N71-24874

NASA CR-118314

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MVT-71-1

FAILURE ACCOMMODATION IN LINEAR SYSTEMS THROUGH SELF-REORGANIZATION

by

Richard Vernon Beard

Degree of Doctor of Philosophy

MAN-VEHICLE LABORATORY
DEPARTMENT OF AERONAUTICS AND ASTRONAUTICS
CENTER FOR SPACE RESEARCH
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139

FAILURE ACCOMMODATION IN LINEAR

SYSTEMS THROUGH SELF-REORGANIZATION

by

RICHARD VERNON BEARD

B. S., Purdue University

(1964)

M. S., Purdue University

(1965)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1971

Signature of Author . . . . Richard V. Beard . . . .
Department of Aeronautics and Astronautics
December 1, 1970

Certified by . . . . Jacob L. Meiry . . . .
Thesis Supervisor

. . . . Wallace E. Vander Velde . . . . . .
Thesis Supervisor

. . . . John J. Deyst Jr. . . . .
Thesis Supervisor

Accepted by . . . . . . Sidney F. Barry . . . .
Chairman, Departmental Graduate Committee

FAILURE ACCOMMODATION IN LINEAR SYSTEMS THROUGH SELF-REORGANIZATION

by

RICHARD VERNON BEARD

Submitted to the Department of Aeronautics and Astronautics on December 1, 1970, in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

ABSTRACT

This research develops methods of self-reorganization which can provide a complex linear dynamic system with the ability to restructure itself to compensate for failures in its effectors and sensors and changes in the linear dynamics. The approach taken is to identify the failure or change in the system and use that information to restructure a feedback control loop to maintain closed-loop stability if possible. Controllability and observability criterion are used to evaluate the potential ability of a system to tolerate failures in its effectors and sensors. A lower bound is established for the number of effectors and sensors a linear time-invariant system requires for complete controllability and observability. The problem of identifying failures and changes in the system is solved through the use of detection filters, which produce error signals indicating the location of a failure or change. It is shown that it is always possible to construct a filter capable of detecting any single failure or change in the observable dynamics of the system. Extensive results are developed on the design of a filter capable of detecting a substantial number of different failures or changes. When the state of the system is fully measurable, a single filter can provide information about all effector and sensor failures and all changes in dynamics. Practical design algorithms are presented. To deal with the feedback restructuring problem several algorithms are presented for determining a linear time-invariant state feedback law. These algorithms can be used on-line to produce any desired closed-loop poles for the controllable portion of the system.

Thesis Advisors:

Jacob L. Meiry
Title: Assistant Professor of Aeronautics and Astronautics

Wallace E. Vander Velde
Title: Professor of Aeronautics and Astronautics

John J. Deyst, Jr.
Title: Assistant Professor of Aeronautics and Astronautics

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to Professor Jacob L. Meiry for his invaluable ideas and guidance throughout this research; to Professor John J. Deyst, Jr., for his unfailing interest, aid, and encouragement through several difficult periods; and to Professor Wallace E. Vander Velde for his excellent advice and editorial assistance.

Special thanks go to Ann Beliveau for her patience, perfection, and exceptional typing skills in the final preparation of this thesis.

Finally, I offer my heartfelt gratitude to my wife, Joann, whose understanding, support, and comfort prevailed through the entire period of my research.

This research was supported in part by a grant from NASA, NGL-22-009-025.

GENERAL NOTATION

I. Lower case letters indicate vector or scalar quantities; upper case letters indicate matrix quantities or Laplace transforms.

II. The following quantities are general n-vectors: w, x, z, z_i.

III. The following quantities are general integers: i, j, k, k_i, ℓ, ρ.

IV. The following are general matrix quantities: Q, S.

V. T is a general coordinate transformation; a subscripted T is a specific coordinate transformation defined in the vicinity of its use. $\hat{T}$ is a general triangular matrix; a subscripted $\hat{T}$ is a specific triangular matrix defined in the vicinity of its use.

