & b'_{h,p-2} - rb'_{h,p-1} & \cdots & -rb'_{h,0} & 0 \\
0 & & \cdots & 0 & b'_{h,p-1} & \cdots & b'_{h,0} - rb'_{h,1} & -rb'_{h,0}
\end{array}\right] \tag{4.19}$$

$$S''_{\varphi}=\left[\begin{array}{cccccccc}
a'_{n-1} & a'_{n-2}-ra'_{n-1} & \cdots & & -ra'_0 & 0 & \cdots & 0\\
\vdots & \ddots & & & & \ddots & & \vdots\\
0 & \cdots & a'_{n-1} & \cdots & \cdots & a'_0-ra'_1 & -ra'_0 & 0\\
0 & \cdots & 0 & a'_{n-1} & \cdots & a'_1-ra'_2 & -ra'_1 & a'_0\\
\hdashline
b'_{1,p-1} & b'_{1,p-2}-rb'_{1,p-1} & & -rb'_{1,0} & 0 & \cdots & & 0\\
\vdots & & & & & & & \vdots\\
0 & \cdots & 0 & b'_{1,p-1} & b'_{1,p-2}-rb'_{1,p-1} & \cdots & -rb'_{1,0} & 0\\
0 & & \cdots & 0 & b'_{1,p-1} & \cdots & rb'_{1,1} & b'_{1,0}\\
\hdashline
 & & & & & & & \\
 & & & \vdots & & & & \\
 & & & & & & & \\
\hdashline
b'_{h,p-1} & b'_{h,p-2}-rb'_{h,p-1} & & -rb'_{h,0} & 0 & \cdots & & 0\\
\vdots & & & & & & & \vdots\\
0 & \cdots & 0 & b'_{h,p-1} & b'_{h,p-2}-rb'_{h,p-1} & \cdots & -rb'_{h,0} & 0\\
0 & & \cdots & 0 & b'_{h,p-1} & \cdots & -rb'_{h,1} & b'_{h,0}
\end{array}\right] \tag{4.20}$$

The column transformation for the previous step is represented by the matrix:

$$Q_{r,n+p}=\begin{array}{c}\xleftrightarrow{\qquad n+p \qquad}\\ \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 & 0\\
0 & 1 & 0 & \cdots & 0 & 0\\
0 & 0 & 1 & \cdots & 0 & 0\\
\vdots & & & \ddots & & \vdots\\
0 & 0 & \cdots & 0 & -\dfrac{1}{r} & 0\\
0 & 0 & \cdots & 0 & -\dfrac{1}{r^2} & -\dfrac{1}{r}
\end{bmatrix}\end{array} \tag{4.21}$$

We proceed backwards with the same column operations and thus finally we obtain the matrix in the form (3.7). The overall reduction procedure is based on steps summarised below in an algorithmic way as:

Reduction Procedure for $(s-r)$ ***factor*** **(Algorithm 4.1):**

For $i=0,\dots,n+p-2$

i) multiply the $m+p-i$ column by $-\dfrac{1}{r}$ and

ii) subtract it from the $m+p-i+1$ column of S. ■

The transformation matrix in every iteration of the above algorithm is of the form:

$$Q_{r,n+p-i}=\begin{array}{cccccccccc} & & & {\scriptstyle m+n-i+1} & {\scriptstyle m+n-i} & & & & \\ \hline 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ & & \ddots & & & & & & \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \frac{1}{r} & -\frac{1}{r} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \end{array} \qquad \text{for } i=0,\ldots,n+p-2 \qquad (4.22)$$

Finally the overall transformation takes the form:

$$\overline{S}_{\varphi'}^{(1)}=S_{\varphi}Q_{r,n+p}Q_{r,n+p-1}\cdots Q_{r,2}=S_{\varphi}Q_r^{n+p}=S_{\varphi}Q_r \qquad (4.23)$$

where Q_r has the form indicated by (4.14), as this follows by using the general formula of (4.22) and (4.23).

■

The matrix Q_r^{n+p} has a structure uniquely defined by the pair $(r,n+p)$, has dimensions $(n+p)\times(n+p)$ and provides a representation of $(s-r)$. It is clear that the above result may be applied for each common factor of multiplicity one and in each step one column of the Sylvester resultant becomes zero. In fact, the nonzero part in (3.7) is an expanded resultant and it will be shown later on that Proposition (3.2) applies also to expanded resultants. The procedure can be repeated for every factor of the gcd and a result may be derived where the corresponding reduced Sylvester matrix has a block of zero columns equal to the degree of the gcd.

The above operations lead to the same result for complex roots too. Then the transformation matrix will have complex elements. In order to avoid that, we can apply a different procedure for the pair of complex conjugate roots. In fact, we consider the quadratic $k_2s^2+k_1s+k_0$ to be a common factor of the polynomials $a(s)$, $b_j(s)$ $\quad j=1,\ldots h$ defined above and let us assume that the roots of the quadratic are complex. The $h+1$ polynomials can be written as:

$$a(s)=\left(k_2s^2+k_1s+k_0\right)\left(a''_{n-2}s^{n-2}+a''_{n-3}{}^{n-3}+\ldots+a''_1s+a''_0\right) \tag{4.24}$$

$$b_j(s)=\left(k_2s^2+k_1s+k_0\right)\left(b''_{j,p-2}s^{n-2}+b''_{j,p-3}s^{n-3}+\ldots+b''_{j,1}s+b''_{j,0}\right),\quad j=1,\ldots,h$$

The matrix, representation of this factorisation, based on the resultant, is established by the following result:

Proposition (4.5): Let $\mathcal{P}=\{a(s),b_1(s),\ldots,b_h(s)\}$ be a set of polynomials such that $\deg a(s)=n$, $\deg b(s)\leq p$, $i=1,\ldots,h$, $p\leq n$ and let $\varphi(s)=k_2s^2+k_1s+k_0$ be an irreducible factor of the GCD of $\mathcal{P}$. There always exists a real transformation C_S and a reduced resultant $\overline{S}^{(2)}_{\mathcal{P}''}$ defined such that

$$\overline{S}^{(2)}_{\mathcal{P}''}=S_{\mathcal{P}}\cdot C_S \tag{4.26}$$

where

$$C_S=\begin{bmatrix} x_0 & 0 & \cdots & & & & & \\ x_1 & x_0 & & & & & & \\ x_2 & x_1 & \ddots & & & & & \\ \vdots & \vdots & \ddots & x_0 & & & & \\ & & & x_1 & x_0 & & & \\ & & & \vdots & \vdots & \ddots & & \\ x_{n+p-2} & x_{n+p-3} & & x_{n+p-j-2} & x_{n+p-j-3} & \cdots & x_0 & 0 \\ x_{n+p-1} & x_{n+p-2} & & x_{n+p-j-1} & x_{n+p-j-2} & \cdots & x_1 & x_0 \end{bmatrix} \tag{4.27a}$$

and the x_i being defined by

$$x_0=\frac{1}{k_0},\ x_1=-\frac{k_1}{k_0^2},\ldots,\ x_\mu=-\frac{k_1}{k_0}x_{\mu-1}-\frac{k_2}{k_0}x_{\mu-2}\ ,\ \mu=2,\ldots,n+p \tag{4.27b}$$

and the reduced resultant is defined by

$$\overline{S}_{\mathcal{P}''}^{(2)}=\begin{bmatrix}
0 & 0 & a''_{n-2} & a''_{n-3} & \cdots & & & & & a''_0 & 0 & & \cdots & 0\\
0 & 0 & 0 & a''_{n-2} & \cdots & & & & & a''_1 & a''_0 & 0 & \cdots & 0\\
\vdots & \vdots & & & \ddots & & & & & & & \ddots & \ddots & \\
 & & & & & & & \cdots & & & & & & 0\\
0 & 0 & & & & & a''_{n-2} & & & & & & a''_1 & a''_0\\
0 & 0 & b''_{1,p-2} & b''_{1,p-3} & \cdots & b''_{1,0} & & & & & & & & \\
\vdots & \vdots & & & & & & & & & & & & \\
0 & 0 & b''_{h,p-2} & b''_{h,p-1} & & b''_{h,0} & & & & & & & & \\
\vdots & \vdots & & & & & & & & & & & & \\
 & & & & & & & & & & & \ddots & & \\
0 & 0 & 0 & 0 & \cdots & & & \cdots & 0 & b''_{h,p-2} & b''_{h,p-1} & & b''_{h,0} & 0\\
0 & 0 & 0 & 0 & \cdots & & & & \cdots & 0 & b''_{h,p-2} & \cdots & b''_{h,1} & b''_{h,0}
\end{bmatrix}$$

(4.28)

where $\mathcal{P}''$ is the reduced set obtained from $\mathcal{P}$ after factorisation of $\varphi(s)$.

Proof:

From the factorisation (4.24) and (4.25) the coefficients of $a(s),b(s)$ can be expressed as:

$$\left.\begin{aligned}
&a_n=k_2a''_{2,n-2}\\
&a_{n-1}=k_2a''_{2,n-3}+k_1a''_{2,n-2}\\
&a_i=k_2a''_{2,i-2}+k_1a''_{2,i-1}+k_0a''_{2,i}\ , \qquad i=2,\dots,m-2\\
&a_1=k_1a''_{2,0}+k_0a''_{2,1}\\
&a_0=k_0a''_{2,0}
\end{aligned}\right\} \tag{4.29}$$

$$\left.\begin{aligned}
&b_{j,p}=k_2b''_{j,p-2}\\
&b_{j,p-1}=k_2b''_{j,p-3}+k_1b''_{j,p-2}\\
&b_{j,i}=k_2b''_{j,i-2}+k_1b''_{j,i-1}+k_0b''_{j,i}\ , \qquad i=2,\dots,p-2\\
&b_{j,1}=k_1b''_{j,0}+k_0b''_{j,1}\\
&b_{j,0}=k_0b''_{j,0}
\end{aligned}\right\}\quad j=1,\dots,h \tag{4.30}$$

Then the Sylvester matrix $S_{\mathcal{P}}$ is expressed as

$$\tilde{S}'_{\wp} = \begin{bmatrix}
k_2a''_{n-2} & & \cdots & k_2a''_{i-2}+k_1a''_{i-1}+k_0a''_i & \cdots & & k_0a''_0 & 0 & \cdots & 0 \\
0 & k_2a''_{n-2} & \cdots & k_2a''_{i-1}+k_1a''_i+k_0a''_{i+1} & & & k_1a''_0+k_0a''_1 & k_0a''_0 & \cdots & 0 \\
\vdots & & & \ddots & & & & & \ddots & \\
0 & \cdots & & 0 & k_2a''_{n-2} & & \cdots & & k_1a''_0+k_0a''_1 & k_0a''_0 \\
k_2b''_{1,p-2} & \cdots & & k_2b''_{1,i-2}+k_1b''_{1,i-1}+k_0b''_{1,i} & \cdots & k_0b''_{1,0} & 0 & \cdots & & 0 \\
0 & k_2b''_{1,p-2} & & & & k_1b''_{1,0}+k_0b''_{1,1} & k_0b''_{1,0} & 0 & \cdots & 0 \\
& & & \ddots & & & & \ddots & & \\
& & & & k_2b''_{1,p-2} & & & \cdots & k_1b''_{1,0}+k_0b''_{1,1} & k_0b''_{1,0} \\
k_2b''_{2,p-2} & & \cdots & & \cdots & k_0b''_{1,0} & & & \cdots & 0 \\
\vdots & & \cdots & & & & & & & \vdots \\
& & & & & & & & & 0 \\
k_2b''_{h,p-2} & \cdots & & k_2b''_{h,i-2}+k_2b''_{h,i-1}+k_2b''_{h,i} & \cdots & k_0b''_{h,0} & 0 & \cdots & & 0 \\
0 & k_2b''_{h,p-2} & & & & k_1b''_{h,0}+k_0b''_{h,1} & k_0b''_{h,0} & 0 & \cdots & 0 \\
& & & \ddots & & & & \ddots & & \\
& & & & k_2b''_{h,p-2} & & & \cdots & k_1b''_{h,0}+k_0b''_{h,1} & k_0b''_{h,0}
\end{bmatrix} \tag{4.31}$$

By applying appropriate column operation, described algorithmically below, the matrix $\tilde{S}'_{\wp}$ can be reduced to the form described by (4.28). The reduction algorithm that leads to this result is described below:

Reduction procedure for $k_2s^2 + k_1s + k_0$ factor (Algorithm 4.2):

i) multiply the $n+p$ column by $\dfrac{1}{k_0}$

ii) from the $n+p-1$ column we subtract the $n+p$ column multiplied by k_1

iii) multiply the $n+p-1$ column by $\dfrac{1}{k_0}$

iv) For $i = 2,\dots,n+p-1$ we

- subtract from the $n+p-i$ column the $n+p-i-2$ column multiplied by k_2
- subtract from the $n+p-i$ column the $n+p-i-1$ column multiplied by k_1
- multiply the $n+p-i$ column by $\dfrac{1}{k_0}$

■

It can be readily shown that the product of the above elementary column operation has the form of C_S described by (4.27).

■

Propositions (4.3), (4.4) and (4.5) express the factorisation of elementary factors of the s^c, $(s-r)$ and $k_2 s^2 + k_1 s + k_0$ quadratic terms when we start from the original resultant. The successive extraction of terms implies that the results have to be extended first to the extraction of factors from the expanded resultant and not just the resultant. This will then allow consideration of extraction of factors in a specific order and thus lead to the generalisation of the previous results to the case of a general factor.

Consider the c-order set $\mathcal{P}_{h+1,n}$ with resultant $S_{\mathcal{P}}$ for which Proposition (4.3) implies that

$$S_{\mathcal{P}} = \left[\mathbf{0}_c \mid \bar{S}_{\tilde{\mathcal{P}}}\right]\begin{bmatrix} \mathbf{0} & I_c \\ I_{n+p-c} & \mathbf{0} \end{bmatrix} = \bar{S}^c_{\tilde{\mathcal{P}}}\hat{Q}_c \qquad (4.32a)$$

or

$$\bar{S}^c_{\tilde{\mathcal{P}}} = \left[\mathbf{0}_c \mid \bar{S}_{\tilde{\mathcal{P}}}\right] = S_{\mathcal{P}} Q_c(0),\ Q_c(0) = \hat{Q}_c^{-1} \qquad (4.32b)$$

where $\tilde{\mathcal{P}}_{h+1,n-c}$ is the reduced set after factorisation of s^c. If $S_{\tilde{\mathcal{P}}}$ is the resultant of $\tilde{\mathcal{P}}_{h+1,n-c}$ then:

- $S_{\tilde{\mathcal{P}}}$ has $h(n-c)+(p-c) = hn+p-(h+1)c$ rows and $(n-c)+(p-c) = n+p-2c$ columns.
- The (n,p)-expanded resultant $\bar{S}_{\tilde{\mathcal{P}}}$ has $n+hp$ rows and $n+p-c$ columns.

The extraction of an $(s-r)$ factor from the gcd of $\mathcal{P}_{h+1,n}$ may be expressed as a factorisation on the different reduced sets as it is shown below:

Proposition (4.6): Let $\mathcal{P}_{h+1,n}$ be a c-order set, $(s-r)$ be a divisor of its gcd and let $\tilde{\mathcal{P}}_{h+1,n-c}$, $\hat{\mathcal{P}}_{h+1,n-c-1}$ be the reduced sets obtained after factorisation of the s^c and

$s^c(s-r)$ factors. If $S_{\mathcal{P}}$, $S_{\tilde{\mathcal{P}}}$, $S_{\hat{\mathcal{P}}}$ are the resultants of the sets $\mathcal{P}$, $\tilde{\mathcal{P}}$, $\hat{\mathcal{P}}$ respectively and Q_r^k denote the k-order transformation associated with $(s-r)$, then the relationship

$$\overline{S}_{\tilde{\mathcal{P}}}^{(1)} = S_{\tilde{\mathcal{P}}} Q_r^{n+p-c} = \left[\mathbf{0} \;\vdots\; \overline{S}_{\hat{\mathcal{P}}} \right] \tag{4.33}$$

where $\overline{S}_{\hat{\mathcal{P}}}$ is an expanded resultant of $\hat{\mathcal{P}}_{h+1,n-c-1}$ implies the following relationship on $\overline{S}_{\tilde{\mathcal{P}}}^{c}$ defined by (4.32b)

$$\overline{S}_{\tilde{\mathcal{P}}}^{c+1} = \left[\mathbf{0}_{c+1} \;\vdots\; \overline{S}'_{\hat{\mathcal{P}}} \right] = \overline{S}_{\hat{\mathcal{P}}}^{c} Q_r^{n+p} \tag{4.34}$$

where $\overline{S}'_{\hat{\mathcal{P}}}$ is also an expanded resultant of $\hat{\mathcal{P}}_{h+1,n-c-1}$.

■

The proof of the above result is a straightforward extension of the proof of proposition (4.4) and it simply states that extending the factorisation of $(s-r)$ from the resultant of the set to the expanded resultant we have to change only the corresponding order of the transformation Q_r^k to the original order $n+p$. The above leads to the following procedure for extracting divisors in the resultant set up.

Remark (4.3): Let $\mathcal{P}_{h+1,n}$ be a c-order set and $(s-r_i)$, $i=1,2,\ldots,\tau$ be a set of divisors of the gcd of $\mathcal{P}_{h+1,n}$. If Q_c, $Q_{r_i}^{n+p}$, $i=1,2,\ldots,\tau$ are the transformations representing the divisors, then the extraction of divisors is represented as sequence of transformations performed on the resultant $S_{\mathcal{P}}$ as:

$$S_{\mathcal{P}} Q_c Q_{r_1}^{n+p} \cdots Q_{r_\tau}^{n+p} = \left[\mathbf{0}_k \;\vdots\; \overline{S}_{\hat{\mathcal{P}}} \right] = \overline{S}_{\hat{\mathcal{P}}}^{\varphi} \tag{4.35}$$

where $\overline{S}_{\hat{\mathcal{P}}}^{\varphi}$ is an expanded resultant of the reduced set $\hat{\mathcal{P}}$ obtained after extraction of $s^c(s-r_1)\cdots(s-r_\tau)=\varphi(s)$ from $\mathcal{P}$.

■

The matrices involved in the extraction of divisors are of the Toeplitz type and their properties are considered next. In the following we shall denote by $\{T_n\}$ the set of non-singular Toeplitz matrices of $n \times n$ dimension of the type

$$A=\begin{bmatrix} a_0 & 0 & 0 & \cdots & & 0 \\ a_1 & a_0 & 0 & & & \\ a_2 & a_1 & a_o & & & \\ \vdots & & \ddots & \ddots & & \\ & & & & & \\ a_{n-2} & a_{n-3} & \cdots & & a_o & 0 \\ a_{n-1} & a_{n-2} & \cdots & & a_1 & a_o \end{bmatrix} = A\left(a_0, a_1, \ldots, a_{n-1}\right) \tag{4.36}$$

Clearly $I_n \in \{T_n\}$. The following properties for the set $\{T_n\}$ are readily established.

Lemma (4.1): The product of any two elements of $\{T_n\}$ is also an element of $\{T_n\}$. If we consider $A, B \in \{\mathrm{T}_n\}$ where $A = A\left(a_0, a_1, \ldots, a_{n-1}\right)$, $B = B\left(b_0, b_1, \ldots, b_{n-1}\right)$ then:

i) $C = A + B \in \{T_n\}$.

ii) $D = [d_{i,j}] = AB = BA \in \{T_n\}$ where

$$\left.\begin{aligned} d_{i,j} &= \sum_{k=0}^{i-j} a_{i-j-k} b_{i-j}, \qquad \text{for } i \geq j \\ d_{i,j} &= 0, \qquad\qquad \textit{otherwise} \end{aligned}\right\} \quad \text{and} \quad d_{i,j} = d_{i+1,j+1}$$

■

Lemma (4.2): For all $A \in T_n$ with $a_0 \neq 0$, $A^{-1} \in T_n$. Then $\widehat{A} = A^{-1}$ is expressed as:

$$\widehat{A}=\begin{bmatrix} l_0 & 0 & 0 & \cdots & & 0 \\ l_1 & l_1 & 0 & & & \\ l_2 & l_1 & l_o & & & \\ \vdots & & \ddots & \ddots & & \\ & & & & & \\ l_{n-2} & l_{n-3} & \cdots & & l_o & 0 \\ l_{n-1} & l_{n-2} & \cdots & & l_1 & l_o \end{bmatrix} \tag{4.37a}$$

where the elements of $A, \widehat{A}$ are related as:

$$l_o = \frac{1}{a_0}\ , \quad l_i = -\frac{1}{a_0} \sum_{j=0}^{i-1} l_j a_{i-j}, \quad i = 1, \ldots, n-1 \tag{4.37b}$$

■

From lemmas (4.1) and (4.2) we have the following result:

Lemma (4.3): The set $\{\mathrm{T}_n\}$ of Toeplitz matrices is a commutative ring under the standard operations of addition and multiplication with units the elements with $a_0 \neq 0$.

■

Some further interesting properties of Toeplitz matrices which are linked to the representation of polynomials are considered next.

Proposition (4.7): Let $\lambda(s) = \lambda_k s^k + \ldots + \lambda_1 s + \lambda_0$ be a polynomial and let $\hat{\Phi} \in \{T_n\}$, $k < n$, be a special Toeplitz matrix representation of $\lambda(s)$ defined by

$$\hat{\Phi}_\lambda^n = \hat{\Phi} = \begin{bmatrix} \lambda_0 & 0 & \cdots & & & & & 0 \\ \lambda_1 & \lambda_0 & 0 & \cdots & & & & 0 \\ \lambda_2 & \lambda_1 & \lambda_0 & & & & & \\ \vdots & \vdots & & \ddots & & & & \\ \lambda_k & & & & \lambda_0 & 0 & & 0 \\ 0 & \ddots & & & & \ddots & \ddots & \vdots \\ \vdots & & & & & \lambda_1 & \lambda_0 & 0 \\ 0 & \cdots & 0 & \lambda_k & \cdots & \lambda_2 & \lambda_1 & \lambda_0 \end{bmatrix} \tag{4.38}$$

Then the inverse $\Phi = \hat{\Phi}^{-1}$ has the Toeplitz form:

$$\Phi = \begin{bmatrix} y_0 & 0 & \cdots & & & & & 0 \\ y_1 & y_0 & \ddots & & & & & \vdots \\ y_2 & y_1 & \ddots & & & & & \\ \vdots & \vdots & \ddots & y_0 & 0 & & & \\ & & & y_1 & y_0 & & & \\ & & & \vdots & \vdots & \ddots & & \\ y_{n-2} & y_{n-3} & \cdots & y_{n-j-2} & y_{n-j-3} & \cdots & y_0 & 0 \\ y_{n-1} & y_{n-2} & \cdots & y_{n-j-1} & y_{n-j-2} & \cdots & y_1 & y_0 \end{bmatrix} \tag{4.39a}$$

where the y_i parameters satisfy the relationships

$$y_0 = \frac{1}{\lambda_0} \,, \quad y_1 = -\frac{\lambda_1}{\lambda_0} y_0 \,, \ldots, \quad y_j = -\frac{1}{\lambda_0} \sum_{i=1}^{\min\{j,k\}} \lambda_i y_{j-i} \,, \; j = 2, \ldots, n-1 \qquad (4.39b)$$

■

Remark (4.4): As a straightforward consequence of the above result we have the matrix Q_r^{n+p} defined in (4.14) is the inverse of the simple Toeplitz matrix $\hat{Q}_r^{n+p}$ defined as

$$\hat{Q}_r^{n+p} = \left\{Q_r^{n+p}\right\}^{-1} = \begin{bmatrix} -r & 0 & 0 & \cdots & 0 & 0 \\ 1 & -r & 0 & \cdots & 0 & 0 \\ 0 & 1 & -r & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -r & 0 \\ 0 & 0 & 0 & \cdots & 1 & -r \end{bmatrix} \in \mathbb{C}^{(n+p)\times(n+p)} \qquad (4.40)$$

■

The above indicates the link of the Q_r^{n+p} matrix with the representation of $(s-r)$ factor in terms of the $(n+p)\times(n+p)$ matrix $\hat{Q}_r^{n+p}$. The significance of the Toeplitz representation of polynomials $\varphi(s)$ as in (3.38) is emphasised by the following result expressing the factorisation of a given polynomial.

Proposition (4.8): Let $f(s) = f_k s^k + \ldots + f_1 s + f_0 \in \mathbb{R}[s]$ and assume a factorisation $f(s) = t(s)\cdot q(s)$ where $t(s) = t_l s^l + \ldots + t_1 s + t_0$, $q(s) = q_m s^m + \ldots + q_1 s + q_0$. Then for any $n \geq k$ the n-th order Toeplitz representations $\hat{\Phi}_f^n$, $\hat{\Phi}_t^n$, $\hat{\Phi}_q^n$ of $f(s)$, $t(s)$, $q(s)$ defined as in (4.38) satisfy the relationship:

$$\hat{\Phi}_f^n = \hat{\Phi}_t^n \, \hat{\Phi}_q^n = \hat{\Phi}_q^n \, \hat{\Phi}_t^n \qquad (4.41)$$

Proof:

Note that the multiplication of the $t(s)$, $q(s)$ polynomials may be expressed as a Toeplitz matrix operation as shown below

$$\begin{bmatrix} t_0 & 0 & \cdots & & 0 \\ t_1 & t_0 & \ddots & & \vdots \\ t_2 & t_1 & \ddots & & 0 \\ \vdots & & \vdots & \ddots & t_0 \\ \vdots & \vdots & & & t_1 \\ t_{l-1} & t_{l-2} & & & \vdots \\ t_l & t_{l-1} & & & \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & & t_l & t_{l-1} \\ 0 & \cdots & & 0 & t_l \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ \vdots \\ q_{m-1} \\ q_m \end{bmatrix} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ \vdots \\ \\ \\ \vdots \\ f_{l+m-2} \\ f_{l+m-1} \\ f_{l+m} \end{bmatrix} \tag{4.42a}$$

which can also extended to the equivalent condition

$$\left[\begin{array}{ccccc|ccccc} t_0 & 0 & \cdots & & 0 & 0 & 0 & \cdots & & 0 \\ t_1 & t_0 & \ddots & & \vdots & 0 & 0 & \cdots & & 0 \\ t_2 & t_1 & \ddots & & 0 & \vdots & & & & \vdots \\ \vdots & & \vdots & \ddots & t_0 & & & & & \\ \vdots & \vdots & & & t_1 & t_0 & 0 & & & \\ t_{l-1} & t_{l-2} & & & \vdots & & t_0 & 0 & & \\ t_l & t_{l-1} & & & & & & & \ddots & \\ 0 & \ddots & & & \vdots & & & & & \\ \vdots & \ddots & & t_l & t_{l-1} & t_{l-2} & & & t_0 & 0 \\ 0 & \cdots & & 0 & t_l & t_{l-1} & t_{l-1} & \cdots & t_1 & t_0 \end{array}\right] \left[\begin{array}{c} q_0 \\ q_1 \\ \vdots \\ \vdots \\ q_{m-1} \\ q_m \\ \hdashline 0 \\ \vdots \\ \vdots \\ 0 \end{array}\right] = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ \vdots \\ \\ \\ \vdots \\ f_{l+m-2} \\ f_{l+m-1} \\ f_{l+m} \end{bmatrix} \tag{4.42b}$$

It can be readily shown by inspection of (4.41) and use of block multiplication that (4.41) is reduced to a set of conditions equivalent to (4.42b) and this establishes the result. Commutativity follows from the corresponding property on polynomials and by writing (4.42a) in a dual way (Toeplitz of $q(s)$).

■

The above result expresses the multiplication of polynomials within their Toeplitz representation framework. It is clear that the above group properties may also be expressed for the inverse transformations $\Phi_f^n = \hat{\Phi}_f^{n-1}$ which are linked to the factorization of resultants. The results so far lead to the following main result:

Theorem (4.4): Let $\mathcal{P}=\{a(s)b_1(s),\dots,b_h(s)\}$ be a 0-order set, $\deg a(s)=n$, $\deg b_i(s)\le p\le n$, $i=1,\dots,h$ be a polynomial set, $S_{\mathcal{P}}$ the respective Sylvester matrix, $\varphi(s)=\lambda_k s^k+\cdots+\lambda_1 s+\lambda_0$ be the greatest common divisor of the set and let k be its degree. Then there exists transformation matrix $\Phi_\varphi\in\mathbf{R}^{(n+p)\times(n+p)}$ such that:

$$\bar{S}^{(k)}_{\mathcal{P}^*}=S_{\mathcal{P}}\Phi_\varphi=\left[\mathbf{0}_k \mid \bar{S}_{\mathcal{P}^*}\right], \tag{4.43}$$

where $\bar{S}^{(k)}_{\mathcal{P}^*}$, Φ_φ are given by:

$$\Phi_\varphi=\begin{bmatrix} y_0 & 0 & \cdots & & & \cdots & & 0 \\ y_1 & y_0 & & & & & & \vdots \\ y_2 & y_1 & \ddots & \ddots & & & & \\ \vdots & \vdots & \ddots & y_0 & 0 & & & \\ & & & y_1 & y_0 & \ddots & & \\ & & & \vdots & \vdots & \ddots & & \\ y_{n+p-2} & y_{n+p-3} & \cdots & y_{n+p-j-2} & y_{n+p-j-3} & \cdots & y_0 & 0 \\ y_{n+p-1} & y_{n+p-2} & \cdots & y_{n+p-j-1} & y_{n+p-j-2} & \cdots & y_1 & y_0 \end{bmatrix} \tag{4.44a}$$

where

$$y_0=\frac{1}{\lambda_0}\ ,\ \ y_1=-\frac{\lambda_1}{\lambda_0}y_0\ ,\ \dots\ ,y_j=-\frac{1}{\lambda_0}\sum_{i=1}^{\min\{j,k\}}\lambda_i y_{j-i}\ ,j=2,\dots,n+p-1 \tag{4.44b}$$

$$
\bar{S}_{\mathcal{P}^*}^{(k)} = \left[\begin{array}{ccc|cccccccc}
0 & \cdots & 0 & a_{n-k}^{(k)} & a_{n-k-1}^{(k)} & \cdots & a_1^{(k)} & a_0^{(k)} & 0 & \cdots & 0 \\
0 & \cdots & 0 & 0 & a_{n-k}^{(k)} & & \cdots & a_1^{(k)} & a_0^{(k)} & \ddots & \vdots \\
\vdots & & \vdots & \vdots & \ddots & \ddots & & & \ddots & \ddots & 0 \\
0 & \cdots & 0 & 0 & \cdots & 0 & & a_{n-k}^{(k)} & \cdots & a_1^{(k)} & a_0^{(k)} \\
0 & \cdots & 0 & b_{1,p-k}^{(k)} & b_{1,p-k-1}^{(k)} & \cdots & b_{1,0}^{(k)} & 0 & \cdots & & 0 \\
0 & \cdots & 0 & 0 & b_{1,p-k}^{(k)} & b_{1,p-k-1}^{(k)} & \cdots & b_{1,0}^{(k)} & 0 & \cdots & 0 \\
\vdots & & \vdots & & & & & & & \vdots & \\
 & & & & & \ddots & \ddots & & & \ddots & \\
0 & \cdots & 0 & 0 & \cdots & & 0 & b_{1,p-k}^{(k)} & \cdots & b_{1,1}^{(k)} & b_{1,0}^{(k)} \\
0 & \cdots & 0 & b_{2,p-k}^{(k)} & b_{2,p-k-1}^{(k)} & \cdots & b_{2,0}^{(k)} & 0 & \cdots & & 0 \\
0 & \cdots & 0 & 0 & b_{2,p-k}^{(k)} & b_{2,p-k-1}^{(k)} & \cdots & b_{2,0}^{(k)} & 0 & \cdots & 0 \\
 & & & & & & & & & & \\
 & & & & & & & & & & \\
0 & \cdots & 0 & b_{h,p-k}^{(k)} & b_{h,p-k-1}^{(k)} & \cdots & b_{h,0}^{(k)} & 0 & \cdots & & 0 \\
0 & \cdots & 0 & 0 & b_{h,p-k}^{(k)} & b_{h,p-k-1}^{(k)} & & b_{h,0}^{(k)} & 0 & \cdots & 0 \\
\vdots & & \vdots & \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & \\
 & & & & & & & & & b_{h,0}^{(k)} & 0 \\
0 & \cdots & 0 & 0 & \cdots & 0 & b_{h,p-k}^{(k)} & b_{h,p-k-1}^{(k)} & \cdots & b_{h,1}^{(k)} & b_{h,0}^{(k)}
\end{array}\right]
\tag{4.45}
$$

where $\left[a_{p-k}^{(k)}, a_{p-k-1}^{(k)}, \ldots, a_0^{(k)}\right]$, $\left[b_{j,p-k}^{(k)}, b_{j,p-k-1}^{(k)}, \ldots, b_{j,0}^{(k)}\right]$ $j = 1, \ldots, h$ are the coefficients of the coprime polynomials obtained from the original set after the division by the gcd, which define the set $\mathcal{P}_{h+1,n-k}^{*}$ and $\bar{S}_{\mathcal{P}^*}$ is the corresponding expanded resultant.

■

Theorem (4.4) in a sense provides a representation in matrix terms of the standard factorisation of the GCD of a set of polynomials and this may be expressed in the following form:

Corollary (4.1): Let $\mathcal{P} = \{a(s), b_1(s), \ldots, b_h(s)\}$ be a 0-order set of polynomials, $\deg a(s) = n$, $\deg b_i(s) \leq p \leq n$, $i = 1, \ldots, h$ and let $\varphi(s)$ be the GCD of $\mathcal{P}$, $\deg \varphi(s) = k$. If

$$a(s) = a'(s)\varphi(s), \; b_i(s) = b_i'(s)\varphi(s), \qquad i = 1, \ldots h \tag{4.46}$$

and $\mathcal{P}^* = \{a'(s), b_1'(s), \ldots, b_h'(s)\}$, $\deg a'(s) = n-k$ $\deg b_i'(s) \le p-k$, $i = 1, \ldots h$ and $S_{\mathcal{P}}$, $\overline{S}_{\mathcal{P}^*}^{(k)}$ are the generalised resultants of $\mathcal{P}$, $\mathcal{P}^*$, where $\overline{S}_{\mathcal{P}^*}^{(k)}$ is structured by the indices of $\mathcal{P}$ (it is assumed that the structuring degrees are (n,p)). Then (4.46) is equivalent to

$$S_{\mathcal{P}} = \overline{S}_{\mathcal{P}^*}^{(k)} \; \hat{\Phi}_{\varphi} = \left[\mathbf{0}_k \;\middle|\; \overline{S}_{\mathcal{P}^*} \right] \hat{\Phi}_{\varphi} \tag{4.47}$$

where $\overline{S}_{\mathcal{P}^*}$ is the (n,p)-expanded resultant of of $\mathcal{P}^*$ and $\hat{\Phi}_{\varphi} = \Phi_{\varphi}^{-1}$ has the form of (4.38) and it is defined by the gcd $\varphi(s)$.

■

For the case of sets which are non-proper the above result may be expressed in the following way.

<u>**Corollary (4.2):**</u> Let $\mathcal{P} = \{a(s) b_1(s), \ldots, b_h(s)\}$ be a c-order set of polynomials, $\deg a(s) = n$, $\deg b_i(s) \le p \le n$, $i = 1, \ldots, h$ and let $\varphi(s) = s^c \omega(s)$, $\omega(0) \ne 0$ be the gcd of $\mathcal{P}$, $\deg \varphi(s) = c + k$. If $\tilde{\mathcal{P}}$, $\mathcal{P}^*$ are the sets obtained by dividing by s^c and $\varphi(s)$ respectively and $S_{\mathcal{P}}$, $\overline{S}_{\mathcal{P}^*}^{(k+c)}$ be the resultants of $\mathcal{P}$ and $\mathcal{P}^*$, where $\overline{S}_{\mathcal{P}^*}^{(k+c)}$ is expressed with respect to (n,p) then

$$\begin{aligned} S_{\mathcal{P}} &= \left[\mathbf{0}_{k+c} \;\middle|\; \overline{S}_{\mathcal{P}^*} \right] \hat{\Phi}_{\omega} \begin{bmatrix} \mathbf{0} & I_c \\ I_{n+p-c} & \mathbf{0} \end{bmatrix} \\ &= \overline{S}_{\mathcal{P}^*}^{(k+c)} \; \hat{\Phi}_{\omega} \; \hat{Q}_c \end{aligned} \tag{4.48}$$

where $\hat{\Phi}_{\omega}$ is the representation of the $\omega(s)$ divisor of the gcd to the $(n+p)$-order.

■

The matrix $\overline{S}_{\mathcal{P}^*}^{(k)}$ for the proper case, or $\overline{S}_{\mathcal{P}^*}^{(k+c)}$ for the non-proper are resultant forms which are referred to the original orders (n,p); we may refer to them as (n,p)- *reduced generalized Sylvester Resultants* and the nonzero part $\overline{S}_{\mathcal{P}^*}$ will be called (n,p)*-expanded resultant* of the $\mathcal{P}^*$ reduced set. The results of this section are now used for developing the classical Generalised Sylvester results within the new framework.

4.4. GENERALISED RESULTANT: RANK PROPERTIES AND THE GCD

The classical result on the link between resultant and gcd of the set has been generalised to the case of many polynomials [Barnett, 1990], [Vardoulakis et al., 1978]. The framework developed in the previous section that allows the linking of the gcd to a canonical factorisation of the Generalised Resultant provides the means for giving a simpler proof to the classical Sylvester result [Barnett, 1990]. The canonical factorisation together with the Generalised Sylvester Result, that links the gcd of the set to the rank properties of the resultant lead to a canonical representation of the gcd which has important implications for the development of robust computations for the gcd. The development of the main result requires the following Lemma [Barnett, 1990].

Lemma (4.4): Let $\varphi(s)$ be the g.c.d. of $\mathcal{P}=\{a(s),\ b(s)\}$ where $\deg a(s)=n$, $\deg b(s)=p$ and $\deg\varphi(s)=k$. Then if $S_{\mathcal{P}}$, $\overline{S}_{\mathcal{P}^*}^{(k)}$ are the resultants of $\mathcal{P}$ and the reduced resultant after extraction of the gcd

$$\operatorname{rank}(S_{\mathcal{P}})=\operatorname{rank}\left(\overline{S}_{\mathcal{P}^*}^{(k)}\right)=n+p-k \tag{4.49}$$

Proof:

Let us denote by $a_k(s)=a(s)/\varphi(s)$, $b_k(s)=b(s)/\varphi(s)$ the polynomials that are derived after division by the gcd, $\mathcal{P}^*=\{a_k(s),b_k(s)\}$ and also denote by $\left[a_{k,n-k},a_{k,n-k-1},\ldots,a_{k,0}\right]$, $\left[b_{k,m-k},b_{k,m-k-1},\ldots,b_{k,0}\right]$ the coefficients in descending order, of $a_k(s)$, $b_k(s)$ respectively. The reduced resultant $\overline{S}_{\mathcal{P}^*}^{(k)}$ has the form as shown in (4.50). We further note that the polynomial s^k cannot be a common factor of $a_k(s)$ and $b_k(s)$ (as they are coprime). Without loss of the generality we can assume that s^k is not a factor of $a_k(s)$. Thus the polynomials of the set $\mathcal{P}_1=\left(a_k(s),s^k b_k(s)\right)$ are coprime. This implies that their resultant is non-singular (classical result) and its corresponding form is given as in (4.51).

$$\bar{S}^{(k)}_{\varphi^*}=\left[\begin{array}{ccc|ccccccccc}
0 & \cdots & 0 & a_{k,n-k} & a_{k,n-k-1} & \cdots & & a_{k,0} & 0 & 0 & \cdots & 0\\
0 & \cdots & 0 & 0 & a_{k,n-k} & \cdots & & a_{k,1} & a_{k,0} & 0 & \cdots & 0\\
\vdots & & \vdots & & & & & \ddots & & & & \\
0 & \cdots & 0 & 0 & & & & 0 & a_{k,n-k} & \cdots & a_{k,1} & a_{k,0}\\
0 & \cdots & 0 & b_{k,p-k} & b_{k,p-k-1} & \cdots & b_{k,0} & 0 & & & & \\
0 & \cdots & 0 & & b_{k,p-k} & \cdots & b_{k,1} & b_{k,0} & 0 & & \cdots & 0\\
\vdots & & \vdots & & & & \ddots & & & & & \\
0 & \cdots & 0 & & & & & b_{k,p-k} & b_{k,p-k-1} & \cdots & b_{k,1} & b_{k,0}
\end{array}\right]$$

$$\longleftarrow\quad n+p-k\quad\longrightarrow$$

(4.50)

$$\begin{array}{cc}
\longleftarrow\quad n+p-k\quad\longrightarrow & \\
S_{\varphi_1}=\left[\begin{array}{cccccc|ccccc}
a_{k,n-k} & \cdots & a_{k,0} & 0 & \cdots & 0 & 0 & \cdots & & & 0\\
0 & \ddots & & \ddots & \ddots & \vdots & \vdots & & & & \vdots\\
\vdots & \ddots & & & & 0 & & & & & \\
0 & \cdots & 0 & a_{k,n-k} & \cdots & a_{k,0} & 0 & & & & \\
\hline
0 & \cdots & & 0 & a_{k,n-k} & & a_{k,0} & 0 & & & 0\\
\vdots & & & & & & & \ddots & \ddots & & \vdots\\
 & & & & & & & \ddots & & a_{k,0} & 0\\
0 & \cdots & & & & 0 & 0 & \cdots & a_{k,n-k} & \cdots & a_{k,0}\\
\hline
b_{k,p-k} & \cdots & b_{k,0} & 0 & \cdots & 0 & 0 & \cdots & & \cdots & 0\\
0 & \ddots & & \ddots & \ddots & \vdots & \vdots & & & & \\
\vdots & \ddots & & & & 0 & 0 & & & & \vdots\\
0 & \cdots & 0 & b_{k,p-k} & \cdots & b_{k,0} & 0 & \cdots & & \cdots & 0
\end{array}\right] & \begin{array}{c} \uparrow \\ p \\ \downarrow \\ \uparrow \\ n-k \\ \downarrow \end{array}\\
\longleftarrow\quad n+p-2k\quad\longrightarrow &
\end{array}$$

(4.51)

This resultant is clearly a submatrix of the reduced resultant $\bar{S}_{\mathcal{P}^*}^{(k)}$. So it is implied that

$$\rho\left(S_{\mathcal{P}}\right)=\rho\left(\bar{S}_{\mathcal{P}^*}^{(k)}\right)\geq S_{\mathcal{P}_1}=(n-k)+p=n+p-k \tag{4.52}$$

However, by the reduction we have proved that $\rho(S)=\rho\left(S_{\mathcal{P}^*}^{(k)}\right)\leq m+n-k$ which together with (4.52) proves the result. ■

An important question linked to the resultants is their rank properties which will be considered next. We consider first a general result linking the rank of resultants to expanded resultants. Let $\mathcal{P}_{h+1,n}$ be a set of polynomials with the two maximal degrees (n,p). We can always assume that the two maximal degrees are $(n'=n+c, p'=p+c)$, $c\geq 0$ by assuming the first c coefficients of the polynomials to be zero. This representation is referred to as c*-extended* and it is denoted by $\mathcal{P}_{h+1,n}^{c}$. If $S_{\mathcal{P}}$ is the generalised resultant of $\mathcal{P}_{h+1,n}$ and $S_{\mathcal{P}^c}$ is the generalised resultant of the c-expanded set then their dimensions are $(p+hn)\times(n+p)$, $\left[p+hn+c(h+1)\right]\times(n+p+2c)$ respectively and $S_{\mathcal{P}^c}$ will be called the c*-extended resultant*. Furthermore, we may express $S_{\mathcal{P}^c}$ as

$$S_{\mathcal{P}^c}=\left[\mathbf{0}_c \mid \bar{S}_{\mathcal{P}}^{c}\right] \tag{4.53}$$

The matrix $\bar{S}_{\mathcal{P}}^{c}$ has dimensions $\left[p+hn+c(h+1)\right]\times(n+p+c)$ and it is the matrix that has been previously defined as expanded or properly(n',p') or c*-expanded resultant*. An important relationship between $\bar{S}_{\mathcal{P}}^{c}$ and $S_{\mathcal{P}}$ is defined below:

<u>Lemma (4.5):</u> Let $\mathcal{P}_{h+1,n}$ be a set and $\mathcal{P}_{h+1,n}^{c}$ its c-extension. If $S_{\mathcal{P}}$, $\bar{S}_{\mathcal{P}}^{c}$ are the resultant and the c-expanded resultant, then

$$\rho\left(\bar{S}_{\mathcal{P}}^{c}\right)=\rho\left(S_{\mathcal{P}}\right)+c \tag{4.54}$$

■

The proof of the above result follows from the way we have constructed $\bar{S}_{\mathcal{P}}^{c}$ and it is thus readily established. Using the above two Lemmas and the link of gcd to the factorisation of the resultant we are led to the Generalised Resultant Theorem [Barnett, 1990], [Vardoulakis et al., 1978]:

Theorem (4.5): (Generalised Resultant Theorem): Given a set of polynomials $\mathcal{P} = \{ a(s) = s^n + a_{n-1}s^{n-1} + \dots a_1 s + a_0,\ b(s) = b_{i,n}s^n + \dots + b_{i,1}s + b_{i,0},\quad i = 1, 2, \dots, h$ $\max\{\deg b_i(s)\} = p \}$ with a generalised resultant $S_{\mathcal{P}}$ the following properties hold true:

i) Necessary and sufficient condition for a set of polynomials to be coprime is that:

$$\rho(S_{\mathcal{P}}) = n + p \tag{4.55}$$

ii) Let $\varphi(s)$ be the g.c.d. of $\mathcal{P}$. Then:

$$\rho(S_{\mathcal{P}}) = n + p - \deg\varphi(s) \tag{4.56}$$

iii) If we reduce $S_{\mathcal{P}}$, by using elementary row operations, to its row echelon form, the last non vanishing row defines the coefficients of the g.c.d..

Proof:

We consider first the case of proper sets and then the non-proper case.

ii) We start by proving part (ii). Thus, we consider the factorisation of (4.47) and denote by $S_{\mathcal{P}}$, $\overline{S}_{\mathcal{P}^*}^{(k)}$ the resultant and the reduced resultant (after division by $\varphi(s)$). We shall prove first that

$$\rho(S_{\mathcal{P}}) = \rho\left(\overline{S}_{\mathcal{P}^*}^{(k)}\right) \geq n + p - \deg\varphi(s) \tag{4.57}$$

We shall use induction and S_m will denote the resultant that refers to a set with m elements. For the case of two polynomials ($m = 2$) Lemma (4.4) establishes already the equality in (4.57) i.e. $\rho(S_2) = n + p - \deg\varphi(s)$.

Using induction we suppose that for the case of $h = m - 1$ the relation of the rank and the degree of the g.c.d. of the set is also given by $\operatorname{rank}(S(\mathcal{P}_{m-1})) = \rho(S_{m-1}) =$ $= n + p - \deg\varphi(s)$. For the investigation of the $h = m$ case we shall denote by $\mathcal{P}_m = \mathcal{P}$, $\mathcal{P}_{m-1}$ the set containing the first $m-1$ polynomials, $S(\mathcal{P}_m), S(\mathcal{P}_{m-1})$ are explicit descriptions of the corresponding resultants and thus we may write:

$\mathcal{P}_m = \mathcal{P}_{m-1} \cup \{b_m\}$, $\deg b_m(s) \leq p$ and we assume that $\operatorname{rank}(S(\mathcal{P}_{m-1})) = n + p - k$, $0 < k \leq p$. We denote by $\varphi_{m-1}(s), \varphi_m(s)$ the corresponding gcds of $\mathcal{P}_{m-1}, \mathcal{P}_m$ respectively where $\deg(\varphi_{m-1}(s)) = k$. Thus, it is implied that

$$\operatorname{rank}(S(\mathcal{P}_{m-1})) \leq \operatorname{rank}(S(\mathcal{P}_m)), \quad \operatorname{rank}(S(\mathcal{P}_m)) = n + p - k', \qquad k' \leq k \tag{4.58}$$

where k' is the right nullity of $S(\mathcal{P}_m)$. Clearly $k' \leq k$ since addition of rows cannot increase the nullity, but only decrease it. Note that the Sylvester matrix of $\mathcal{P}_{m-1}$ is equivalent to its row echelon form which is:

$$E(\mathcal{P}_{m-1}) = \left[\begin{array}{c} \tilde{P}_{m-1} \\ \hdashline \mathbf{0} \end{array}\right] \tag{4.59}$$

where $\tilde{P}_{m-1} \in \mathbb{R}^{(n+p-k)\times(n+p)}$ and all rows of $\tilde{P}_{m-1}$ are produced by elementary row operations, in other words are linear combinations of the rows of the Sylvester matrix $S(\mathcal{P}_{m-1})$. If we denote $\mu = k - k'$ then a base for $\mathcal{P}_m$ consists of the independent (non-zero) rows of $\tilde{P}_{m-1}$ and μ rows of the last block of $S(\mathcal{P}_m)$ associated with $b_m(s)$. Then, if we choose the last $\mu + 1$ rows of $S(\mathcal{P}_m)$, one of them has to be dependent on the row set made up from the rows of $E(\mathcal{P}_{m-1})$ and the rest μ rows chosen from S_m. This means that for some $l \in \{1, 2, \ldots, \mu + 1\}$ the l-th from the bottom row, can be eliminated under suitable elementary row operations. In terms of the polynomials corresponding to these rows, this elimination expresses the following polynomial relationship:

$$s^{l-1} b_m(s) = \sum_{\substack{i=0 \\ i \neq l-1}}^{\mu} v_{h,i} s^i b_h(s) + \sum_{i=0}^{p-1} v_{0,i} s^i a(s) + \sum_{j=1}^{m} \sum_{i=0}^{n-1} v_{j,i} s^i b_j(s) \tag{4.60}$$

where $v_{i,j} \in \mathbb{R}$ express the elementary row operations for the elimination of the columns. The next equation is readily derived from (4.60)

$$b_m(s) \sum_{i=0}^{\mu} v'_{h,i} s^i = a(s) \sum_{i=0}^{p-1} v_{0,i} s^i + \sum_{j=1}^{m} \left[b_j(s) \sum_{i=0}^{n-1} v_{j,i} s^i \right] \tag{4.61}$$

where

$$\left.\begin{array}{ll} v_m(s) = s^{l-1} - \sum\limits_{\substack{i=0\\ i\neq l}}^{\mu} v_{h,i}s^i = \sum\limits_{i=0}^{\mu} v'_{h,i}s^i, & \deg v_m(s) = \mu \\ v_j(s) = \sum\limits_{i=0}^{n-1} v_{j,i}s^i, \ \ j = 1,\dots,p-1, & \deg v_j(s) = n-1 \\ v_0(s) = \sum\limits_{i=0}^{p-1} v_{0,i}s^i, & \deg v_0(s) = p-1 \end{array}\right\} \quad (4.62)$$

and from (4.59), (4.62) it follows that

$$v_m(s)b_m(s) = v_0(s)a(s) + v_1(s)b_1(s) + \cdots + v_{m-1}(s)b_{m-1}(s) \quad (4.63)$$

and (4.63) implies that the gcd of $\mathcal{P}_{m-1}$ divides the LHS of the equation. Thus we may write $\varphi_{m-1}(s)|v_m(s)b_m(s)$. If we express $\varphi_{m-1}(s)$ in factors that divide $v_m(s)$, $b_m(s)$ respectively, i.e

$$\varphi_{m-1}(s) = \varphi_v(s)\cdot\varphi_b(s) \quad (4.64).$$

such that $\varphi_v(s)$, $b_m(s)$ are coprime, then it is obvious that $\varphi_b(s)$ *is the g.c.d of the set* $\mathcal{P}_m$, i.e. $\varphi_m(s) \equiv \varphi_b(s)$ and from the definition of the polynomials we have that: $\deg\varphi_v(s) \leq \deg v_m(s) = \mu$. By this and knowing that by definition $\deg\varphi_{m-1}(s) = k$ we finally obtain that $\deg\varphi_m(s) \geq k = k' - \mu$ and this implies that

$$\rho(S(\mathcal{P}_m)) \geq n + p - \deg\varphi(s) \quad (4.65)$$

However, by the factorisation of the resultant, as this is expressed by (4.47) it follows that

$$\rho(S(\mathcal{P}_m)) = \rho(S_{\mathcal{P}}) = \rho(\bar{S}_{\varphi^*}) \leq n + p - \deg\varphi(s) \quad (4.66)$$

and thus by combining (4.65) and (4.66) part (ii) of the result is established. Clearly, part (i) follows from part (ii) in a straightforward way.

iii) The proof of the last part makes use of the factorisation of the resultant as this is described by (4.47) i.e.

$$S_{\mathcal{P}} = \left[\mathbf{0}_k \mid \bar{S}_{\varphi^*}\right]\hat{\Phi}_{\varphi} \quad (4.67)$$

where $\rho\left(\bar{S}_{\varphi^*}\right) = n + p - \deg\varphi(s) = n + p - k$, $k = \deg\varphi(s)$ and $\hat{\Phi}_\varphi$ as given by (4.48). Since $\rho\left(\bar{S}_{\varphi^*}\right) = n + p - k$ and $\bar{S}_{\varphi^*}$ has $n + p - k$ columns, there exists $R \in \mathbb{R}^{\tau\times\tau}$, $\tau = p + hn$ such that $|R| \neq 0$ and

$$R\bar{S}_{\varphi^*} = \left[\begin{array}{c} I_{n+p-k} \\ \hdashline \mathbf{0} \end{array}\right] \tag{4.68}$$

By (4.55) and (4.56) we have

$$RS_\varphi = R\left[\mathbf{0}_k \;\vdots\; \bar{S}_{\varphi^*}\right]\hat{\Phi}_\varphi = \left[\begin{array}{c:c} \mathbf{0}_k & \begin{array}{c} I_{n+p-k} \\ \hdashline \mathbf{0} \end{array} \end{array}\right]\hat{\Phi}_\varphi \tag{4.69}$$

Taking into account the structure of $\hat{\Phi}_\varphi$ (as in (4.38)) for the gcd

$$\varphi(s) = \lambda_k s^k + \cdots + \lambda_1 s + \lambda_0 \tag{4.70}$$

leads to

$$RS_\varphi = \begin{array}{cl} \left[\begin{array}{cccccccccc} \lambda_k & \lambda_{k-1} & \lambda_{k-2} & \cdots & \cdots & \lambda_0 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_k & \lambda_{k-1} & \lambda_{k-2} & \cdots & \cdots & \lambda_0 & 0 & & 0 \\ \vdots & \ddots & \ddots & & & & & \ddots & & \\ \vdots & & & \ddots & & & & & \ddots & \\ 0 & \cdots & \cdots & 0 & \lambda_k & \lambda_{k-1} & \lambda_{k-2} & \cdots & \cdots & \lambda_0 \\ \hdashline & & & & & & & & & \\ \multicolumn{10}{c}{\mathbf{0}} \\ & & & & & & & & & \end{array}\right] & \begin{array}{l} \Bigg\updownarrow n+p-k \\ \\ \\ \Bigg\updownarrow \tau-n-p+k \end{array} \end{array} \tag{4.71}$$

The structure of (4.71) suggests that the echelon form is obtained by making the leading coefficient of each row 1, starting from the first and then making all elements in the column above the leading coefficient zero. This process starts from the first row in (4.71) and progressively reaches the last and leads to the echelon form

$$R'RS_{\mathcal{P}}=\left[\begin{array}{cccccccccc}1 & 0 & 0 & \cdots & 0 & * & * & \cdots & \cdots & * \\ 0 & 1 & 0 & \cdots & 0 & * & * & \cdots & \cdots & * \\ \vdots & \ddots & \ddots & & \vdots & \vdots & \vdots & & & \vdots \\ \vdots & & & \ddots & 0 & * & * & \cdots & \cdots & * \\ 0 & \cdots & \cdots & 0 & 1 & \lambda'_{k-1} & \lambda'_{k-2} & \cdots & \cdots & \lambda'_0 \\ \hdashline \\ \multicolumn{10}{c}{\mathbf{0}} \\ \\ \end{array}\right] \qquad (4.72)$$

where $\{1,\lambda'_{k-1},\lambda'_{k-2},\ldots,\lambda'_1,\lambda'_0\}$ are the coefficients of the monic gcd and this completes the proof of part (iii).

For the case of non-proper sets we can use factorisation (4.32) and Lemma (4.5) and part (ii) follows. Part (iii) follows along similar lines to that of the proper sets.

■

An important corollary of the above result that provides an alternative representation of the gcd of the set is given below:

Corollary (4.3): Let $\mathcal{P}=\{a(s),b_i(s),i=1,\ldots,h,\deg a(s)=n,\max\{b_i(s)\}=p,n\geq p\}$ be a set of polynomials with a $(p+hn)\times(n+p)$ generalised resultant $S_{\mathcal{P}}$ and let $\varphi(s)=\lambda_k s^k+\lambda_1 s+\cdots+\lambda_0$, $\lambda_k\neq 0$ be the gcd of $\mathcal{P}$. If $a(s)=a'(s)\varphi(s)$, $b_i(s)=b_i'(s)\varphi(s)$, $i=1,\ldots,h$ and $\mathcal{P}^*=\{a'(s),b_i'(s),i=1,\ldots,h\}$ is the corresponding reduced coprime set with a generalized resultant $S_{\mathcal{P}^*}$, then $S_{\mathcal{P}}$ may be expressed as

$$S_{\mathcal{P}}=\overline{S}_{\mathcal{P}^*}\Theta_{\varphi} \qquad (4.73)$$

where $\overline{S}_{\mathcal{P}^*}$ is the $(p+hn)\times(n+p-k)$ (n,p)-expanded resultant of $\mathcal{P}^*$, $\rho\left(\overline{S}_{\mathcal{P}^*}\right)=n+p-k$ and Θ_{φ} is the $(n+p-k)\times(n+p)$ matrix

$$\Theta_{\varphi}=\begin{bmatrix}\lambda_k & \lambda_{k-1} & \lambda_{k-2} & \cdots & \cdots & \lambda_0 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_k & \lambda_{k-1} & \lambda_{k-2} & \cdots & \cdots & \lambda_0 & 0 & & 0 \\ \vdots & \ddots & \ddots & & & & & \ddots & & \\ \vdots & & & \ddots & & & & & \ddots & \\ 0 & \cdots & \cdots & 0 & \lambda_k & \lambda_{k-1} & \lambda_{k-2} & \cdots & \cdots & \lambda_0\end{bmatrix} \qquad (4.74)$$

Proof:

The set of polynomials $\mathcal{P}^*$ is by definition coprime and thus its resultant $S_{\mathcal{P}^*}$ which has dimensions $\left[p+hn-k(h+1)\right]\times(n+p-2k)$ has rank $(n+p-2k)$. By lemma (4.5) it is clear that $\rho\left(\overline{S}_{\mathcal{P}^*}\right)=n+p-2k+k=n+p-k$. Consider first the case of a proper set $\mathcal{P}$. Then by (4.47)

$$S_{\mathcal{P}}=\left[\mathbf{0}_k \;\middle|\; \overline{S}_{\mathcal{P}^*}\right]\hat{\Phi}_{\varphi} \tag{4.75a}$$

and from (4.48) form of $\hat{\Phi}_{\varphi}$ we have that

$$\hat{\Phi}_{\varphi}=\left[\begin{array}{c}\Theta'_{\varphi}\\ \hdashline \Theta_{\varphi}\end{array}\right] \tag{4.75b}$$

where Θ_{φ} has the (4.74) structure and thus

$$S_{\mathcal{P}}=\left[\mathbf{0}_k \;\middle|\; \overline{S}_{\mathcal{P}^*}\right]\left[\begin{array}{c}\Theta'_{\varphi}\\ \hdashline \Theta_{\varphi}\end{array}\right]=\overline{S}_{\mathcal{P}^*}.\Theta_{\varphi} \tag{4.75c}$$

To show that the same factorisation holds true for the non-proper case we start from the factorisation (4.48) described by corollary (4.5), that is

$$S_{\mathcal{P}}=\left[\mathbf{0}_{k'+c} \;\middle|\; \overline{S}_{\mathcal{P}^*}\right]\hat{\Phi}_{\varphi'}\begin{bmatrix}\mathbf{0} & I_c\\ I_{n+p-c} & \mathbf{0}\end{bmatrix} \tag{4.76}$$

where now k' denotes the degree of the gcd after extracting s^c factor i.e.

$$\varphi(s)=\varphi'(s)s^c=s^c\lambda_{k'}s^{k'}+\cdots+\lambda_1 s+\lambda_0 \tag{4.77}$$

and $\hat{\Phi}_{\varphi'}$ is the $(n+p)\times(n+p)$ formed from $\varphi'(s)$. We can always partition $\hat{\Phi}_{\varphi'}$ as

$$\hat{\Phi}_{\varphi'}=\left[\begin{array}{c}\Theta'_{\varphi'}\\ \hdashline \Theta_{\varphi'}\end{array}\right]\begin{array}{l}\updownarrow\; k'+c\\ \updownarrow\; n+p-k'-c\end{array} \qquad \in\mathbb{R}^{(n+p)\times(n+p)} \tag{4.78a}$$

where $\Theta_{\varphi'}$ has the form

$$\Theta_{\varphi'}=\begin{array}{cc}\left[\begin{array}{ccc|cccccccc}
0 & \cdots & 0 & \lambda_{k'} & \cdots & & \lambda_1 & \lambda_0 & 0 & \cdots & 0\\
\vdots & & \vdots & 0 & \lambda_{k'} & \cdots & & \lambda_1 & \lambda_0 & \ddots & \vdots\\
 & & & \vdots & & \ddots & & & \ddots & \ddots & 0\\
0 & \cdots & 0 & 0 & \cdots & 0 & \lambda_{k'} & \cdots & & \lambda_1 & \lambda_0
\end{array}\right] & \begin{array}{c}\uparrow\\ n+p-k'-c\\ \downarrow\end{array}\\
\begin{array}{cc}\xleftrightarrow{\quad c\quad} & \xleftrightarrow{\qquad\qquad n+p-c\qquad\qquad}\end{array} & \end{array} \tag{4.78b}$$

and by introducing (4.78a), (4.78b) into (4.76) we have

$$S_{\varphi} = \left[\mathbf{0}_{k'+c} \mid \bar{S}_{\varphi^*} \right] \begin{bmatrix} \Theta'_{\varphi'} \\ \hdashline \hat{\Theta}_{\varphi'} \end{bmatrix} \begin{bmatrix} \mathbf{0} & I_c \\ I_{n+p-c} & \mathbf{0} \end{bmatrix}$$

$$= \bar{S}_{\varphi^*}.\hat{\Theta}_{\varphi'} \begin{bmatrix} \mathbf{0} & I_c \\ I_{n+p-c} & \mathbf{0} \end{bmatrix} = \bar{S}_{\varphi^*}.\tilde{\Theta}_{\varphi} \qquad (4.78c)$$

where

$$\tilde{\Theta}_{\varphi} = \left[\begin{array}{cccccccccc:ccc} \lambda_{k'} & \cdots & & \lambda_1 & \lambda_0 & 0 & & \cdots & 0 & & 0 & \cdots & 0 \\ 0 & \lambda_{k'} & \cdots & & \lambda_1 & \lambda_0 & & \ddots & \vdots & & \vdots & & \vdots \\ \vdots & & \ddots & & & \ddots & & \ddots & 0 & & & & \\ 0 & \cdots & 0 & \lambda_{k'} & \cdots & & & \lambda_1 & \lambda_0 & & 0 & \cdots & 0 \end{array} \right]$$
$$\hphantom{\tilde{\Theta}_{\varphi} = \left[\lambda_{k'} \cdots \lambda_1 \lambda_0 0 \cdots 0 \right.} \underset{c}{\longleftrightarrow}$$

The above however is the Θ_{φ} matrix that corresponds to the gcd $\varphi(s) = s^c \left(\lambda_{k'} s^{k'} + \cdots + \lambda_1 s + \lambda_0 \right)$ and this completes the proof.

■

The significance of the above factorization is that unifies the resultant factorization for the proper and the non-proper case since (4.73), (4.74) are valid for both cases and emphasises the minimality of this factorization since $\mathcal{P}^*$ is obtained by the division of the set by the gcd. Such a representation will be called *canonical representation* of gcd and involves the minimal number of parameters.

Example 4.4: Let $a(s) = s^3 + 4s^2 + 4s + 3$, $b_1(s) = s^2 + s - 6$ and $b_2(s) = s^2 + 5s + 6$ then the associated resultant is:

$$S = \begin{bmatrix} S_0 \\ \hdashline S_1 \\ \hdashline S_2 \end{bmatrix} = \left[\begin{array}{ccccc} 1 & 4 & 4 & 3 & 0 \\ 0 & 1 & 4 & 4 & 3 \\ \hdashline 1 & 1 & -6 & 0 & 0 \\ 0 & 1 & 1 & -6 & 0 \\ 0 & 0 & 1 & 1 & -6 \\ \hdashline 1 & 5 & 6 & 0 & 0 \\ 0 & 1 & 5 & 6 & 0 \\ 0 & 0 & 1 & 5 & 6 \end{array} \right]$$

The rank of the resultant is 4 which means that a gcd of order one is expected. Indeed if we reduce the resultant to its row echelon form using only row transformation gives:

$$S_{ech}=\begin{bmatrix}1&0&0&0&-81\\0&1&0&0&27\\0&0&1&0&-9\\0&0&0&1&3\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\end{bmatrix}$$

and from the last non-vanishing row it follows that the g.c.d. is $\varphi(s)=s+3$. This implies that the transformation matrix Φ, given by:

$$\Phi=\hat{\Phi}^{-1}=\begin{bmatrix}3&0&0&0&0\\1&3&0&0&0\\0&1&3&0&0\\0&0&1&3&0\\0&0&0&1&3\end{bmatrix}^{-1}=\begin{bmatrix}\frac{1}{3}&0&0&0&0\\-\frac{1}{9}&\frac{1}{3}&0&0&0\\\frac{1}{27}&-\frac{1}{9}&\frac{1}{3}&0&0\\-\frac{1}{81}&\frac{1}{27}&-\frac{1}{9}&\frac{1}{3}&0\\\frac{1}{243}&-\frac{1}{81}&\frac{1}{27}&-\frac{1}{9}&\frac{1}{3}\end{bmatrix}$$

and then the factorization of the resultant is expressed as:

$$S=\begin{bmatrix}0&1&1&1&0\\0&0&1&1&1\\0&1&-2&0&0\\0&0&1&-2&0\\0&0&0&1&-2\\0&1&2&0&0\\0&0&1&2&0\\0&0&0&1&2\end{bmatrix}\begin{bmatrix}3&0&0&0&0\\1&3&0&0&0\\0&1&3&0&0\\0&0&1&3&0\\0&0&0&1&3\end{bmatrix}$$

■

4.5. RESULTANT SETS FOR POLYNOMIALS AND A SYSTEM THEORETIC CHARACTERISATION OF GCD

The matrix pencil theory [Gantmacher, 1988] has been used in [Karcanias et al., 1994] to provide a procedure for evaluation of the gcd of a set of polynomials. The procedure in [Karcanias et al., 1994] makes no special assumption on the nature of the set of polynomials and provides also a system theoretic characterization of the gcd. In this section we simplify the computational procedure for the gcd evaluation and provide a more direct system theoretic characterization of the gcd by using the properties of the Sylvester set of a given set of polynomials. We first define:

Definition (4.7): Consider the set $\mathcal{P}=\{a(s),b_i(s),i=1,\dots,h,\deg a(s)=n,$ $\max\{b_i(s)\}=p,n\geq p\}$ and let $S_{\mathcal{P}}$ be its $(p+hn)\times(n+p)$ resultant. If $\underline{e}_k(s)=\left[s^k,\dots,s,1\right]^t$ then the polynomial vector

$$\underline{p}_{\mathcal{P}}^{s}(s)=S_{\mathcal{P}}\underline{e}_{n+p}(s) \tag{4.79}$$

is the polynomial vector representative of a set

$\mathcal{P}^s=\{\tilde{a}(s),\tilde{b}_i(s),i=1,\dots,p+hn,\deg\{\tilde{a}(s)\}=n+p\}$ which will be called the *Sylvester Resultant set* of $\mathcal{P}$.

■

It is clear that the basis matrix contains information for the gcd (see Theorem (2.1)). An alternative way for expressing gcd related properties is through some appropriate matrix pencil established by the following result:

Lemma (4.6): [Karcanias et al., 1994]: For a general set of polynomials $\mathcal{P}$ of maximal degree d and a basis matrix, $P\in\mathbb{R}^{k\times(d+1)}$, $\operatorname{rank}(P)=\rho<d+1$ we define a basis $M\in\mathbf{R}^{(d+1)\times\mu}$, $\mu=d-\rho+1$ for $\mathcal{N}_r\{\mathcal{P}\}$ and denote by M_1,M_2 the matrices obtained from M by deleting the last, first row of M respectively. The pencil $Z(s)=M_1-sM_2$ is known as the *GCD pencil* of the set and has the following properties:

i) The set of Kronecker invariants consists of row minimal indices (rmi) and possibly finite elementary divisors (fed).

ii) The zero polynomial of $Z(s)$ (product of all fed) is the g.c.d. of the set $\mathcal{P}$. ■

The above result is the initial step for the matrix pencil approach to gcd computation [Karcanias et al., 1994]. The fact that $Z(s)$ may have nonzero rmi led to a further investigation on the evaluation of the zero polynomial for the case of a general set. In the following we shall explore the resultant properties to avoid the difficulties associated with the presence of non-zero rmi, which may arise with general sets of polynomials. The use of resultant sets leads to the following result.

Theorem (4.6): Let $\mathcal{P}=\{a(s),b_i(s),i=1,...,h,\deg a(s)=n,\ \max\{b_i(s)\}=p,n\geq p\}$ be a general set of polynomials $S_{\mathcal{P}}$ be its generalized resultant and $\mathcal{P}^s$ the corresponding resultant set. Then, the following properties hold true:

i) The set $\mathcal{P}$ is coprime, if and only if $\mathcal{N}_r\{S_{\mathcal{P}}\}=\{0\}$.

ii) The set $\mathcal{P}$ has a non-trivial gcd $\varphi(s)$ if and only if $\mathcal{N}_r\{S_{\mathcal{P}}\}\neq\{0\}$. In this case $\deg\varphi(s)=\dim\mathcal{N}_r\{S_{\mathcal{P}}\}=k$.

iii) If $Z(s)$ is the gcd pencil of $\mathcal{P}^s$ resultant set then

a) $Z(s)$ is characterized by fed and possibly only zero rmi.

b) $Z(s)$ may be expressed as

$$Z(s)=M_1-sM_2=\bar{M}\left(sI-\bar{Z}\right),\ \bar{Z}\in\mathbb{R}^{k\times k} \tag{4.80}$$

where the characteristic polynomial of $\bar{Z}$ defines the monic g.c.d. $\varphi(s)$ of $\mathcal{P}$.

Proof:

(i) and (ii). These parts follow directly from the resultant theorem.

(iii) The g.c.d. is nontrivial when $\mathcal{N}_r\{S_{\mathcal{P}}\}\neq\{0\}$. In this case the gcd pencil $Z(s)=sM_1-M_2$ as defined by Lemma (4.6) [Karcanias et al., 1994]has rmi and possibly fed. Then the GCD pencil has the following Kronecker decomposition [Gantmacher, 1988]

$$Z(s) = R \left[\begin{array}{c|c} \cdots & \cdots \\ 0 & 0 \\ \cdots & \cdots \\ L_2(s) & 0 \\ 0 & sI - A \end{array}\right] Q \tag{4.81}$$

where $L_2(s)$ is the set of blocks associated with the nonzero rmi and $sI - A$ characterize the finite zeros. Clearly, if $L_2(s)$ exists, then $\deg\{\varphi(s)\} = \deg\{|sI - A|\} < \dim \mathcal{N}_r\{S_{\mathcal{P}}\}$ and by the resultant theorem that leads to a contradiction. Thus the GCD pencil $Z(s)$ has no nonzero rmi and thus its structure, as expressed by the Kronecker decomposition, becomes

$$Z(s) = R \begin{bmatrix} 0 \\ \vdots \\ sI - A \end{bmatrix} Q \tag{4.82}$$

By partitioning R according to the partitioning of the Kronecker form we have

$$Z(s) = [R', R] \begin{bmatrix} 0 \\ \vdots \\ sI - A \end{bmatrix} Q = \overline{R}(sI - A)Q \tag{4.83}$$

■

<u>**Corollary (4.4):**</u> If $\mathcal{N}_r\{S_{\mathcal{P}}\} \neq \{0\}$ and $Z(s)$ is the GCD pencil of $\mathcal{P}$ then any minor of maximal order of $Z(s)$ which is not identically zero defines the gcd of the set $\mathcal{P}$.

■

The above results clearly indicate that use of the resultant set $\mathcal{P}^s$ overcomes the difficulties of the presence of rmi in the pencil and reduces the evaluation of gcd to a problem of computing a basis for the right null space of the resultant $S_{\mathcal{P}}$. An interesting corollary that follows from the Theorem and which emphasises this alternative computational procedure is given below:

<u>**Corollary (4.5):**</u> Consider the set $\mathcal{P}$ for which $\mathcal{N}_r\{S_{\mathcal{P}}\} \neq \{0\}$ and let M be a basis of $\mathcal{N}_r\{S_{\mathcal{P}}\}$. If we write

$$M = \begin{bmatrix} M_1 \\ \hdashline \underline{\hat{m}}^t \end{bmatrix} = \begin{bmatrix} \underline{\tilde{m}}^t \\ \hdashline M_2 \end{bmatrix} \tag{4.84}$$

then $\operatorname{colsp}\{M_1\} \subseteq \operatorname{colsp}\{M_2\}$ and thus

$$M_1 = M_2\bar{Z}\ ,\ Z \in \mathbb{R}^{k\times k}\ ,\ \dim \mathcal{N}_r\{S_{\mathcal{P}}\} = k \tag{4.85}$$

where $\bar{Z}$ is the matrix that has as characteristic polynomial the monic gcd of $\mathcal{P}$.

Proof:

From condition (4.80) of Theorem (4.6) it follows that

$$M_1 = -\bar{M}\bar{Z}\ ,\ M_2 = -\bar{M} \tag{4.86}$$

and thus

$$M_1 = M_2\bar{Z}\ ,\ Z \in \mathbb{R}^{k\times k}\ ,\ \dim \mathcal{N}_r\{S_{\mathcal{P}}\} = k$$

which clearly implies that $\operatorname{colsp}\{M_1\} \subseteq \operatorname{colsp}\{M_2\}$ with equality holding only when the gcd has no zero divisors.

■

The above characterization provides the means for deriving numerical procedures which may improve the overall performance of the matrix pencil framework for the gcd computation.

Example 4.5: Lets examine again the polynomials $a(s) = s^3 + 4s^2 + 4s + 3$, $b_1(s) = s^2 + s - 6$ and $b_2(s) = s^2 + 5s + 6$ then

$$S_3 = \left[\begin{array}{ccccc} 1 & 4 & 4 & 3 & 0 \\ 0 & 1 & 4 & 4 & 3 \\ \hdashline 1 & 1 & -6 & 0 & 0 \\ 0 & 1 & 1 & -6 & 0 \\ 0 & 0 & 1 & 1 & -6 \\ \hdashline 1 & 5 & 6 & 0 & 0 \\ 0 & 1 & 5 & 6 & 0 \\ 0 & 0 & 1 & 5 & 6 \end{array}\right]$$

The right null space of S_3 is:

$$\mathrm{N}_r\{S_3\} = \left\langle \begin{bmatrix} 81 \\ -27 \\ 9 \\ -3 \\ 1 \end{bmatrix} \right\rangle$$

The corresponding GCD Pencil is:

$$Z_3(s) = \begin{bmatrix} 81 \\ -27 \\ 9 \\ -3 \end{bmatrix} - s \begin{bmatrix} -27 \\ 9 \\ -3 \\ 1 \end{bmatrix} = \begin{bmatrix} 81+27s \\ -27-9s \\ 9+3s \\ -3-s \end{bmatrix}$$

and it is obvious that the monic g.c.d. of the set $\{a(s)b_1(s)b_2(s)\}$ is $\varphi(s) = s+3$ ■

The current extention of the Matrix Pencil algorithm for evaluating the gcd has the advantage that reduces the gcd computation to determing a matrix pencil, which has the property that any nonzero maximal order minor defines the gcd. This is a considerable simplification and it is due to the properties of the Sylvester resultant which establishes equivalence between right kernel dimension of Sylvester Resultant and degree of gcd.

4.6. DISCUSSION

This chapter dealt with two matrix based methods for the evaluation of the greatest common divisor of polynomials. The first one was based on the Sylvester resultant matrix. A new proof for the known properties of the Sylvester matrix was described based on properties of polynomials. The approach provides a new insight in the structure of resultants and their properties, such as those of defining g.c.d. in terms of reduced resultants. A useful result from this proof is that by applying column operations on any basis matrix we can obtain the set of coprime polynomials using specially structured column operations. This provides a factorization of resultants into resultants corresponding to coprime polynomials and square structured transformations and provides a representation in matrix terms of the factorization of a set of polynomials in terms of the gcd and a coprime set. In this representation, the gcd is represented by a Toeplitz matrix characterized entirely by the coefficient vector of the gcd and a reduced Sylvester resultant parameterised by the degree set of the original set of polynomials and by the coefficients of the factor polynomials in the gcd factorization.

Next we have considered the significance of the Sylvester set of polynomials of the original set in the Matrix Pencil based computation procedure. After the description of the initial Matrix Pencil algorithm, we examined a new approach that combines the Sylvester matrix and the Matrix Pencil algorithm. The benefits of this combination is that eliminates the nonzero row minimal indices problem in the original Matrix Pencil algorithm (that necessitates further transformations) and thus simplifies the gcd computation.

The new resultant factorization opens the way for establishing results on the approximate gcd and this will be considered in the following chapters.

Chapter 5:

CHARACTERISATION OF APPROXIMATE GREATEST COMMON DIVISORS OF POLYNOMIALS OF DIFFERENT ORDERS

5.1. INTRODUCTION

In the previous chapter we have dealt with the notion of the gcd of many polynomials and we have introduced new properties related to the Sylvester Resultant matrix and the matrix representation of the gcd. The notion of gcd of many polynomials is characterised by the property that its computation is a non-generic problem [Karcanias et al., 1999]. However, the need for defining notions such as "almost zeros" and "approximate gcd" has been recognised as important in many applications. The notion of a zero of a set of polynomials $\mathcal{P}$ with vector representative $\underline{p}(s)$ has been extended to that of "almost zero" [Karcanias et al., 1983] as a problem of minimisation of the function $\left\|\underline{p}(\sigma + j\omega)\right\|$.

Methods for computing the gcd of the set $\mathcal{P}$, which deploy relaxation of the exact conditions for gcd evaluation, such as ERES method [Mitrouli et al., 1993] lead to expressions for the "approximate gcd" [Mitrouli et al., 1993]. Recently, [Noda et al., 1991], [Emiris et al., 1997], the problem of the "approximate gcd" has been considered in the context of the Euclidean division algorithm and for the case of two polynomials and the approximate gcd concept has been related to the accuracy of the approximation. The definition of the accuracy indicates how good the characterisation of the approximate gcd is. The essence of current methods for introduction of "approximate gcd" is the relaxation of conditions characterising the exact notion. The difficulty with many of the current methods is in quantifying how good is the approximation that is offered. The Euclidean approach addresses this problem, but it is limited to the case of two polynomials. The efficient numerical method based on ERES Algorithm, defines approximate solutions, but does not characterise the strength of approximation.

The problem which is addressed in this chapter is to introduce formally the notion of the "approximate gcd" and then develop a computational procedure that allows the evaluation of how good is the given "approximate gcd". We will define the strength or quality of a given "approximate gcd" by the size of the minimal perturbation required to make a chosen "approximate gcd" an exact gcd of a perturbed set of polynomials. The results that were introduced in Chapter 4 based on the representation of the gcd in terms of the generalised resultant and its factorisation into a reduced resultant and a Toeplitz matrix, allow the parameterisation of all perturbations which are required to make a selected "approximate gcd", an exact gcd of a perturbed set.