VI. Subscripted vector and matrix quantities not appearing explicitly in the table of symbols are partitions or elements of the unscripted quantity, e.g., A_{ij} is a partition of A. A lower case letter is used when the partition is a vector or scalar quantity, e.g., b_i is the i^{th} column of B. Underscores are used occasionally to indicate a vector quantity which may be confused with a scalar quantity.

VII. The following notational rules apply to any quantities not appearing explicitly in the table of symbols:

1. $(\ )^T$ indicates a transposed quantity.

2. $\overline{(\ )}$ and $\widehat{(\ )}$ indicate transformed quantities resulting from coordinate transformations.

3. $\widetilde{(\ )}$ indicates an augmented matrix.

TABLE OF SYMBOLS

Symbol	Defined or First Used
A	(2-1)
A'	(4-133)
A''	(4-411)
$a_{ij\ell}$	(4-405)
a^*_{ij}	(4-430)
$\overline{a}^*_{ij\ell}$	(C-62)
B	(2-1)
B_f	(4-5)
b_i	(2-6)
$b_{ij\ell}$	(4-426)
C	(2-2)
C'	(4-134)
C'_2	(4-320)
$\breve{C}$	(B-5)
c_i	(2-46)
$\breve{c}_i$	(B-5)
$c(t)$	Section 3.2
D	(4-5)
D'	(4-124)
D''	(4-409)
d_i	(4-455)
E	Section 4.3.7
E_1	Section 2.4.1
$E'(s)$	(4-220)
e_{ij}	(C-8)
$\hat{e}_i$	(4-41)
$\hat{e}_{ri}$	(3-7)
$\hat{e}_{mi}$	(3-8)
F	(4-242)
F_2	(4-318)
$F(s)$	(6-34)
$F_i(s)$	(6-25)
$F_{ii}(s)$	(6-44)
f	(4-69)
f_i	Section 4.3.2
G	(4-5)
g	(4-100)
g_i	(4-244)

Symbol	Defined or First Used
g'	(4-464)
$g(t)$	(2-18)
H	(6-43)
H'	(6-45)
$H(s)$	(4-238)
$H_{bi}(s)$	(4-398)
$H_i(s)$	(4-401)
h	(2-47)
$h_{bi}(t)$	(4-398)
$h_i(t)$	(4-401)
$h_{\ell i}(t)$	(4-421)
$i_j(s)$	(2-38)
J_i	(4-297)
K	(4-167)
K_2	(4-319)
k_e	(4-274)
k_{ei}	(B-11)
L	(6-6)
L_c	(6-6)

Symbol	Defined or First Used
M	(2-25)
M'	(4-182)
M'_2	(4-321)
M_i	(2-51)
M_h	(2-50)
M_T	(A-19)
M_{Tj}	(4-432)
M_{TK}	(A-14)
$\breve{M}$	(B-2)
$M_i(t_1, t_2)$	(5-25)
m_{min}	Section 2.5.1
N'	(4-138)
N'_c	(4-142)
N_i	(2-54)
$N(s)$	(5-58)
$\overline{N}(s)$	(5-61)
$\overline{n}_{ij}$	(5-74)
$n(t)$	(3-6)
P	(4-94)
P'	(4-467)
P_i	(4-374)
$P_{\alpha i}$	(4-268)
p_i	(4-112)

Symbol	Defined or First Used
p_{ij}	(4-244)
p'_i	(4-464)
Q_d	(4-245)
Q_j	(4-349)
q'	(4-183)
q_i	Theorem 4.7
q'_i	(A-20)
$q(t)$	(5-43)
R_i	(6-57)
$R(t_1, t_2)$	(5-74)
r_{min}	(2-46)
s	(4-112)
s_j	(4-408)
$U_{dc}(s)$	(6-23)
$u(t)$	Section 3.2
$u_d(t)$	Section 3.2
$u_{dc}(t)$	(6-20)
$\hat{u}_i$	(4-298)
V	(A-4)
$V_\epsilon(s)$	(4-221)