5.2. CHARACTERISATION OF APPROXIMATE GCD USING PROPERTIES OF THE RESULTANT MATRIX

Let us denote by $\mathcal{T}(n,p;h+1)$ the set of all polynomial sets $\mathcal{P}_{h+1,n}$ having $h+1$ elements, and with the two higher degrees (n,p), $n \geq p$; that is if $\mathcal{P}_{h+1,n} = \{p_i(s),\ i=0,1,...,h\} \in \mathcal{T}(n,p;h+1)$ then $\deg\{p_0(s)\} = n$, $\deg\{p_1(s)\} = p$, $\deg\{p_i(s)\} \leq p$, $i = 2,...,h$. If we denote

$$p_0(s) = a_n s^n + a_{n-1}s^{n-1} + ... + a_1 s + a_0 = \underline{a}^t \underline{e}_n(s)$$

$$p_i(s) = b_{i.p}s^p + b_{i,p-1}s^{p-1} + ... + b_{i,1}s + b_{i,0} = \underline{b}_i^t \underline{e}_p(s) \tag{5.1a}$$

where $\underline{e}_k(s) = \left[s^k s^{k-1},...,s,1\right]$, then to the set $\mathcal{P}_{h+1,n}$ we may correspond the vector

$$\underline{p}_{h+1,n} = \left[\underline{a}^t \quad \underline{b}_1^t \quad \cdots \quad \underline{b}_n^t\right]^t \in \mathbb{R}^N \tag{5.1b}$$

where $N = (n+1) + h(p+1)$, or alternatively a point $P_{h+1,n}$ in the projective space $\mathbf{P}^{N-1}$. The set $\mathcal{T}(n,p;h+1)$ is clearly isomorphic with $\mathbb{R}^N$, or $\mathbf{P}^{N-1}$. An important question relates to the characterisation of all points of $\mathbf{P}^{N-1}$, which correspond to sets of polynomials with a given degree gcd. Such sets of polynomials correspond to certain varieties of $\mathbf{P}^{N-1}$, which are defined below. We first note that an alternative representation of $\mathcal{P}_{h+1,n}$ is provided by the generalised Sylvester resultant $S_{\mathcal{P}} \in \mathbb{R}^{(p+hn)\times(n+p)}$ which is a matrix defined by the vector of coefficients $\underline{p}_{h+1,n}$. If we denote by $C_k(\cdot)$ the k-th compound of S_φ [Marcus et al., 1969], then the varieties characterising the sets having, a given degree d, gcd are defined below:

Proposition 5.1: Let $\mathcal{T}(n,p;h+1)$ be the set of all polynomial sets $\mathcal{P}_{h+1,n}$ with $h+1$ elements and with the two higher degrees (n,p), $n \geq p$ and let S_φ be the Sylvester resultant of the general set $\mathcal{P}_{h+1,n}$. The variety of $\mathbf{P}^{N-1}$ which characterise all sets $\mathcal{P}_{h+1,n}$ having a gcd with degree d, $0 < d \leq p$ is defined by the set of equations

$$C_{n+p-d+1}(S_\varphi) = 0 \tag{5.2}$$

Proof:

By Theorem 4.5 $\rho(S_\varphi) = n + p - \deg \varphi(s)$. Thus if $\deg \varphi(s) = d$, then all $n+p-d+1$ minors of S_φ are zero and there is at least one minor of $n+p-d$ order which is nonzero. These conditions are clearly derived by (5.2)

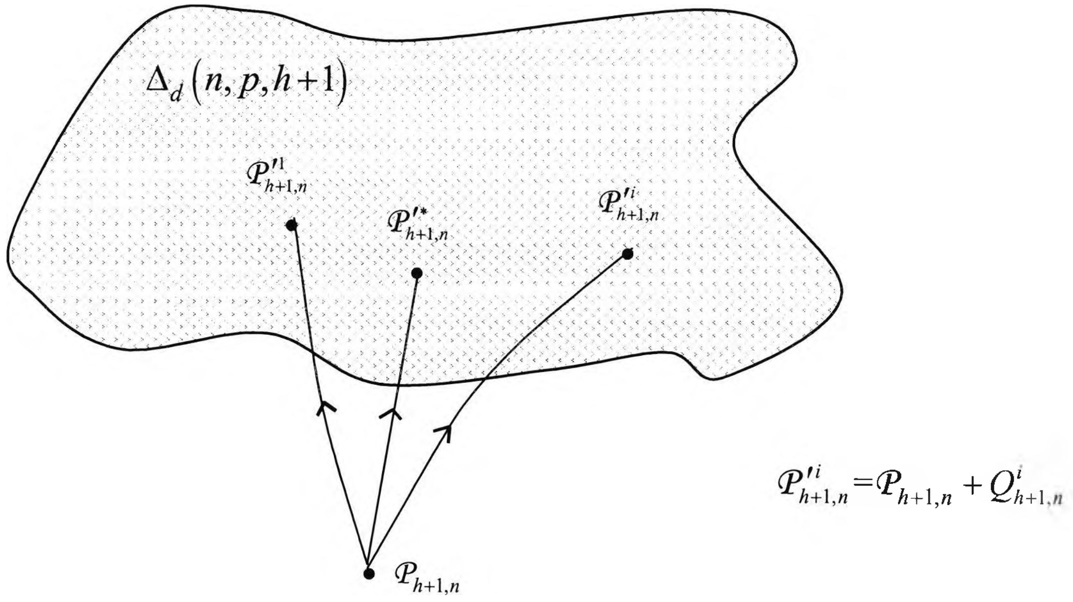
■

Conditions (5.2) define polynomial equations in the parameters of the vector $\underline{p}_{h+1,n}$, or the point $P_{h+1,n}$ of $\mathbf{P}^{N-1}$ (note that the gcd of $\mathcal{P}_{h+1,n}$ is not affected by scaling uniformly all coefficients by a constant). The set of equations in (5.2) define a variety of $\mathbf{P}^{N-1}$, which will be denoted by $\Delta_d(n,p;h+1)$ and will be referred to as the *d-gcd variety* of $\mathbf{P}^{N-1}$. $\Delta_d(n,p;h+1)$ characterises all sets in $\mathcal{T}(n,p;h+1)$ which have a gcd with degree d. Clearly all roots of the GCDs may be any.

Remark (5.1): The sets $\Delta_d(n,p;h+1)$ have measure zero [Hirsch et al., 1974] and thus the existence of a nontrivial gcd $d > 0$ is a nongeneric property.

■

From the above, a generic set $\mathcal{P}_{h+1,n}$ does not belong to $\Delta_d(n,p;h+1)$. The important question that is posed, is how close the given set $\mathcal{P}_{h+1,n}$ is to given variety $\Delta_d(n,p;h+1)$. Being able to define a distance in that sense is the key to defining the notion of the "approximate gcd". The following diagram illustrates the notion of "approximate gcd" we are trying to define.



$\Delta_d(n,p,h+1)$
$\mathcal{P}'^{1}_{h+1,n}$
$\mathcal{P}'^{*}_{h+1,n}$
$\mathcal{P}'^{i}_{h+1,n}$
$\mathcal{P}'^{i}_{h+1,n} = \mathcal{P}_{h+1,n} + \mathcal{Q}^{i}_{h+1,n}$
$\mathcal{P}_{h+1,n}$

Figure 5.1 The d-gcd variety and the distance problem.

In fact, if $\mathcal{Q}^{i}_{h+1,n}$ is some perturbation set (to be properly defined) and assuming that $\mathcal{P}'^{i}_{h+1,n} = \mathcal{P}_{h+1,n} + \mathcal{Q}^{i}_{h+1,n}$ such that $\mathcal{P}'^{i}_{h+1,n} \in \Delta_d(n,p;h+1)$, then the gcd of $\mathcal{P}'^{i}_{h+1,n}$, $\varphi(s)$, with degree d defines the notion of the approximate gcd" and its strength is defined by the "size" of the perturbation $\mathcal{Q}^{i}_{h+1,n}$. Numerical procedures such as ERES, produce estimates of an "approximate gcd". Estimating the size of the corresponding perturbations provides the means to evaluate how good such approximations are. By letting the parameters of the gcd free and searching for that with the minimal size of the corresponding perturbations is a distance problem and this leads to the notion of the "optimal approximate gcd". The key questions which have to be considered for such studies are:

i) Existence of perturbations that produce from $\mathcal{P}_{h+1,n}$ an element $\mathcal{P}'_{h+1,n}=\mathcal{P}_{h+1,n}+\mathcal{Q}_{h+1,n}\in\Delta_d(n,p;h+1)$.

ii) Parameterisations of all such perturbations.

iii) Minimal distance of $\mathcal{P}_{h+1,n}$ from an element of $\Delta_d(n,p;h+1)$ with a given gcd $\varphi(s)$, and thus evaluation of strength of $\varphi(s)$.

iv) Minimal distance of $\mathcal{P}_{h+1,n}$ from $\Delta_d(n,p;h+1)$ and thus computation of the "optimal approximate gcd.

The characterisaton of the $\Delta_d(n,p;h+1)$ variety in a parametric form, as well as subvarieties of it, is a crucial issue for the further development of the topic. The subset of $\Delta_d(n,p;h+1)$, characterised by the property that all $\mathcal{P}_{h+1,n}$ in it have a given gcd $\upsilon(s)\in\mathbb{R}[s]$, $\deg\{\upsilon(s)\}=d$, it can be shown to be a subvariety of $\Delta_d(n,p;h+1)$ and shall be denoted by $\Delta_d^{\upsilon}(n,p;h+1)$. In fact $\Delta_d^{\upsilon}(n,p;h+1)$ is characterised by the equations of $\Delta_d(n,p;h+1)$ and a set of additional linear relations amongst the parameters of the vector $\underline{p}_{h+1,n}$. The parametric description of $\Delta_d(n,p;h+1)$ and $\Delta_d^{\upsilon}(n,p;h+1)$ follows from the results of Chapter 4 on the factorisation of resultants and it is expressed bellow:

<u>**Proposition 5.2:**</u> Consider the set $\mathcal{T}(n,p;h+1)$, $\mathbf{P}^{N-1}$ be the associated projective space, $\mathcal{P}_{h+1,n}\in\mathcal{T}(n,p;h+1)$ and let $S_{\mathcal{P}}$ be the associated resultant. Then,

i) The variety $\Delta_d(n,p;h+1)$ of $\mathbf{P}^{N-1}$ is expressed parametrically by the generalised resultant

$$S_{\mathcal{P}} = \left[\mathbf{0}_d \mid \bar{S}_{\mathcal{P}^*}\right]\hat{\Phi}_\upsilon \tag{5.3}$$

where $\hat{\Phi}_\upsilon$ is the $(n+p)\times(n+p)$ Toeplitz representation of an arbitrary $\upsilon(s)\in\mathbb{R}[s]$ with $\deg\{\upsilon(s)\}$ and $\bar{S}_{\mathcal{P}^*}\in\mathbb{R}^{(p+hn)\times(n+p-d)}$ is the (n,d)-expanded resultant of an arbitrary set of polynomials $\mathcal{P}^*\in\Pi(n-d,p-d;h+1)$.

ii) The variety $\Delta_d^u(n,p;h+1)$ of $\mathbf{P}^{N-1}$ is defined by (5.3) with the additional constraint that $\upsilon(s)\in\mathbb{R}[s]$ is given.

■

Clearly, the free parameters in $\Delta_d(n,p;h+1)$ are the coefficients of the polynomials of $\Pi(n-d,p-d;h+1)$. Having defined the description of these varieties we consider next the perturbations that transfer a general set $\mathcal{P}_{h+1,n}$ on them. If $\mathcal{P}_{h+1,n}\in\Pi(n,p;h+1)$ we can define an (n,p)-ordered perturbed set by:

$$\mathcal{P}'_{h+1,n} = \mathcal{P}_{h+1,n} - \mathcal{Q}_{h+1,n} \in \Pi(n,p;h+1) \ :$$

$$\mathcal{P}'_{h+1,n} = \left\{p'_i(s) = p_i(s) - q_i(s) : \deg\{q_i(s)\} \le \deg\{p_i(s)\},\ i=0,1,\dots,h\right\} \tag{5.4}$$

Using the set of perturbations defined above we may now show that any polynomial from a certain class may become an exact gcd of a perturbed set under a family of perturbations.

Proposition (5.3): Given a set $\mathcal{P}_{h+1,n}$ with maximal degrees (n,p), $n\ge p$ and a polynomial $\omega(s)\in\mathbb{R}[s]$ with $\deg\{\omega(s)\}\le p$. There always exists a family of (n,p)-

ordered perturbations $Q_{h+1,n}$ such that for every element of this family $\mathcal{P}'_{h+1,n} = \mathcal{P}_{h+1,n} - Q_{h+1,n}$ has a gcd which is divisible by $\omega(s)$.

Proof:

Given $\mathcal{P}_{h+1,n}$, consider $\bar{Q}_{h+1,n} = \{\bar{q}_i(s)\}$ as an arbitrary (n,p)-ordered perturbation and let $\bar{\mathcal{P}}_{h+1,n} = \mathcal{P}_{h+1,n} - \bar{Q}_{h+1,n} = \{\bar{p}_i(s),\ i=0,1,\ldots,h\}$. Consider now the division of every $\bar{p}_i(s)$ by $\omega(s)$ i.e.

$$\bar{p}_i(s) = \bar{t}_i(s)\omega(s) + \bar{r}_i(s), \qquad i=0,1,\ldots,h \tag{5.5a}$$

Then clearly, by selecting $Q'_{h+1,n} = \{\bar{r}_i(s),\ \ i=0,1,\ldots,h\}$ we have that

$$\bar{p}_i(s) - \bar{r}_i(s) = \bar{t}_i(s)\omega(s), \qquad i=0,1,\ldots,h \tag{5.5b}$$

and thus $Q_{h+1,n} = q_i(s) = \bar{q}_i(s) + \bar{r}_i(s),\ i=0,1,\ldots,h$ is a perturbation that has the above property.

■

The above result establishes the existence of perturbations making $\omega(s)$ an exact GCD of the perturbed set and motivates the following definition, which defines $\omega(s)$ as an approximate gcd in an optimal sense.

Definition (5.1): Let $\mathcal{P}_{h+1,n} \in \mathcal{II}(n,p;h+1)$ and $\omega(s) \in \mathbb{R}[s]$ be a given polynomial with $\deg\{\omega(s)\} = r \le p$. Furthermore, let $\Sigma_\omega = \{Q_{h+1,n}\}$ be the set of all (n,p)-order perturbations such that

$$\mathcal{P}'_{h+1,n} = \mathcal{P}_{h+1,n} - Q_{h+1,n} \in \mathcal{II}(n,p;h+1) \tag{5.6}$$

with the property that $\omega(s)$ is a common factor of the elements of $\mathcal{P}'_{h+1,n}$. If $Q^{*}_{h+1,n}$ is the minimal norm element of the set Σ_{ω}, then $\omega(s)$ is referred as an *r-order almost common factor* of $\mathcal{P}_{h+1,n}$, and the norm of $Q^{*}_{h+1,n}$, denoted by $\|Q^{*}\|$ is defined as the *strength* of $\omega(s)$. If $\omega(s)$ is the gcd of

$$\mathcal{P}^{*}_{h+1,n} = \mathcal{P}_{h+1,n} - Q^{*}_{h+1,n} \tag{5.7}$$

then $\omega(s)$ will be called an *r-order almost gcd* of $\mathcal{P}_{h+1,n}$ with strength $\|Q^{*}\|$.

■

The above definition suggests that any polynomial $\omega(s)$ may be considered as an "approximate gcd", as long $\deg\{\omega(s)\} \leq p$. The best choice of "approximate gcd" is an issue that is addressed in the following chapter. In this chapter we consider the problem of determining the minimal norm perturbation and through that the strength of a given $\omega(s)$ selection. This is a distance problem from a specific subvariety of $\Delta_{d}(n,p;h+1)$ which has $\omega(s) \in \mathbb{R}[s]$ as a given gcd. The solution of this problem involves the following important issues:

- Parameterisation of the Σ_{ω} set.
- Definition of an appropriate metric for $Q_{h+1,n}$.
- Solution of an optimization problem to define $Q^{*}_{h+1,n}$.

These problems may be considered within the framework of the resultant representation of $\mathcal{P}_{h+1,n}$ set, which also permits the gcd representation through the factorisation, as we have established in Chapter 4. Note that, the representation of $\mathcal{P}_{h+1,n}$ through the resultant, implies that the degrees of the polynomials are structured by the maximal two values (n,p), $n \geq p$, which define the structure of the resultant $S_{\mathcal{P}}$.

Furthermore the perturbations $Q_{h+1,n}$ and the perturbed sets $\mathcal{P}'_{h+1,n}$ are also structured by the (n,p) pair and thus their corresponding generalized resultants are structured in a compatible way. It is worth pointing that the elements in $Q_{h+1,n}$ have nominal degrees (n,p), whereas the effective values of degrees may be less than those values. The set of all resultants corresponding to $h+1$ polynomials and with maximal nominal degrees (n,p), that is those corresponding to $\mathcal{T}(n,p;h+1)$, will be denoted by $\Psi(n,p;h+1)$. From (5.4) and the compatibility of resultants of $\mathcal{T}(n,p;h+1)$ set we have:

Remark (5.2): If $\mathcal{P}_{h+1,n}$, $Q_{h+1,n}$, $\mathcal{P}'_{h+1,n} \in \mathcal{T}(n,p;h+1)$ are sets of polynomials in (5.6) and $S_{\mathcal{P}}$, S_Q, $\overline{S}_{\mathcal{P}'}$ denote their generalised resultants, then these resultants are elements of $\Psi(n,p;h+1)$ and (5.6) is equivalent to

$$S_{\mathcal{P}'} = S_{\mathcal{P}} - S_Q \tag{5.8}$$

■

The above remark together with the factorisation of resultants described in Chapter 4 leads to the following result establishing the parameterization of perturbations which lead to that $\omega(s)$ becomes exact gcd of the perturbed set.

Theorem (5.1): For $\mathcal{P}_{h+1,n} \in \mathcal{T}(n,p;h+1)$, let $S_{\mathcal{P}} \in \Psi(n,p;h+1)$ be the corresponding generalized resultant and let $\upsilon(s) \in \mathbb{R}[s]$, $\deg\{\upsilon(s)\} = r \le p$. The following properties hold true:

a) Any perturbation set $Q_{h+1,n} \in \mathcal{T}(n,p;h+1)$ that leads to $\mathcal{P}'_{h+1,n} = \mathcal{P}_{h+1,n}$ - $Q_{h+1,n}$, which has $\upsilon(s)$ as common divisor, has a generalized resultant $S_Q \in \Psi(n,p;h+1)$ that is expressed as shown below:

i) If $\upsilon(0) \neq 0$ then

$$S_{\mathcal{Q}} = S_{\mathcal{P}} - \bar{S}^{(r)}_{\mathcal{P}^*}\hat{\Phi}_{\upsilon} = S_{\mathcal{P}} - \left[0_r \mid \bar{S}_{\mathcal{P}^*}\right]\hat{\Phi}_{\upsilon} \tag{5.9}$$

where $\hat{\Phi}_{\upsilon}$ is the $(n+p)\times(n+p)$ Toeplitz representation of $\upsilon(s)$ as defined by (4.38) and $\bar{S}_{\mathcal{P}^*} \in \mathbb{R}^{(p+hn)\times(n+p-r)}$ is the (n,p)-expanded resultant of an arbitrary set of polynomials $\mathcal{P}^* \in \mathcal{T}(n-r, p-r; h+1)$.

ii) If $\upsilon(s)$ has k zeros at $s=0$, then

$$S_{\mathcal{Q}} = S_{\mathcal{P}} - \bar{S}_{\mathcal{P}^*}\Theta_{\upsilon}$$

where $\bar{S}_{\mathcal{P}^*}$ is again the (n,p)-expanded resultant of an arbitrary set of polynomials $\mathcal{P}^* \in \mathcal{T}(n-r, p-r; h+1)$ and Θ_{υ} is the $(n+p-k)\times(n+p)$ representation of $\upsilon(s)$ defined by (4.74).

b) If the parameters of $\bar{S}_{\mathcal{P}^*}$ are constrained such that $\bar{S}_{\mathcal{P}^*}$ has full rank, then $\upsilon(s)$ is a gcd of the perturbed set $\mathcal{P}'_{h+1,n}$.

<u>Proof:</u>

By Proposition (5.3), any arbitrary polynomial $\upsilon(s) \in \mathbb{R}[s]$, $\deg\{\upsilon(s)\} = r \le p$ may be consider as the gcd of some perturbed set of polynomials $\mathcal{P}'_{h+1,n} \in \mathcal{T}(n, p; h+1)$ with some perturbation $\mathcal{Q}_{h+1,n} \in \mathcal{T}(n, p; h+1)$, that is $\mathcal{P}'_{h+1,n} = \mathcal{P}_{h+1,n} - \mathcal{Q}_{h+1,n}$, which implies for the corresponding resultants (Remark 5.2) that

$$S_{\mathcal{P}'} = S_{\mathcal{P}} - S_{\mathcal{Q}} \tag{5.10}$$

Given that $\mathcal{P}'_{h+1,n}$ has $\upsilon(s)$ as divisor then:

i) If $\upsilon(0) \neq 0$, then Theorem (3.1) implies that

$$S_{\mathcal{P}'} = \bar{S}^{(r)}_{\mathcal{P}^*}\hat{\Phi}_{\upsilon} = \left[\mathbf{0}_r \mid \bar{S}_{\mathcal{P}^*}\right]\hat{\Phi}_{\upsilon} \tag{5.11}$$

where $\bar{S}_{\mathcal{P}^*}$ is the (n,p)-expanded resultant of some $\mathcal{P}^* \in \mathcal{J}(n-r,p-r;h+1)$ and $\hat{\Phi}_\upsilon$ is the $(n+p)\times(n+p)$ representation of $\upsilon(s)$. From (5.8) and (5.11) it follows that:

$$S_Q = S_{\mathcal{P}} - S_{\mathcal{P}'} = S_{\mathcal{P}} - \left[\mathbf{0}_r \mid \bar{S}_{\mathcal{P}^*}\right]\hat{\Phi}_\upsilon \tag{5.12}$$

ii) If $\upsilon(s)$ has k zeros at $s=0$, then by Corollary (4.3) we have that

$$S_{\mathcal{P}'} = \bar{S}_{\mathcal{P}^*}\Theta_\upsilon$$

where $\bar{S}_{\mathcal{P}^*}$ is the (n,p)-expanded resultant of some $\mathcal{P}^* \in \mathcal{J}(n-r,p-r;h+1)$ and Θ_υ is the $(n+p-k)\times(n+p)$ Toeplitz representation of $\upsilon(s)$ defined by (4.74) From (5.6) and (5.11) we have that

$$S_Q = S_{\mathcal{P}} - \bar{S}_{\mathcal{P}^*}\Theta_\upsilon \tag{5.13}$$

Clearly in both cases, if $\mathcal{P}^*$ is coprime, i.e. $\bar{S}_{\mathcal{P}^*}$ has full column rank, then the matrix $\bar{S}_{\mathcal{P}^*}$ cannot be further reduced (Theorem (4.5)) and the polynomial $\upsilon(s)$ is a gcd of $\mathcal{P}'_{h+1,n}$. The above holds for every perturbation $Q_{h+1,n}$ thet leads to a perturbed $Q_{h+1,n}$ with $\upsilon(s)$ a divisor and this completes the necessity of proof. The fact that the perturbed polynomials have $\upsilon(s)$ as divisor is obvious.

■

Remark (5.3): The above result provides a parameterisation of all perturbations $Q_{h+1,n} \in \mathcal{J}(n,p;h+1)$ which lead to sets $\mathcal{P}'_{h+1,n}$ having a gcd with degree at least r and divided by the given polynomial $\upsilon(s)$. The set of free parameters is the set of coefficients of the polynomials $\mathcal{P}^*_{h+1,n-r} \in \mathcal{J}(n-r,p-r;h+1)$. For a given selection of free parameters, $\upsilon(s)$ is a divisor of the elements of $\mathcal{P}'_{h+1,n}$ and if the polynomials are generic, then $\upsilon(s)$ is a gcd of $\mathcal{P}'_{h+1,n}$.

■

Having established a parameterisation of the perturbations generating sets with $\upsilon(s)$ common divisor we consider now a metric that can be used for evaluation of strength of "approximate gcd". Given that such a metric has to relate in a direct way to the set of polynomials, the Frobenius norm seems to be an appropriate choice.

Lemma (5.1): If $\mathcal{P}_{h+1,n} \in \mathcal{II}(n,p;h+1)$, then the Frobenius norm of the generalized resultant $S_{\mathcal{P}}$ is given by

$$\left\|S_{\mathcal{P}}\right\|_{\mathrm{F}}^{2} = p\left\|\underline{p}_0\right\|_2^2 + n\sum_{i=1}^{h}\left\|\underline{p}_i\right\|_2^2 \tag{5.13}$$

where $\underline{p}_i$ are the coefficients vectors of the polynomials $p_i(s) \in \mathcal{P}_{h+1,n}$ defined by $p_i(s) = \underline{e}_n(s)^t\,\underline{p}_i$, $\underline{e}_n(s)^t = [s^h,\ldots,s,1]$, $p_0(s) = a(s)$, $p_i(s) = b_i(s)$, $i = 1,2,\ldots,h$.

■

The above result follows readily from the definition of the Frobenius norm and the structure of $S_{\mathcal{P}}$. Using this norm and the parameterisation Theorem (5.1) we can define the strength of a given r-order almost common factor of $\mathcal{P}_{h+1,n}$.

Corollary (5.1): Let $\mathcal{P}_{h+1,n} \in \mathcal{II}(n,p;h+1)$ and $\upsilon(s) \in \mathbb{R}[s]$, $\deg\{\upsilon(s)\} = r \le p$. The polynomial $\upsilon(s)$ is an r-order almost common divisor of $\mathcal{P}_{h+1,n}$ and its strength is defined as a solution of the following minimization problems:

i) If $\upsilon(0) \neq 0$, then its strength is defined by the global minimum of

$$f\left(\mathcal{P},\mathcal{P}^*\right) = \min_{\forall \mathcal{P}^*}\left\|S_{\mathcal{P}} - \left[\mathbf{0}_r \;\middle|\; \overline{S}_{\mathcal{P}^*}\right]\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}} \tag{5.14}$$

ii) If $\upsilon(s)$ has k zeros at $s = 0$, then its strength is defined by the global minimum of

$$f\left(\mathcal{P},\mathcal{P}^*\right)=\min_{\forall\mathcal{P}^*}\left\|S_{\mathcal{P}}-\bar{S}_{\mathcal{P}^*}\Theta_{\upsilon}\right\|_F \tag{5.15}$$

where $\mathcal{P}^*$ takes values from the set $\mathcal{J\!T}(n,p;h+1)$. Furthermore $\upsilon(s)$ is an r-order almost gcd of $\mathcal{P}_{h+1,n}$ if the minimal corresponds to a coprime set $\mathcal{P}^*$ or to full rank $S_{\mathcal{P}^*}$.

■

Example (5.1): We consider the set of two polynomials $\mathcal{P}_{2,2}$,

$$\mathcal{P}_{2,2}=\left\{a_0(s)=(s-1)(s-2)=s^2-3s+2,\ b_0(s)=s-0.99999\right\}$$

We have $n=2,\ p=1$ and

$$S_{\mathcal{P}}=\begin{bmatrix}1 & -3 & 2\\ 1 & -0.99999 & 0\\ 0 & 1 & -0.99999\end{bmatrix}$$

An approximate gcd of the set using ERES method [Mitrouli et al., 1993] is $\upsilon(s)=s-1$. Then

$$\hat{\Phi}=\begin{bmatrix}-1 & 0 & 0\\ 1 & -1 & 0\\ 0 & 1 & -1\end{bmatrix},\quad \bar{S}_{\mathcal{P}^*}=\begin{bmatrix}a & b\\ c & 0\\ 0 & c\end{bmatrix}$$

Solving the minimization problem $f\left(\mathcal{P},\mathcal{P}^*\right)=\min_{\forall\mathcal{P}^*}\left\|S_{\mathcal{P}}-\left[\mathbf{0}_r \mid \bar{S}_{\mathcal{P}^*}\right]\hat{\Phi}_{\upsilon}\right\|_F$ with MAPLE we get:

$$a=1,\ b=-2,\ c=0.999995,\ f\left(\mathcal{P},\mathcal{P}^*\right)=0.4\cdot10^{-11}.$$

The above demonstrates that the approximation is very good

■

The computation of the strength of approximation for any given $\upsilon(s)$ is considered next.

5.3. THE STRUCTURED SINGULAR VALUE APPROACH FOR THE APPROXIMATE GCD EVALUATION

A recent result [Halikias et al., 2003] shows that the optimisation problem of finding the smallest perturbation in the coefficients of two coprime polynomials, so that they have a common root, can be solved in the context of structured singular values and $H-\infty$ control applications. The calculation of the minimal perturbation is shown next to correspond to the distance of a structured matrix from singularity or equivalently to the calculation of the *structured singular value* of a matrix [Young et al., 1996], [Zhou et al., 1998].

Definition (5.2): [Young et al., 1996]: Let $M \in \mathbb{R}^{m\times m}$ and define the "structured" set:

$$\mathcal{D} = \left\{ \operatorname{diag}\left(\delta_1 I_{r_1}, \delta_2 I_{r_2}, \ldots, \delta_s I_{r_s}\right) : \delta_i \in \mathbb{R},\ i = 1, 2, \ldots, s \right\} \tag{5.17}$$

where the r_i are positive integers such that $\sum_{i=1}^{s} r_i = m$. The structured singular value of M (relative to "structure" $\mathcal{D}$) is defined as:

$$\mu_{\mathcal{D}}(M) = \frac{1}{\min\left\{ \|\Delta\| : \Delta \in \mathcal{D},\ \det(I_m - M\Delta) = 0 \right\}} \tag{5.18}$$

unless no $\Delta \in \mathcal{D}$ makes $I_n - M\Delta$ singular, in which case $\mu_{\mathcal{D}}(M) = 0$.

■

Provided that $\mu_{\mathcal{D}}(M) \neq 0$, the problem of the calculation of the structured singular value, is to find a $\Delta \in \mathcal{D}$ of minimal norm such that $\det(I_m - M\Delta) = 0$. We will show that this problem is equivalent to the problem of the minimal perturbation for an approximate factor.

Theorem (5.2) [Halikias et al., 2003]: Let us consider the set of two polynomials

$$\mathcal{P} = \left\{ a(s) = s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0,\ b(s) = s^p + b_{p-1}s^{p-1} + \ldots + b_1 s + b_0,\ n \geq p \right\} \tag{5.19a}$$

and denote by $S_{\mathcal{P}}$ the corresponding resultant of the set and assume that $S_{\mathcal{P}}$ is non singular. The evaluation (in magnitude) of the smallest real perturbation in the coefficients so that the perturbed polynomials have a common root is equivalent to the finding of the structured singular value of the matrix M where:

$$M = -Z \cdot S_{\mathcal{P}}^{-1} \cdot W \tag{5.19b}$$

where

$$W = \left(\begin{array}{cccc|cccc} I_p & I_p & \cdots & I_p & \mathbf{0}_{p,n} & \mathbf{0}_{p,n} & \cdots & \mathbf{0}_{p,n} \\ \hline \mathbf{0}_{n,p} & \mathbf{0}_{n,p} & \cdots & \mathbf{0}_{n,p} & I_n & I_n & \cdots & I_n \end{array} \right) \in \mathbb{R}^{(n+p)\times(2np)} \tag{5.19c}$$

and

$$Z = \left(\begin{array}{cccc|cccc} \left(Z_{p,n}^{0}\right)' & \left(Z_{p,n}^{1}\right)' & \cdots & \left(Z_{p,n}^{p-1}\right)' & \left(Z_{n,p}^{0}\right)' & \left(Z_{n,p}^{1}\right)' & \cdots & \left(Z_{n,p}^{n-1}\right)' \end{array} \right)' \in \mathbb{R}^{(2np)\times(n+p)} \tag{5.19d}$$

in which

$$Z_{p,n}^{k} = \left(\mathbf{0}_{n.k+1} \quad I_n \quad \mathbf{0}_{n,p-k-1} \right),\quad k = 0,1,\ldots,n-1 \tag{5.19e}$$

Proof:

Since $a(s)$, $b(s)$ are coprime, their Sylvester matrix $S_{\mathcal{P}}$ is non-singular. The perturbed polynomials will have the form:

$$\tilde{a}(s) = s^n + \left(a_{n-1} + \delta_{n-1}\right)s^{n-1} + \ldots + \left(a_1 + \delta_1\right)s + \left(a_0 + \delta_0\right) \tag{5.20a}$$

$$\tilde{b}(s) = s^p + \left(b_{p-1} + \varepsilon_{n-1}\right)s^{n-1} + \ldots + \left(b_1 + \varepsilon_1\right)s + \left(b_0 + \varepsilon_0\right) \tag{5.20b}$$

Let also denote

$$\gamma = \max\left\{ \left|\delta_{n-1}\right|\left|\delta_{n-2}\right|,\ldots,\left|\delta_0\right|\left|\varepsilon_{p-1}\right|\left|\varepsilon_{p-2}\right|,\ldots,\left|\varepsilon_0\right| \right\} = \left\|\Delta\right\| \tag{5.20c}$$

and the corresponding Sylvester matrix can be decomposed as $S_{\mathcal{P}'} = S_{\mathcal{P}} + E$ where E denotes the perturbation matrix:

$$E=\begin{bmatrix} 0 & \delta_{n-1} & \delta_{n-2} & & \delta_0 & 0 & \cdots & 0 \\ 0 & 0 & \delta_{n-1} & & \delta_1 & \delta_0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \ddots & \\ 0 & \cdots & & 0 & \delta_{n-1} & \cdots & \delta_1 & \delta_0 \\ 0 & \varepsilon_{p-1} & \varepsilon_{p-2} & & \varepsilon_0 & 0 & \cdots & 0 \\ 0 & 0 & \varepsilon_{p-1} & & \varepsilon_1 & \varepsilon_0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & & \ddots & \\ 0 & \cdots & & 0 & \varepsilon_{p-1} & \cdots & \varepsilon_1 & \varepsilon_0 \end{bmatrix} \tag{5.21}$$

Matrix E can be factorised as $E = W \cdot \Delta \cdot Z$ where $W,\ Z$ are defined in (5.19b) and (5.19c) respectively and

$$\Delta = \operatorname{diag}\left(\delta_{n-1}I_p, \delta_{n-2}I_p, \ldots, \delta_0 I_p, \varepsilon_{p-1}I_n, \varepsilon_{p-2}I_n, \ldots, \varepsilon_0 I_n\right) \tag{5.22}$$

Clearly $\Delta \in \mathcal{D}$, i.e. it has a block-diagonal structure, $m = n + p$, $r_i = p$ for $1 \le i \le n$ and $r_i = n$ for $n+1 \le i \le n+p$. Since the resultant matrix $S_{\mathcal{P}}$ loses rank if and only if there is a common factor between $\tilde{a}(s)$ and $\tilde{b}(s)$. Then the problem of the minimum strength is equivalent to:

$$\min\|\Delta\| \text{ such that } \det\left(S_{\mathcal{P}} + W \cdot \Delta \cdot Z\right) = 0 \text{ and } \Delta \in \mathcal{D} \tag{5.23}$$

Using the matrix property that $\det(I + BC) = \det(I + CB)$ for any matrices B, C of appropriate dimensions, and the fact that the resultant is non-singular, we have that:

$$\det\left(S_{\mathcal{P}} + W \cdot \Delta \cdot Z\right) = 0 \Leftrightarrow \det\left(I + Z \cdot S_{\mathcal{P}}^{-1} \cdot W \cdot \Delta\right) \Leftrightarrow \det\left(I - M\Delta\right) = 0 \tag{5.24}$$

which establishes the equivalence. The minimum strength is then

$$\min\gamma = \mu_{\mathcal{D}}^{-1}(M) \tag{5.25}$$

■

Example 5.2 Let $a(s) = s^3 + a_2 s^2 + a_1 s + a_0$ and $b(s) = s^2 + b_1 s + b_0$. The Sylvester matrix of the perturbed set is:

$$S_{\mathcal{P}^*} = \begin{bmatrix} 1 & a_2+\delta_2 & a_1+\delta_1 & a_0+\delta_0 & 0 \\ 0 & 1 & a_2+\delta_2 & a_1+\delta_1 & a_0+\delta_0 \\ 1 & b_1+\varepsilon_1 & b_0+\varepsilon_0 & 0 & 0 \\ 0 & 1 & b_1+\varepsilon_1 & b_0+\varepsilon_0 & 0 \\ 0 & 0 & 1 & b_1+\varepsilon_1 & b_0+\varepsilon_0 \end{bmatrix}$$

which can be written as:

$$S_{\mathcal{P}^*} = S_{\mathcal{P}} + E = \begin{bmatrix} 1 & a_2 & a_1 & a_0 & 0 \\ 0 & 1 & a_2 & a_{11} & a_0 \\ 1 & b_1 & b_0 & 0 & 0 \\ 0 & 1 & b_1 & b_0 & 0 \\ 0 & 0 & 1 & b_1 & b_0 \end{bmatrix} + \begin{bmatrix} 0 & \delta_2 & \delta_1 & \delta_0 & 0 \\ 0 & 0 & \delta_2 & \delta_1 & \delta_0 \\ 0 & \varepsilon_1 & \varepsilon_0 & 0 & 0 \\ 0 & 0 & \varepsilon_1 & \varepsilon_0 & 0 \\ 0 & 0 & 0 & \varepsilon_1 & \varepsilon_0 \end{bmatrix}$$

The "perturbation" matrix E can be factored as:

$$E = \left[\begin{array}{cc|cc|cc|ccc|ccc} 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \cdot \begin{bmatrix} \delta_2 I_2 & 0 & 0 & 0 & 0 \\ 0 & \delta_2 I_2 & 0 & 0 & 0 \\ 0 & 0 & \delta_2 I_2 & 0 & 0 \\ 0 & 0 & 0 & \delta_3 I_3 & 0 \\ 0 & 0 & 0 & 0 & \delta_3 I_3 \end{bmatrix} \cdot \left[\begin{array}{ccccc} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

which is of the required form $E = W\Delta Z$ with $\Delta \in \mathcal{D}$. The minimum coefficient perturbation is the inverse of the structured singular value of the matrix:

$$M=-\left[\begin{array}{ccccc}0&1&0&0&0\\0&0&1&0&0\\\hline 0&0&1&0&0\\0&0&0&1&0\\\hline 0&0&0&1&0\\0&0&0&0&1\\\hline 0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\\hline 0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{array}\right]\cdot\begin{bmatrix}1&a_2&a_1&a_0&0\\0&1&a_2&a_1&a_0\\1&b_1&b_0&0&0\\0&1&b_1&b_0&0\\0&0&1&b_1&b_0\end{bmatrix}^{-1}\cdot\left[\begin{array}{cc|cc|cc|ccc|ccc}1&0&1&0&1&0&0&0&0&0&0&0\\0&1&0&1&0&1&0&0&0&0&0&0\\\hline 0&0&0&0&0&0&1&0&0&1&0&0\\0&0&0&0&0&0&0&1&0&0&1&0\\0&0&0&0&0&0&0&0&1&0&0&1\end{array}\right]$$

The structured singular value of M can be computed numerically using efficient existing techniques [Zhou et al., 1998], [Young et al., 1996].

■

The technique described above may be used to develop an algorithm to calculate the “best” approximate factor of the two polynomials after finding the perturbation matrix that corresponds to the structured singular value. A sequence of iterations consists of approximate factor calculation and extraction of the factor, repeated until a termination condition is met. The extracted factors will be then the factors of an approximate gcd of order equal to the number of iterations. A more elegant non-iterative procedure that will estimate the approximate gcd requires rank constraints in order to generalise the notion of the structured singular value:

<u>Definition (5.3)</u> [Halikias et al., 2003]: Let $M \in \mathbb{R}^{m\times m}$ and define the “structured” set:

$$\mathcal{D}=\left\{\operatorname{diag}\left(\delta_1 I_{r_1},\delta_2 I_{r_2},\ldots,\delta_s I_{r_s}\right) : \delta_i\in\mathbb{R},\ i=1,2,\ldots,s\right\} \tag{5.26}$$

where the r_i are positive integers such that $\sum_{i=1}^{s} r_i = m$. The *generalised structured singular value* of M (relative to “structure” $\mathcal{D}$) and for a non-negative integer k is defined as:

$$\hat{\mu}_{\mathcal{D},k}(M) = \frac{1}{\min\{\|\Delta\| : \Delta \in \mathcal{D},\ \operatorname{null}(I_m - M\Delta) > k\}} \tag{5.27}$$

unless there does not exist $\Delta \in \mathcal{D}$ such that $\operatorname{null}(I_m - M\Delta) > k$, in which case $\hat{\mu}_{\mathcal{D},k}(M) = 0$.

■

It follows from the definition that $\hat{\mu}_{\mathcal{D},0}(M) = \mu_{\mathcal{D}}(M)$ and that $\hat{\mu}_{\mathcal{D},k}(M) \geq \hat{\mu}_{\mathcal{D},k+1}(M)$ for all $k \geq 0$. The following theorem establishes the framework for an approximate gcd algorithm:

<u>Theorem (5.3)</u> [Halikias et al., 2003]: Let $a(s)$, $b(s)$ be two coprime polynomials as defined in (5.18) and let $\tilde{a}(s)$, $\tilde{b}(s)$ be the perturbed polynomials as defined in (5.20a) and (5.20b) respectively and set

$$\gamma = \max\{|\delta_{n-1}||\delta_{n-2}|,\ldots,|\delta_0||\varepsilon_{p-1}||\varepsilon_{p-2}|,\ldots,|\varepsilon_0|\} = \|\Delta\| \tag{5.28}$$

where $\{\delta_i\}$, $\{\varepsilon_i\}$ denote the perturbed coefficients of $\tilde{a}(s)$, $\tilde{b}(s)$ respectively. Further, let $\gamma^*(k)$ denote the minimum value of γ such that $a(s)$, $b(s)$ have a common factor $\varphi(s)$ of degree $\deg\{\varphi(s)\} > k$ ($k = 0,1,2,\ldots,n-1$). Then

$$\gamma^*(k) = \frac{1}{\mu_{\mathcal{D},k}(M)} \tag{5.29}$$

where $\mu_{\mathcal{D},k}(M)$ denotes the generalised structured singular value of $M = -Z \cdot S_{\varphi}^{-1} \cdot W$ with respect to "structure" $\mathcal{D}$ defined in (5.16) and Z, W the matrices defined in (5.19c), (5.19b) respectively.

■

The proof of the above theorem is identical to the proof of Theorem (5.2), on noting that the transformations in (5.24) do not affect the nullity of the corresponding matrices.

5.4. DISCUSSION

Based on the factorisation property of the resultant, a new characterisation has been given for the approximate gcd. The new definitions are based on the "distance" between the Sylvester matrix of the initial set and the resultant of the perturbed set we have constructed and it is a distance problem that will be subsequently considered in detail.

An interesting result for the case of two polynomials links the problem to the equivalent of the structured singular values of a matrix. This equivalence establishes a solution of the optimisation for the "best" common factor of a specific order. An open issue is related with possible generalisation of this theory in order to cover the case of many polynomials which is still open.

The study of the optimisation problem that leads to the evaluation of the strength of a given approximate gcd and the computation of the optimal gcd of a given order is considered in the following section. Although it seems that such an optimisation is rather hard, it will be proved that it may be reformulated in a way that leads to a standard solution.

Chapter 6:

EVALUATION OF THE BEST APPROXIMATE GCD OF A POLYNOMIAL SET USING RESULTANT SETS

6.1. INTRODUCTION

In the previous chapter, we have defined the notion of approximate GCD as a distance problem and we have examined the way we can qualify the quality of the approximate gcd by defining the strength of the approximation as the value of the distance between the set and the selected gcd on the given d-gcd variety. So far we have considered approximate solutions derived from different methods. For a given set of polynomials a question that naturally arises is the definition of the best approximate gcd of a given degree. It has be shown in the previous chapter that almost every polynomial may be considered as an approximate gcd. The key issues addressed here are:

i) Define the "best" approximate gcd of a given degree.

ii) Qualify the "best" selection of degree for the approximate gcd

The theoretic characterisation of the "best" approximate gcd is based on further investigation of the optimisation results of Chapter 5. This problem is equivalent to defining the distance of the given set from the d-gcd variety. The algorithm focuses on the case where the approximate gcd does not have a root at zero. Otherwise we can extract first the s^k factor and then proceed with the reduced set.

An analytic investigation of the optimisation problem established in Chapter 5 is the key to the definition of the approximate gcd and its evaluation. The degree of the gcd is considered fixed or can be determined with numerical criteria such as the numerical rank [Foster, 1986], [Mitrouli et al., 1991] which will be briefly described. An algorithm for the evaluation of the approximate gcd, based again on the resultant sets, will be introduced in the present chapter and will be demonstrated by examples.

6.2. EVALUATION OF DEGREE OF THE APPROXIMATE GCD

The existence of a non trivial gcd is determined by the rank properties of the Sylvester Resultant. In fact, Theorem (4.5) states that the right nullity of the Sylvester Resultant defines the degree of the gcd. In numerical terms, the right nullity of resultant is determined by the number of zero singular values. In defining the appropriate degree of the approximate gcd we need to use the notion of numerical rank and thus effective right nullity of the Sylvester Resultant. The definition of the effective right nullity is based on the notion of the numerical rank of the Sylvester matrix, which in turn is based on the following property of the *Singular Value Decomposition* [Horn et al., 1985]:

Proposition (6.1): Let us consider the matrices $A \in \mathbb{R}^{m \times n}$, $U \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$ so that $U \Sigma V$ be the Singular Value Decomposition of A, i.e. $U \Sigma V = A$. Then the matrices Σ, A have the same rank, $\rho(\Sigma) = \rho(A)$.

■

Considering the structure of Σ, with the singular values on the diagonal and zeros everywhere else, the rank of a matrix and its Nullity are linked with the number of its non-zero singular values. Thus, in the case of very "small" singular values we can introduce the approximate version of the concept of the Nullity:

Definition (6.1) [Foster, 1986]: The numerical ε-rank of a matrix $A \in \mathbb{R}^{m \times n}$ is defined by

$$\rho_{\varepsilon}(A) = \min_{B} \left\{ \rho(B) : \ \|A - B\| \leq \varepsilon, \ \varepsilon > 0 \right\}$$

and the numerical ε-right nullity

$$\mathcal{N}_{\varepsilon}(A) = \max_{B} \left\{ \mathcal{N}(B) : \ \|A - B\| \leq \varepsilon, \ \varepsilon > 0 \right\} = n - \rho_{\varepsilon}(A)$$

■

The generalisations of the notions of rank and of matrix nullity into their numerical versions provide the criterion for the degree of the approximate gcd. In the case of the exact gcd (generalised resultant theorem) the degree of the gcd is equal to the right nullity of the resultant. When we investigate the approximate gcd, we seek a polynomial of degree equal to the numerical ε-nullity. The ε-nullity can be evaluated with the use of the following theorem:

Theorem (6.1): [Foster, 1986]: For a matrix $A \in \mathbb{R}^{m\times n}$, $\rho_{\varepsilon}(A) =$ number of singular values of A that are $> \varepsilon$

■

The above theorem determines the criterion for the degree of the Optimal GCD. The degree is chosen to be equal to the numerical loss of rank of the resultant. Th
e determination of the desirable degree of the approximate gcd is based on the selection of a given accuracy ε. Before we define the ε-accuracy approximate gcd degree we note:

Remark (6.1): The singular values of a matrix A are affected by multiplication of A by scaling of A on the left or right by a diagonal nonsingular matrix.

■

The above implies that defining the given accuracy ε approximate gcd requires some standardization for the set of polynomials and thus for the corresponding resultant. This can be done by assuming that the coefficient vector has unit length, or by considering sets of polynomials which are monic. In the following we will assume the standardization based on polynomials being monic.

Definition (6.2): Let $\mathcal{P}_{h+1,n}$ be a set of monic polynomials with maximum degrees (n,p), $S_{\mathcal{P}}$ being the $(p+hn)\times(n+p)$ generalized resultant and let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n+p}$, be the set of singular values of $S_{\mathcal{P}}$. For a given accuracy ε, $\varepsilon \leq \sigma_1$ we define

i) If $\sigma_{n+p} > \varepsilon$, then the set $\mathcal{P}_{h+1,n}$ will be said to be ε-*coprime*.

ii) If $\sigma_1 \geq \cdots \geq \sigma_k > \varepsilon \geq \sigma_{k+1} \geq \cdots \geq \sigma_{n+p}$, then we say that the set $\mathcal{P}_{h+1,n}$ has an approximate gcd of degree $\delta = n + p - k$

■

In the following we will assume the above definition of the degree of the approximate gcd and we will consider its optimal evaluation.

6.3. SOLUTION AND COMPUTATION OF THE MINIMAL DISTANCE PROBLEM

In this section we consider the computation of the distance of a given set of polynomials from the d-gcd variety. We will solve this problem by developing the investigation on the strength of the approximation by considering the optimisation problem in some detail. Let us consider again the case of a given approximate gcd $\upsilon(s)$, $\deg \upsilon(s) = k$ of a set $\mathcal{P}_{h+1,n} \in \mathcal{T}(n,p;h+1)$ where first we assume that we do not approximate or have exact roots at 0. Let us also denote by $S_{\mathcal{P}} \in \Psi(n,p;h+1)$ the corresponding generalized resultant. If we denote by $Q_{h+1,n} \in \mathcal{T}(n,p;h+1)$ the perturbation set that leads to $\mathcal{P}'_{h+1,n} = \mathcal{P}_{h+1,n} - Q_{h+1,n}$, which has $\upsilon(s)$ as common divisor, then this set has a generalized resultant $S_Q \in \Psi(n,p;h+1)$, expressed as in (5.7) which leads us to the optimization problem of (5.14) defined by:

$$S_Q = S_{\mathcal{P}} - \bar{S}^{(k)}_{\mathcal{P}^*}\hat{\Phi}_\upsilon = S_{\mathcal{P}} - \left[0_k \mid \bar{S}_{\mathcal{P}^*}\right]\hat{\Phi}_\upsilon \tag{5.7}$$

$$f\left(\mathcal{P},\mathcal{P}^*\right) = \min_{\forall \mathcal{P}^*}\left\| S_{\mathcal{P}} - \left[\mathbf{0}_k \mid \bar{S}_{\mathcal{P}^*}\right]\hat{\Phi}_\upsilon \right\|_{\mathrm{F}} \tag{5.14}$$

The following property of the Frobenius norm simplifies the optimisation problem and motivates the use of this norm for the study of the optimization problem.

Lemma (6.1): [Horn et al., 1]: The Frobenius norm has the following property: $\|A \cdot B\|_{\mathrm{F}} = \|A\|_{\mathrm{F}} \cdot \|B\|_{\mathrm{F}}$

■

Using Lemma 6.1 in (5.14) we obtain:

$$f\left(\mathcal{P},\mathcal{P}^*\right) = \min_{\forall \mathcal{P}^*}\left\{ \left\| S_{\mathcal{P}}\Phi_\upsilon - \left[\mathbf{0}_k \mid \bar{S}_{\mathcal{P}^*}\right] \right\|_{\mathrm{F}} \cdot \left\|\hat{\Phi}_\upsilon\right\|_{\mathrm{F}} \right\} \tag{6.1a}$$

and if we denote by $\tilde{S}_{\mathcal{P}} \triangleq S_{\mathcal{P}}\Phi_{\upsilon}$, then (6.1) can be expressed as

$$f\left(\mathcal{P},\mathcal{P}^{*}\right)=\min_{\forall\mathcal{P}^{*}}\left\{\left\|\tilde{S}_{\mathcal{P}}-\left[\mathbf{0}_{k} \mid \bar{S}_{\mathcal{P}^{*}}\right]\right\|_{\mathrm{F}}\cdot\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}}\right\} \tag{6.1b}$$

where $\hat{\Phi}_{\upsilon}$ is described by a Toeplitz structure from $\upsilon(s)$ and $\Phi_{\upsilon},\mathcal{P}$ have forms given in the (6.2) and (6.3) respectively below:

$$\Phi_{\upsilon}=\begin{bmatrix} y_0 & 0 & \cdots & & & & & 0 \\ y_1 & y_0 & \ddots & & & & & \vdots \\ y_2 & y_1 & \ddots & & & & & \\ \vdots & \vdots & \ddots & y_0 & 0 & & & \\ & & & y_1 & y_0 & & & \\ & & & \vdots & \vdots & \ddots & & \\ y_{n+p-2} & y_{n+p-3} & \cdots & y_{n+p-j-2} & y_{n+p-j-3} & \cdots & y_0 & 0 \\ y_{n+p-1} & y_{n+p-2} & \cdots & y_{n+p-j-1} & y_{n+p-j-2} & \cdots & y_1 & y_0 \end{bmatrix} \tag{6.2}$$

$$S_{\mathcal{P}}=\left[\begin{array}{cccccccc} b_{0,n} & b_{0,n-1} & \cdots & & b_{0,0} & 0 & \cdots & 0 \\ \vdots & \ddots & & & & \ddots & & \vdots \\ 0 & \cdots & b_{0,n} & \cdots & \cdots & b_{0,1} & b_{0,0} & 0 \\ 0 & \cdots & 0 & b_{0,n} & \cdots & \cdots & b_{0,1} & b_{0,0} \\ \hline b_{1,p} & b_{1,p-1} & & b_{1,0} & 0 & \cdots & & 0 \\ \vdots & & & & & & & \vdots \\ 0 & \cdots & 0 & b_{1,p} & b_{1,p-1} & \cdots & b_{1,0} & 0 \\ 0 & & \cdots & 0 & b_{1,p-1} & \cdots & b_{1,1} & b_{1,0} \\ \hline & & & & & & & \\ & & \vdots & & \vdots & & & \\ & & & & & & & \\ \hline b_{h,p} & b_{h,p-1} & & b_{h,0} & 0 & \cdots & & 0 \\ \vdots & & & & & & & \vdots \\ 0 & \cdots & 0 & b_{h,p} & b_{h,p-1} & \cdots & b_{h,0} & 0 \\ 0 & & \cdots & 0 & b_{h,p} & \cdots & b_{h,1} & b_{h,0} \end{array}\right] \tag{6.3}$$

and thus the matrix $\tilde{S}_{\varphi} = S_{\varphi}\Phi_{\upsilon}$ in (6.1) will have the form:

$$\tilde{S}_{\varphi} = \left[\begin{array}{ccccccccc}
\bar{c}_{0,n} & \bar{c}_{0,n-1} & \cdots & & \bar{c}_{0,0} & 0 & 0 & \cdots & 0 \\
\bar{e}_{0,1} & \bar{c}_{0,n} & \bar{c}_{0,n-1} & \cdots & & \bar{c}_{0,0} & 0 & & \vdots \\
\vdots & \ddots & \ddots & & & & \ddots & \ddots & \\
\bar{e}_{0,p-2} & \cdots & \bar{e}_{0,1} & \bar{c}_{0,n} & \bar{c}_{0,n-1} & & & \bar{c}_{0,0} & 0 \\
\bar{e}_{0,p-1} & \bar{e}_{0,p-3} & \cdots & \bar{e}_{0,1} & \bar{c}_{0,n} & \bar{c}_{0,n-1} & \cdots & & \bar{c}_{0,0} \\
\hdashline
 & & & & \vdots & \vdots & & & \\
\hdashline
\bar{c}_{i,p} & \bar{c}_{i,p-1} & \cdots & \bar{c}_{i,0} & 0 & \cdots & & & 0 \\
\bar{e}_{i,1} & \bar{c}_{i,p} & \bar{c}_{i,p-1} & \cdots & \bar{c}_{i,0} & 0 & & & 0 \\
\bar{e}_{i,2} & \ddots & \bar{c}_{i,p} & \bar{c}_{i,p-1} & \cdots & \bar{c}_{i,0} & 0 & & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & & \ddots & \ddots & \vdots \\
\bar{e}_{i,n-2} & & \bar{e}_{i,2} & \bar{e}_{i,1} & \bar{c}_{i,p} & \bar{c}_{i,p-1} & & \bar{c}_{i,0} & 0 \\
\bar{e}_{i,n-1} & \bar{e}_{i,n-2} & & \bar{e}_{i,2} & \bar{e}_{i,1} & \bar{c}_{i,p} & \bar{c}_{i,p-1} & \cdots & \bar{c}_{i,0} \\
\hdashline
 & & & & \vdots & \vdots & & & \\
\hdashline
\bar{c}_{h,p} & \bar{c}_{h,p-1} & \cdots & \bar{c}_{h,0} & 0 & \cdots & & & 0 \\
\bar{e}_{h,1} & \bar{c}_{h,p} & \bar{c}_{h,p-1} & & \bar{c}_{h,0} & 0 & & & 0 \\
\bar{e}_{h,2} & \bar{e}_{h,1} & \bar{c}_{h,p} & \bar{c}_{h,p-1} & & \bar{c}_{h,0} & 0 & & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & & \ddots & \ddots & \vdots \\
\bar{e}_{h,n-2} & & \bar{e}_{h,2} & \bar{e}_{n,1} & \bar{c}_{h,p} & \bar{c}_{h,p-1} & \cdots & \bar{c}_{h,0} & 0 \\
\bar{e}_{h,n-1} & \bar{e}_{h,n-2} & \cdots & \bar{e}_{h,2} & \bar{e}_{h,1} & \bar{c}_{h,p} & \bar{c}_{h,p-1} & \cdots & \bar{c}_{h,0}
\end{array}\right] \tag{6.4a}$$

where

$$\bar{e}_{0,\mu} = \sum_{j=0}^{n} b_{0,n-j} y_{\mu+j},\ \mu = 1,\ldots,n-1, \qquad \bar{e}_{i,\nu} = \sum_{j=0}^{p} b_{i,p-j} y_{\nu+j},\ i = 1,\ldots h,\ \nu = 1,\ldots,n-1$$

$$\bar{c}_{0,n-\theta} = \sum_{j=\theta}^{n} b_{0,n-j} y_{j-\theta},\ \theta = 0,\ldots,n, \qquad \bar{c}_{i,p-\sigma} = \sum_{j=\sigma}^{p} b_{i,p-j} y_{j-\sigma},\ i = 1,\ldots h,\ \sigma = 0,\ldots,p \tag{6.4b}$$

We may split $\tilde{S}_{\varphi}$ as follows:

$$\tilde{S}_{\varphi} = \tilde{S}'_{\varphi} + \tilde{S}''_{\varphi} \tag{6.5}$$

such that $\tilde{S}''_{\varphi} = \left[\mathbf{0}_k \mid \hat{S}^{(2)}_{\varphi}\right]$ has the same structure with the reduced resultant $\left[\mathbf{0}_r \mid \bar{S}_{\varphi^*}\right]$ in (5.14), corresponding to the perturbed set. Thus (6.1) takes the form

$$f\left(\mathcal{P},\mathcal{P}^{*}\right)=\min_{\forall\mathcal{P}^{*}}\left\{\left\|\left(\tilde{S}'_{\mathcal{P}}+\tilde{S}''_{\mathcal{P}}\right)-\left[\mathbf{0}_{r}\mid\bar{S}_{\mathcal{P}^{*}}\right]\right\|_{\mathrm{F}}\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}}\right\} \tag{6.6a}$$

$$f\left(\mathcal{P},\mathcal{P}^{*}\right)=\min_{\forall\mathcal{P}^{*}}\left\{\left\|\tilde{S}'_{\mathcal{P}}+\left[\mathbf{0}_{r}\mid\hat{S}^{(2)}_{\mathcal{P}}-\bar{S}_{\mathcal{P}^{*}}\right]\right\|\cdot\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}}\right\} \tag{6.6b}$$

Using the standard norm inequalities we have:

$$f\left(\mathcal{P},\mathcal{P}^{*}\right)\leq\min_{\forall\mathcal{P}^{*}}\left\{\left\|\tilde{S}'_{\mathcal{P}}\right\|_{\mathrm{F}}\cdot\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}}\right\}+\min_{\forall\mathcal{P}^{*}}\left\{\left\|\left[\mathbf{0}_{r}\mid\hat{S}^{(2)}_{\mathcal{P}}-\bar{S}_{\mathcal{P}^{*}}\right]\right\|_{\mathrm{F}}\cdot\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}}\right\} \tag{6.7}$$

and this splits the optimisation problem into two independent parts involving different set of free variables. In fact, the first part depends only on the selection of the approximate gcd $\upsilon(s)$ (expressed by the parameters in $\hat{\Phi}_{\upsilon}$, whereas the second can get the absolute minimum value zero with appropriate selection of the parameters in the Toeplitz parameterization of perturbations. This natural splitting of the optimization problem provides the means for reducing the original problem and lead to explicit solutions for the optimal approximation based on functions defined on the original set of polynomials. Note that $\tilde{S}'_{\mathcal{P}}$ and $\tilde{S}''_{\mathcal{P}}$ matrices preserve the Toeplitz structure of the blocks, i.e.

$$\begin{array}{cc}
 & \longleftarrow\ k\ \longrightarrow \\
\tilde{S}'_{\mathcal{P}}= & \left[\begin{array}{cccccccccccc}
\bar{c}_{0,n} & \cdots & \bar{c}_{0,n-k+1} & 0 & \cdots & & & & & 0 & \cdots & 0\\
\bar{e}_{0,1} & \bar{c}_{0.n} & \cdots & \bar{c}_{0,n-k+1} & 0 & & & & & & & \\
\bar{e}_{0,2} & \bar{e}_{0,1} & \ddots & & \ddots & & \ddots & & & & & \\
 & \ddots & \ddots & \bar{c}_{0,n} & \cdots & & \bar{c}_{0,n-k+1} & & 0 & & \cdots & 0\\
\bar{e}_{0,p-1} & & \bar{e}_{1,2} & \bar{e}_{0,1} & \bar{c}_{0,n} & & \cdots & & \bar{c}_{0,n-k+1} & 0 & \cdots & 0\\
\hline
 & & \vdots & & & & & \vdots & & & & \\
\hline
\bar{c}_{i,p} & \cdots & \bar{c}_{i,p-k+1} & 0 & \cdots & & & & & 0 & \cdots & 0\\
\bar{e}_{i,1} & \bar{c}_{i,p} & & \bar{c}_{i,p-k+1} & 0 & & & & & & & \\
\bar{e}_{i,2} & \bar{e}_{i,1} & \bar{c}_{i,p} & & \bar{c}_{i,p-k+1} & & \ddots & & & & & \\
\bar{e}_{i,3} & \bar{e}_{i,2} & \ddots & \ddots & & & \ddots & & & & & \\
\vdots & & \ddots & \bar{e}_{i,1} & \bar{c}_{i,p} & & \cdots & \bar{c}_{i,p-k+1} & 0 & & \cdots & 0\\
\bar{e}_{i,n-1} & \cdots & \bar{e}_{i,3} & \bar{e}_{i,2} & \bar{e}_{i,1} & \bar{c}_{i,p} & \cdots & & \bar{c}_{i,p-k+1} & 0 & \cdots & 0\\
\hline
 & & \vdots & & & & & \vdots & & & & \\
\hline
\bar{c}_{h,p} & \cdots & \bar{c}_{h,p-k+1} & 0 & \cdots & & 0 & & \cdots & 0 & \cdots & 0\\
\bar{e}_{h.0} & \bar{c}_{h.p} & & \bar{c}_{i,p-k+1} & 0 & & & & & & & \\
\bar{e}_{h,1} & \bar{e}_{h,0} & \bar{c}_{h,p} & \cdots & \bar{c}_{i,p-k+1} & & \ddots & & & & & \\
\vdots & \ddots & & \ddots & & & \ddots & & & & & \\
\bar{e}_{h,n-3} & \cdots & \bar{e}_{h,1} & \bar{e}_{h,0} & \bar{c}_{h,p} & \cdots & \bar{c}_{i,p-k+1} & & 0 & & & \\
\bar{e}_{h,n-2} & \cdots & \bar{e}_{h,2} & \bar{e}_{h,1} & \bar{e}_{h,0} & \bar{c}_{h,p} & \bar{c}_{h,p} & \cdots & \bar{c}_{i,p-k+1} & 0 & \cdots & 0
\end{array}\right]
\end{array} \tag{6.8a}$$

and thus $\tilde{S}''_{\mathcal{P}}$ is defined as

$$
\begin{array}{cc}
 & \leftarrow \; k \; \rightarrow \\
\tilde{S}''_{\mathcal{P}} = & \left[\begin{array}{ccc|cccccccc}
0 & \cdots & 0 & \bar{c}_{0,n-k} & \bar{c}_{0,n-k-1} & \cdots & \bar{c}_{0,0} & 0 & \cdots & & 0 \\
0 & \cdots & 0 & 0 & \bar{c}_{0,n-k} & \bar{c}_{0,n-k-1} & \cdots & \bar{c}_{0,0} & 0 & & 0 \\
\vdots & & \vdots & & \ddots & \ddots & \ddots & & & \ddots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 & \bar{c}_{0,n-k} & \bar{c}_{0,n-k-1} & & \bar{c}_{0,0} & 0 \\
0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \bar{c}_{0,n-k} & \cdots & \bar{c}_{0,1} & \bar{c}_{0,0} \\
\hline
 & \vdots & & & & \vdots & & & \vdots & & \\
\hline
0 & \cdots & 0 & \bar{c}_{i,p-k} & \cdots & \bar{c}_{i,0} & & & & & \\
0 & & 0 & 0 & \bar{c}_{i,p-k} & \cdots & \bar{c}_{i,0} & & & & \\
\vdots & & \vdots & & & \ddots & & \ddots & & & \\
0 & & 0 & & & & \bar{c}_{i,n-k} & \cdots & \bar{c}_{i,0} & 0 & 0 \\
0 & & 0 & 0 & \cdots & & 0 & \bar{c}_{i,n-k} & \cdots & \bar{c}_{i,0} & 0 \\
0 & & 0 & 0 & \cdots & & & 0 & \bar{c}_{i,n-k} & \cdots & \bar{c}_{i,0} \\
\hline
 & \vdots & & & & \vdots & & & \vdots & & \\
\hline
0 & \cdots & 0 & \bar{c}_{h,p-k} & \cdots & \bar{c}_{h,0} & 0 & \cdots & & & 0 \\
0 & \cdots & 0 & 0 & \bar{c}_{h,p-k} & \cdots & \bar{c}_{h,0} & 0 & \cdots & & 0 \\
\vdots & & \vdots & \vdots & \ddots & \ddots & & \ddots & \ddots & & \\
0 & & 0 & 0 & \cdots & 0 & \bar{c}_{h,p-k} & \cdots & \bar{c}_{h,0} & 0 & 0 \\
0 & \cdots & 0 & 0 & \cdots & & 0 & \bar{c}_{h,p-k} & \cdots & \bar{c}_{h,0} & 0 \\
0 & \cdots & 0 & 0 & \cdots & & & 0 & \bar{c}_{h,p-k} & \cdots & \bar{c}_{h,0}
\end{array}\right]
\end{array}
\tag{6.8b}
$$

<u>Theorem (6.2)</u>: Consider the set of polynomials $\mathcal{P} \in \mathcal{T}(n,p;h+1)$, and let $S_{\mathcal{P}}$ be the corresponding Sylvester matrix , then the following hold true:

a) The problem of defining the optimal approximate gcd is equivalent to two optimisation problems involving different sets of independent variables that is minimise the functions

$$f_1\left(\mathcal{P},\mathcal{P}^*\right)=\left\|\tilde{S}'_{\mathcal{P}}\right\|_{\mathrm{F}}\cdot\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}}, \tag{6.9a}$$

in terms of the parameters in $\upsilon(s)$,

$$f_2\left(\mathcal{P},\mathcal{P}^*\right)=\left\|\left[\mathbf{0}_r \;\middle|\; \hat{S}^{(2)}_{\mathcal{P}}-\bar{S}_{\mathcal{P}^*}\right]\right\|_{\mathrm{F}}\cdot\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}} \tag{6.9b}$$

in terms of the rest of the free parameters.

b) For a given approximate gcd $\upsilon(s)$ of degree k, then:

i) The perturbed set $\tilde{\mathcal{P}}$ corresponding to minimal perturbation applied on $\mathcal{P}$, such that $\upsilon(s)$ becomes the exact gcd of of the perturbed set, is obtained from:

$$S_{\tilde{\mathcal{P}}} = \tilde{S}''_{\mathcal{P}}\hat{\Phi}_{\upsilon} = \left[\mathbf{0}_k \mid \hat{S}^{(2)}_{\mathcal{P}}\right]\hat{\Phi}_{\upsilon} \tag{6.10}$$

ii) The strength of the approximate gcd $\upsilon(s)$ of degree k is independent of the coefficients of the perturbation and it is given by:

$$f_1\left(\mathcal{P},\mathcal{P}^*\right) = \left\|\tilde{S}'_{\mathcal{P}}\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}} \tag{6.11}$$

Proof:

a) The splitting of the optimisation problem follows from the inequality (6.7) and the observation that $f_1\left(\mathcal{P},\mathcal{P}^*\right)$ involves only the parameters of $\upsilon(s)$, whereas $f_2\left(\mathcal{P},\mathcal{P}^*\right)$ can be minimised and take the value of the absolute minimum, i.e. the 0 value, by choosing only the free parameters in the perturbation matrix.

b) i) For a given approximate gcd $\upsilon(s)$ the matrices Φ_{υ}, $\hat{\Phi}_{\upsilon}$ are fixed and that implies that matrices $\tilde{S}'_{\mathcal{P}}$, $\tilde{S}''_{\mathcal{P}}$ are also fixed. Thus in the optimization problem (6.6) $f\left(\mathcal{P},\mathcal{P}^*\right) = \min\limits_{\forall\mathcal{P}^*}\left\{\left\|\tilde{S}'_{\mathcal{P}} + \left[\mathbf{0}_r \mid \hat{S}^{(2)}_{\mathcal{P}} - \bar{S}_{\mathcal{P}^*}\right]\right\|_{\mathrm{F}} \cdot \left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}}\right\}$, the only free parameters are the elements of the resultant $\bar{S}_{\mathcal{P}^*}$. Thus $\bar{S}_{\mathcal{P}^*} \equiv \hat{S}^{(2)}_{\mathcal{P}}$ is an obvious solution of (6.6)

ii) The proof follows in a straight forward way from the result of part (b) (i) and given that $f_1\left(\mathcal{P},\mathcal{P}^*\right)$ is fixed, we have only to compute the norm of the corresponding matrix.

■

Theorem 6.2 provides an explicit formula for computing the strength of a given approximate gcd in terms of computing the Frobenius norm of a given matrix (condition

(6.11)). Furthermore, the separation of the general optimization into two separate problems leads to the following characterisation of the "best" approximate gcd.

Corollary (6.1): For any polynomial set $\mathcal{P} \in \mathcal{T}(n, p; h+1)$ the *Optimal Approximate GCD* of degree k is a polynomial $\varphi(s)$ that corresponds to the solution of

$$\min_{\forall \mathcal{P}^*} \left\{ f_1\left(\mathcal{P}, \mathcal{P}^*\right) \right\} = \min_{\substack{\forall \varphi(s): \\ \deg\{\varphi(s)\}=k}} \left\{ \left\| \tilde{S}'_{\mathcal{P}} \hat{\Phi}_{\upsilon} \right\|_{\mathrm{F}} \right\} \tag{6.12}$$

Proof:

Theorem 6.2 states that the overall optimisation problem is divided into two independent problems depending on different set of free parameters. When minimisation of $f_1\left(\mathcal{P}, \mathcal{P}^*\right)$ is achieved (a problem independent from the parameterisation of perturbations), then minimisation of $f_2\left(\mathcal{P}, \mathcal{P}^*\right)$ is always possible for any selection of $\upsilon(s)$. Thus, if $\varphi(s)$ is the optimal solution, this must be the global minimum of $f\left(\mathcal{P}, \mathcal{P}^*\right) = \min_{\forall \mathcal{P}^*} \left\| \tilde{S}'_{\mathcal{P}} \hat{\Phi}_{\upsilon} \right\|_{\mathrm{F}}$. Thus the gcd of the specific degree that corresponds to the minimum strength corresponds to the solution of (6.12) and it is the Optimal Approximate GCD.

■

The above result provides an elegant solution to the distance problem of a set from the d-gcd variety and provides the means for its computation as the solution of a standard minimisaton problem based on a set of polynomials defined from the original set. In fact, Corollary (6.1) identifies the "best" common factor of a specific degree with the one that corresponds to the solution of the optimization problem expressed in (6.11). The degree of this factor can be chosen by SVD of the initial Sylvester matrix. The matrix $\hat{S}_{\mathrm{Q}} \triangleq \tilde{S}'_{\mathcal{P}} \hat{\Phi}_{\varphi}$ that corresponds to the optimal solution $\varphi(s) = \lambda_k s^k \cdots \lambda_1 s + \lambda_0$ has a Toeplitz form as it will be shown in the following analysis. From (6.2) and (6.8) the structure $\hat{S}_{\mathrm{Q}} \triangleq \tilde{S}'_{\mathcal{P}} \hat{\Phi}_{\varphi}$ is of the type indicated below:

$$\hat{S}_{Q}=\left[\begin{array}{cccccccccccc}
z_{0,n} & \cdots & z_{0,n-k+1} & 0 & \cdots & & & & & 0 & \cdots & 0\\
f_{0,1} & z_{0,n} & \cdots & z_{0,n-k+1} & 0 & & & & & & & \\
f_{0,2} & f_{0,1} & \ddots & & \ddots & & \ddots & & & & & \\
 & \ddots & \ddots & z_{0,n} & \cdots & & z_{0,n-k+1} & & 0 & \cdots & & 0\\
f_{0,p-1} & & f_{1,2} & f_{0,1} & z_{0,n} & & \cdots & & z_{0,n-k+1} & 0 & \ldots & 0\\
\hline
 & & \vdots & & & & & \vdots & & & & \\
\hline
z_{i,p} & \cdots & z_{i,p-k+1} & 0 & \cdots & & & & & 0 & \cdots & 0\\
f_{i,1} & z_{i,p} & & z_{i,p-k+1} & 0 & & & & & & & \\
f_{i,2} & f_{i,1} & z_{i,p} & & z_{i,p-k+1} & & \ddots & & & & & \\
f_{i,3} & f_{i,2} & \ddots & \ddots & & & \ddots & & & & & \\
\vdots & & \ddots & \overline{e}_{i,1} & z_{i,p} & & \cdots & z_{i,p-k+1} & 0 & \cdots & & 0\\
f_{i,n-1} & \cdots & f_{i,3} & f_{i,2} & f_{i,1} & z_{i,p} & \cdots & & z_{i,p-k+1} & 0 & \cdots & 0\\
\hline
 & & \vdots & & & & & \vdots & & & & \\
\hline
z_{h,p} & \cdots & z_{h,p-k+1} & 0 & \cdots & & 0 & & \cdots & 0 & \cdots & 0\\
f_{h,0} & z_{h,p} & & z_{i,p-k+1} & 0 & & & & & & & \\
f_{h,1} & f_{h,0} & z_{h,p} & \cdots & z_{i,p-k+1} & & \ddots & & & & & \\
\vdots & \ddots & & \ddots & & & \ddots & & & & & \\
f_{h,n-3} & \cdots & f_{h,1} & f_{h,0} & z_{h,p} & \cdots & z_{i,p-k+1} & & 0 & & & \\
f_{h,n-2} & \cdots & f_{h,2} & f_{h,1} & f_{h,0} & \overline{c}_{h,p} & z_{h,p} & \cdots & z_{i,p-k+1} & 0 & \cdots & 0
\end{array}\right]$$

(6.13a)

where

$$f_{0,r}=\sum_{\xi=0}^{r-1}\overline{e}_{0,r-\xi}\lambda_{\xi}+\sum_{\xi=r}^{k}\overline{c}_{0,n+r-\xi}\lambda_{\xi},\quad f_{i,r}=\sum_{\xi=0}^{r-1}\overline{e}_{i,r-\xi}\lambda_{\xi}+\sum_{\xi=r}^{k}\overline{c}_{i,p+r-\xi}\lambda_{\xi} \tag{6.13b}$$

and by (6.4b) and (6.13a) it follows that

$$\begin{aligned}
z_{0,n-\theta}&=\sum_{j=\theta}^{k-1}\overline{c}_{0,n-j}\lambda_{j}=\sum_{j=\theta}^{k-1}\left[\left(\sum_{\mu=j}^{n}b_{0,n-\mu}y_{\mu-j}\right)\lambda_{j}\right]=\sum_{j=\theta}^{k-1}\sum_{\mu=1}^{n}\left(b_{0,n-\mu}y_{\mu-j}\lambda_{j}\right)\\
z_{i,n-\theta}&=\sum_{j=\theta}^{k-1}\sum_{\mu=1}^{p}\left(b_{0,p-\mu}y_{\mu-j}\lambda_{j}\right),\qquad\qquad \theta=0,\ldots,k-1
\end{aligned} \tag{6.13c}$$

By Definition (4.4) of the Sylvester resultant it follows that the elements the i-block of S_{P} below the diagonal of the block are all zero. Combining that with the fact that $S_{\varphi}=\tilde{S}_{\varphi}\hat{\Phi}_{\upsilon}$ and using the structure of the related matrices we have that the $(\nu+r,\nu)$ element of the i-block:

$$\begin{array}{c}\left[\begin{array}{cccc|ccc|ccc|c}\overline{e}_{i,r} & \overline{e}_{i,r-1} & \cdots & \overline{e}_{i,1} & \overline{c}_{i,p} & \cdots & \overline{c}_{i,p-(k-r)} & \overline{c}_{i,p-(k-r)-1} & \cdots & \overline{c}_{i,0} & 0\cdots 0\end{array}\right] \\ \longleftarrow \quad n+p \quad \longrightarrow\end{array} \cdot \begin{array}{r}\\ \\ \\ \\ \\ \\ \\ \\ \\ \uparrow \\ n+p-k \\ \downarrow\end{array} \left[\begin{array}{c}\lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_{r-1} \\ \hline \lambda_r \\ \lambda_{r+1} \\ \vdots \\ \lambda_k \\ \hline 0 \\ \vdots \\ 0 \\ \\ \end{array}\right] = 0 \tag{6.14}$$

which implies

$$f_{i,r} = \sum_{\xi=0}^{r-1} \overline{e}_{i,r-\xi}\lambda_\xi + \sum_{\xi=r}^{k} \overline{c}_{i,p+r-\xi}\lambda_\xi = 0 \tag{6.15a}$$

and in similar way for the top block:

$$f_{0,r} = 0 \tag{6.15b}$$

The above results simplify the optimisation problem which may now be expressed in the following way:

Theorem (6.3): The Optimal GCD is defined by the minimisation of $\left\|\hat{S}_Q\right\|_F$ where $\hat{S}_Q = \tilde{S}'_{\mathcal{P}}\hat{\Phi}_\varphi$ that corresponds to the polynomial $\varphi(s) = \lambda_k s^k + \ldots + \lambda_1 s + \lambda_0$ where the matrix $\hat{S}_Q$ has a Toeplitz structure with nonzero elements defined by

$$\begin{aligned} z_{0,n-\theta} &= \sum_{j=\theta}^{k-1} \overline{c}_{0,n-j}\lambda_j = \sum_{j=\theta}^{k-1}\left[\left(\sum_{\mu=j}^{n} b_{0,n-\mu}y_{\mu-j}\right)\lambda_j\right] = \sum_{j=\theta}^{k-1}\sum_{\mu=1}^{n}\left(b_{0,n-\mu}y_{\mu-j}\lambda_j\right) \\ z_{i,n-\theta} &= \sum_{j=\theta}^{k-1}\sum_{\mu=1}^{p}\left(b_{0,p-\mu}y_{\mu-j}\lambda_j\right), \qquad\qquad \theta = 0,\ldots,k-1 \end{aligned} \tag{6.16a}$$

$$\hat{S}_{Q} = \left[\begin{array}{cccccc}
z_{0,n} & \cdots & z_{0,n-k+1} & 0 & \cdots & 0 \\
0 & z_{0,n} & \cdots & z_{0,n-k+1} & \ddots & \vdots \\
\vdots & \ddots & \ddots & & \ddots & 0 \\
0 & \cdots & 0 & z_{0,n} & \cdots & z_{0,n-k+1} \\
\hline
 & & & & & \\
\hline
z_{i,p} & \cdots & z_{i,p-k+1} & 0 & & 0 \\
0 & z_{i,p} & \cdots & z_{i,p-k+1} & \ddots & \\
\vdots & \ddots & & & \ddots & 0 \\
0 & \cdots & 0 & z_{i,p} & \cdots & z_{i,p-k+1} \\
\hline
 & & & & & \\
\hline
z_{h,p} & \cdots & z_{h,p-k+1} & 0 & \cdots & 0 \\
0 & z_{h,p} & & z_{h,p-k+1} & \ddots & \vdots \\
\vdots & \ddots & \ddots & & \ddots & 0 \\
0 & \cdots & 0 & z_{h,p} & \cdots & z_{h,p-k+1}
\end{array}\right] \qquad (6.16b)$$

We may demonstrate the structure of the resulting optimisation problem by an example of a generic type.