Symbol	Defined or First Used
$V_h(s)$	(6-22)
$V_H(s)$	(6-23)
$\tilde{V}_i$	(C-78)
v_i	(A-4)
$\tilde{v}_{ij}$	(C-72)
$v_\epsilon(t)$	(4-69)
$v_h(t)$	(2-12)
$v_H(t)$	(6-42)
W	(2-3)
W_i	(2-8)
W_f	(4-77)
W_{fT}	(4-89)
W_{gi}	(4-284)
W_g	(4-470)
W'_g	(4-471)
w_i	(A-7)
w_{fi}	(A-20)
$\breve{w}_{fi}$	(B-12)
w_{Ij}	(2-55)
w_{di}	(4-247)
X_I	(2-55)
$x(t)$	(2-1)

Symbol	Defined or First Used
Y_e	(4-356)
y_d	(4-276)
$y(t)$	(2-2)
Z'	(4-293)
Z_d	(4-458)
Z_e	(4-287)
Z'_e	(4-357)
Z_I	(2-58)
z_d	(4-170)
z_{di}	(4-458)
z_E	(4-465)
z'_E	(4-484)
z_{ei}	(4-287)
z_{fi}	(6-94)
z'_i	(4-293)
$z(t)$	(4-5)
α_i	(4-106)
α_{ij}	(4-259)
α_{bij}	(2-9)
α'_{ij}	(4-346)
α_{Ei}	(4-278)
α_{ei}	(4-382)
α_{Iij}	(2-38)

Symbol	Defined or First Used
β_I, β_{Ij}	(2-58)
β	(4-90)
β'	(4-184)
β_i	(4-333)
β_e	(4-333)
Γ_i	(4-306)
$\Gamma(s)$	(5-59)
γ_{ij}	(4-290)
γ'_{ij}	(4-294)
$\tilde{\gamma}$	(4-442)
$\gamma_{ij}(s)$	(5-55)
$\epsilon(t)$	(4-6)
$\epsilon'(t)$	(4-19)
ζ_i	(C-51)
η	(6-28)
η'	(6-98)
$\eta_i(s)$	(5-36)
$\eta(s)$	(5-63)
Θ	(4-315)
θ_i	(4-310)
θ'_i	(4-363)

Symbol	Defined or First Used
κ	(6-17)
κ_i	(6-54)
Λ	(4-306)
Λ'	(4-360)
λ	(5-34)
μ	(4-108)
μ_i	(4-243)
$\mu_i(s)$	(5-40)
ν	(4-183)
ν_i	Section 4.3.3
$\nu_{ij}(s)$	(5-53)
$\nu_o(s)$	(5-64)
$\xi_i(t)$	(5-21)
$\xi_{\epsilon i}$	(5-24)
π_i	(5-20)
π_o	(5-27)
$\rho(s)$	(5-64)
$\rho_i(s)$	(5-37)

Symbol	Defined or First Used
σ_f	(4-14)
$\Phi(t, t_o)$	(2-14)
$\phi(t)$	(4-38)
$\phi_{ij\ell}(t)$	(4-424)
$\chi_j(s)$	(5-56)
$\psi(t)$	(4-39)
$\psi_{ij\ell}(t)$	(4-428)
$\omega_{ij\ell}$	(C-4)
ω^*_{ij}	(C-4)
$\tilde{\omega}_{ij\ell}$	(C-70)
$\bar{\omega}_{ij\ell}$	(C-60)

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

CHAPTER 1

INTRODUCTION

1.1 Background

A self-reorganizing system is a system capable of altering its own internal structure in order to maintain a satisfactory performance level in spite of changes or failures in its components or changes in the environment. The goal of self-reorganization is reliability. As engineering systems become more complex, the problem of achieving reliability becomes increasingly difficult. When a large number of components is involved, the chance that one or more of them will fail can be significant even if the components are highly reliable as individuals. One way of increasing overall reliability is to increase the reliability of individual components. Often such improvements must await technological developments and scientific advances in areas related to the theory, design, construction of specific components. Usually the systems engineer is concerned with another approach to achieving reliability, which is the use of redundancy. Redundancy can take many forms, but basically it may be regarded as "padding", or providing somewhat more than is necessary for the system to function satisfactorily. In this way certain component failures can be tolerated without causing the failure of the system as a whole.