■

Example (6.1): Let us now consider the set of two polynomials $\mathcal{P}_{2,4} = \{p_0(s), p_1(s)\}$ where $p_0(s) = s^4 + a_3 s^3 + a_2 s^2 + a_1 s + a_0$, $p_1(s) = s^3 + b_2 s^2 + b_1 s + b_0$ and we search for the optimal approximate gcd of degree 2 with representation $\upsilon(s) = k_2 s^2 + k_1 s + k_0$. Then the Sylvester resultant is defined by

$$S_{\mathcal{P}} = \left[\begin{array}{ccccccc}
1 & a_3 & a_2 & a_1 & a_0 & 0 & 0 \\
0 & 1 & a_3 & a_2 & a_1 & a_0 & 0 \\
0 & 0 & 1 & a_3 & a_2 & a_1 & a_0 \\
\hline
1 & b_2 & b_1 & b_0 & 0 & \cdots & 0 \\
0 & 1 & b_2 & b_1 & b_0 & \ddots & \vdots \\
\vdots & \ddots & 1 & b_2 & b_1 & b_0 & 0 \\
0 & \cdots & 0 & 1 & b_2 & b_1 & b_0
\end{array}\right] \qquad (6.17a)$$

and the $\hat{\Phi}_{\upsilon}$, Φ_{υ} matrices are given by:

$$\hat{\Phi}_{\upsilon}=\begin{bmatrix} k_0 & 0 & 0 & 0 & 0 & 0 & 0 \\ k_1 & k_0 & 0 & 0 & 0 & 0 & 0 \\ k_2 & k_1 & k_0 & 0 & 0 & 0 & 0 \\ 0 & k_2 & k_1 & k_0 & 0 & 0 & 0 \\ 0 & 0 & k_2 & k_1 & k_0 & 0 & 0 \\ 0 & 0 & 0 & k_2 & k_1 & k_0 & 0 \\ 0 & 0 & 0 & 0 & k_2 & k_1 & k_0 \end{bmatrix},\ \Phi_{\upsilon}=\begin{bmatrix} l_0 & 0 & 0 & 0 & 0 & 0 & 0 \\ l_1 & l_0 & 0 & 0 & 0 & 0 & 0 \\ l_2 & l_1 & l_0 & 0 & 0 & 0 & 0 \\ l_3 & l_2 & l_1 & l_0 & 0 & 0 & 0 \\ l_4 & l_3 & l_2 & l_1 & l_0 & 0 & 0 \\ l_5 & l_4 & l_3 & l_2 & l_1 & l_0 & 0 \\ l_6 & l_5 & l_4 & l_3 & l_2 & l_1 & l_0 \end{bmatrix} \qquad (6.17b)$$

where $\Phi_{\upsilon}=\hat{\Phi}_{\upsilon}^{-1}$ and thus its elements are defined by

$$l_0=\frac{1}{k_0},\ l_i=-l_{i-1}\frac{k_1}{k_0}-l_{i-2}\frac{k_2}{k_0},\ \ i=1,\ldots,6 \qquad (6.17c)$$

and thus are given as:

$$\begin{aligned} l_0&=\frac{1}{k_0} \\ l_1&=-\frac{k_1}{k_0} \\ l_2&=\frac{k_1^{\ 2}}{k_0^{\ 3}}-\frac{k_2}{k_0^{\ 2}} \\ l_3&=-\frac{k_1^{\ 3}}{k_0^{\ 4}}+2\frac{k_1k_2}{k_0^{\ 3}} \\ l_4&=\frac{k_1^{\ 4}}{k_0^{\ 5}}-3\frac{k_1^{\ 2}k_2}{k_0^{\ 4}}+\frac{k_2^{\ 2}}{k_0^{\ 3}} \\ l_5&=-\frac{k_1^{\ 5}}{k_0^{\ 6}}+4\frac{k_1^{\ 3}k_2}{k_0^{\ 5}}-3\frac{k_1k_2^{\ 2}}{k_0^{\ 4}} \\ l_6&=\frac{k_1^{\ 6}}{k_0^{\ 7}}-5\frac{k_1^{\ 4}k_2}{k_0^{\ 6}}+6\frac{k_1^{\ 2}k_2^{\ 2}}{k_0^{\ 5}}-\frac{k_2^{\ 3}}{k_0^{\ 4}} \end{aligned} \qquad (6.17.d)$$

The perturbation resultant that introduces the optimisation problem is

$$S_Q=S_{\varphi}\Phi_{\upsilon}\hat{\Phi}_{\upsilon}-\left[\mathbf{0}_2 \mid S_W\right]=\left\{S_{\varphi}\Phi_{\upsilon}-\left[\mathbf{0}_2 \mid S_W\right]\right\}\hat{\Phi}_{\upsilon} \qquad (6.17.e)$$

where $\left[\mathbf{0}_2 \mid S_{\mathcal{W}}\right]$ is the matrix containing the free parameters associated with the perturbation and has the form

$$\left[\mathbf{0} \mid S_{\mathcal{W}}\right]=\left[\begin{array}{cc:ccccc} 0 & 0 & x_2 & x_1 & x_0 & 0 & 0 \\ 0 & 0 & 0 & x_2 & x_1 & x_0 & 0 \\ 0 & 0 & 0 & 0 & x_2 & x_1 & x_0 \\ \hdashline 0 & 0 & y_1 & y_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & y_1 & y_0 & 0 & 0 \\ 0 & 0 & 0 & 0 & y_1 & y_0 & 0 \\ 0 & 0 & 0 & 0 & 0 & y_1 & y_0 \end{array}\right] \tag{6.17.f}$$

Note that the matrix $S_{\varphi}\Phi_{\upsilon}$ has the form

$$S_{\varphi}\Phi_{\upsilon}=\tilde{S}_{\varphi}=\left[\begin{array}{ccccccc} \bar{c}_{0,4} & \bar{c}_{0,3} & \bar{c}_{0,2} & \bar{c}_{0,1} & \bar{c}_{0,0} & 0 & 0 \\ \bar{e}_{0,1} & \bar{c}_{0,4} & \bar{c}_{0,3} & \bar{c}_{0,2} & \bar{c}_{0,1} & \bar{c}_{0,0} & 0 \\ \bar{e}_{0,2} & \bar{e}_{0,1} & \bar{c}_{0,4} & \bar{c}_{0,3} & \bar{c}_{0,2} & \bar{c}_{0,1} & \bar{c}_{0,0} \\ \hdashline \bar{c}_{1,3} & \bar{c}_{1,2} & \bar{c}_{1,1} & \bar{c}_{1,0} & 0 & 0 & 0 \\ \bar{e}_{1,1} & \bar{c}_{1,3} & \bar{c}_{1,2} & \bar{c}_{1,1} & \bar{c}_{1,0} & 0 & 0 \\ \bar{e}_{1,2} & \bar{e}_{1,1} & \bar{c}_{1,3} & \bar{c}_{1,2} & \bar{c}_{1,1} & \bar{c}_{1,0} & 0 \\ \bar{e}_{1,3} & \bar{e}_{1,2} & \bar{e}_{1,1} & \bar{c}_{1,3} & \bar{c}_{1,2} & \bar{c}_{1,1} & \bar{c}_{1,0} \end{array}\right] \tag{6.17.g}$$

where

$$\begin{array}{ll} \bar{c}_{0,4}=l_0+a_3l_1+a_2l_2+a_1l_3+a_0l_4 & \bar{c}_{1,3}=l_0+b_2l_1+b_1l_2+b_0l_3 \\ \bar{c}_{0,3}=a_3l_0+a_2l_1+a_1l_2+a_0l_3 & \bar{c}_{1,2}=b_2l_0+b_1l_1+b_0l_2 \\ \bar{c}_{0,2}=a_2l_0+a_1l_1+a_0l_2 & \bar{c}_{1,1}=b_1l_0+b_0l_1 \\ \bar{c}_{0,1}=a_1l_0+a_0l_1 & \bar{c}_{1,0}=b_0l_0 \\ \bar{c}_{0,0}=a_0l_0 & \\ & \\ \bar{e}_{0,1}=l_1+a_3l_2+a_2l_3+a_1l_4+a_0l_5 & \bar{e}_{1,1}=l_1+b_2l_2+b_1l_3+b_0l_4 \\ \bar{e}_{0,2}=l_2+a_3l_3+a_2l_4+a_1l_5+a_0l_6 & \bar{e}_{1,2}=l_2+b_2l_3+b_1l_4+b_0l_5 \end{array} \tag{6.17h}$$

and thus $\tilde{S}_{\varphi}$ may be split as $\tilde{S}_{\varphi}=\tilde{S}'_{\varphi}+\tilde{S}''_{\varphi}$ which is demonstrated below:

$$\tilde{S}_{\varphi} = \underbrace{\left[\begin{array}{ccccccc} \bar{c}_{0,4} & \bar{c}_{0,3} & 0 & 0 & 0 & 0 & 0 \\ \bar{e}_{0,1} & \bar{c}_{0,4} & \bar{c}_{0,3} & 0 & 0 & 0 & 0 \\ \bar{e}_{0,2} & \bar{e}_{0,1} & \bar{c}_{0,4} & \bar{c}_{0,3} & 0 & 0 & 0 \\ \hline \bar{c}_{1,3} & \bar{c}_{1,2} & 0 & 0 & 0 & 0 & 0 \\ \bar{e}_{1,1} & \bar{c}_{1,3} & \bar{c}_{1,2} & 0 & 0 & 0 & 0 \\ \bar{e}_{1,2} & \bar{e}_{1,1} & \bar{c}_{1,3} & \bar{c}_{1,2} & 0 & 0 & 0 \\ \bar{e}_{1,3} & \bar{e}_{1,2} & \bar{e}_{1,1} & \bar{c}_{1,3} & \bar{c}_{1,2} & 0 & 0 \end{array}\right]}_{\triangleq \tilde{S}'_{\varphi}} + \underbrace{\left[\begin{array}{ccccccc} 0 & 0 & \bar{c}_{0,2} & \bar{c}_{0,1} & \bar{c}_{0,0} & 0 & 0 \\ 0 & 0 & 0 & \bar{c}_{0,2} & \bar{c}_{0,1} & \bar{c}_{0,0} & 0 \\ 0 & 0 & 0 & 0 & \bar{c}_{0,2} & \bar{c}_{0,1} & \bar{c}_{0,0} \\ \hline 0 & 0 & \bar{c}_{1,1} & \bar{c}_{1,0} & 0 & 0 & 0 \\ 0 & 0 & 0 & \bar{c}_{1,1} & \bar{c}_{1,0} & 0 & 0 \\ 0 & 0 & 0 & 0 & \bar{c}_{1,1} & \bar{c}_{1,0} & 0 \\ 0 & 0 & 0 & 0 & 0 & \bar{c}_{1,1} & \bar{c}_{1,0} \end{array}\right]}_{\triangleq \tilde{S}''_{\varphi}} \tag{6.17.i}$$

Then $\tilde{S}''_{\varphi} = \left[\mathbf{0}_2 \;\middle|\; \hat{S}''_{\varphi}\right]$ and thus the optimisation equation can be slit as

$$S_{Q} = \tilde{S}'_{\varphi}\hat{\Phi}_{\upsilon} - \left[\mathbf{0}_2 \;\middle|\; \tilde{S}''_{\varphi} - S_{W}\right]\hat{\Phi}_{\upsilon} \tag{6.17.j}$$

where by Corollary (6.3) the optimal approximate gcd is obtained from the minimisation of the norm of the first part $\tilde{S}'_{\varphi}\hat{\Phi}_{\upsilon}$. By Theorem (6.3) it follows that the elements below the main diagonal are zero. Finally $\tilde{S}'_{\varphi}\hat{\Phi}_{\upsilon}$ has the form

$$\tilde{S}'_{\varphi}\hat{\Phi}_{\upsilon} = \left[\begin{array}{ccccccc} z_{0,4} & z_{0,3} & 0 & 0 & 0 & 0 & 0 \\ 0 & z_{0,4} & z_{0,3} & 0 & 0 & 0 & 0 \\ 0 & 0 & z_{0,4} & z_{0,3} & 0 & 0 & 0 \\ \hline z_{1,3} & z_{1,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & z_{1,3} & z_{1,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & z_{1,3} & z_{1,2} & 0 & 0 & 0 \\ 0 & 0 & 0 & z_{1,3} & z_{1,2} & 0 & 0 \end{array}\right] \tag{6.17.k}$$

where from (6.17.h) and (6.17.b) we have:

$$z_{0,4} = \bar{c}_{0,4}k_0 + \bar{c}_{0,3}k_1 = \left(l_0 + a_3 l_1 + a_2 l_2 + a_1 l_3 + a_0 l_4\right)k_0 + \left(a_3 l_0 + a_2 l_1 + a_1 l_2 + a_0 l_3\right)k_1$$

$$z_{0,3} = \left(a_3 l_0 + a_2 l_1 + a_1 l_2 + a_0 l_3\right)k_0$$

$$z_{1,3} = \bar{c}_{1,3}k_0 + \bar{c}_{1,2}k_1 = \left(l_0 + b_2 l_1 + b_1 l_2 + b_0 l_3\right)k_0 + \left(b_2 l_0 + b_1 l_1 + b_0 l_2\right)k_1$$

$$z_{1,2} = \left(b_2 l_0 + b_1 l_1 + b_0 l_{23}\right)k_0$$

The Frobenius norm of $\tilde{S}_{\varphi}$ is then expressed as:

$$f_1\left(\mathcal{P},\mathcal{P}^*\right)=\left\|\tilde{S}'_{\mathcal{P}}\hat{\Phi}_{\upsilon}\right\|_F^2=3\left(z_{0,4}^2+z_{0,3}^2\right)+4\left(z_{1,3}^2+z_{1,2}^2\right)$$

where $f_1\left(\mathcal{P},\mathcal{P}^*\right)=f\left(k_0,k_1,k_2\right)$ and thus the minimisation of the resulting function leads to the definition of the optimal gcd.

■

Remark (6.2): The above example demonstrates the nature of the function $f_1\left(\mathcal{P},\mathcal{P}^*\right)$ as a function of the coefficients of the given degree polynomial. In fact, the conditions (6.16a) together with the relationships between y and λ parameters allow the definition of $f_1\left(\mathcal{P},\mathcal{P}^*\right)$ as an explicit function on the set $\mathcal{P}$.

■

We may further illustrate the procedure in terms of a numerical example as shown below:

Example (6.2): Let the set of three polynomials

$$\mathcal{P}_{3,2}=\left\{b_0(s)=s^3-1.99s^2-s+2.01,\ b_1(s)=s^2-3s+2,\ b_2(s)=s-0.99\right\}$$

the corresponding resultant is

$$S_{\mathcal{P}}=\left[\begin{array}{ccccc}1 & -1.99 & -1 & 2.01 & 0\\ 0 & 1 & -1.99 & -1 & 2.01\\ \hdashline 1 & -3 & 2 & 0 & 0\\ 0 & 1 & -3 & 2 & 0\\ 0 & 0 & 1 & -3 & 2\\ \hdashline 1 & -0.99 & 0 & 0 & 0\\ 0 & 1 & -0.99 & 0 & 0\\ 0 & 0 & 1 & -0.99 & 0\end{array}\right]$$

and the singular values of $S_{\mathcal{P}}$ are

$$\{5.7200,\ 5.0483,\ 3.0328,\ 0.7343,\ 0.0088\}$$

Thus from Definition (6.1) and Theorem (6.1) for $\varepsilon = 0.0088$ (or higher) we have numerical ε-nullity $\mathcal{N}_{\varepsilon}\left(S_{\mathcal{P}}\right)=1$. Thus the approximate factor we will estimate will be of order 1. Let $\varphi(s)=s+k_0,\ k_0 \neq 0$ be the the approximate factor then $\tilde{S}'_{\mathcal{P}}$ defined in (6.5) and the matrix Φ_{φ} are respectively given by:

$$\tilde{S}'_{\mathcal{P}} = \left[\begin{array}{ccccc}
\frac{k_0^3+1.99k_0^2-k_0-2.01}{k_0^4} & 0 & 0 & 0 & 0 \\
-\frac{k_0^3+1.99k_0^2-k_0-2.01}{k_0^5} & \frac{k_0^3+1.99k_0^2-k_0-2.01}{k_0^4} & 0 & 0 & 0 \\
\hline
\frac{k_0^2+3k_0+2}{k_0^3} & 0 & 0 & 0 & 0 \\
-\frac{k_0^2+3k_0+2}{k_0^4} & \frac{k_0^2+3k_0+2}{k_0^3} & 0 & 0 & 0 \\
\frac{k_0^2+3k_0+2}{k_0^5} & -\frac{k_0^2+3k_0+2}{k_0^4} & \frac{k_0^2+3k_0+2}{k_0^3} & 0 & 0 \\
\hline
\frac{k_0+0.99}{k_0^2} & 0 & 0 & 0 & 0 \\
-\frac{k_0+0.99}{k_0^3} & \frac{k_0+0.99}{k_0^2} & 0 & 0 & 0 \\
\frac{k_0+0.99}{k_0^4} & -\frac{k_0+0.99}{k_0^3} & \frac{k_0+0.99}{k_0^2} & 0 & 0
\end{array}\right]$$

$$\hat{\Phi}_{\varphi} = \begin{bmatrix} k_0 & 0 & 0 & 0 & 0 \\ 1 & k_0 & 0 & 0 & 0 \\ 0 & 1 & k_0 & 0 & 0 \\ 0 & 0 & 1 & k_0 & 0 \\ 0 & 0 & 0 & 1 & k_0 \end{bmatrix}$$

and thus $\tilde{S}'_{\mathcal{P}}\Phi_{\varphi}$ is expressed as shown below:

$$\tilde{S}'_{\varphi}\Phi_{\varphi}=\left[\begin{array}{ccccc}
\dfrac{k_0^3+1.99k_0^2-k_0-2.01}{k_0^3} & 0 & 0 & 0 & 0\\
0 & \dfrac{k_0^3+1.99k_0^2-k_0-2.01}{k_0^3} & 0 & 0 & 0\\
\hdashline
\dfrac{k_0^2+3k_0+2}{k_0^2} & 0 & 0 & 0 & 0\\
0 & \dfrac{k_0^2+3k_0+2}{k_0^2} & 0 & 0 & 0\\
0 & 0 & \dfrac{k_0^2+3k_0+2}{k_0^2} & 0 & 0\\
\hdashline
\dfrac{k_0+0.99}{k_0} & 0 & 0 & 0 & 0\\
0 & \dfrac{k_0+0.99}{k_0} & 0 & 0 & 0\\
0 & 0 & \dfrac{k_0+0.99}{k_0} & 0 & 0
\end{array}\right]$$

It is now clear that k_0 can be obtained from the optimisation problem of Definition (6.2) which in the present example is expressed as

$$f_1^2\left(\mathcal{P},\mathcal{P}^*\right)=\min\left\{\left\|\tilde{S}'_{\varphi}\Phi_{\varphi}\right\|_{\mathrm{F}}^2\right\}=$$

$$=\min\left\{2\left(\frac{k_0^3+1.99k_0^2-k_0-2.01}{k_0^3}\right)^2+3\left(\frac{k_0^2+3k_0+2}{k_0^2}\right)^2+3\left(\frac{k_0+0.99}{k_0}\right)^2\right\}$$

$$=\min\left\{\frac{2k_0{}^6+7.96k_0{}^5+3.9202k_0{}^4-16k_0{}^3-13.9996k_0{}^2+8.04k_0+8.0802}{k_0{}^6}+\right.$$

$$\left.+\frac{3k_0{}^4+18k_0{}^3+39k_0{}^2+36k_0+12}{k_0{}^4}+\frac{3k_0{}^2+5.94k_0+2.9403}{k_0{}^2}\right\}$$

By setting $w=\dfrac{1}{k_0}$ and substituting in the above expression, the problem is reduced to the equivalent problem below:

$$f^2\left(\mathcal{P},\mathcal{P}^*\right)=\min\left\{\left\|\tilde{S}_{\mathcal{P}}'\Phi_{\varphi}\right\|_F^2\right\}=\min\left\{8.0802w^6+8.04w^5-1.9996w^4+20w^3+45.86w^2+31.9w+8\right\}$$

Using Matlab we may find the solution of the last minimisation problem to be $w=-0.9964$ and thus $k_0\cong-1.0036$ which means that the best approximate factor order 1 for the polynomials of our example is $\varphi(s)=s-1.0036$

■

Let us now consider the case where s is an approximate factor of the initial set of polynomials. This can be characterized by a norm applied on the last set of columns of the matrix. This is expressed by the next definition.

Definition 6.3: Let $\mathcal{P}_{h+1,n}$ be a polynomial set and $S_{\mathcal{P}}=\left[\underline{p}_{n+p},\underline{p}_{n+p-1},\ldots,\underline{p}_2,\underline{p}_1\right]$ the corresponding Sylvester matrix expressed in terms of its columns, then

i) s is an $\varepsilon(0)$-approximate factor of $\mathcal{P}_{h+1,n}$ if $\left\|\underline{p}_1\right\|\leq\varepsilon(0)$, $\varepsilon(0)\geq0$

ii) s^k is an $\varepsilon_k(0)$-approximate factor of $\mathcal{P}_{h+1,n}$ if $\left\|\left[\underline{p}_k,\underline{p}_{k-1},\ldots,\underline{p}_1\right]\right\|\leq\varepsilon_k(0)$, $\varepsilon_k(0)\geq0$

The numbers $\varepsilon(0)$, $\varepsilon_k(0)$ will be referred as the *strength of the approximate factors* s and s^k respectively

■

6.4. ALGORITHM FOR OPTIMAL APPROXIMAL GCD

We can now design a global algorithm for the evaluation of the approximate gcd that will include a check for possible s^k factors and elimination of the last k rows.

Evaluation of Approximate Common Factor (Algorithm 6.1):

i) Investigation of existence and extraction of s^{k_1} factor

- Evaluation of numerical nullity $\mathcal{N}_{\varepsilon}(S_{\mathcal{P}}) = k$
- For a specified $\varepsilon(0)$ find the maximum positive integer k_1 such that s^{k_1} is an $\varepsilon(0)$-approximate common factor of $\mathcal{P}_{h+1,n}$, obviously $k_1 \leq k$
- Eliminate the last k_1 columns of the resultant $S_{\mathcal{P}}$ that lead to the $S_{\mathcal{P},1}$ matrix

ii) Extraction of proper approximate common factor from $S_{\mathcal{P},1}$

- Construction of the transformation free-variable matrices Φ_{φ}, $\hat{\Phi}_{\varphi}$, $\deg\{\varphi(s)\} = k - k_1 = k_2$
- Construction of the matrices $\tilde{S}_{\mathcal{P},1} = S_{\mathcal{P},1}\Phi_{\varphi}$, $\tilde{S}'_{\mathcal{P},1}$
- Construction of $\tilde{S}'_{\mathcal{P},1}\hat{\Phi}_{\varphi}$
- Minimisation of $\left\| \tilde{S}'_{\mathcal{P},1}\hat{\Phi}_{\varphi} \right\|$ leads to the definition of $\varphi(s)$

Best approximate common factor is $s^{k_1}\varphi(s)$

■

6.5. DISCUSSION

The investigation of the approximate common factor and gcd for many polynomials has been completed in this chapter. The overall approach has been based on the formulation of the "approximate gcd" as a distance problem. This has been achieved by a combination of the results on Chapter 3 related to the representation theory, the definition of Chapter 4 of the strength of the approximation and the study of the optimisation properties of the defined problem. A new algorithm has been established, which is based on standard optimisation of functions constructed from the original sets of polynomials. New open issues may arise on this, related to the choice of the appropriate rank for the optimisation problem and the relation between the accuracy of the numerical nullity and the strength of the approximation.

The explicit form of the reduced optimisation, based only on the free parameters of the approximate gcd, simplifies a lot the optimisation and generalises the results on almost zeros previously defined [Karcanias et al., 1983]. Furthermore the current results provide the means for studying concrete problems such as those considered in the following chapters and dealing with issues of root clustering to the case of linear systems. The latter extension motivates the need to extend the approach to matrix polynomials.

Chapter 7:

APPROXIMATE FACTORISATION OF POLYNOMIALS

7.1.INTRODUCTION

The algebra of polynomials provides the basis for the development of algebraic control approaches and issues such as computation of Smith forms, solvability of Diophantine equations, etc. Problems such as factorisations of polynomials play a key role within this area. Of special interest is the problem of factorisation of polynomials without resorting to root finding, as well as handling issues of approximate factorisations and grouping roots which are close. The latter is important especially when there is uncertainty on the exact values of the coefficients. Recently [Karcanias et al., 2000], some special factorisation of polynomials has been introduced, which is within the general factorisation theory, and which can be performed without resorting to procedures based on finding roots. This factorisation is referred to as "normal factorisation" and its derivation is based on gcd algorithms. The results on the approximate gcd, introduced in the previous chapters, are now used to extend the normal factorisation in an approximate sense, which in turn provide the means for handling issues of "root clustering" of polynomials in a systematic way, using the notion of approximate gcd.

The essence of the normal factorisation is that we use the original polynomial and its derivatives, defined explicitly (and not numerically) and then gcd algorithms provide the tools for working this factorisation. The extension of these results is given here to the case of approximate factorisation that is linked to the "root clustering problem"

7.2. BACKGROUND RESULTS: NORMAL FACTORISATION

The investigation of the approximate factorization is based on resent results related with the *normal factorization of polynomials* [Karcanias et al., 2000]. The approximate factorization is a generalization of the existing techniques for the exact case based on elementary divisors.

Remark 7.1: For every polynomial $b(s) \in \mathbb{R}[s]$ there exist positive integers $d_1,...,d_\sigma$ where $d_1 > d_2 > ... > d_\sigma \geq 1$ such that $b(s)$ may be expressed as

$$b(s) = f_1(s)^{d_1} f_2(s)^{d_2} \cdots f_\sigma(s)^{d_\sigma} \tag{7.1}$$

where the polynomials $f_1(s), f_2(s),..., f_\sigma(s)$ are pairwise coprime and the polynomial $\hat{b}(s) = f_1(s) f_2(s) \cdots f_\sigma(s)$ has distinct roots.

■

The factorization introduced above is referred to as normal factorisation [Karcanias et al., 2000] and it is characterised by the following properties:

Proposition (7.1): Consider an n-degree polynomial $b(s) \in \mathbb{R}[s]$, $b(s) = s^n + b_{n-1}s^{n-1} + ... + b_1 s + b_0$. We assume that $(s+\lambda)^\tau$ is an elementary divisor of $b(s)$ over C. The following properties hold true:

i) The first derivative $b^{(1)}(s)$ has $(s+\lambda)^{\tau-1}$ as elementary devisor.

ii) The k-th derivative $b^{(k)}(s)$, $k < \tau$, has $(s+\lambda)^{\tau-k}$ as elementary divisor.

iii) The $b^{(\tau)}(s)$ derivative is the smallest order derivative that has no roots at $s = -\lambda$.

■

The proof of the above is obvious. A respective generalised result for the structure of the derivatives is the following:

Theorem (7.1): [Karcanias et al., 2000]: Let $b(s) \in \mathbb{R}[s]$ assumed in the irreducible factorised, ordered form

$$b(s) = (s+\lambda_1)^{\tau_1} \cdots (s+\lambda_\sigma)^{\tau_\sigma} (s+\lambda_{\sigma+1}) \cdots (s+\lambda_\mu) \tag{7.2}$$

where $\tau_1 \geq ... \geq \tau_\sigma > 1$. The following properties hold true:

i) The first derivative of $b(s)$ is expressed as

$$b^{(1)}(s) = (s+\lambda_1)^{\tau_1-1} \cdots (s+\lambda_\sigma)^{\tau_\sigma-1} g_0(s)$$

where $g_0(s)$ has no roots from the set $\{-\lambda_1,...,-\lambda_\mu\}$.

ii) The k-th derivative, $b^{(k)}(s)$, where $\tau_1 \geq ... \geq \tau_v \geq \tau_{v+1}$ is expressed as

$$b^{(k)}(s) = (s+\lambda_1)^{\tau_1-k} \cdots (s+\lambda_v)^{\tau_v-k} g_k(s)$$

where $g_k(s)$ has no common roots with $g_{k-1}(s)$.

iii) The integer $k=\tau_1$ is the smallest for which the polynomials $b(s)$, $b^{(1)}(s)$,..., $b^{(k)}(s)$ become coprime

■

Proposition (7.1) and Theorem (7.1) provide the means to link the problem of the normal factorisation of a polynomial and the algorithms for the gcd evaluation of many polynomials. Thus the problem is transformed to the evaluation of gcd of the set consists of the initial polynomial and its derivatives.

Corollary (7.1): The existence of a non trivial gcd $\upsilon(s)$ for the set $\{b(s), b^{(1)}(s),...,b^{(k)}(s)\}$ implies that $\upsilon(s)$ is a polynomial factor of $b(s)$ of multiplicity k, i.e.:

$$b(s) = \upsilon(s)^k \hat{b}(s) \tag{7.3}$$

■

Corollary (7.1) follows from Proposition (7.1) and states that the factorisation of a polynomial can be expressed in terms of polynomials with distinct roots in a procedure that does not involve root finding. The Matrix Factorisation theory, developed in Chapter 4, can be applied in the factorisation of a polynomial with respect to Theorem (7.1) or Corollary (7.1). This is demonstrated by the following example which motivates the developments in the following section.

Example (7.1): Consider the polynomial $b(s) = s^7 - s^6 - 3s^3 + 3s^2 + 2s - 2$ and its corresponding first and second order derivatives $b'(s) = 7s^6 - 6s^5 - 9s^2 + 6s + 2$ and

$b''(s) = 42s^5 - 30s^4 - 18s + 6$. The resultant matrix, $S\left(b(s), b'(s), b''(s)\right)$ is given in (7.4a) and its row echelon form in (7.4b):

$$S\left(b(s),b'(s),b''(s)\right)=\left[\begin{array}{ccccccccccccc}
1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 \\
\hline
7 & -6 & 0 & 0 & -9 & 6 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 7 & -6 & 0 & 0 & -9 & 6 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 7 & -6 & 0 & 0 & -9 & 6 & 2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 7 & -6 & 0 & 0 & -9 & 6 & 2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 7 & -6 & 0 & 0 & -9 & 6 & 2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 7 & -6 & 0 & 0 & -9 & 6 & 2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 7 & -6 & 0 & 0 & -9 & 6 & 2 \\
\hline
42 & -30 & 0 & 0 & -18 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 42 & -30 & 0 & 0 & -18 & 6 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 42 & -30 & 0 & 0 & -18 & 6 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 42 & -30 & 0 & 0 & -18 & 6 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 42 & -30 & 0 & 0 & -18 & 6 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 42 & -30 & 0 & 0 & -18 & 6 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -18 & 6 & 0
\end{array}\right] \tag{7.4a}$$

The row echelon form of the resultant is given in (7.4b). From its last non-vanishing row it is implied that the gcd of $\left\{b(s), b'(s), b''(s)\right\}$ is $f_1(s) = s - 1$.

$$S \sim \begin{bmatrix} 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -3 & 3 & 2 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 12 & -15 & -12 & 14 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -2 & 2 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 7 & -9 & -7 & 8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \longleftarrow \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \qquad (7.4b)$$

Corollary 7.1 implies that $f_1(s)$ is a factor of multiplicity 1 of the initial polynomial $f(s)$. By applying Euclidean division we take:

$$b(s) = f_1(s)^3 \cdot \hat{b}(s) = (s-1)^3\left(s^4 + 2s^3 + 3s^2 + 4s + 2\right) \qquad (7.4c)$$

Investigating the resultant of $\hat{f}(s)$ and its first and second order derivatives, it follows that the set $\left\{\hat{f}(s), \hat{f}'(s), \hat{f}''(s)\right\}$ is coprime and thus we reduce the investigation to the resultant of $\left\{\hat{f}(s), \hat{f}'(s)\right\}$. The latter has the form:

$$S\left(\hat{f}(s), \hat{f}'(s)\right) = \left[\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 2 & 0 & 0 \\ 0 & 1 & 2 & 3 & 4 & 2 & 0 \\ 0 & 0 & 1 & 2 & 3 & 4 & 2 \\ \hdashline 0 & 0 & 0 & 1 & -3 & -6 & -2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad (7.4d)$$

and the echelon form is

$$S\left(\hat{f}(s),\hat{f}'(s)\right)\sim\begin{bmatrix}1&2&3&4&2&0&0\\0&1&2&3&4&2&0\\0&0&1&2&3&4&2\\0&0&0&1&-3&-6&0\\0&0&0&0&1&1&0\\0&0&0&0&0&1&1\\0&0&0&0&0&0&0\end{bmatrix}\tag{7.4e}$$

Thus $f_2(s)^2=(s+1)^2$ is a factor of $\hat{f}(s)$ and thus the polynomial $f(s)$ is factorised as

$$f(s)=f_1(s)^3f_2(s)^2\cdot\tilde{f}(s)=(s-1)^3(s+1)^2(s^2+2)$$

■

7.3. APPROXIMATE FACTORISATION OF POLYNOMIALS USING THE RESULTANT ALGORITHM

The previous discussion links the problem of the normal factorisation with the algorithms for the evaluation gcd of polynomials. In Example 7.1 the background theory of the resultants has been applied. The combination of the definitions and the properties on normal factorisation [Karcanias et al., 2000], with the analysis in Chapter 5 and Chapter 6 on the definition and evaluation of the approximate common factor and the approximate gcd of many polynomials lead to the notion of the "approximate factorization" which is now considered below:

Definition (7.1): Consider a monic polynomial $p(s)\in\mathbb{R}[s]$ with $\deg\{p(s)\}=n$ and the family of perturbed polynomials $\tilde{p}(s)$ expressed by

$$\tilde{p}(s)=p(s)+\varepsilon(s),\tag{7.5a}$$

where

$$\varepsilon(s)=\varepsilon_{n-1}s^{n-1}+\varepsilon_{n-2}s^{n-2}+\ldots+\varepsilon_1 s+\varepsilon_0=\underline{\varepsilon}^t\cdot\underline{e}_{n-1}(s),\ \deg\{\varepsilon(s)\}\le n-1\tag{7.5b}$$

We define:

i) $p(s)$ has an approximate multiple root at $s=-\lambda$ of multiplicity τ with error $\|\underline{\varepsilon}\|$, if $(s+\lambda)^{\tau}$ is an elementary divisor of $\tilde{p}(s)$ with some perturbation $\varepsilon(s)$ with bound $\|\underline{\varepsilon}\|<\delta$.

ii) $p(s)$ has an approximate $\|\underline{\varepsilon}\|$ factorisation defined by the elementary divisors $e_{\mu}(s)^{\tau_{\mu}}=(s+\lambda_{\mu})^{\tau_{\mu}}$, $\mu=1,\ldots,\nu$, $\tau_{\mu}\geq 1$, if there exists a perturbation $\varepsilon(s)\in Q_{\delta}$ such that $\|\underline{\varepsilon}\|<\delta$ and

$$\tilde{p}(s)=e_1(s)^{\tau_1}\cdot e_2(s)^{\tau_2}\cdots e_{\nu}(s)^{\tau_{\nu}}$$

■

An iterated algorithm can be constructed on this basis. First we examine some properties of the set of the derivatives and its corresponding resultant

Proposition (7.2): Let $\mathcal{D}_n^{h+1}=\{b_0(s)=s^n+b_{0,n-1}s^{n-1}+\ldots+b_{0,1}s+b_{0,0},\ b_i(s)=b_0^{(i)}(s),\ i=1,\ldots,h,\ h\leq n\}$ be a polynomial set that consists of a n-th degree polynomial and its derivatives of order $1,\ldots,h$. The $b_i(s)$ polynomial of the set are then given by:

$$b_i(s)=b_0^{(i)}(s)=\sum_{j=i}^{n}\frac{j!}{(j-i)!}b_{0,j}s^{j-i} \tag{7.6}$$

■

The proof of Proposition 7.2 is a straightforward result of the basic properties of polynomial derivatives. Differentiation properties of polynomials (linearity) imply the following result:

Proposition 7.3: Let $\tilde{b}(s)$, $b(s)$, $\varepsilon(s)$ be polynomials such that $\tilde{b}(s)=b(s)+\varepsilon(s)$. Then, for every h, $h=1,\ldots$ we have:

$$\frac{d^h\tilde{b}(s)}{ds^h}=\frac{d^hb(s)}{ds^h}+\frac{d^h\varepsilon(s)}{ds^h} \tag{7.7}$$

■

Remark (7.2): The polynomial sets of the type $\mathcal{D}_n^{h+1}$, consists of a polynomial $b(s)$ with degree n and its derivatives of order $1,\dots,h$ define a family $\mathcal{T}_{\mathcal{D}}^{n,h}$ that is clearly a subset of $\mathcal{T}(n,n-1;h+1)$. The perturbation set has to be redefined with respect to the restrictions of sub-family that are described in Proposition (7.2) and (7.3). In fact, althought the perturbation on $b_0(s)$ can be arbitrary the following perturbations on its derivatives are functions of the original perturbation.

■

The Sylvester matrix $S_{\mathcal{D}}$ of the $\mathcal{D}_n^{h+1}$ set will be a $\left[(h+1)n-1\right]\times\left(2n-1\right)$ matrix of the form:

$$S_{\mathcal{D}} = \begin{bmatrix} S_{\mathcal{D}}^{(0)} \\ \vdots \\ S_{\mathcal{D}}^{(i)} \\ \vdots \\ S_{\mathcal{D}}^{(h)} \end{bmatrix} \tag{7.7a}$$

where

$$S_{\mathcal{D}}^{(0)} = \begin{bmatrix} 1 & b_{0,p-1} & b_{0,p-2} & \cdots & b_{0,1} & b_{0,0} & 0 & \cdots & 0 \\ 0 & 1 & b_{0,p-1} & \cdots & b_{0,2} & b_{0,1} & b_{0,0} & \cdots & 0 \\ \vdots & & \ddots & & & & & & \vdots \\ 0 & \cdots & 0 & 1 & b_{0,p-1} & \cdots & & b_{0,1} & b_{0,0} \end{bmatrix} \tag{7.7b}$$

$$S_{\mathcal{D}}^{(i)} = \left[\begin{array}{ccccccc|ccc} b_{i,p-i} & \cdots & b_{i,1} & b_{i,0} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \ddots & & & & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & b_{i,p-i} & b_{i,p-i-1} & \cdots & b_{i,0} & 0 & 0 & & 0 \\ 0 & \cdots & 0 & b_{i,p-i} & b_{i,p-i-1} & \cdots & b_{i,0} & 0 & \cdots & 0 \end{array}\right] \tag{7.7c}$$

from (7.6) it follows that the elements of $S_{\mathcal{P}^D}$ are:

$$b_i(s) = b_0^{(i)}(s) = \sum_{j=i}^{n} \frac{j!}{(j-i)!} b_{0,j} s^{j-i} = \sum_{k=0}^{n-i} b_{i,k} s^k \tag{7.8a}$$

which in terms of the coefficients is expressed as:

$$b_{i,k} = \frac{(k+i)!}{(k)!} b_{0,k+i} \ , \quad k = 0,1,\ldots,n-i \ , \quad i \le h \tag{7.8b}$$

or

$$b_{i,k} = f_{k+i,k} b_{0,k+i} \ , \quad k = 0,1,\ldots,n-i \ , \quad i \le h \tag{7.8c}$$

where by $F_{a,b}$ we denote the factorial

$$f_{a,b} = \frac{a!}{(a-b)!} \ , \quad a \ge b \ge 0 \tag{7.8d}$$

Remark (7.3): An alternative and less complicated expression to the above can be derived if we express the terms of the coefficients of the i-th order derivative in terms of the the $(i-1)$-th order one, $i = 1,\ldots,h$. Then we have the equivalent to (7.8) recursive formulas:

$$b_{1,j-1} = j b_{0,j} \ , \quad j = 1,\ldots,n \tag{7.9a}$$

$$b_{i,j-1} = j b_{i-1,j} \ , \quad j = i,\ldots,n \ , \quad i \le n \tag{7.9b}$$

■

Remark (7.4): The above demonstrates that the set $\mathcal{D}_n^{h+1}$ is a special set, generated by a single polynomial $b_0(s)$, and thus the general results for approximate gcd developed before also apply here. However, because the elements are generated by the derivatives of the perturbations of the nominal polynomial $b_0(s)$, this imposes some structure and introduces fewer perturbation parameters in the study of the resulting distance problem.

■

From the above analytic description of the elements of the derivative polynomials follows that the Sylvester resultant matrix $S_{\mathcal{D}}$ will have the form:

$$S_{\mathcal{D}} = \left[\begin{array}{ccccccccc}
b_{0,n} & \cdots & b_{0,j} & \cdots & b_{0,1} & b_{0,0} & 0 & \cdots & 0 \\
0 & \ddots & & & & & & \ddots & \vdots \\
\vdots & \ddots & & & & & & & 0 \\
0 & \cdots & 0 & b_{0,n} & \cdots & b_{0,j} & \cdots & b_{0,1} & b_{0,0} \\
\hdashline
nb_{0,n} & \cdots & jb_{0,j} & \cdots & 1b_{0,1} & 0 & 0 & \cdots & 0 \\
0 & & & & & & \ddots & \ddots & \vdots \\
\vdots & \ddots & & & & & & 0 & 0 \\
0 & \cdots & 0 & nb_{0,n} & \cdots & jb_{0,j} & & 1b_{0,1} & 0 \\
\hdashline
 & \vdots & & & \vdots & & & & \\
 & \vdots & & & \vdots & & & & \\
\hdashline
f_{n,h}b_{0,n} & \cdots & f_{h+k,h}b_{0,h+k} & \cdots\ f_{h,h}b_{0,h} & 0\ \ \cdots & 0 & \cdots & & 0 \\
0 & & & & \ddots\ \ \ddots & & & & \vdots \\
\vdots & \ddots & & & f_{h,h}b_{0,h} & 0 & \cdots & & 0 \\
0 & \cdots\ \ 0 & f_{n,h}b_{0,n} & \cdots & f_{h+1,h}b_{0,h} & f_{h,h}b_{0,h} & 0 & \cdots & 0
\end{array}\right] \tag{7.10}$$

Let us now denote by $\mathcal{Q}_{h+1,n}$ the set of all the perturbation polynomials applied on $\mathcal{D}_n^{h+1}$ so that $\mathcal{D}_n^{h+1} + \mathcal{Q}_{h+1,n}$, will have a common factor of degree h. Then the perturbation is:

$$\mathcal{Q}_{h+1,n} = \left\{ q_0(s) = q_{0,n}s^n + \ldots + q_{0,1}s + q_{0,0},\ q_i(s) = q_{i,n-i}s^{n-i} + \ldots + q_{i,1}s + q_{i,0},\ i = 0,1,\ldots h \right\} \tag{7.11}$$

and we shall also denote by S_{Q} the corresponding resultant matrix of $\mathcal{Q}_{h+1,n}$. The polynomials of the perturbed set will have the form

$$\bar{\mathcal{D}}_n^{h+1} = \left\{ \bar{p}_i(s) = p_i(s) + q_i(s),\ i = 0,1,\ldots,h \right\} \tag{7.12}$$

The investigation for the approximate factors of the polynomial $p(s)$ is reduced with the above notation to the problem of approximate common factors of a polynomial set and the theoretical algorithm introduced in Chapter 6 can be applied to this case too. The difference in this case is that the polynomials in $\mathcal{D}_n^{h+1}$ are "dependent" in the sense that they are defined as the derivatives of $p(s)$. Thus the perturbations on each polynomial (and the corresponding resultant block) have to be related in a similar way. The structure of these perturbations is a straight forward and it is based on the fact that the k-order derivative (where defined) is a linear function. This is expressed in (7.13) and according to this expression, (7.14) defines the form of the perturbation set $Q_{h+1,n}$.