One of the simplest kinds of redundancy is what might be called standby redundancy. This type of redundancy is seen in the use of spare components and backup systems. In case of failure, the malfunctioning component or system is simply replaced by the spare component or backup system. When this replacement process is carried out automatically, the system exhibits an elementary form of self-reorganization.

One of the appealing features of standby redundancy is its relative simplicity, both in design and implementation. Design of a spare component, for example, may be a simple matter of duplicating the primary component. Implementation is normally accomplished by isolating a defective component and switching in a spare. Seldom is it necessary to significantly alter other parts of the system to obtain compatibility with the spare component. Therefore, no extensive logical capacity is necessary to implement a replacement. However, even in this elementary form of reorganization, one part of the process which is not always simple is the detection or localization of a failure in time to deal with it before it causes the failure of the entire system. Some kinds of failure can be detected and located immediately by simple sensory information; for example, loss of pressure in a hydraulic system. In other cases the problem of locating a defective component is circumvented by grouping a number of components into a single unit whose failure can be detected easily. Then, instead of trying to locate a particular defective component in the unit, the entire unit is replaced. A backup system is an extreme example of this approach. It is a rather inefficient use of hardware, since a number of good components are discarded along with the defective one.

Although it can be an effective means of achieving reliability, standby redundancy with replacement reorganization has certain limitations. In many cases, providing spare components is not the most efficient use of hardware. Better performance can often be achieved by making simultaneous use of all redundant components instead of allowing them to remain idle until failure of the primary component. For instance, a number of redundant sensors measuring the same quantity can produce a more accurate estimate (i.e., a smaller variance) than a single sensor. A number of devices whose total output is the sum of individual outputs (such as force-applying devices or parallel connected amplifiers) can also be used more effectively in concert than individually. Not only is the total capacity or saturation level increased, but the average operating level of each device is reduced. A lower operating level may yield a longer average lifetime for each device. The same argument applies to a group of components whose total output is the product of individual outputs, such as cascaded amplifiers. Admittedly, in the case of components with limited lifetimes which are not much affected by operating levels, standby redundancy may still be the most effective way to achieve acceptable reliability.

A second limitation of standby redundancy is that it provides little protection against degradation of performance due to changes in operating characteristics; for example, changes in dynamic behavior such as might be caused by environmental conditions. If the changes can be predicted prior to putting the system into operation, and they are not too numerous, it may be possible to incorporate several operating modes in the system. As changes occur, the system could be switched to the mode appropriate for existing conditions. However, determining when such changes occur may still be a significant problem. If the changes are not known ahead of time, then a more general restructuring capability will be necessary to deal with them.

The motivation, then, for turning to more sophisticated self-reorganization schemes is to produce a system with a greater capability for coping with changes in the system and in the environment, and to make more efficient use of redundancy. With greater restructuring capabilities it becomes possible to employ a kind of redundancy which is more active than the standby redundancy described above. Instead of providing spare components, redundancy is obtained by designing the active components to supplement each other, or to serve overlapping functions. Then when a component fails it is not replaced by a spare, but its function is taken over by other active components.

An important special case of this kind of redundancy is seen in the use of redundant multi-dimensional arrays of like components which measure or control a vector quantity. For example, the inertial angular velocity of a body can be measured by three orthogonal single-degree-of-freedom inertial reference gyros. By arranging more than three such gyros in a three-dimensional array, a certain degree of supplementary redundancy among the sensors is obtained. This example is a simple illustration of the more efficient use of hardware afforded by supplementary redundancy as opposed to standby redundancy. If a single redundant gyro were added to a set of three orthogonal gyros to be used purely as a replacement, it would be mounted with its input axis colinear with that of one of the first three gyros. It could then serve as a backup to that gyro only. But if it were mounted so that its input axis had a nonzero projection on all three input axis for the first gyros, then it would be supplementary to all three and complete information would be retained if any one of the gyros failed. However, the required data processing is more complex than in the standby case. Gilmore [8] has investigated such redundant gyro arrays. Another example of redundant like-component arrays can be found in multi-jet reaction control systems. Crawford [6] has considered the design and implementation of redundant reaction jet arrays in spacecraft control systems.

The use of supplementary components requires more restructuring capability than standby redundancy, because when a component fails the system must reorganize itself to function with fewer active components. Having been provided with an expanded capacity for reorganization, a system then has a potential for dealing with other changes in the system or in the environment. Some changes might be similar to a failure in that a component becomes unusable; for instance, the target of a star tracker being occulted by another body. Other changes, such as in dynamic behavior, are more subtle.

In order to administer the more sophisticated restructuring schemes, greater logical and computational capacities are required. These greater capacities have become feasible with the rapidly growing capabilities of special purpose computers. This growth has stimulated an increasing interest in various on-line restructuring schemes, exemplified by "adaptive", "self-organizing", and "self-optimizing" systems. It is difficult to make sharp distinctions among these terms, so a definitive categorization will not be attempted here. All the terms suggest a certain restructuring capability, and therefore such systems may exhibit some of the characteristics which have been used to describe self-reorganizing behavior. The approaches to restructuring used in these systems frequently bear on some of the same kinds of problems encountered in self-reorganization. Chapter 3 discusses some of the fundamental concepts on which many of the restructuring methods are based.

1.2 General Problem Description

The basic system considered in this research is a linear plant with feedback. Control forces are applied by effectors which are subject to failure. The outputs of the plant are measured by sensors which are also subject to failure. The linear dynamics are assumed to be either piecewise time-invariant or slowly time-varying. A completely reliable data processing capability is presumed. The problem is to maintain satisfactory closed-loop performance in spite of failures in the effectors and sensors and changes in the linear dynamics. Satisfactory performance means at least closed-loop stability. Some additional properties of the closed-loop dynamic behavior are also considered in situations where time is available for more extensive computation.

The sensors and effectors are assumed to be supplementary, so there are no spare components (although some of the results on failure detection can be used with standby redundancy). In case of failure, the system is expected to function with a reduced number of effectors or sensors. Chapter 2 introduces some concepts for describing more specifically the idea of supplementation as applied to sensors and effectors for a linear plant. A quantitative measure for the degree of supplementation among these components is also suggested.

The remaining chapters deal with the problem of implementing a self-reorganization scheme assuming the basic plant is given. Chapter 3, in addition to discussing some basic approaches to reorganization, presents a detailed formulation of the problem, describes the method of approach used in this research, and introduces the subject matter of the remaining chapters.

CHAPTER 2

COMPONENT SUPPLEMENTATION

2.1 General Discussion

The concept of supplementary redundancy was discussed in Chapter 1. Supplementary components were described in a general way as those which perform overlapping functions so that when one component fails its function can be taken over by others. Before one can proceed to construct systems with supplementary components, it is necessary to have more specific definitions of the properties of supplementation. This chapter investigates the supplementary properties of effectors and sensors for a linear time-invariant system.

To discuss supplementation one must first define the functions of the various components. Effectors are control devices, so it is natural to define their function in terms of controllability. Sensors are measuring devices, so it is likewise natural to define their function in terms of observability. Fortunately controllability and observability are already well-established concepts in the theory of linear systems. Sections 2.2 and 2.3 apply these concepts to individual effectors and sensors. They illustrate how the function of an effector, for example, can be defined in terms of that portion of the state space which the effector can control. A similar definition can be applied to a sensor. The remaining sections in the chapter use these results to develop several ways of defining more specifically the idea of supplementation as applied to effectors and sensors. Attention is given to the problem of how to measure degrees of supplementation among components. Such ideas provide a measure of the potential ability of a system to cope with failures of its effectors and sensors.