Proposition 7.4: The set of perturbed polynomials consists of the perturbed polynomial $\overline{p}_0(s) = p(s) + q_0(s)$ and of its first h derivatives.

$$\overline{p}_0^{(k)}(s) = p^{(k)}(s) + q_0^{(k)}(s) \tag{7.13}$$

i.e.

$$Q_{h+1,n} = \left\{ q_0(s),\ q_i(s):\ q_i(s) = q_0^{(i)}(s) = \sum_{j=i}^{n} f_{j,i} q_{0,j} s^{j-i},\ i = 1,\ldots,h \right\} \tag{7.14}$$

■

The problem we investigate in this Chapter, that is the approximate factorisation of a polynomial, can now be examined with in the framework of the distance problem of the initial resultant set $\mathcal{D}_n^{h+1}$ and a perturbed set $\mathcal{D}^{*\,h+1}_{n}$ that has an exact nontrivial factor. The optimal solution will correspond to the minimal distance of $\mathcal{D}_n^{h+1}$ from the family of all perturbations that satisfy the derivative constrains defined above. This is described by the following result:

Definition 7.2: Consider the set $\mathcal{D}_n^{h+1} = \left\{ b_0(s), b_i(s) = \frac{d^i b_0(s)}{ds^i},\ i \in \underset{\sim}{h},\ h < n \right\} \in \mathcal{T}_{\mathcal{D}}^{n,h}$ and let $f(s) \in \mathbb{R}[s]$ be a given polynomial with $\deg\{f(s)\} = r \leq n - h,\ rh \leq n$. Furthermore, let $\Sigma_{\mathcal{D}}^{h+1,n} = \{Q_{h+1,n}\}$ be the set of all perturbations $Q_{h+1,n} \in \mathcal{T}_{\mathcal{D}}^{n-1,h}$ such that

$$\mathcal{D}^{*\,h+1}_{n} = \mathcal{D}_n^{h+1} - Q_{h+1,n} \in \mathcal{T}_{\mathcal{D}}^{n,h} \tag{7.15}$$

with the property that $f(s)$ is a common factor of the elements of $\mathcal{D}^{*\,h+1}_{n}$. If $Q_{h+1,n}^{*}$ is the minimal norm element of the set $\Sigma_{\mathcal{D}}^{h+1,n}$, then $f(s)$ is referred as an *h-order almost common factor* of $\mathcal{D}_n^{h+1}$, $f(s)^h$ is referred as an approximate factor of $b_0(s)$ and the norm of $Q_{h+1,n}^{*}$, denoted by $\|Q^{*}\|$ is defined as the *strength* of $f(s)^h$. ■

With the use of the above definition the problem now is to find the optimal (minimum distance) $\|Q^{*}\|$. This is equivalent to the minimum distance between the resultants of the two sets. This distance is described by equation (7.16a) and their equivalent forms (7.16b) and (7.16c). Φ_{ω} is transformation with low triangular Toeplitz matrix corresponding to the polynomial $\omega(s)$ and $\hat{\Phi}_{\omega}=\Phi_{\omega}^{-1}$. The properties of these matrices are described thoroughly in chapter 4. The problem under consideration is then:

$$\min\left\|S_{Q^{*}}\right\|=\min\left\|S_{\mathcal{D}^{*}}-S_{\mathcal{D}}\right\| \tag{7.16a}$$

$$\mu\left(\mathcal{D},\mathcal{D}^{*}\right)=\min_{\forall\mathcal{D}^{*}}\left\{\left\|S_{\mathcal{D}}\Phi_{\omega}-\left[\mathbf{0}_{k}\;\middle|\;\bar{S}_{\mathcal{D}^{*}}\right]\right\|_{\mathrm{F}}\cdot\left\|\hat{\Phi}_{\omega}\right\|_{\mathrm{F}}\right\} \tag{7.16b}$$

and if we denote by $\tilde{S}_{\mathcal{D}}\triangleq S_{\mathcal{D}}\Phi_{\omega}$, then (6.1) can be expressed as

$$\mu\left(\mathcal{P},\mathcal{P}^{*}\right)=\min_{\forall\mathcal{D}^{*}}\left\{\left\|\tilde{S}_{\mathcal{D}}-\left[\mathbf{0}_{k}\;\middle|\;\bar{S}_{\mathcal{D}^{*}}\right]\right\|_{\mathrm{F}}\cdot\left\|\hat{\Phi}_{\omega}\right\|_{\mathrm{F}}\right\} \tag{7.16c}$$

We may split $\tilde{S}_{\mathcal{D}}$ as follows:

$$\tilde{S}_{\mathcal{D}}=\tilde{S}'_{\mathcal{D}}+\tilde{S}''_{\mathcal{D}} \tag{7.17}$$

such that $\tilde{S}''_{\mathcal{D}}=\left[\mathbf{0}_{k}\;\middle|\;\hat{S}_{\mathcal{D}}^{(2)}\right]$ has the same structure with the reduced resultant $\left[\mathbf{0}_{r}\;\middle|\;\bar{S}_{\mathcal{D}^{*}}\right]$ in (7.16), corresponding to the perturbed set and thus the optimisation is now equivalent to:

$$f\left(\mathcal{P},\mathcal{P}^{*}\right)=\min_{\forall\mathcal{D}^{*}}\left\{\left\|\tilde{S}'_{\mathcal{D}}+\left[\mathbf{0}_{r}\;\middle|\;\hat{S}_{\mathcal{D}}^{(2)}-\bar{S}_{\mathcal{D}^{*}}\right]\right\|\cdot\left\|\hat{\Phi}_{\omega}\right\|_{\mathrm{F}}\right\}$$

The following Theorem (7.2) is a specialisation of Theorem 6.2 in the case of sets of the $\mathcal{T}_{\mathcal{D}}^{n,h}$ family:

Theorem 7.2: Consider the set of polynomials $\mathcal{D}_n^{h+1} \in \Pi_{\mathcal{D}}^{n,h}$, and let $S_{\mathcal{D}}$ be the corresponding Sylvester matrix , then the following hold true:

a) The problem of defining the optimal approximate gcd is equivalent to the solution of two optimisation problems involving different sets of independent variables that is minimise the functions in terms of the free variables $\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}}$ and $\bar{S}_{\mathcal{D}^*}$ i.e.:

$$\mu_1\left(\mathcal{D},\mathcal{D}^*\right)=\left\|\tilde{S}'_{\mathcal{D}}\right\|_{\mathrm{F}}\cdot\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}} \tag{7.18a}$$

$$\mu_2\left(\mathcal{D},\mathcal{D}^*\right)=\left\|\left[\mathbf{0}_k \;\middle|\; \hat{S}_{\mathcal{D}}^{(2)}-\bar{S}_{\mathcal{D}^*}\right]\right\|_{\mathrm{F}}\cdot\left\|\hat{\Phi}_{\upsilon}\right\|_{\mathrm{F}} \tag{7.18b}$$

b) For a given approximate gcd $\upsilon(s)$ of degree k, then:

i) The perturbed set $\tilde{\mathcal{D}}$ corresponding to minimal perturbation applied on $\mathcal{P}$, such that $\upsilon(s)$ becomes the exact gcd of of the perturbed set, is obtained from:

$$\mathcal{D}_{\tilde{\mathcal{P}}}=\tilde{\mathcal{D}}''_{\mathcal{P}}\hat{\Phi}_{\upsilon}=\left[\mathbf{0}_k \;\middle|\; \hat{S}_{\mathcal{D}}^{(2)}\right]\hat{\Phi}_{\upsilon}=\left[\mathbf{0}_k \;\middle|\; \bar{S}_{\mathcal{D}^*}\right]\hat{\Phi}_{\upsilon} \tag{7.19}$$

ii) The strength of the approximate gcd $\upsilon(s)$ of degree k is independent of the coefficients of the perturbation and it is given by:

$$\mu_1\left(\mathcal{D},\mathcal{D}^*\right)=\left\|\tilde{S}'_{\mathcal{D}}\hat{\Phi}_{\omega}\right\|_{\mathrm{F}} \tag{7.20}$$

■

Note that (7.19) result is obtained by setting:

$$\mu_2\left(\mathcal{D},\mathcal{D}^*\right)=0 \tag{7.19b}$$

Theorem (7.2) and the related analysis are applications of the general theory for “approximate factors” of polynomial sets to the case of the restricted structure of the $\Pi_{\mathcal{D}}^{n,h}$ family. Propositions (7.3) and (7.4) denote the invariance of the structure under the solution of optimisation. The question of approximate factorisation is considered next and we will focus on the generalisation of the factorisation formula of Theorem (7.1) to its "approximate" extension. In other words will work on the extraction of almost common factors of order 1 from the $\mathcal{D}_n^{h+1}$.

Assume that "0" is not an approximate root of the initial polynomial. Then by applying the procedure of algorithm (6.1) for the set $\mathcal{D}_n^{h+1}$, a transformation matrix, defined by a polynomial of variable coefficients, $\Phi \in \mathbb{R}^{(2n-1)\times(2n-1)}$ will have the form

$$\Phi = \begin{bmatrix} y_0 & 0 & \cdots & & & & & 0 \\ y_1 & y_0 & \ddots & & & & & \vdots \\ y_2 & y_1 & \ddots & & & & & \\ \vdots & \vdots & \ddots & y_0 & 0 & & & \\ & & & y_1 & y_0 & & & \\ & & & \vdots & \vdots & \ddots & & \\ y_{n-2} & y_{n-3} & \cdots & y_{n-j-2} & y_{n-j-3} & \cdots & y_0 & 0 \\ y_{n-1} & y_{n-2} & \cdots & y_{n-j-1} & y_{n-j-2} & \cdots & y_1 & y_0 \end{bmatrix} \tag{7.21a}$$

where in the case of the transformation matrix Φ_{e_j} for the extraction of the factor $e_j = (s + \lambda_j)$ the y_i parameters satisfy the relationships

$$y_0 = \frac{1}{\lambda_j},\ y_1 = -\left(\frac{1}{\lambda_j}\right)^2, \dots,\ y_\theta = (-1)^{\theta-1}\left(\frac{1}{\lambda_j}\right)^\theta = -\left(-\frac{1}{\lambda_j}\right)^\theta \quad \theta = 2, \dots, n+p-1, \quad \lambda_j \neq 0 \tag{7.21b}$$

Φ_{e_j} is also the inverse of the matrix $\hat{\Phi}_{e_j}$ that is defined by

$$\hat{\Phi}_{e_j} = \begin{bmatrix} \lambda_j & 0 & 0 & \cdots & 0 & 0 \\ 1 & \lambda_j & 0 & \cdots & 0 & 0 \\ 0 & 1 & \lambda_j & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & \cdots & 0 & 1 & \lambda_j & 0 \\ 0 & \cdots & 0 & 0 & 1 & \lambda_j \end{bmatrix} \in \mathbb{R}^{(2n-1)\times(2n-1)} \tag{7.22}$$

The product of the resultant $S_{\mathcal{D}}$ in (7.10) by Φ_{e_j} is given in (7.23)

$$\tilde{S}_{\varphi^{\mathcal{D}}} \triangleq S_{\varphi^{\mathcal{D}}}\Phi_{e_j} = \left[\begin{array}{ccccccccc}
\overline{c}_{0,n} & \overline{c}_{0,n-1} & \cdots & & \overline{c}_{0,0} & 0 & 0 & \cdots & 0 \\
\overline{e}_{0,1} & \overline{c}_{0,n} & \overline{c}_{0,n-1} & \cdots & & \overline{c}_{0,0} & 0 & & \vdots \\
\vdots & \ddots & \ddots & & & & \ddots & \ddots & \\
\overline{e}_{0,p-2} & \cdots & \overline{e}_{0,1} & \overline{c}_{0,n} & \overline{c}_{0,n-1} & & & \overline{c}_{0,0} & 0 \\
\overline{e}_{0,p-1} & \overline{e}_{0,p-3} & \cdots & \overline{e}_{0,1} & \overline{c}_{0,n} & \overline{c}_{0,n-1} & \cdots & & \overline{c}_{0,0} \\
\hdashline
 & & & & \vdots & \vdots & & & \\
\hdashline
\overline{c}_{i,p} & \overline{c}_{i,p-1} & \cdots & \overline{c}_{i,0} & 0 & \cdots & & & 0 \\
\overline{e}_{i,1} & \overline{c}_{i,p} & \overline{c}_{i,p-1} & \cdots & \overline{c}_{i,0} & 0 & & & 0 \\
\overline{e}_{i,2} & \ddots & \overline{c}_{i,p} & \overline{c}_{i,p-1} & \cdots & \overline{c}_{i,0} & 0 & & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & & \ddots & \ddots & \vdots \\
\overline{e}_{i,n-2} & & \overline{e}_{i,2} & \overline{e}_{i,1} & \overline{c}_{i,p} & \overline{c}_{i,p-1} & & \overline{c}_{i,0} & 0 \\
\overline{e}_{i,n-1} & \overline{e}_{i,n-2} & & \overline{e}_{i,2} & \overline{e}_{i,1} & \overline{c}_{i,p} & \overline{c}_{i,p-1} & \cdots & \overline{c}_{i,0} \\
\hdashline
 & & & & \vdots & \vdots & & & \\
\hdashline
\overline{c}_{h,p} & \overline{c}_{h,p-1} & \cdots & \overline{c}_{h,0} & 0 & \cdots & & & 0 \\
\overline{e}_{h,1} & \overline{c}_{h,p} & \overline{c}_{h,p-1} & & \overline{c}_{h,0} & 0 & & & 0 \\
\overline{e}_{h,2} & \overline{e}_{h,1} & \overline{c}_{h,p} & \overline{c}_{h,p-1} & & \overline{c}_{h,0} & 0 & & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & & \ddots & \ddots & \vdots \\
\overline{e}_{h,n-2} & & \overline{e}_{h,2} & \overline{e}_{n,1} & \overline{c}_{h,p} & \overline{c}_{h,p-1} & \cdots & \overline{c}_{h,0} & 0 \\
\overline{e}_{h,n-1} & \overline{e}_{h,n-2} & \cdots & \overline{e}_{h,2} & \overline{e}_{h,1} & \overline{c}_{h,p} & \overline{c}_{h,p-1} & \cdots & \overline{c}_{h,0}
\end{array}\right] \tag{7.23a}$$

where

$$\overline{c}_{0,n-\sigma} = -\sum_{\theta=0}^{n-\sigma}\left[b_{n-\sigma-\theta}\left(-\frac{1}{\lambda_j}\right)^{\theta+1}\right] = -\sum_{\theta=0}^{n-\sigma}\left[b_{n-\sigma-\theta}\left(-\lambda_j\right)^{-(\theta+1)}\right], \quad \sigma = 0,1,\ldots,n,$$

$$\overline{e}_{0,\mu} = -\sum_{\theta=0}^{n}\left[b_{n-\theta}\left(-\lambda_j\right)^{-(\mu+\theta)}\right], \quad \mu = 1,\ldots,n-1$$

$$\overline{c}_{i,n-i-\sigma} = -\sum_{\theta=0}^{(n-i)-\sigma}\left[b_{i,(n-i)-\sigma-\theta}\left(-\lambda_j\right)^{-(\theta+1)}\right] = -\sum_{\theta=0}^{(n-i)-\sigma}\left[f_{(n-\sigma-\theta),i}b_{n-\sigma-\theta}\left(-\lambda_j\right)^{-(\theta+1)}\right], \quad \sigma = 0,1,\ldots,n-i,$$

$$\overline{e}_{i,\nu} = -\sum_{\theta=0}^{n-i}\left[f_{(n-\theta),i}b_{n-\theta}\left(-\lambda_j\right)^{-(\nu+\theta)}\right], \quad \nu = 1,\ldots,n-2 \tag{7.23b}$$

Using the optimisation results of Theorem (6.2), we split $\tilde{S}_{\mathcal{D}}$ as follows:

$$\tilde{S}_{\mathcal{D}} = \tilde{S}'_{\mathcal{D}} + \tilde{S}''_{\mathcal{D}} \tag{7.24}$$

such that

$$\tilde{S}'_{\mathcal{D}} = \left[\begin{array}{ccccccc}
\bar{c}_{0,n} & 0 & \cdots & 0 & 0 & \cdots & 0 \\
\bar{e}_{0,1} & \bar{c}_{0,n} & 0 & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & & \ddots & \vdots \\
\bar{e}_{0,p-1} & \cdots & \bar{e}_{0,1} & \bar{c}_{0,n} & 0 & \cdots & 0 \\
\hdashline
 & & \vdots & & \vdots & & \\
\hdashline
\bar{c}_{h,p} & 0 & \cdots & 0 & 0 & \cdots & 0 \\
\bar{e}_{h,1} & \bar{c}_{h,p} & 0 & \cdots & 0 & & 0 \\
\vdots & \ddots & & \ddots & & \ddots & \vdots \\
\bar{e}_{h,n-1} & \cdots & \bar{e}_{h,1} & \bar{c}_{h,p} & 0 & \cdots & 0
\end{array}\right] \tag{7.24a}$$

and

$$\tilde{S}'_{\mathcal{D}} = \left[\begin{array}{ccccccc}
0 & \bar{c}_{0,n-1} & \cdots & \bar{c}_{0,0} & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \bar{c}_{0,n-1} & & \bar{c}_{0,0} & 0 \\
0 & \cdots & 0 & 0 & \bar{c}_{0,n-1} & \cdots & \bar{c}_{0,0} \\
\hdashline
 & & \vdots & & \vdots & & \\
\hdashline
0 & \bar{c}_{h,p-1} & \cdots & \bar{c}_{h,0} & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \bar{c}_{h,p-1} & \cdots & \bar{c}_{h,0} & 0 \\
0 & \cdots & 0 & 0 & \bar{c}_{h,p-1} & \cdots & \bar{c}_{h,0}
\end{array}\right] \tag{7.24b}$$

Using the results of Chapter 6, and more specifically those of Theorem (6.3), the best approximate factor of $\mathcal{D}_n^{h+1}$ can be found by the minimisation of $\left\| \tilde{S}'_{\mathcal{D}} \hat{\Phi}_{e_j} \right\|_F$. Theorem (6.4) and in particular conditions (7.22), (7.23) and (7.24b) imply that the specific form of the optimisation problem is described by the following theorem:

Theorem (7.3): The best approximate factor $e_j = \left(s + \lambda_j\right)$ of multiplicity h is obtained from the minimisation of $\min \left\| \tilde{S}'_{\mathcal{D}} \hat{\Phi}_{e_j} \right\|$, where

$$\tilde{S}'_{\mathcal{D}}\hat{\Phi}_{e_j} = \left[\begin{array}{ccccccc} z_{0,n} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & z_{0,n} & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \vdots \\ 0 & \cdots & 0 & z_{0,n} & 0 & \cdots & 0 \\ \hline & & \vdots & & \vdots & & \\ \hline z_{h,p} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & z_{h,p} & 0 & \cdots & 0 & & 0 \\ \vdots & \ddots & & \ddots & & \ddots & \vdots \\ 0 & \cdots & 0 & z_{h,p} & 0 & \cdots & 0 \end{array}\right] \tag{7.25a}$$

and

$$z_{0,n} = \lambda_j \bar{c}_{0,n} = -\lambda_j \sum_{\theta=0}^{n-\sigma} \left[b_{n-\sigma-\theta} \left(-\lambda_j\right)^{-(\theta+1)} \right] = \sum_{\theta=0}^{n} \left(-\lambda_j\right)^{-\theta} b_{n-\theta} \ ,$$

$$z_{i,p} = \lambda_j \bar{c}_{i,p} = -\lambda_j \sum_{\theta=0}^{n-\sigma} \left[b_{n-\sigma-\theta} \left(-\lambda_j\right)^{-(\theta+1)} \right] = \sum_{\theta=0}^{p} = \lambda_j^{-\theta} f_{(n-\theta),i} b_{n-\theta} \tag{7.25b}$$

■

Note that the above result in (7.14) is for the case of the extraction of the monic 1-order factor $e_j = \left(s + \lambda_j\right)$. The above analysis is sufficient in order to relate the "approximate factorisation" of $b(s)$ with the approximate common factors of the set $\mathcal{D}_n^{h+1}$. Thus the problem can be solved using Algorithm (6.1).

In Chapter 5, we have introduced the *strength* of the approximation as a quality criterion. The strength refers to all perturbations applied on the polynomials. In the case of the approximate normal factorisation, the perturbations are dependent on the perturbation on the initial polynomial. Thus the accuracy ε_{app} is related to the strength and is an alternative criterion.

The first step in an algorithm for the extraction of an approximate factor from $\mathcal{D}_n^{h-1}$ is to determine h, in other words the number of the, additional to the initial, Toeplitz blocks of the resultant corresponding to $1, 2, ..., h$-order derivatives. This is chosen to be the higher order of derivative for which there exists loss of numerical rank [Foster, 1986]. The evaluation of the numerical rank is based on the following theorem.

Definition (7.3) [Foster, 1986]: For a matrix $A \in \mathbb{R}^{m\times n}$

$\rho_\varepsilon(A) =$ number of singular values of A that are $\leq \varepsilon$

■

The results are demonstrated by the following example:

Example 7.1: Let us consider a set $\mathcal{D}^h(p(s))$ generated of 6-degree polynomial

$$p(s) = s^6 + 1.01s^5 - 4.97s^4 - 1.07s^3 + 7.93s^2 - 3.82s + 0.08$$

and its h first derivatives, where we now investigate the clustering of a number of roots. Note that the derivatives have the form:

$$p'(s) = 6s^5 + 5.05s^4 - 19.88s^3 - 3.21s^2 + 15.86s - 3.82 \ ,$$

$$p''(s) = 30s^4 + 20.2s^3 - 59.64s^2 - 6.42s - 15.86 \ ,$$

$$p^{(3)}(s) = 120s^3 + 60.6s^2 - 119.28s - 6.42 \ , \ldots$$

Evaluation of the approximate factor with higher multiplicity

For the following operations the appropriate functions of Matlab 6 are used. For the polynomial set $\mathcal{D}^3(p(s)) = \{p(s), p'(s), p''(s), p^{(3)}(s)\}$ the respective Sylvester matrix has minimum singular value $\sigma = 1.36$ which implies that the resultant has clearly full column rank and thus an approximate factor of degree 4 cannot be established. The next step is now to examine the set $\mathcal{D}^2(p(s)) = \{p(s), p'(s), p''(s)\}$. The resultant of this set has the form

$$S_{\mathcal{D}^2}=\left[\begin{array}{ccccccccccc}
1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 & 0 & 0 & 0 & 0 \\
0 & 1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 & 0 & 0 & 0 \\
0 & 0 & 1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 & 0 & 0 \\
0 & 0 & 0 & 1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 & 0 \\
0 & 0 & 0 & 0 & 1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 \\
\hline
6 & 5.05 & -19.88 & -3.21 & 15.86 & -3.82 & 0 & 0 & 0 & 0 & 0 \\
0 & 6 & 5.05 & -19.88 & -3.21 & 15.86 & -3.82 & 0 & 0 & 0 & 0 \\
0 & 0 & 6 & 5.05 & -19.88 & -3.21 & 15.86 & -3.82 & 0 & 0 & 0 \\
0 & 0 & 0 & 6 & 5.05 & -19.88 & -3.21 & 15.86 & -3.82 & 0 & 0 \\
0 & 0 & 0 & 0 & 6 & 5.05 & -19.88 & -3.21 & 15.86 & -3.82 & 0 \\
0 & 0 & 0 & 0 & 0 & 6 & 5.05 & -19.88 & -3.21 & 15.86 & -3.82 \\
\hline
30 & 20.2 & -59.64 & -6.42 & 15.86 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 30 & 20.2 & -59.64 & -6.42 & 15.86 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 30 & 20.2 & -59.64 & -6.42 & 15.86 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 30 & 20.2 & -59.64 & -6.42 & 15.86 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 30 & 20.2 & -59.64 & -6.42 & 15.86 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 30 & 20.2 & -59.64 & -6.42 & 15.86 & 0
\end{array}\right] \tag{7.26a}$$

and the singular values of $S_{\mathcal{D}^2}$ are approximately:

$$\left\{106.06,\ 103.44,\ 76.52,\ 69.07,\ 44.26,\ 31.89,\ 15.86,\ 8.41,\ 1.06,\ 0.71,\ 5\cdot 10^{-15}\right\}$$

where from Theorem (7.4) for a very small accuracy $\varepsilon = 10^{-14}$ we have numerical ε-nullity $\mathcal{N}_{\varepsilon}\left(S_{\mathcal{D}^2}\right)=1$. This implies that for the above accuracy there exists an approximate factor of order 1 of the polynomial set $\mathcal{D}^2\left(p(s)\right)=\left\{p(s),p'(s),p''(s)\right\}$. In other words there exists exactly one elementary factor $e_1(s)=(s+\lambda_1)$ of multiplicity 3 with accuracy $\varepsilon=10^{-14}$. That means that $p(s)=(s+\lambda_1)^3\,p_1(s)+q_1(s)$.

The factor $e_1(s)$ is obtained from the minimisation of $\tilde{S}'_{\mathcal{D}}\hat{\Phi}_{e_1}$. By (7.25b) and (7.23b) it is implied that $\left\|\tilde{S}'_{\mathcal{D}}\hat{\Phi}_{e_1}\right\|_{\mathrm{F}}^{2}$ has the form:

$$\left\|\tilde{S}'_{\mathcal{D}}\hat{\Phi}_{e_1}\right\|_F^2 = 5\left(1-\frac{1.01}{\lambda_1}-\frac{4.97}{\lambda_1^2}+\frac{1.07}{\lambda_1^3}+\frac{7.93}{\lambda_1^4}+\frac{3.82}{\lambda_1^5}-\frac{0.08}{\lambda_1^6}\right)^2 +$$

$$+6\left(6-\frac{5.05}{\lambda_1}-\frac{19.88}{\lambda_1^2}+\frac{3.21}{\lambda_1^3}+\frac{15.86}{\lambda_1^4}+\frac{3.82}{\lambda_1^5}\right)^2 +$$

$$+6\left(30-\frac{20.2}{\lambda_1}-\frac{59.64}{\lambda_1^2}+\frac{6.42}{\lambda_1^3}+\frac{15.86}{\lambda_1^4}\right)^2$$

or

$$\begin{aligned}\left\|\tilde{S}'_{\mathcal{D}}\hat{\Phi}_{e_1}\right\|_F^2 = \big(&5w^{12}+10.1w^{11}+171.4005w^{10}+302.703w^9+4313.6525w^8+\\ &+5931.224w^7-16131.2581w^6-15211.8874w^5+21901.0151w^4+\\ &+8437.4316w^3-9380.3844w^2-1945.8208w+1596.824\big)\end{aligned} \tag{7.26b}$$

where $w_1 = (\lambda_1)^{-1}$ and the global minimum of this polynomial is estimated using a standard minimisation procedure at $w_1 = -1$. This implies that the best approximate root is $r_1 = -\lambda_1 = 1$. This is an approximate triple root of the initial polynomial. Now the factor has to be extracted for the first polynomial and an investigation for possible other factors of multiplicity 2 will follow. We do not have to investigate for possible triple factors because their existence will contradict the fact that $\mathcal{N}_\varepsilon\left(S_{\mathcal{D}^2}\right)=1$.

Extraction of the factor $(s-1)^3 = s^3-3s^2+3s-1$

The extraction can be carried out by Euclidean division or by using the equivalent methodology described in Chapter 6 for the approximate factorisation of the resultant. Note here that the factorisation involves only the first block of the resultant that corresponds to the initial polynomial. That is:

$$S_{\mathcal{D}^0} = \begin{bmatrix} 1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1.01 & -4.97 & -1.07 & 7.93 & -3.82 & -0.08 \end{bmatrix}$$

(7.26c)

Using Theorem (6.2) and condition (6.10) the transformation matrices associated with the factor $(s-1)^3 = s^3 - 3s^2 + 3s - 1$ have the form

$$S_{\tilde{\mathcal{D}}^0} = \tilde{S}''_{\mathcal{D}^0}\hat{\Phi}_{e_1} = \left[\mathbf{0}_1 \;\middle|\; \hat{S}^{(2)}_{\mathcal{D}^0}\right]\hat{\Phi}_{e_1} \tag{7.26d}$$

where

$$\hat{\Phi}_{e_1} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -3 & 3 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -3 & 3 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -3 & 3 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -3 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 3 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -3 & 3 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -3 & 3 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -3 & 3 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -3 & 3 & -1 \end{bmatrix} \tag{7.26e}$$

and

$$\Phi_{e_1} = \left(\hat{\Phi}_{e_1}\right)^{-1} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -3 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -6 & -3 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -10 & -6 & -3 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -15 & -10 & -6 & -3 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -21 & -15 & -10 & -6 & -3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -28 & -21 & -15 & -10 & -6 & -3 & -1 & 0 & 0 & 0 & 0 \\ -36 & -28 & -21 & -15 & -10 & -6 & -3 & -1 & 0 & 0 & 0 \\ -45 & -36 & -28 & -21 & -15 & -10 & -6 & -3 & -1 & 0 & 0 \\ -55 & -45 & -36 & -28 & -21 & -15 & -10 & -6 & -3 & -1 & 0 \\ -66 & -55 & -45 & -36 & -28 & -21 & -15 & -10 & -6 & -3 & -1 \end{bmatrix} \tag{7.26f}$$

from Theorem (6.2) and using the matrix operations we have

$$\tilde{S}''_{\mathcal{D}^0} = \left[\begin{array}{ccc|cccccccc} 0 & 0 & 0 & 1 & 4.01 & 4.06 & 0.08 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 4.01 & 4.06 & 0.08 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 4.01 & 4.06 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4.01 & 4.06 & 0.08 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4.01 & 4.06 & 0.08 \end{array}\right] \tag{7.26g}$$

The last row of (7.26g) gives the quotient polynomial after the extraction of $e_1(s)^3$. The next step is to investigate and extract possible multiple factors from the quotient polynomial $f_1(s) = s^3 + 4.01s^2 + 4.06s + 0.08$.

Let us denote $\tilde{\mathcal{D}}_h$ as the set of derivatives, i.e. $\{f_1(s), f_1'(s), f_1''(s), ..., f_1^{(h)}(s)\}$. For $\tilde{\mathcal{D}}_1$ the respective resultant is

$$S_{\tilde{\mathcal{D}}_1} = \left[\begin{array}{ccccc} 1 & 4.01 & 4.06 & 0.08 & 0 \\ 0 & 1 & 4.01 & 4.06 & 0.08 \\ \hline 3 & 8.02 & 4.06 & 0 & 0 \\ 0 & 3 & 8.02 & 4.06 & 0 \\ 0 & 0 & 3 & 8.02 & 4.06 \end{array}\right]$$

and its singular values are approximately $\{15.12,\ 9.48,\ 4.11,\ 1.03,\ 2.5\cdot 10^{-6}\}$. Thus with an accuracy $\varepsilon = 10^{-5}$ we have an approximate common factor of $\tilde{\mathcal{D}}_1$, i.e. there is an approximate factor $e_2(s) = s + \lambda_2$ of $f_1(s)$ of multiplicity 2. Then,

$$\left\| \tilde{S}'_{\mathcal{D}} \hat{\Phi}_{e_2} \right\|_F^2 = 2\left(1 - \frac{4.01}{\lambda_2} + \frac{4.06}{\lambda_2^2} - \frac{0.08}{\lambda_3^3}\right)^2 + 3\left(3 - \frac{8.02}{\lambda_1} + \frac{4.06}{\lambda_1^2}\right)^2 = \tag{7.26h}$$
$$\simeq 2w_2^6 + 16w_2^5 + 75.4w_2^4 - 209.8w_2^3 + 300.3w_2^2 - 196.7w_2 + 49.5$$

where $w_2 = \dfrac{1}{\lambda_2}$ and when(7.26h) is minimised, then the approximate solution is at $w_2 = 0.54$ or $\lambda_2 = 1.85$. The last implies that the approximate factor is $e_2(s) = s + 1.8367$.

Extraction of $e_2(s) = s^2 + 3.6736s + 3.3738$ *from* $f_1(s)$

Note that

$$\hat{\Phi}_{e_2} = \begin{bmatrix} 3.3738 & 0 & 0 & 0 & 0 \\ 3.7 & 3.3738 & 0 & 0 & 0 \\ 1 & 3.7 & 3.3738 & 0 & 0 \\ 0 & 1 & 3.7 & 3.3738 & 0 \\ 0 & 0 & 1 & 3.7 & 3.42 \end{bmatrix}, \quad \Phi_{e_2} = \begin{bmatrix} 0.292 & 0 & 0 & 0 & 0 \\ -0.316 & 0.292 & 0 & 0 & 0 \\ 0.257 & -0.316 & 0.292 & 0 & 0 \\ -0.186 & 0.257 & -0.316 & 0.292 & 0 \\ 0.125 & -0.186 & 0.257 & -0.316 & 0.292 \end{bmatrix}$$

and

$$\tilde{S}''_{\mathcal{D}^0} = \left[\begin{array}{cc|ccc} 0 & 0 & 1.177 & 0.024 & 0 \\ 0 & 0 & 0 & 1.177 & 0.024 \end{array}\right]$$

and this defines the last quotient polynomial $f_2(s) = 1.177s + 0.024$

Thus the approximate normal factorisation of $p(s)$ is

$$\tilde{p}(s)=(s-1)^3(s+1.8367)^2(1.177s+0.024)$$

■

A theoretic algorithm may now be defined. Note that this algorithm is not optimised in terms of complexity, but it is just a theoretic evaluation approach for the optimal approximate factorisation.

Evaluation of Approximate Factor (Algorithm 7.1):

Given a polynomial $b(s)\in\mathbb{R}[s]$ the approximate factorisation follows the steps:

i) Investigation of the existence and extraction of s^{k_0} factor

- For a specified accuracy $\varepsilon_0>0$ find the maximum non-negative integer k_0 such that s^{k_0} is an ε_0-approximate factor of $b(s)$, $f_0(s)=b_n s^{n-k_0}+\ldots+b_{n-k_0+1}s+b_{n-k_0}$
- Eliminate the last k_0 elements of $b(s)$
- $\|q_0(s)\|=\left\|b(s)-f_0(s)s^{k_0}\right\|=\left\|b_{k_0-1}s^{k_0-1}+\ldots+b_1 s+b_0\right\|$

ii) Extraction of proper approximate common factor from the set consists of f_0 and its derivatives $\mathcal{P}^{\mathcal{D}}(f_0)$

- Construction of the transformation free-variable matrices Φ_{e_1}, $\hat{\Phi}_{e_1}$, corresponding to $e_1(s)$, where $\deg\{e_1(s)\}=1$
- Construction of $S_{h_1}\triangleq S_{\mathcal{P}^{\mathcal{D}}(f_0),h_1}$ where $h_1=\max\left\{0,1,\ldots,n-k_0:\ \mathcal{N}_{\varepsilon}\left(S_{h_1}\right)\geq 1\right\}$
- Construction of the matrices $\tilde{S}_{h_1}=S_{h_1}\Phi_{e_1}$, $\tilde{S}'_{h_1}$, $\tilde{S}'_{h_1}\hat{\Phi}_{e_1}$
- Minimisation of $\left\|\tilde{S}'_{h_1}\hat{\Phi}_{e_1}\right\|$ implies $e_1(s)$

- Derive $f_1(s)$ from $f_0(s)=\left(e_1(s)\right)^{\tau_1} f_1(s)+q_1(s)$ with extraction of $\left(e_1(s)\right)^{\tau_1}$ from the first polynomial, using factorisation properties (Chapter 4 and Chapter 5) on the first block.
- $\left\| \tilde{S}'_{h_1} \hat{\Phi}_{e_1} \right\|$ is the strength of the approximation. Alternatively the error can be found by $\left\|q_1(s)\right\| = f_0(s)-e_1(s)^{\tau_1} f_1(s)$

iii) Repeat the above procedure for $f_i(s)$ for $i=1,\dots,\sigma$ where σ is the maximum positive integer for which $f_i(s)$ and its first derivative $f_i'(s)$ have approximate common factor, i.e. $\sigma+1=\min\left\{k\in Z_+ : \mathcal{N}_\varepsilon\left(S\left(f_k(s),f_k'(s)\right)\right)=0\right\}$ (termination criterion)

iv) Best approximate factorization of is $s^{k_0} e_1(s)^{\tau_1}\cdots e_\sigma(s)^{\tau_\sigma} f_\sigma(s)$

v) Final Error polynomial: $\varepsilon(s)=b(s)-s^{k_0} e_1(s)^{\tau_1}\cdots e_\sigma(s)^{\tau_\sigma} f_\sigma(s)$

■

7.4. DISCUSSION

The results on the Optimal Approximate GCD in Chapter 6 have been combined with the theory of Normal Factorisation of polynomials [Karcanias et al., 2000] to create an algorithm for the computation of approximate factorisation of a polynomial. The algorithm introduced in this chapter includes several steps and iterations that increase its complexity. This can be reduced in practice with simplification of the procedure of change of the operational order.

The importance of the algorithm is that provides a general theoretical framework for the approximate factorisation. The results on the "best approximate factorisation" provide the means for defining polynomials close to the original one with clustered roots. This can be the basis for many calculations in the algebra of polynomials, where root clustering may be the vehicle for defining approximate algebraic forms, such as “approximate Smith forms” etc, where standard procedures of the algebraic nature will automatically lead to coprimeness, since they will fail to define the notions of approximate common factors.

Chapter 8:

APPROXIMATE DECOUPLING ZERO POLYNOMIALS AND APPROXIMATE ZEROS OF LINEAR SYSTEMS

8.1 INTRODUCTION

The notion of almost zeros and almost decoupling zeros for a linear system has been introduced in [Karcanias et al., 1983] and their properties have been linked to mobility of poles under compensation. The basis of that definition has been the use of Grassman polynomial vectors to define system invariant and the definition of "almost zeros" of a set of polynomials as the minima of a function associated with the polynomial vector [Karcanias et al., 1983]. In this chapter we use the exterior algebra framework introduced in [Karcanias et al., 1983], expand it by introducing some new invariants and then use the results on the approximate gcd defined before to define the notion of "approximate input-output decoupling zero polynomial" and "approximate zero polynomial". The current framework allows the characterisation of strength of the given order approximate zero polynomial, as well as permits the characterisation of the optimal approximate solutions of a given order. As such, the current approach extends the result in [Karcanias et al., 1983] by introducing approximate polynomials, rather than simple frequencies (roots) and by defining the "strength" of such approximate solutions.

The results allow the definition of new measures of distance of systems from uncontrollability, unobservabillity using the "strength" associated with a given approximate polynomial, and this is another advantage of the current approach. The use of Grassmann vectors, that is polynomial vectors in a projective space, implies that the general results on the "strength" of approximation yield upper bounds for the corresponding approximate polynomials, when these are defined in the affine space set up. The current approach makes a distinction between the Grassmann input-state and state-output polynomial vectors introduced using the controllability, observability pencils and their restricted versions [Karcanias et al., 1994]. This distinction permits the definition of a classification of almost input-decoupling (output-decoupling) zero polynomials into strong and weak. This classification reflects their property of being dependent, or invariant under feedback (state, respectively output injection).

The chapter is structured as follows: We review first the exterior algebra framework that leads to the Grassmann invariants characterising system properties and then we use the previous results to define the approximate zero polynomial notions and the corresponding properties. The background on the DAP problems in Control and the definition of Grassmann invariants follow [Karcanias et al., 1983].

8.2 DETERMINAL ASSIGNMENT PROBLEMS IN CONTROL THEORY

Consider the linear system described by

$$S(A,B,C,D)\ :\ \begin{array}{ll} \dot{\underline{x}} = A\underline{x}+B\underline{u} \ , & A\in\mathbb{R}^{n\times n} \ , \ B\in\mathbb{R}^{n\times p} \\ \underline{y} = C\underline{x}+D\underline{u} \ , & C\in\mathbb{R}^{m\times n} \ , \ D\in\mathbb{R}^{m\times p} \end{array} \tag{8.1}$$

where (A,B) is controllable, (A,C) is observable, or by the transfer function matrix $G(s)=C(sI-A)^{-1}B+D$, where $\operatorname{rank}\{G(s)\}=\min\{m,p\}$. In terms of left, right coprime matrix fraction descriptions (LCMFD, RCMFD), $G(s)$may be represented as

$$G(s)=D_l(s)^{-1}\ N_l(s)\ =\ N_r(s)\ D_r(s)^{-1} \tag{8.2}$$

where $N_l(s)\ N_r(s)\in\mathbb{R}^{m\times p}[s]$, $D_l(s)\in\mathbb{R}^{m\times m}[s]$ and $D_r(s)\in\mathbb{R}^{p\times p}[s]$The system will be called *square* if $m=p$and *nonsquare* if $m\neq p$. Within the state space framework we may define the following frequency assignment problems :

(i) Pole assignment by state feedback: Consider $L\in\mathbb{R}^{n\times p}$, where L is a state feedback applied on system (8.1). The closed loop characteristic polynomial is then given by

$$P_L(s)=\det\{sI-A-BL\}=\det\ \{B(s)\,\tilde{L}\} \tag{8.3}$$

where $B(s)=[sI-A,-B]$ and $\tilde{L}=\left[I_n,L^t\right]^t$.

■

(ii) Design of an n-state observer: Consider the problem of designing an n-state observer for the system of (8.1). The characteristic polynomial of the observer is then defined by

$$P_T(s)=\det\{sI-A-TC\}=\det\tilde{T}C(s) \tag{8.4}$$

where $T\in\mathbb{R}^{n\times m}$ is a feedback, $\tilde{T}=[I_n,T]$ and $C(s)=\left[sI-A^t,-C^t\right]^t$ ■

(iii) <u>Pole assignment by constant output feedback</u>: Consider the system described by (8.2) under an output feedback $K \in \mathbb{R}^{m \times p}$. The closed loop characteristic polynomial $P_K(s)$ is given by [Kailath, 1980]:

$$P_K(s) = \det\{D_l(s) + N_l(s)K\} = \det\{D_r(s) + KN_r(s)\} \tag{8.5}$$

By defining the matrices

$$T_l\ (s) = \left[D_l(s),\ N_l(s)\right] \in \mathbb{R}^{m \times (m+l)}[s],\ T_r(s) = \begin{bmatrix} D_r(s) \\ N_r(s) \end{bmatrix},\ \tilde{K}_1 = \left[I_m, K^t\right]^t \in \mathbb{R}^{(m+p) \times m},$$

$$\tilde{K}_r = \left[I_1, K\right] \in \mathbb{R}^{p \times (m+p)} \tag{8.6}$$

then,

$$P_K(s) = \det\{T_1(s)K_1\} = \det\{\tilde{K}_r T_r(s)\} \tag{8.7}$$

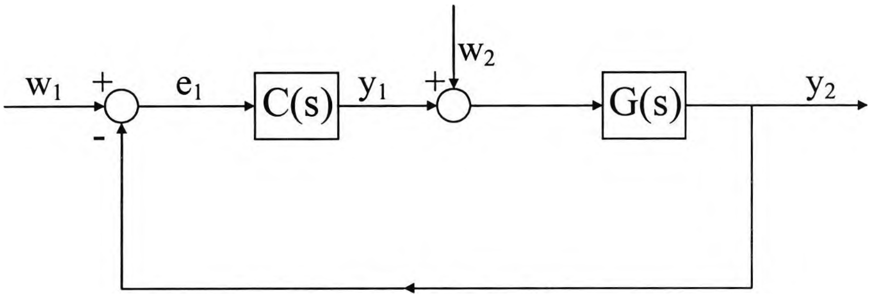
■

(iv) <u>Zero assignment by squaring down</u>: For a system with $m > p$ we can expect to have independent control over at most p linear combinations of m outputs. If $\underline{c} \in \mathbb{R}^p$ is the vector of the variables which are to be controlled, then $\underline{c} = Hy$ where $H \in \mathbb{R}^{p \times m}$ is a *squaring down postcompensator*, and $G'(s) = HG(s)$ is the *squared down transfer function matrix* [Karcanias, 1989], [Kouvaritakis et al., 1976]. A right MFD for $G'(s)$ is defined by $G'(s) = HN_r(s)D_r(s)^{-1}$ where $G(s) = N_r(s)D_r(s)^{-1}$. Finding H such that $G'(s)$ has assigned zeros is defined as the *zero assignment by squaring down* problem. The zero polynomial of $S(A, B, HC, HD)$ is given by

$$z_K(s) = \det\{HN_r(s)\} \tag{8.8}$$

Note that the above problem belongs to the general class of model projection problems [Kar.1]. A larger family of problems of determinantal type problemsare associated with dynamic compensation and are considered next.

(v) **<u>Dynamic Compensation Problems</u>:** Consider the standard feedback configuration [Kuc.1] below



If $G(s) \in \mathbb{R}_{\text{pr}}^{m\times p}[s]$, $C(s) \in \mathbb{R}^{p\times m}[s]$, and assume coprime MFD's as in (8.2) and

$$C(s) = A_1(s)^{-1} B_1(s) = B_r(s) A_{-1}(s)^{-1} \tag{8.9}$$

Then, the closed loop characteristic polynomial may be expressed as

$$f(s) = \det\left\{\left[D_l(s),\ N_l(s)\right]\begin{bmatrix} A_r(s) \\ B_r(s) \end{bmatrix}\right\} \tag{8.10}$$

$$f(s) = \det\left\{\left[A_l(s),\ B_l(s)\right]\begin{bmatrix} D_r(s) \\ N_r(s) \end{bmatrix}\right\} \tag{8.11}$$

i) if $p \le m$, then $C(s)$ may be interpreted as FEEDBACK COMPENSATOR and we will use the expression of the closed loop polynomial described by (8.11)

ii) if $p \geq m$, the $C(s)$ may be interpreted as PRECOMPENSATOR and we will use the expression of the closed loop polynomial described by (8.10)

The above general dynamic formulation covers a number of important families of C(s) compensators as :

(a) Constant, (b) PI, (c) PD, (d) PID, (e) Bounded degree

In fact

(a) Constant Controllers: If $p \leq m$, $A_1 = I_p$, $B_1 = K \in \mathbb{R}^{p\times m}$, then (8.11) expresses the constant output feedback case, whereas if $p \geq m$, $A_r = I_m$, $B_r = K \in \mathbb{R}^{p\times m}$ expresses the constant precompensation formulation of the problem.

(b) Proportional plus Integral Controllers: Such controllers are defined by

$$C(s) = K_0 + \frac{1}{s}K_1 = \left[sI_p\right]^{-1}\left[sK_0 + K_1\right] \tag{8.12}$$

where $K_0, K_1 \in \mathbb{R}[s]^{p\times m}$ and the left MFD for $C(s)$ is coprime, iff $\operatorname{rank}\{K_1\} = p$. From the above and (2.11), the determinantal problem for the output feedback PI design is expressed as :

$$f(s) = \det\left\{[sI_p,\ sK_0 + K_1]\begin{bmatrix} D_r(s) \\ N_r((s)) \end{bmatrix}\right\} = \det\left\{[I_p,\ K_0,\ K_1]\begin{bmatrix} sD_r(s) \\ sN_r(s) \\ N_r(s) \end{bmatrix}\right\} \tag{8.13}$$

(c) Proportional plus Derivative Controllers: Such controllers are expressed as

$$C(s) = sK_0 + K_1 = \left[I_p\right]^{-1}\left[sK_0 + K_1\right] \tag{8.14}$$

where $K_0, K_1 \in \mathbb{R}^{p\times m}[s]$ and the left MFD for $C(s)$ is coprime for finite s and also for $s = \infty$ if $\operatorname{rank}(K_0) = p$. From the above and (8.11) the determinantal output PD feedback is expressed as:

$$f(s) = \det\left\{\left[I_p,\ sK_0 + K_1\right]\begin{bmatrix} D_r(s) \\ N_r(s) \end{bmatrix}\right\} = \det\left\{\left[I_p,\ K_1,\ K_0\right]\begin{bmatrix} D_r(s) \\ N_r(s) \\ sN_r(s) \end{bmatrix}\right\} \tag{8.15}$$

(d) PID Controllers: These controllers are expressed as

$$C(s) = K_0 + \frac{1}{s}K_1 + sK_2 = \left[sI_p\right]^{-1}\left[s^2K_2 + sK_0 + K_1\right] \tag{8.16}$$

where $K_0, K_1 \in \mathbb{R}^{p\times m}$ and the left MFD is coprime with the only exception possibly at $s=0$, $s=\infty$ (coprimeness at $s=0$ is guaranteed by $\operatorname{rank}(K_1)=p$ and at $s=\infty$ by $\operatorname{rank}(K_2)=p$). From (8.11), the determinantal output PID feedback is expressed as :

$$f(s)=\det\left\{\left[sI_p,\ s^2K_2+sK_0+K_1\right]\begin{bmatrix}D_r(s)\\N_r(s)\end{bmatrix}\right\}=$$
$$=\det\left\{\left[I_p,\ K_0,\ K_1,\ K_2\right]\begin{bmatrix}sD_r(s)\\sN_r(s)\\N_r(s)\\s^2N_r(s)\end{bmatrix}\right\} \tag{8.17}$$

(e) <u>Observability Index Bounded Dynamics (OBD) Controllers</u>: These are defined by the property that their McMillan degree is equal to pk, where k is the observability index of the controller. Such controllers are expressed as in (2.9) where

$$\left[A_1(s),B_1(s)\right]=T_ks^k+\ldots+T_0 \tag{8.18}$$

$T_k+T_{k-1}+\ldots+T_0\in\mathbb{R}^{p\times(p\times m)}$ and $T_k=\left[I_p,X\right]$. Note that the above representation is not always coprime, and coprimeness has to be guaranteed first for McMillan degree to be pk; otherwise, the McMillan degree is less than pk. The dynamic determinantal OBD output feedback problem is expressed from (8.11) as

$$f(s)=\det\left\{[T_ks^k\ +\ldots\ +T_0]\begin{bmatrix}D_r(s)\\N_r(s)\end{bmatrix}\right\}$$

or

$$f(s)=\det\left\{\left(T_ks^k\ +\ \ldots\ +T_0\right)M(s)\right\}=\det\left\{[T_k,\ T_{k\text{-}1},\ \ldots,\ T_0]\begin{bmatrix}s^kM(s)\\s^{k\text{-}1}M(s)\\.\\.\\M(s)\end{bmatrix}\right\} \tag{8.19}$$

Remark (8.1): The above formulation of the determinantal dynamic assignment problems is based on the assumption that $p \leq m$ and thus output feedback configuration is used. If $p \geq m$, we can similarly formulate the corresponding problems as determinantal dynamic precompensation problems and use right coprime MFDs for C(s).

■

(vi) Decentralised Determinantal Problems: The problems considered above assume that the controller is of the centralised type. The decentralisation assumption implies that the controller is not a full matrix, but it has a block diagonal structure and thus reduced degrees of freedom. For the constant output feedback case ($m \geq p$) and assuming that we have μ channels, then the decentralised output feedback is expressed as

$$K_{dec} = \begin{bmatrix} K & 0 & \dots & 0 \\ 0 & K & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & \dots & K\mu \end{bmatrix} \tag{8.20}$$

where $K_i \in \mathbb{R}^{p_i \times m_i}$, $\sum p_i = p$ and $\sum m_i = m$ and thus

$$\mathrm{f}(s) = \det\left\{ \begin{bmatrix} I_p, & K_{\mathrm{dec}} \end{bmatrix} \begin{bmatrix} \mathrm{D_r}(s) \\ \mathrm{N_r}(s) \end{bmatrix} \right\} \tag{8.21}$$

The decentralised problems will be considered separately and their main feature is that the controller has a partially fixed structure which introduces some special characteristics to the analysis of the problem.

8.3 THE ABSTRACT DETERMINANTAL ASSIGNMENT PROBLEM

All the problems introduced in the previous sections belong to the same problem family i.e. the determinantal assignment problem (DAP). This problem is to solve the following equation with respect t o polynomial matrix $H(s)$:

$$\det\left(H(s)N(s)\right)=f(s) \qquad (8.22)$$

where $f(s)$ is the polynomial of an appropriate degree d. The difficulty for the solution of DAP is mainly due to the multilinear nature of the problem as this is described by its determinantal character. We should note, however, that in all cases mentioned previously, all dynamics can be shifted from $H(s)$ to $N(s)$, which, in turn, transforms the problem to a constant DAP. This problem may be described as follows:

Let $M(s)\in\mathbb{R}^{p\times r}[s]$, $r\leq p$ such that rank(M(s)) = r $\operatorname{rank}\{M(s)\}=r$ and let $\mathcal{H}$ be a family of full rank $r\times p$ constant matrices having a certain structure. Solve with respect to $H\in\mathcal{H}$ the equation:

$$f_M(s,H)=\det\left(H\cdot M(s)\right)=f(s) \qquad (8.23)$$

where $f(s)$ is a real polynomial of an appropriate degree d.

Remark 8.2: The degree of the polynomial $f(s)$ depends firstly upon the degree of $M(s)$ and secondly, upon the structure of H. However, in most of our problems the degree of $p(s)$ is equal to the degree of $M(s)$.

■

The determinantal assignment problem has two main aspects. The first has to do with the solvability conditions for the problem and the second, whenever this problem is solvable, to provide methods for constructing these solutions. We classify the solutions to two classes : exact and generic solvability conditions. The characterisation of the generic solvability conditions is linked to the problem of system parametrisation.

Notation: Let $Q_{k,n}$ denote the set of lexicographically ordered, strictly increasing sequences of k integers from 1, 2, ..., n. If $\{\underline{x}_{i_1},\dots,\underline{x}_{i_k}\}$ is a set of vectors of V, ω=$(\mathrm{i}_1,\dots,\mathrm{i}_\mathrm{k})\in Q_{k,n}$, then $\underline{x}_{i_1}\wedge\ \dots\ \wedge\underline{x}_{i_k}=\underline{x}_\omega$ $\wedge$denotes the exterior product and by $\wedge^r V$

we denote the r-th exterior power of V. If $H \in F^{m\times n}$ and $r \leq \min\{m,n\}$, then by $C_r(H)$ we denote the r-th compound matrix of H [Marcus et al., 1969].

■

If $\underline{h}_i^t$, $\underline{m}_i(s)$, $i \in \underset{\sim}{r}$, we denote the rows of H, columns of $M(s)$ respectively, then

$C_r(M) = \underline{h}_1^t \wedge \ldots \wedge \underline{h}_r^t = \underline{h}^t \wedge \ \in \mathbb{R}^{l\times\sigma}$

and

$$C_r(M(s)) = \underline{m}_1(s) \wedge \ \ldots \ \wedge \underline{m}_r(s) = \underline{m}\wedge \in \mathbb{R}^{\sigma}[s],\ \sigma = \binom{p}{r} \tag{8.24}$$

and by Binet-Cauchy theorem [Marcus et al., 1969] we have that [Karcanias et al., 1984]:

$$f_M(s,H) = C_r(H)\cdot C_r(M(s)) = \langle \underline{h}\wedge, \underline{m}(s)\wedge \rangle = \sum_{\omega\in Q_{r,p}} h_\omega m_\omega(s) \tag{8.25}$$

where $\langle * \ , \ * \rangle$ denotes inner product, $\omega = (i_1,\ldots,i_r) \in Q_{r,p}$, and h_ω, $m_\omega(s)$ are the coordinates of $\underline{h}\wedge$, $m(s)\wedge$ respectively. Note that h_ω is the $r\times r$ minor of H which corresponds to the ω set of columns of H and thus h_ω is a multilinear alternating function of the entries h_{ij} of H. The multilinear, skew symmetric nature of DAP suggests that the natural framework for its study is that of exterior algebra. The essence of exterior algebra is that it reduces the study of multilinear skew-symmetric functions to the simpler study of linear functions. The study of the zero structure of the multilinear function $f_M(s,H)$ may thus be reduced to a linear subproblem and a standard multilinear algebra problem as it is shown below.

(i) **Linear subproblem of DAP**: Set $\underline{m}(s)\wedge \underline{p}(s) \in \mathbb{R}^{\sigma}[s]$ Determine whether there exists a $\underline{k} \in \mathbb{R}^{\sigma}$, $\underline{k} \neq 0$, such that

$$f_M(s,H) = \underline{k}^t \underline{p}(s) = \sum k_i p_i(s) = f(s),\ i \in \underset{\sim}{\sigma}\ \ f(s) \in \mathbb{R}[s] \tag{8.26}$$

(ii) <u>Multilinear subproblem of DAP</u>: Assume that K is the family of solution vectors $\underline{k}$ of (8.26). Determine whether there exists $H^t = [h_1, \ldots, h_r]$, where $H^t \in \mathbb{R}^{p \times r}$, such that

$$\underline{h}_1 \wedge \ldots \wedge \underline{h}_r = \underline{h} \wedge = \underline{k}, \ k \in K \tag{8.27}$$

■

Polynomials defined by Eqn.(8.26) are called *polynomial combinants* [Karcanias et al., 1984] and the zero assignability of them provides necessary conditions for the solution of the DAP. The solution of the exterior equation (8.27) is a standard problem of exterior algebra and it is known as *decomposability* of multivectors. Note that notions and tools from exterior algebra play also an important role in the linear subproblem, since $f_M(s,H)$ is generated by the decomposable multivector $\underline{m}(s) \wedge$.

The essence of our approach is *projective*, that is we use a natural embedding for determinantal problems to embed the space of the unknown, $\mathcal{H}$, of DAP, into an appropriate projective space. In this way we can see our problem as search for common solutions of some set of linear equations and another set of second order polynomial equations. This also allows us to compactify $\mathcal{H}$ into $\bar{\mathcal{H}}$ and then use *algebraic geometric*, or *topological intersection theory* methods to determine existence of solutions for the above sets of equations. The characteristic of the current framework is that it allows the use of algebraic geometry and topological methods [Leventides, 1993] for the study of solvability conditions but also computations. Central to the latter is the solution of the linear system derived by (8.26) with the quadratics characterising the solvability of (8.27), which are known as Quadratic Plücker Relations (QPR). The importance of the DAP framework is that it uses the natural embedding of a Grassmannian into a projective space and this in turn defines new sets of invariants characterising the solvability of the different DAP problems [Karcanias et al., 1984]. We may summarise the results here, because they provide the basis for a variety of structural indicators and diagnostics and form the basis for defining the "approximate zero polynomial".

Let $T(s) \in \mathbb{R}^{p \times r}[s]$, $T(s) = \left[\underline{t}_1(s), \ldots, \underline{t}_r(s)\right]$, $p \geq r$, $\operatorname{rank}\{T(s)\} = r$ and let $X_t = \mathrm{R}_{\mathbb{R}[s]}(T(s))$. If $T(s) = M(s)D(s)^{-1}$ is a RCMFD of $T(s)$, then $M(s)$ is a polynomial basis for X_t. If $Q(s)$ is a greatest right divisor of $M(s)$ then $T(s) = \tilde{M}(s)Q(s)D(s)^{-1}$, where $\tilde{M}(s)$ is a least degree polynomial basis of X_t [Rosenbrock, 1979], [Kailath, 1980]. A GR for X_t is defined by [Karcanias et al., 1984]

$$\underline{t}(s)\wedge = \underline{t}_1(s)\wedge\ldots\wedge\underline{t}_r(s) = \underline{\tilde{m}}_1(s)\wedge\ldots\wedge\underline{\tilde{m}}_r(s)\ \cdot\frac{z_t(s)}{p_t(s)} \tag{8.28}$$

where $z_t(s) = \det\{Q(s)\}$, $p_t(s) = \det\{D(s)\}$ are the zero, pole polynomials of $T(s)$ and $\underline{\tilde{m}}(s) = \underline{m}_1(s)\wedge\ldots\wedge\underline{\tilde{m}}_r(s)\in\mathbb{R}^\sigma[s]$, $\sigma = \binom{p}{r}$, is also a GR of X_t. Since $\tilde{M}(s)$ is a least degree polynomial basis for X_t, the polynomials of $\underline{\tilde{m}}(s)\wedge$ are coprime and $\underline{\tilde{m}}(s)\wedge$ will be referred to as a *reduced polynomial* GR (R $-\mathbb{R}[s]$ -GR) of X_t. If $\delta = \deg\{\underline{\tilde{m}}(s)\wedge\}$, then δ is the Forney dynamical order [For.1] of X_t . $\underline{\tilde{m}}(s)\wedge$ may always be expressed as

$$\underline{\tilde{m}}(s)\wedge = p(s) = p_0 + p_1 s + \ldots + p_\delta s^\delta = P_\delta\cdot\underline{e}_\delta(s)\ ,\ \ P_\delta\in\mathbb{R}^{\sigma\times(\delta+1)} \tag{8.29}$$

where P_δ is a basis matrix for $\underline{\tilde{m}}(s)\wedge$ and $\underline{e}_\delta(s) = [1, s, \ldots, s^\delta]'$. It can be readily shown that all $\mathbb{R}[s]$-GRs of X_t differ only by a nonzero scalar factor $a\in\Re$. By choosing an $\underline{\tilde{m}}(s)\wedge$ for which $\left\|\underline{p}_\delta\right\| = 1$, a *monic reduced Grassman representative* (R $-\mathbb{R}[s]$ -GR) of X_t and shall be denoted by $\underline{g}(X_t)$ and is referred to as canonical $\mathbb{R}[s]$ -GR; the basis matrix P_δ of $\underline{g}(X_t)$ is defined as the *Plücker matrix* of X_t [Karcanias et al., 1984]. The importance of the above is established by the following results [Karcanias et al., 1984]:

Result (4.1): $\underline{g}(X_t)$, or the associated Plücker matrix P_δ, is a complete (basis free) invariant of X_t. ■

Result (4.2): Let $T(s) \in \mathbb{R}^{p \times r}[s]$, $p \geq r$ rank$\{T(s)\} = r$, $z_t(s)$, $p_t(s)$ be the monic zero, pole polynomials of $T(s)$ and let $\underline{g}(X_t) = \underline{p}(s)$ be the C -$\mathbb{R}[s]$ -GR of the column space X_t of $T(s)$. $t(s)\wedge$ may be uniquely decomposed as

$$t(s)\wedge = c\underline{p}(s)\frac{z_t(s)}{p_t(s)}, \text{ where } c \in \mathbb{R} - \{0\} \tag{8.30}$$

If $M(s) \in \mathbb{R}^{p \times r}[s]$, $p \geq r$, rank$\{M(s)\} = r$, then $M(s) = \tilde{M}(s)Q(s)$, where $\tilde{M}(s)$ is a least degree basis and $Q(s)$ is a greatest right divisor of the rows of M(s) and thus

$$\underline{m}(s)\wedge = \underline{\tilde{m}}(s)\wedge \cdot \det(Q(s)) = \underline{p}(s)z_m(s) = P_\delta \cdot e_\delta(s) \cdot z_m(s) \tag{8.31}$$

The linear part of DAP is thus reduced to

$$f_M(s,\underline{k}) = \underline{k}'\underline{p}(s)z_m(s) = \underline{k}'P_\delta \cdot e_\delta(s) \cdot z_m(s) \tag{8.32}$$

Result 8.3: The zeros of $M(s)$ are fixed zeros of all combinants of $\underline{m}(s)\wedge$

■

The zeros of $f_M(s,\underline{k})$ which may be freely assigned are those of the combinant $f_{\tilde{M}}(s,\underline{k}) = \underline{k}'\underline{\tilde{m}}(s)\wedge$, where $\underline{\tilde{m}}(s)\wedge$ is reduced. Given that the zeros of $f_{\tilde{M}}(s,\underline{k})$ are not affected by scaling with constants, we may always assume that $\underline{\tilde{m}}(s)\wedge = P_\delta \cdot e_\delta(s)$. In the following, the case of combinants generated by reduced $\underline{\tilde{m}}(s)\wedge$ will be considered. If $a(s) \in \mathbb{R}[s]$ is the polynomial which has to be assigned, then $\max\{\deg a(s)\} = \delta$, where δ is the Forney dynamical order of X_t. If $a(s) = \underline{a}_\delta^t \underline{e}_\delta(s) = a_0 + a_1 s + \ldots + a_\delta s^\delta$, where $\underline{a}_\delta \in \mathbb{R}^{\delta+1}$, then the problem of finding $\underline{k} \in \mathbb{R}^\sigma$ such that $f_{\tilde{M}}(s,\underline{k}) = a(s)$ is reduced to the solution of

$$P_{\delta}^{t}\underline{k} = \underline{a}\,,\; P_{\delta}^{t} \in \mathbb{R}^{(\delta+1)\times\sigma}\,,\; \sigma = \begin{pmatrix} p \\ r \end{pmatrix} \tag{8.33}$$

The matrix $M(s) \in \mathbb{R}^{p\times r}[s]$ generating DAP will be called *linearly assignable* (LA), if Eqn.(8.33) has a solution for all $\underline{a}$; otherwise, it will be called *linearly nonassignable* (LNA). $M(s)$ will be called *completely assignable* (CA), if it is LA and Eqn (8.27) has a solution for at least a solution of the linear problem defined by Eqn (8.33). An important family of nonassignable $M(s)$ matrices are those for which there is no $\underline{k}$ such that $f_M(s,\underline{k}) = c,\ c \in \mathbb{R}$; such $M(s)$ are called strongly nonassignable (SNA) and they imply that they cannot assign all zeros at $s = \infty$. Note that strong nonassignability implies nonassignability, but not vice versa. Some results characterising the above properties are stated below [Karcanias et al., 1984]:

Remark 8.1: Necessary condition for $M(s)$ to be LA is that $M(s)$ is a least degree matrix (i.e. has coprime rows).

■

Result 8.2: Let $M(s) \in \mathbb{R}^{p\times r}[s]$ be a least degree matrix, P_δ be the Plücker matrix of X_m and let $\xi = \operatorname{rank}\{P_\delta\}$. Then,

(i) Necessary and sufficient condition for $M(s)$ to be LA is that $\xi = \delta + 1$

(i.e. $\begin{pmatrix} p \\ r \end{pmatrix} \ge \delta + 1$ and $\xi = \delta + 1$).

(ii) Let $M(s)$ be LA and let $\tilde{P}^{\dagger}$ be the right inverse of P_δ^t and $\tilde{P}_\delta^{\perp}$ a basis for $\mathcal{N}_r(P_\delta^t)$. For every $\underline{a} \in \mathbb{R}^{\delta+1}$, the solution of (8.33) is given by

$$\underline{k} = \tilde{P}^{\dagger}\underline{a} + \tilde{P}_\delta^{\perp}\underline{c}\text{, where } \underline{c} \in \mathbb{R}^{\sigma-\delta-1} \text{ arbitrary} \tag{8.34}$$

(iii) Let $P_\delta = \left[\underline{p}_0, \underline{p}_1, \dots, \underline{p}_\delta\right] \in \mathbb{R}^{\sigma\times(\delta+1)}$ and let $\xi = \operatorname{rank}\{P_\delta\}$, $\bar{\xi} = \operatorname{rank}\{\bar{P}_\delta\}$. $M(s)$ is strongly nonassignable, iff $\bar{\xi} = \sigma$.

■

The solvability of DAP involves examination of the exterior equation (8.27). We summarise next some of the fundamentals which affect the study of approximate zero polynomials". A proper treatment may be found in [Leventides, 1993], [Karcanias et al., 1984], [Karcanias et al., 1989], [Karcanias et al., 1984] etc. A necessary and sufficient condition for a solution of DAP is that from the family of solutions of Eqn. (8.26) there exists at least a $\underline{k}$ for which Eqn. (8.27) has a solution. A vector $\underline{k} \in \mathbb{R}^{\sigma}$ defines a point in the projective space $P_{\sigma-1}(\mathbb{R})$; the points of which satisfy for some $H \in \mathbb{R}^{r\times p}$ Eqn. (8.27) are those which belong to the *Grassmann variety* of $P_{\sigma-1}(\mathbb{R})$ which is [Hodge, 1952] characterised by the following result:

<u>**Result 8.5:**</u> Let $\underline{k} \in \mathbb{R}^{\sigma}$, $\sigma = \binom{p}{r}$ and let k_ω, $\omega = (i_1, \dots, i_r) \in Q_{r,p}$ be the coordinates of $\underline{k}$ (seen as the Plücker coordinates of a point in $P_{\sigma-1}(\mathbb{R})$). Necessary and sufficient condition for an $H \in \mathbb{R}^{r\times p}$, $H = \left[\underline{h}_1, \dots, \underline{h}_r\right]^t$, to exist such that

$$\underline{h}\wedge = \underline{h}_1 \wedge \dots \wedge \underline{h}_r = \underline{k} = \left[\dots, k_\omega, \dots\right]^t \tag{8.35}$$

is that the coordinates k_ω satisfy the following quadratic relations

$$\sum_{k=1}^{r+1} (-1)^{\nu-1} k_{i_1, \dots, i_{r-1}, j_\nu^k j_1, \dots, j_{\nu-1}, j_{\nu+1}, j_{r+1}} = 0 \tag{8.36}$$

where $1 \le i_1 < i_2 < \cdots < i_{r-1} \le n$ and $1 \le j_1 < j_2 < \cdots < j_{r+1} \le n$

■

The set of quadratics defined by Eqn (8.36) are known [Hodge, 1952], [Marcus et al., 1969] as the *Quadratic Plücker Relations* (QPR) and they define the Grassmann variety of $P_{\sigma-1}(\mathbb{R})$. Conditions (8.36) clearly reveal the nonlinear nature of DAP. The two main questions which naturally arise are: (1) Given a decomposable $\underline{k}$, which satisfies (8.35), construct the matrix H. (2) Parameterise the set of conditions (8.36). the second question is crucial for the study of compensators which considerably simplify DAP.

Example 8.1: Let $p=5,\ r=3$ and let $(k_0, k_1, k_2, \ldots, k_g)$ be the coordinates of a vector defining a point in the projective space P_g. The set of QPRs describing the Grassman variety of P_g (usually denote by $\Omega(3,5)$) is given by

$$k_0k_5 - k_1k_4 + k_2k_3 = 0,\ k_0k_8 - k_1k_7 + k_2k_6 = 0,\ k_0k_9 - k_3k_7 + k_4k_6 = 0 \tag{8.37}$$

$$k_1k_9 - k_3k_8 + k_5k_6 = 0,\ k_2k_9 - k_4k_8 + k_5k_7 = 0 \tag{8.38}$$

It may be readily shown that the above set of equations is not minimal; in fact, the set (8.38) may be obtained from the set (8.37) and thus (8.37) is a minimal set of quadratics describing the Grassmann variety $\Omega(3,5)$.

The above example makes clear the need for the definition of a minimal set of quadratics describing $\Omega(r,p)$. The problem of reconstructed H from the decomposable $\underline{k}$ is examined first [Giannacopoulos et al., 1984]

Result 8.6: Let $\underline{k} = [\ldots, k_\omega, \ldots]^t \in \mathbb{R}^\sigma$, $\sigma = \begin{pmatrix} p \\ r \end{pmatrix}$, be a decomposable vector (satisfying the set of QPRs) and let $k_{a_1,\ldots,a_r}$ be a nonzero coordinate of $\underline{k}$. If we define by

$$h_{ij} = k_{a_1,\ldots,a_{r-1},j,a_{i+1},\ldots,a_r}\ ,\quad i \in \underset{\sim}{r}\ ,\quad j \in \underset{\sim}{p} \tag{8.39}$$

then for the matrix $H = [h_{ij}]$, $C_r(H) = \underline{k}$.

■

Remark 8.4: Let $\underline{k} = [\ldots, k_\omega, \ldots]^t \in \mathbb{R}^\sigma$, $\sigma = \begin{pmatrix} p \\ r \end{pmatrix}$, be a decomposable vector and let the first coordinate of $\underline{k}$ be nonzero. The H matrix defined by Result 8.5 has the form

$$H=\left[k_a I_r X_t\right]^t \in \mathbb{R}^{p\times r}\text{, where } k_a = k_{1,2,\dots,r} \neq 0 \qquad (8.40)$$

■

The reconstruction of H from a decomposable vector is demonstrated by the following example.

Example 8.2: Let $\underline{k}=\left[k_0,k_1,k_2,k_3,k_4,k_5\right]^t$ be a point of the Grassmann variety $\Omega(2,4)$ of the projective space $P^5(\mathbb{R})$. A basis matrix for the vector space V whose Plücker coordinates are coordinates of the given point is defined by:

$$H_0=\begin{bmatrix} k_0 & 0 \\ 0 & k_0 \\ -k_3 & k_1 \\ -k_4 & k_2 \end{bmatrix} \text{ if } k_0 \neq 0 \text{ and } H_0=\begin{bmatrix} k_2 & 0 \\ k_4 & k_0 \\ k_5 & k_1 \\ 0 & k_2 \end{bmatrix} \text{ if } k_2 \neq 0$$

it may be readily shown that $H_2 = H_0 \cdot Q$ where

$$Q=\begin{bmatrix} k_2\cdot k_0^{-1} & 0 \\ k_4\cdot k_0^{-1} & 1 \end{bmatrix},\ |Q|\neq 0$$

We may verify that $C_2\left(H_0\right)=\left[k_0^2,\ k_0k_1,\ k_0k_2,\ k_0k_3,\ k_0k_4,\ k_1k_4,\ -k_2k_3\right]^t$ and since $k_0k_5-k_1k_4+k_2k_3=0$ we have that $k_1k_4-k_2k_3=k_0k_5$ and thus $C_2\left(H\right)=K_0\cdot\underline{k}$.

■

The above procedure for constructing H out of a decomposable $\underline{k}$ also suggests a procedure for writing down an independent set of QPRs which completely describes $\Omega(r,p)$; this set is referred to as the *Reduced Quadratic Plücker Relations* (RQPR) [Giannacopoulos et al., 1984]. Alternative ways for characterising decomposability and reconstructing the space from a decomposable vectoris given in terms of the Grassmann matrix [Karcanias et al., 1984, 2], which reduces the study of decomposability to a rank problem.

8.4 GRASSMANN AND PLUCKER INVARIANTS FOR LINEAR SYSTEMS AND THE NOTION OF ALMOST ZEROS

For the control problems discussed in section (8.2), the matrix $M(s)$ has a special structure; thus the matrix coefficient of $\underline{m}(s)\wedge$ has important properties which stem from the properties of the corresponding control problem. A number of Plücker type matrices associated with a linear system are defined below:

(a) Controllability Plücker Matrix:

For the pair (A,B), $\underline{b}(s)^t\wedge$ denotes the exterior product of the rows of $B(s)=[sI-A,-B]$ and $P(A,B)$ is the $(n+1)\times\binom{n+p}{n}$ basis matrix of $\underline{b}(s)^t\wedge$. $P(A,B)$ will be called the *controllability Plücker matrix* and its rank properties characterise the system controllability.

Result (4.5) [Karcanias et al., 1996]: The system $S(A,B)$ is controllable iff $P(A,B)$ has full rank.

■

The singular values of $P(A,B)$ characterise the degree of controllability, affect state feedback design and are primary controllability indicators associated with state feedback design.

(b) Observability Plücker Matrix :

For the pair (A,C), $\underline{c}(s)\wedge$ denotes the exterior product of the columns of $C(s)=\left[sI-A^t,-C^t\right]^t$ and $P(A,C)$ is the $\binom{n+m}{n}\times(n+1)$ basis matrix of $\underline{c}(s)\wedge$.

$P(A,C)$ will be called *observability Plücker matrix* and its rank properties characterise system observability.

<u>Result 8.8</u> [Karcanias et al., 1996]: The system $S(A,C)$ is observable, iff $P(A,C)$ has full rank

■

$P(A,C)$ is important for observer design and its singular values are prime indicators in the solution of such problems.

(c) <u>Transfer Function Matrix Plücker Matrices</u>

For the transfer function matrix $G(s)$ represented by the RCMFD, LCMFD of Eqn.(8.2) we define by $\underline{t}_r(s)\wedge$, $\underline{t}_l(s)^t\wedge$ the exterior product of the columns of $T_r(s)$, rows of $T_l(s)$ respectively, where $T_r(s)$, $T_l(s)$ are defined by Eqn.(8.6). By $P(T_r)$ we denote the $\begin{pmatrix} m+p \\ p \end{pmatrix}\times(n+1)$ basis matrix of $\underline{t}_r(s)\wedge$, and by $P(T_l)$ the $(n+1)\times\begin{pmatrix} m+p \\ p \end{pmatrix}$ basis matrix for $\underline{t}_l(s)^t\wedge$. $P(T_r)$, $P(T_l)$ will be referred to as *right*, *left fractional representation Plücker matrices* respectively. Such matrices provide the prime indicators for the solution of the output feedback, or constant pre-compensation problem.

<u>Result 8.9</u> [Leventides et al., 1995]: For a generic system with $mp>n$, then the corresponding Plücker matrices $P(T_r)$, $P(T_l)$ have full rank.

■

The full rank of these matrices is a necessary condition for the solvability of pole assignment problems and their singular values characterise the norm properties of the corresponding solutions. Given that $T_r(s)$, $T_l(s)$ uniquely characterise (modulo unimodular equivalence) the transfer function $G(s)$, we may also refer to $P(T_r)$, $P(T_l)$

as the *right-, left- transfer function Plücker matrices* and we denote them simply by $P_r(G)$, $P_l(G)$.

(d) Column, Row Plücker Matrices

For the transfer function $G(s)$, $m \geq p$, we denote by $\underline{n}(s)\wedge$ the exterior product of the columns of the numerator $N_r(s)$, of a RCMFD and by $P(N)$ the $\binom{m}{l} \times (d+1)$ basis matrix of $\underline{n}(s)\wedge$. Note that $d = \delta$, the Forney order of X_g, if $G(s)$ has no finite zeros and $d = \delta + k$, where k is the number of finite zeros of $G(s)$, otherwise. If $N_r(s)$ is least degree (has no finite zeros), then $P_c(N)$ will be called the *column space Plücker matrix* of the system. For this case the *row space Plücker matrix* may be similarly defined and it is $P_r(N) = 1$. For systems with $m \leq p$ and full rank transfer functions $P_c(N) = 1$, whereas $P_r(N)$ is a nontrivial matrix. Such matrices play a key role in problems such as the squaring down, or more general model projection problems.

Result 8.10: For a generic system with $m > p$, for which $p(m-p) > \delta + 1$, where δ is the Forney order, $P_c(N)$ has full rank.

■

(e) Dynamic Compensation Transfer Function Matrices

For the transfer function $G(s)$ $m \geq p$ we may define the matrices

$$T_r(s) = \begin{bmatrix} D_r(s) \\ N_r(s) \end{bmatrix} = M(s) \ , \quad T_r^{PI}(s) = \begin{bmatrix} sM(s) \\ \cdots\cdots \\ N_r(s) \end{bmatrix} \ , \quad T_r^{PD}(s) = \begin{bmatrix} M(s) \\ sN_r(s) \end{bmatrix} \qquad (8.41)$$

$$T_r^{PID}(s)=\begin{bmatrix} sM(s) \\ N_r(s) \\ s^2N_r(s) \end{bmatrix}, \quad T_r^{k,OBD}(s)=\begin{bmatrix} s^kM(s) \\ s^{k\text{-}1}M(s) \\ \vdots \\ \vdots \\ M(s) \end{bmatrix} \tag{8.42}$$

By taking the exterior products of the columns of such matrices $P_r^{PI}(G)$, $P_r^{PD}(G)$, $P_r^{PID}(G)$, $P_r^{k,OBD}(G)$ which are referred to as *right-, PI-, PD, PID, k-order OBD transfer function Plücker matrices.* Such matrices enter the solvability of the corresponding assignment problems and their singular values are prime design indicators for the corresponding problem.

Remark 8.5: The matrices $P_r(G)$, $P_l(G)$ for a given $G(s)$ are not independent, but they are related by [Karcanias et al., 1984, 2]

$$P_l(G)^t = U \cdot P_r(G),\ U \text{ invertible} \tag{8.43}$$

■

Properties of the rank of the above matrices for the generic case may be established in a similar manner.

With a linear system it is clear from the above that we may associate new types of invariants, which characterise completely respective vector spaces and they are known as Plücker matrices, or $\mathbb{R}[s]$- canonical Grassmann representatives. For every Plücker matrix, there is a polynomial vector associated with it and thus we have the following Grassmann representatives:

$\underline{g}(A,B)^t = \underline{e}_n(s)^t\, P(A,B)$: Controllability - GR

$\underline{g}(A,C) = P(A,C)\underline{e}_n(s)$: Observability - GR

$\underline{g}_r(G) = P_r(G)\underline{e}_n(s)$: Right Tranfer function - GR

$\underline{g}_l(G)^t = \underline{e}_n^t(s) \cdot P_l(G)$: Left Tranfer function - GR

$\underline{g}_c(N) = P_c(N)\underline{e}_{\delta_c}(s)$: Column Space – GR

$\underline{g}_r(N) = \underline{e}_{\delta_r}(s)P_r(N)$: Row Space - GR

where $\underline{e}_k(s) = \left[1, s, \ldots, s^k\right]$, n is the McMillan degree and δ_c, are the Forney orders of the column, row rational vector spaces associated with $G(s)$. Similar Grassmann representatives are defined for dyadic compensation schemes based on the $P_r^{PI}(G)$, $P_r^{PD}(G)$, $P_r^{PID}(G)$, $P_r^{k,OBD}(G)$ Plücker matrices when $m \geq p$ (respective matrices based on left MFD when $m < p$). The $\mathbb{R}[s]$- GRs are generating combinants and the full rank property of the corresponding Plücker matrix is necessary condition for assignability. The problem we examine next is the investigation of consequences of the rank deficiency of the Plücker matrix and especially on the distribution of zeros. Our discussion is presented for a general set of polynomials $\mathcal{P}$ expressing the coordinates of any of the above system $\mathbb{R}[s]$-GRs.

Let $\mathcal{P} = \left\{p_i(s) : p_i(s) \in \mathbb{R}[s],\ i \in \underline{m},\ d_i = \deg\left(p_i(s)\right)\right\}$ and let $d = \max\left\{d_i,\ i \in \underline{m}\right\}$. With the set $\mathcal{P}$ we may always associate a polynomial vector $\underline{p}(s) = \in \mathbb{R}^m[s]$ where

$$\underline{p}_{h+1}(s) = \begin{bmatrix} p_1(s) \\ p_2(s) \\ \vdots \\ p_m(s) \end{bmatrix} = \left[\underline{p}_0, \underline{p}_1, \ldots, \underline{p}_d\right]\underline{e}_d(s) = P_d\underline{e}_d(s) \tag{8.44}$$

where $P_d \in \mathbb{R}^{m \times (d+1)}$ and $\underline{e}_d(s) \in \mathbb{R}^{d+1}[s]$. The polynomial vector $\underline{p}(s)$ is defined as a *vector representative* of $\mathcal{P}$ and $d = \deg\left(\underline{p}(s)\right)$ will be referred to as the *degree* of $\mathcal{P}$. The matrix P_d characterises the properties of $\mathcal{P}$ and it is defined as a *basis matrix* of $\mathcal{P}$. The set $\mathcal{P}$ will be called *reduced*, if the polynomials $\underline{p}_i(s)$ are coprime; otherwise it will be called *nonreduced*. Finally, $\mathcal{P}$ will be called monic, if $\left\|\underline{p}_d\right\|=1$ ($\|\cdot\|$ denotes the usual Euclidean norm). The polynomial function defined for any $\underline{k} \in \Re^m$

$$f\left(s,\mathcal{P},\underline{k}\right)=\underline{k}^t P_d \underline{e}_d\left(s\right)=\sum_{i=1}^{m} k_i p_i\left(s\right) \tag{8.45}$$

is called a $\underline{k}$- *polynomial combinant* of $\mathcal{P}$ and shall be denoted in short by $f\left(s,\mathcal{P},\underline{k}\right)$. The families of $\mathcal{P}$ polynomials may be classified according to the properties of P matrix to linearly assignable (LA), nonassignable (NA) and strongly nonassignable sets.

Remark (8.6): If $s=z$ is a zero of the set $\mathcal{P}$, then for all $\underline{k} \in \mathbb{R}^m$, $f\left(s,\mathcal{P},\underline{k}\right)$ has a fixed zero at $s=z$.

■

The presence of zeros in $\mathcal{P}$, implies the P$_d$ is rank deficient and $\mathcal{P}$ is nonassignable. The concept of almost zeros introduced in [Karcanias et al., 1984] provides an analytic extension of the algebraic concept and regarding distribution of zeros, extends the fixed zero of combinants to discs trapping zeros of combinants. The almost zero and its properties are introduced on a general $\mathcal{P}$ set and their application to systems is obvious.

When $s \in \mathbb{C}$, the vector representative $\underline{p}\left(s\right)$ of $\mathcal{P}$ defines a vector analytic function with domain and $\mathbb{C}$ codomain $\mathbb{C}^m$; we define the norm of $\underline{p}\left(s\right)$ (or norm of $\mathcal{P}$) as

$$\left\|\underline{p}(s)\right\|=\phi\left(\sigma,\omega\right)=\sqrt{\underline{p}\left(s^*\right)^t\,\underline{p}\left(s\right)}=\sqrt{\underline{e}_d\left(s^*\right)^t P_d^t P_d \underline{e}_d(s)} \tag{8.46}$$

where s^* is the complex conjugate of s ($s=\sigma+j\omega$). Note that if $q\left(s\right)=s+a$ is a divisor of $\mathcal{P}$, then $p\left(-a\right)=0$ and thus $\left\|\underline{p}\left(-a\right)\right\|=0$. This observation leads to the following definition.

Definition 8.1: Let $\mathcal{P}$ be a reduced set of polynomials. If $s=z$, $z \in \mathbb{C}$, is a local minimum of $\left\|\underline{p}\left(s\right)\right\|$, then z will be called an *almost zero* (AZ) of $\mathcal{P}$ and the value

$\left\|\underline{p}(z)\right\| = \varepsilon$ will be referred to as the *order* of the AZ. If $s = \bar{z}$ is the global minimum of $\left\|\underline{p}(s)\right\|$, then $\bar{z}$ will be called the *prime almost zero* (PAZ) of the set $\mathcal{P}$.

■

Clearly, if $\mathcal{P}$ is not reduced, then the set of AZs, which have order $\varepsilon = 0$, defines the zeros of $\mathcal{P}$. Thus, the definition unifies the notions of exact and 'approximate' zeros, since both emerge as minima of a norm function of $\mathcal{P}$. The order ε of an AZ indicates how well z may be considered as an 'approximate' zero of $\mathcal{P}$; we should note, however, that scaling of the polynomials of $\mathcal{P}$ by $c \in \mathbb{R}$, $c \neq 0$, affects the order ε of an AZ, but not its location. In the following, $\mathcal{P}$ will be assumed monic ($\underline{p}(s)$is assumed monic). The properties of distribution of AZs in the complex plane and their computation is discussed in [Karcanias et al., 1984]; their significance in the distribution of zeros of polynomial combinants is discussed below for the case of SNA polynomial sets.

<u>**Result 8.10:**</u> Let $\mathcal{P}$ be a SNA set, $a \in \mathbb{C}$, and let $\underline{p}(w) = \underline{b}_0 + \underline{b}_1 w + \ldots + \underline{b}_d w^d$, $w = s - a$, be the Taylor expansion of the vector representative of $\mathcal{P}$ at $s = a$.

i) For every $\underline{k} \in \mathbb{R}^m$, $f(s, \mathcal{P}, \underline{k})$ has at least one zero representative in the finite, minimal radius disk $D_m\left[a, R_m(a, \underline{k})\right] = \left\{s : |s - a| \leq R_m(a, \underline{k})\right\}$, where $R_m(a, \underline{k})$ is defined by

$$R_m(a, \underline{k}) = \min\left\{ \binom{d}{i} \frac{\left|\underline{k}^t \underline{b}_0\right|}{\left|\underline{k}^t \underline{b}_i\right|^{1/i}} \;, i \in \underline{d} \right\} \tag{8.47}$$

ii) For every $a \in \mathbb{C}$ there exists a finite, minimal radius disk $\tilde{D}_m\left[a, \tilde{R}_m(a)\right]$, where $\tilde{\mathrm{R}}_m = \max\left\{R_m(a, \underline{k}) \;, \underline{k} \in \mathbb{R}^m\right\}$, which contains at least one zero of all combinants of $\mathcal{P}$.

■

A number of upper bounds for $R_m(a, \underline{k})$ are given in [Karcanias et al., 1983]. Result 8.10 makes no distinction between a general point $a \in \mathbb{C}$ and an AZ z of the polynomial set $\mathcal{P}$. The feature that distinguish an AZ of $\mathcal{P}$ from all other points of the complex plane is that "strong activity" summarised by the following property [Karcanias et al., 1983]:

Remark 8.11: For a family of upper bounds for the radius $R_m(a,\underline{k})$, $\hat{R}_i(a,\underline{k})$ the AZ z has the property that for all $a \in \mathbb{C}: |z-a| < \varepsilon$, $\varepsilon > 0$, $\hat{R}_i(z,\underline{k}) < \hat{R}_i(a,\underline{k})$ for all $\underline{k}$.

■

The above suggests that AZs emerge as "strong poles" of attraction for the zeros of all combinants $f(s,\underline{k})$. The SNA property makes the radii of such disks finite

The notion os AZs has motivated the study of approximate gcd of polynomials undertaken in this thesis. It has been recognised that dealing with single frequencies only provides a partial extension of the almost zero notion, since extension to a set requires finding a set of AZs that involves a set of minima. Such an approach daoes not easily provide estimates of the strength of the approximation, when we refer to approximate solutions of order higher that one. The notion of approximate gcd is now used to define defferent notion of approximate zero polynomial for linear systems.

8.5 GRASSMANN INVARIANT OF LINEAR SYSTEMS AND APPROXIMATE ZERO POLYNOMIALS

The results in Chapters 5 and 6 are now used to define some important system concepts related to optimal approximate gcd. We first summarise the results by defining them on an arbitrary set of polynomials using the already introduced notation.

Definition 8.2: Let $\mathcal{T}(n,p;h+1)$ be a set of polynomial sets, $\Delta_k(n,p;h+1)$ be the k-gcd variety of $\mathbf{P}^{N-1}$ and $\mathcal{P}_{h+1,n} \in \mathcal{T}(n,p;h+1)$. We shall denote by

$$d(\mathcal{P},\Delta) = \min_{\forall \mathcal{P}^*,\varphi} \left\| S_{\mathcal{P}} - \left[\mathbf{0}_k \mid \bar{S}_{\mathcal{P}^*} \right] \hat{\Phi}_{\varphi} \right\|_F \tag{8.48}$$

where $\varphi(s) \in \mathbb{R}[s]$, $\deg(\varphi(s)) = k$, $\mathcal{P}^* \in \mathcal{T}(n-k, p-k; h+1)$ the k-distance of $\mathcal{P}_{h+1,n}$ from the the k-gcd variety $\Delta_k(n,p;h+1)$. The polynomial $\tilde{\varphi}(s)$ emerging as the solution of the optimisation problem will be called the k*-optimal approximate gcd* (k-OAGCD) and the corresponding value $d(\mathcal{P},\Delta)$ the k*-strenght* of $\mathcal{P}_{h+1,n}$.

Remark 8.12: If $d(\mathcal{P},\Delta) = 0$, then the k-OAGCD $\tilde{\varphi}(s)$ becomes the exact gcd of the set $\mathcal{P}_{h+1,n}$, which is not now coprime. In all other cases, $d(\mathcal{P},\Delta) \neq 0$, the optimal approximate gcd notion will be used.

■

In section 8.3, it has been shown that with any polynomial matrix $T(s) = [\underline{t}_1(s), \underline{t}_2(s), \ldots, \underline{t}_r(s)] \in \mathbb{R}^{q \times r}[s]$, $q \geq r$ we can always associate the multivector

$$\underline{t}(s)\wedge = \underline{t}_1(s) \wedge \underline{t}_2(s) \wedge \ldots \wedge \underline{t}_r(s) \tag{8.49}$$

where $\underline{t}(s)\wedge \in \mathbb{R}^{\sigma}[s]$, $\sigma = \begin{pmatrix} q \\ r \end{pmatrix}$. Such vectors $\underline{t}(s)\wedge$ define sets of polynomials $\mathcal{P}^{\wedge}_{\sigma,n}$ which are introduced by decomposable multivectors [Marcus et al., 1969] and this is denoted by the "$\wedge$" symbol. From the discussion in section 8.3, the following important property follows:

Lemma (8.1): Consider the set $\mathcal{T}(n,p;h+1)$ and let $\mathcal{T}^{\wedge}(n,p;h+1)$ denote its subset defined by polynomial vectors $\underline{p}(s) \in \mathbb{R}^{\sigma}[s]$, $\sigma = \begin{pmatrix} q \\ r \end{pmatrix}$, which are decomposable correspond to the $q \times r$ polynomial matrices with order n. The following properties hold true:

i) The set $\mathcal{T}^{\wedge}(n,p;h+1)$ is defined by the Grassmann variety $G(q,r;\mathbb{R}[s])$ of the projective space $\mathbf{P}^{\sigma-1}(\mathbb{R}[s])$.

ii) $\mathcal{T}^{\wedge}(n,p;h+1)$ is proper subset $\mathcal{T}(n,p;h+1)$ if $r \neq 1$ and $q \neq r-1$.

iii) $\mathcal{T}^{\wedge}(n,p;h+1)=\mathcal{T}(n,p;h+1)$ if either $r=1$ or $q=r-1$

Proof:

The result is a direct implication of the decomposability conditions for multivectors [Marcus et al., 1969], which state that the Grassmann variety of a projective space coincides with the projective space if either $r=1$ and/or $q=r-1$; In all other cases $r\neq 1$ and $q\neq r-1$, then it is a proper subset of the projective space.

■

The above has important implications on the definition of k-OAGCD defined on sets generated by decomposable multivectors obtained from polynomial matrices. The analysis is based on parameterising the perturbations that move a general set $\mathcal{P}_{\sigma,n}$, $\sigma=\begin{pmatrix}q\\r\end{pmatrix}$ to a set $\mathcal{P}'_{\sigma,n}=\mathcal{P}_{\sigma,n}+Q_{\sigma,n}\in\Delta_k(n,p;\sigma)$ and then determine k-OAGCD by minimising some norm of $Q_{\sigma,n}$. In this analysis $Q_{\sigma,n}$ and $\mathcal{P}'_{\sigma,n}$ are free.

However, any analysis that has to be interpreted back to the original polynomial matrix, has to be based on sets $\mathcal{P}'_{\sigma,n}$ generated by decomposable multivectors and it is the set denoted by $\mathcal{T}^{\wedge}(n,p;\sigma)$. The latter will be referred to as the *n-order subset* of the Grassmann variety $G(q,r;\mathbb{R}[s])$ and the sets $\mathcal{P}'_{\sigma,n}$ must be such that

$$\mathcal{P}'_{\sigma,n}\in\mathcal{T}(n,p;\sigma)\bigcap\Delta_k(n,p;\sigma)=\Delta_k^{\wedge}\mathcal{T}(n,p;\sigma) \qquad (8.50)$$

where $\Delta_k^{\wedge}\mathcal{T}(n,p;\sigma)$ will be called the decomposable subset of $\Delta_k(n,p;\sigma)$. Note:

Lemma 8.2: The set $\Delta_k^{\wedge}\mathcal{T}(n,p;\sigma)$ is always nonempty.

Proof:

The above is trivial to show and the proof is based on constructing a matrix $T(s)\in\mathbb{R}^{qxr}[s]$ which has a nontrivial right divisor.

■

We may illustrate the formulation of the problem as shown in Figure (8.1) below:

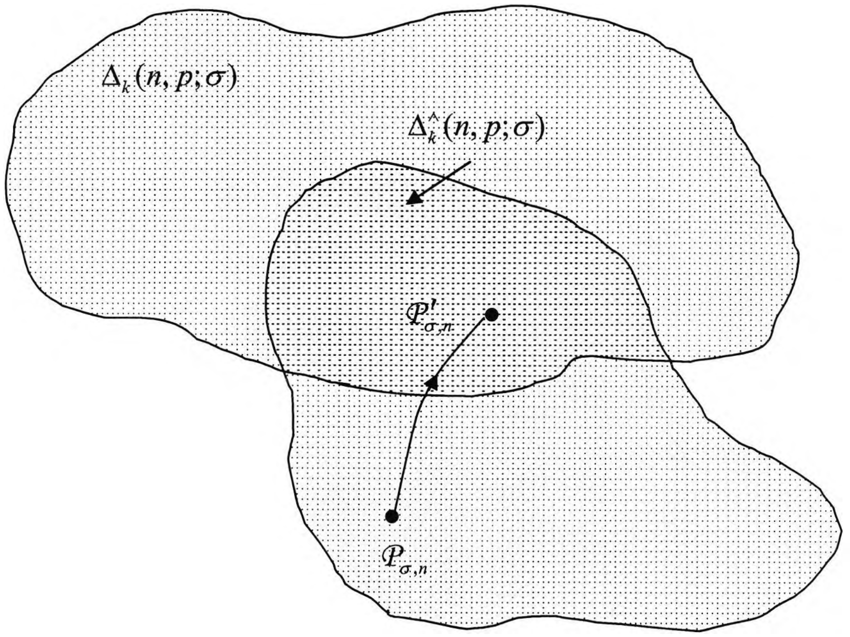

$\Delta_k(n,p;\sigma)$
$\Delta_k^{\wedge}(n,p;\sigma)$
$\mathcal{P}'_{\sigma,n}$
$\mathcal{P}_{\sigma,n}$
$\mathcal{P}_{\sigma,n} \in \Pi^{\wedge}(n,p;\sigma)$
$\mathcal{P}'_{\sigma,n} \in \Delta_k^{\wedge}(n,p;\sigma)$

Figure (8.1): Distance problem under decomposability.

The above suggests that the study of the distance problem takes now a constrained form. This is expressed by (8.48) with the additional condition that the set $\mathcal{P}'$ corresponding to $\mathcal{P}^*$ and $\hat{\Phi}_\varphi$ is decomposable. This new problem will be referred to as the *Grassmann distance problem* (GDP). The following main result is deduced from the above analysis.

Theorem 8.1: Let $\mathcal{P}_{\sigma,n} \in \Pi^{\wedge}(n,p;\sigma)$ and denote by $d(\mathcal{P},\Delta_k)$, $d(\mathcal{P},\Delta_k^{\wedge})$ the distance from $\Delta_k(n,p;\sigma)$ and $\Delta_k^{\wedge}(n,p;\sigma)$ respectively. The following hold true:

i) If $q=r-1$ or $r=1$, then the solutions of the two optimisation problems are identical and $d(\mathcal{P},\Delta_k)=d(\mathcal{P},\Delta_k^{\wedge})$

ii) If $q \neq r-1$ and $r \neq 1$, then $d(\mathcal{P},\Delta_k) \leq d(\mathcal{P},\Delta_k^{\wedge})$

Proof

i) By Lemma 8.1, part (iii) $\Pi^{\wedge}(n,p;\sigma)=\Pi(n,p;\sigma)$ that is $\mathbf{P}^{\sigma-1}$ is characterised by decomposable set and the result follows.

ii) Assuming $q \neq r-1$ and $r \neq 1$ implies again that $\Pi^{\wedge}(n,p;\sigma)$ is a proper subset of $\Pi(n,p;\sigma)$. We can always define an element of $\Delta_k(n,p;\sigma)$ which corresponds to a nondecomposable set. This may be proved by constructing simply an example. The latter then proves that if $q \neq r-1$, $r \neq 1$, then $\Delta_k^{\wedge}(n,p;\sigma)$ is a proper subset of $\Delta_k(n,p;\sigma)$ and (851) then is obvious.

■

The above suggests that the Grassmann distance problem has to be considered only when $q \neq r-1$ and $r \neq 1$. The Grassmann distance problem requires the study of some additional topicslinked to algebraic geometry and exterior algebra such as:

- Parameterisation of all decomposable sets $\mathcal{P}$ with a fixed order n.
- Characterisation of the set $\Delta_k^{\wedge}(n,p;\sigma)$ and its properties.

The above issues are central for the solution of the GDP and are topics for further research. We shall now use the distance $d(\mathcal{P},\Delta_k)$ to define the notion of *k-order almost GCD* for decomposable sets $\mathcal{P}_{\sigma,n}$, which in the special cases $r=1$, $q=r-1$ also define the following classes of approximate zero polynomials:

<u>Definition 8.3:</u> Let $T(s)=\left[\underline{t}_1(s),\underline{t}_2(s),\ldots,\underline{t}_r(s)\right]\in\mathbb{R}^{q\times r}[s]$, $q \geq r$ and let $\underline{t}(s)\wedge C_r(T(s))$ be the corresponding Grassmann vector that defines the set $\mathcal{P}_{\sigma,n}^{\wedge}$. We define:

i) The *k-almost zero polynomial* (k-AZP) of $T(s)$ the polynomial, $\tilde{\varphi}(s)$, solution of the distance problem of $\mathcal{P}_{\sigma,n}^{\wedge}$ from $\Delta_k(n,p;\sigma)$ defined by

$$d\left(\mathcal{P}_{\sigma,n}^{\wedge},\Delta_k\right)=\min_{\forall\mathcal{P}^*,\varphi}\left\|S_{\mathcal{P}^{\wedge}}-\left[\mathbf{0}_k \mid \bar{S}_{\mathcal{P}^*}\right]\hat{\Phi}_{\varphi}\right\|_{\mathrm{F}} \qquad (8.51)$$

where $\varphi(s)\in\mathbb{R}[s]$, $\deg(\varphi(s))=k$, $\mathcal{P}^{*}\in\mathcal{T}(n-k,p-k;\sigma)$; the value $d\left(\mathcal{P}_{\sigma,n}^{\wedge},\Delta_{k}\right)$ will be referred to as its *strength*.

ii) The k-optimal approximate zero polynomial (k-OAZP) of $T(s)$, $\hat{\varphi}(s)$, solution of the Grassmann distance problem of $\mathcal{P}_{\sigma,n}^{\wedge}$ from $\Delta_{k}^{\wedge}(n,p;\sigma)$ defined by

$$d\left(\mathcal{P}_{\sigma,n}^{\wedge},\Delta_{k}^{\wedge}\right)=\min_{\forall\mathcal{P}^{*},\varphi}\left\|S_{\mathcal{P}^{\wedge}}-\left[\mathbf{0}_{k}\;\vdots\;\bar{S}_{\mathcal{P}^{*}}\right]\hat{\Phi}_{\varphi}\right\|_{\mathrm{F}} \tag{8.52}$$

where $\varphi(s)\in\mathbb{R}[s]$, $\deg(\varphi(s))=k$, $\mathcal{P}^{*}\in\mathcal{T}(n-k,p-k;\sigma)$ and with the polynomial set $\mathcal{P}^{*}$ being decomposable; the value $d\left(\mathcal{P}_{\sigma,n}^{\wedge},\Delta_{k}^{\wedge}\right)$ will be referred to as its *strength*.

■

The above definition may now be applied to the various polynomial Grassmann representatives defined on a linear system and this leads to the following definition.

<u>Definition 8.4:</u> For a linear system described in a state space form $S(A,B,C,D)$, or transfer function $G(s)=N_{r}(s)D_{r}(s)=D_{l}(s)^{-1}N_{l}(s)$ with dimensions (n,p,m) we define:

i) For the pair (A,B) with controllability -GR $\underline{g}(A,B)^{t}=\underline{e}_{n}(s)^{t}P(A,B)$ its k-AZP will be called the k*-almost input decoupling zero polynomial* (k-AID-ZP) and its k-OAZP will be called the k*-Optimal Approximate input decoupling zero polynomial* (k-OAID-ZP).

ii) For the pair (A,C) with observability -GR $\underline{g}(A,B)=P(A,C)\underline{e}_{n}(s)$ its k-AZP will be called the k*-almost output decoupling zero polynomial* (k-AOD-ZP) and its k-OAZP will be called the k*-Optimal Approximate output decoupling zero polynomial* (k-OAOD-ZP).

iii) For the transfer function with column space -GR $\underline{g}_c(N) = P_c(N)\underline{e}_{\delta_c}(s)$, row space - GR $\underline{g}_r(N) = \underline{e}_{\delta_r}(s)' P_r(N)$, the corresponding k-AZP will be called the *k-almost column zero polynomial* (k-AC-ZP), *k-almost row zero polynomial* (k-AR-ZP); Similarly the k-OAZP of $\underline{g}_c(N)$, $\underline{g}_r(N)$ will be called the *k-Optimal Approximate column zero polynomial* (k-OAC-ZP), *k-Optimal Approximate row zero polynomial* (k-OAR-ZP).

■

The definition above may be extended to "almost" and "approximate" notions introduced for the many other Grassmann polynomial representatives introduced in section 8.4 and which may cover dynamic and/or decentralised control problems. The notions of k-AZP, k-OAZP introduce new invariants for linear systems, when the polynomial sets are introduced by Grassmann vectors associated with a system. The system theoretic significance of such notions stems from their definition as solutions of distance problems and thus express the most likely system properties to emerge under model parameter perturbations, which lead to perturbations in the corresponding Grassmann vector generating the polynomial set.

Remark 8.13: The notions of k-AZP, k-OAZP introduced on polynomial sets generated by polynomial-GRs defined by a system are system invariants. The k-OAZP express system consepts obtained under minimal parameter variations on the original system model and the k-AZP provide estimates for such problems, if we do not solve the Grassman distance problem. For the case where the decomposability of polynomial sets is guaranteed, the two notions coincide.

■

The above discussion, together with the decomposability property expressed by Lemma 8.1 leads to the following results:

Corollary 8.1: For the linear system $S(A,B,C,D)$ with transfer function $G(s)$ having n-states p-inputs and m-outputs we have the following properties:

i) The k-OAID-ZP of (A,B), $\varphi_{\text{OAID}}(s)$, defines the input decoupling zero polynomial obtained by minimal perturbations of the corresponding system. The respective value of the Grassmann distance, $d^{\wedge}(A,B)$ expresses the distance of the original system from the set of uncontrollable systems with n states and p inputs.

ii) The k-OAOD-ZP of (A,C), $\varphi_{\text{OAOD}}(s)$, defines the output decoupling zero polynomial obtained by minimal perturbations of the corresponding system. The respective value of the Grassmann distance, $d^{\wedge}(A,C)$ expresses the distance of the original system from the set of uncontrollable systems with n states and p inputs.

iii) The k-OAC-ZP, $\varphi_{\text{OC}}(s)$, (k-OAR-ZP, $\varphi_{\text{OR}}(s)$)defines for the case $m > p$ ($m < p$) the zero polynomial obtained by minimal perturbations of the corresponding system and the respective value of the Grassmann distance, $d^{\wedge}(N_r)$, ($d^{\wedge}(N_l)$) expresses the distance of the original system from the set of systems which have a zero polynomial with degree k.

■

Althought the solution of the Grassmann distance problem is still an open issue, the decomposability results and the solution of the general unstructured distance problem (without the decomposability constraint) leads to the following important special results.

<u>Corollary 8.2:</u> For the linear system $S(A,B,C,D)$ with dimensions (n,p,m) and transfer function $G(s)$ the following properties hold true:

i) If $p=1$, or $p=n-1$, the k-OAID-ZP of (A,B), $\varphi_{\text{OAID}}(s)$ and the corresponding distance, $d^{\wedge}(A,B)$ are given by the solution of the general distance problem, that is they are the k-AID-ZP of (A,B), $\varphi_{\text{AID}}(s)$ and the distance, $d(A,B)$.

ii) If $m=1$ or $m=n-1$, the k-OAOD-ZP of (A,C), $\varphi_{\text{OAOD}}(s)$ and the corresponding distance, $d^{\wedge}(A,C)$ are defined by the solution of the general distance problem, that is they are the k-AOD-ZP of (A,C), $\varphi_{\text{AOD}}(s)$ and the distance, $d(A,C)$.

iii) If $m>p$, and $p=1$ or $p=m-1$ the k-OAC-ZP $\varphi_{\text{OC}}(s)$ and the corresponding distance, $d^{\wedge}(N_r)$ are given by the solution of the general distance problem.

iv) If $m<p$, and $m=1$ or $m=p-1$ the k-OAR-ZP $\varphi_{\text{OR}}(s)$ and the corresponding distance, $d^{\wedge}(N_l)$ are given by the solution of the general distance problem.

■

The cases which are not covered by the decomposability of all vectors require the study of Grassmann distance problem defined before which is an issue for further research. Although we have restricted ourselves here to three different types of polynomial-GRs and the corresponding approximate notions, the approach may be followed for other system properties associated with other sets of polynomials. Thus, notions, such as optimal approximate almost fixed pole polynomial may be introduced for decentralised control problems, as well as similar notions for systems associated with dynamic control.

8.6 DISCUSSION

The results of the previous chapters on the optimal approximate gcd have been applied to linear systems introduce new system invariants with significance in defining system properties under parameter variations on the corresponding model. The natural way for introducing such notions has been the notion of the polynomial Grassmann representative [Karcanias et al., 1984], which introduces new sets of polynomials. The fact that such vectors are exterior products of the columns of a polynomial matrix implies that the distance problem has to be considered from a subvariety of the general k-gcd variety that is the intersection with the Grassmann variety of the corresponding projective space.

This is an open issue that may be handled within the current framework, but requires some basic research on the properties and parameterisation of dynamic Grossmann varieties. Solutions to this problem were given only for special cases where the Grassmann variety coincides with the projective space.

Chapter 9:
CONCLUSIONS AND FURTHER WORK

The main objective of this thesis has been the development of a mathematical framework for the characterisation of the optimal approximate GCD of a set of polynomials and the characterisation of associated system properties. The results provide a complete solution to the optimal gcd, of a given set of polynomials. Evaluation of strength of this approximation allows evaluation of strength of approximate GCDs worked out with numerical procedures, provides a solution to root clustering and allows the characterisation of a number of gcd-dependent properties in an approximate sense. In particular, the new results, given in the thesis are:

- Parameterisation of the family of the proper controllers of scalar polynomial Diophantine equations.
- New proof of the classical Sylvester resultant theorem, based on the structure of polynomials and matrix properties.
- Characterisation of the GCD of a set of polynomials in terms of a canonical factorisation of original resultant into a product of a reduced resultant and a canonical Toeplitz matrix defined by the coefficients of the gcd.
- The characterisation of the “approximate” gcd based on the resultant factorisation and evaluation of the “optimal” approximate gcd and its “strength”.
- A theoretical framework for root clustering based on the results for the approximate gcd
- Application of the approximate GCD framework to the case of Linear System properties and introduction of metrics measuring distances from fundamental properties.

The work of this thesis provides the basis for a proper development of the approximate algebraic computations framework and by no means finishes the large number of topics which are still open. Future work has to deal with a number of issues which relate the algebraic theory, numerical analysis and implementation of the current results, and extensions to polynomial matrices and their interpretation in system theoretic way. In particular the following issues are left for future work:

- Extension of the parameterisation of proper solution of matrix Diophantine equations, their McMillan degree characterisation and definition of the minimal McMillan degree.
- Extension of the resultant factorisation to the case of characterisation of matrix divisors as a step for studying approximate factorisations for matrix polynomials development of numerical algorithms, Error and Complexity Analysis of the algorithmic procedures linked to the approximate gcd and related applications.
- Investigation of the system significance of the approximate gcd in the solution of Diophantine equations. This involves an extension of the results characterising the properties of almost zeros as disks containing at least one root of polynomial combinant
- Further development of root clustering by considering factorisations of different order work to the computation of approximate LCMs.
- Extension of the approximate gcd notion using the already introduced frame work based on the exterior algebra on [Karcanias et al., 1984] to the cases where multivectors are not necessarily always decomposable. This involves the solution of the Grassmann distance problem that is evaluation of the distance of the given decomposable set of polynomials from the subvariety defined as the intersection of the d -gcd variety.

The above problems are direct extensions of the work done in the Thesis and form an outline of major topics left for further work.

Appendix

Implementations on MATLAB 5.3 for the for the construction of the resultant of a polynomial set.

```
('Construction of the GENERALIZED SYLVESTER MATRIX')

disp(' input number of polynomials');
h=input('');

disp('  Input  the  coefficients  of  the  greater  order  polynomial  in
accending order. (Type [a0 a1 .. an])    ');
b0=input('coeff1 : ');
[l,m]=size(b0);
m=m-1;

disp(' input the coefficients of the second greater order polynomial in
accending order      ');
b1=input('coeff2 : ');
[l,n]=size(b1);
n=n-1;
i=1;

  S=[];
  i=1;
  while i<=n
     S0=zeros(1,m+n);
     j=i;
     while j<=m+i
        S0(1,j)=b0(1,j-i+1);
        j=j+1;
     end
     S=[S;S0];
     i=i+1;
  end
  S1=zeros(1,m+n);
  i=1;
  while i<=m
     S1=zeros(1,m+n);
     j=i;
     while j<=n+i
        S1(1,j)=b1(1,j-i+1);
        j=j+1;
     end
     S=[S;S1];
     i=i+1;
  end

  k=2

  while k<h
     disp(' input the coefficients of the next polynomial in accending
order      ');
     bk=input('coeff : ');
     [l,nk]=size(bk);
     nk=nk-1;
     i=1;
     while i<=m
        SK=zeros(1,m+n);
        j=i;
        while j<=nk+i
           SK(1,j)=bk(1,j-i+1);
           j=j+1;
        end
        S=[S;SK];
        i=i+1;
     end
     k=k+1;
  end
  disp(S);
  r=rank(S);
end
```

REFERENCES

[Barnett, 1990] BARNETT, S. (1990): "Matrices Methods and Applications", *Clarendon Press*, Oxford

[Barnett, 1983] BARNETT, S. (1983): "Polynomials and Linear Control Systems", *Marcel Dekker Inc.*, New York

[Corless et al, 1987] CORLESS, R. M., GIANNI, B. M., TRAGER, B. M. and WATT, S. M. "The Singular Value Decomposition for Polynomial Systems, *Int. Journ. Control*, **46** (1987), 1751-1760.

[Emiris et al, 1997] EMIRIS, I. Z., GALLIGO, A. and LOMBARDI, H. (1996): "Certified approximate polynomial gcd" *J. Pure Appl. Algebra* **117/118**, 229-251

[Fatouros et al, 2002] FATOUROS, S, KARCANIAS, N and MITROULI, M (2002): "Approximate Greatest Common Divisor of many polynomials and generalised resultants" *ACA' 2002 Conference*, June 25-28, 2002, Volos, Greece

[Fatouros et al, 2003] FATOUROS, S. and KARCANIAS, N. (2003): "The GCD of many polynomials and the factorisation of generalised resultants", Submitted for publication (2003), Control Engineering Centre, City University London

[Fatouros et al, 2003, 2] FATOUROS, S., KARCANIAS, N. AND MITROULI, M. (2003): "Approximate Solutions to Root Clustering Problem", *11th Mediterranean Conference on Control and Automation MED'03*, June 18-20, 2003, Rhodes Greece.

[Foster, 1986] FOSTER, L. (1986): "Rank AND Null Space calculations using matrix decomposition without column interchanges", *Lin. Alg. And its Appl.*, **74**, 47-71.

[Gantmacher, 1988] GANTMACHER, F.R. (1998): “Matrix Theory”, *American Mathematical Society*, USA

[Giannacopoulos et al, 1984] GIANNACOPOULOS, C., KARCANIAS, N. and KALOGEROPOULOS, G. (1984): “The theory of polynomial combinants in linear system problems, in Multivariable Control: New concepts and Tools”, *Ed. S.G. Tzafestas*, 27-41.

[Halikias et al, 2003] HALIKIAS, G., FATOUROS, S. and KARCANIAS, N. (2003): "Approximate Greatest Common Divisor of Polynomials and the Structured Singular Value" *European Control Conference 2003*, University of Cambridge, UK: 1- 4 September 2003

[Hirsch et al, 1974] HIRSCH, M.,W. and SMALE, S. (1974): “Differential equations, dynamical systems & linear algebra”, Academic Press, New York &London.

[Hodge, 1952] HODGE, W.V.D and PEDOE, P.D (1952): “Methods of Algebraic Geometry”, Vol.2, *Cambridge Univ. Press*, U.K.

[Horn et al, 1] HORN, R. and JOHNSON, C. (1985): “Matrix Analysis”, *Cambridge University Press*, U.K.

[Kailath, 1980] KAILATH, T. (1980): “Linear Systems”, *Pentice Hall*, Englewood Cliffs N.J.

[Karcanias et al, 1983] KARCANIAS, N., GIANNAKOPOULOS, C. and HUBBARD, M (1983): “Amost zeros of a set of polynomials of R[s]”, *Int. J. Control*, **38**(6), 1213-1238.

[Karcanias et al, 1984] KARCANIAS, N., GIANNAKOPOULOS, C. (1984): "On Grassmann invariants and almost zeroes of linear systems and the determinantal zero, pole assignment problem". *Int. J. Control*, **40**, 673.

[Karcanias et al, 1984, 2] KARCANIAS, N., GIANNAKOPOULOS, C. (1984): "Frequency assignment problems in linear multivariable systems: New concepts and Tools". *Ed. S.G.Tzafestas, D.Reidel Co.* (Holland), 211-232

[Karcanias, 1987] KARCANIAS, N. (1987): "Invariance properties, and characterization of the greatest common divisor of a set of polynomials", *Int. J. Control*, 46(5), 1751-1760.

[Karcanias et al, 1989] KARCANIAS, N., GIANNAKOPOULOS, C. (1989): "Necessary and Sufficient Conditions for Zero Assignment by Constant Squaring Down". *Linear Algebra and Its Applications, Special Issue on Control Theory*, Vol.122/123/124, 415-446

[Karcanias et al, 1994] KARCANIAS, N., and MITROULI, M., (1994): "A Matrix Pencil Based Method for Computation of the GCD of Polynomials", *IEEE Transactions on Automatic Control*, **39**(5), 977-981.

[Karcanias et al, 1996] KARCANIAS, N and LEVENTIDES, J. (1996): "Grassman Invariants, Matrix pencils and Linear System Properties". *Linear Algebra and its Applications*, **241-243**, 705-731.

[Karcanias et al, 1999] KARCANIAS, N., and MITROULI, M., (1999): "Approximate algebraic computations of algebraic invariants" *Symbolic methods in control systems analysis and design, IEE Control Engin. Series*, **56**, 135-168

[Karcanias et al, 2000] KARCANIAS, N., and MITROULI, M., (2000): "Numerical computation of the least common multiple of a set of polynomials", *Reliable Computing*, **4**(6), 439-457.

[Karcanias et al, 2001] KARCANIAS, N., MITROULI, M., and FATOUROS, S. (2001): "Computation of normal factorisation of polynomials using resultant sets", *IFAC Symposium on System Structure and Control*, Prague, Sep. 2001.

[Karcanias et al, 2002] KARCANIAS, N. and MITROULI, M. (2002): "Normal factorisation of polynomials and computational issues", *An Inter. Journal of Computers and Mathematics with Applications, Submitted 2002.*

[Karcanias et al, 2003] KARCANIAS, N., FATOUROS, S., N.MITROULI, M. and HALIKIAS, G. (2003): "Approximate Greatest Common Divisor of many polynomials, generalised resultants and strength of approximation", submitted to *Journal of Symbolic Computations.*

[Kouvaritakis et al, 1976] KOUVARITAKIS, B & McFARLANE, A.G.J., (1976): "Geometric approach to analysis and synthesis of system zeros. Part II: Non-Square Systems*". Int.J.Control*, **23**, 167-181.

[Kucera, 1979] KUCERA, V. (1979): "Discrete Linear Control: The Polynomial Equation Approach", *Willey*, Chichester. UK.

[Leventides, 1993]: LEVENTIDES J. (1993): "Algebrogeometric and Topological methods in Control Theory", PhD thesis, *Control Eng. Centre, City University* London.

[Leventides et al, 1995]: LEVENTIDES, J. and KARCANIAS, N. (1995): "Global asymptotic linearisation of the pole placement map: A closed form solution for the output feedback problem", *Automatica*, **31**, 1303-1309.

[Marcus et al, 1969] MARCUS, M. and MINC, H. "A Survey of Matrix Theory and Matrix Inequalities", *Dover Publications, Inc.*, New York

[Marcus, 1973] MARCUS, M (1973): "Finite dimensional multilinear algebra (in two parts)". *Marcel Deker*, New York.

[Mitrouli et al, 1993] MITROULI, M., and KARCANIAS, N., (1993): "Computation of the GCD of polynomials using Gaussian transformation and shifting", *Int. Journ. Control*, **58**, 211-228.

[Mitrouli et al, 1995] MITROULI, M., and KARCANIAS, N., (1995): "Computational aspects of almost zeros and related properties". Proceedings of the 3rd European Control Conference, Rome, Italy, Sept 1995, 2094-2099.

[Mitrouli et al, 1996] MITROULI, M., KARCANIAS, N., and KOUKOUVINOS (1996): "Further numerical aspects of the ERES algorithm for the computation of the greatest common divisor of polynomials and comparison with other existing methodologies" *Utilitas Mathematica*, **50** (1996), 335-351.

[Mitrouli et al, 1997] , MITROULI, M., KARCANIAS, N., and KOUKOUVINOS (1997): "Numerical aspects for nongeneric computations in control problems and related applications" *Congressus Numerantium*, 1997, **126**, 5-19.

[Noda et al, 1991] NODA M.-T. and SASAKI T. (1991): "Approximate GCD and its application to ill-conditioned algebraic equations" *J. Comput. Appl. Math.* **38**, 335-351

[Pace et al, 1973] PACE I.S. & BARNETT S.(1973): "Comparison of algorithms for calculation of GCD of polynomials*", Int. Journ. System Scien*,**4**, 211-226

[Rosenbrock, 1970] ROSENBROCK, H.H. (1970): "State Space and Multivariable Theory", *Nelson*, London

[Rosenbrock, 1979] ROSENBROCK, H.H. (1979): “Order, degree and complexity”. *Int. J.Control*, **19**, 323-331.

[Vardoulakis et al, 1978] VARDOULAKIS, A.I.G. and STOYLE, P.N.R (1978): “Generalised Resultant Theorem”, *J. Inst. of Maths. and its Appl.*, **22**, 331-335.

[Young et al, 1996] YOUNG, P.M. and DOYLE J.C. (1996): “Properties of the mixed μ problem and its bounds”, *IEEE Tran. Auto. Contr.*, **41**(1), 155-159

[Zhou et al, 1998] ZHOU and DOYLE J.C. (1998): “Essentials of Robust Control” *Pentice Hall Inc.*, Upper Saddle River, NJ, USA