2.2 Partial Controllability

In this section some results concerning the concept of controllability are reviewed. The primary purpose is to illustrate how these results can be used to describe the control function of each individual effector. The ideas presented here will be used in the later sections of this chapter and also in Chapters 4 and 6 in a different context.

Consider the linear time-invariant system described by

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2-1}$$

$$y(t) = Cx(t) \tag{2-2}$$

where $x(t)$ is an n-dimensional state vector, $u(t)$ is an r-dimensional control vector, and $y(t)$ is an m-dimensional sensor output vector. The matrices A, B, and C are of dimension $n \times n$, $n \times r$, and $m \times n$ respectively. Employing the definition used by Athans and Falb [1], a state x_o is defined to be controllable at time t_o if the state of the system can be driven from $x(t_o) = x_o$ to the origin in a finite time interval by some control $u(t)$. Athans and Falb show that for the system described by (2-1) the set of controllable states is a subspace of the state space, R^n. Moreover, this subspace is spanned by the columns (considered as vectors in R^n) of the matrix

$$W = [B, AB, \ldots, A^{n-1} B] \tag{2-3}$$

or equivalently, the controllable subspace is the range space of W. Hereafter the range space of W will be referred to as the controllable space of B (with respect to A). Since by definition a state trajectory joins every state in the controllable space with the origin, this space can also be viewed as that portion of the state space which is reachable by some control u(t) starting from the origin. The matrix W has dimension $n \times (n \cdot r)$, so the number of independent columns in W, and thus the rank of W, can be no greater than n. If the rank of W is n, the system given by (2-1) is said to be controllable and (A, B) is a controllable pair. If the rank of W is less than n, the system is only partially controllable.

Each component of the control vector in (2-1) is considered to be the control force applied by one effector. To clearly indicate the action of each of the r effectors (2-1) can be written as

$$\dot{x}(t) = Ax(t) + b_1 u_1(t) + \ldots + b_r u_r(t) \tag{2-4}$$

where $u_i(t)$ is the i^{th} component of u(t) and b_i is the i^{th} column of B

$$u(t) = \begin{bmatrix} u_1(t) \\ \cdot \\ \cdot \\ \cdot \\ u_r(t) \end{bmatrix} \tag{2-5}$$

$$B = [b_1, \ldots, b_r] \tag{2-6}$$

Now suppose the system is being controlled by only one effector, say the i^{th} effector. Then the state equation is

$$\dot{x}(t) = Ax(t) + b_i u_i(t) \tag{2-7}$$

The statements concerning the controllability of (2-1) with the full control vector can be applied to (2-7) as well by simply replacing B with b_i. Define

$$W_i = [b_i, Ab_i, \ldots, A^{n-1}b_i] \tag{2-8}$$

The range space of W_i is that part of the state space which is controllable by the i^{th} effector. This means that acting alone the i^{th} effector can drive any state in the range space of W_i to the origin, or can reach any state in that space starting from the origin. The range space of W_i is the controllable space of b_i.

The matrix W_i has several important properties which are due to the manner in which the columns of W_i are generated. If the rank of W_i is k, then the first k columns of W_i (from the left) are independent and form a basis for the range space of W_i. This is verified by noting that if any column of W_i is linearly dependent on the previous columns, say

$$A^k b_i = \sum_{j=1}^{k} \alpha_{bij} A^{j-1} b_i \tag{2-9}$$

(where the α_{bij} are scalars) then by premultiplying (2-9) repeatedly by A it can be shown that $A^j b_i$ for any $j \geq k$ is also dependent on the first k columns, $\{b_i, \ldots, A^{k-1} b_i\}$. It can also be shown from (2-9) that the range space of W_i is an invariant subspace with respect to A. A subspace is invariant with respect to A if for any vector x in that subspace, Ax is also in the subspace. A subspace which has a set of basis vectors of the form $\{b_i, Ab_i, \ldots, A^{k-1} b_i\}$ is called a cyclic subspace because of the cyclic manner in which the basis is generated from b_i. The vector b_i is called the generator of the subspace. A cyclic subspace is always invariant. The concept of cyclic subspaces and their generators play an important role in the study of the structure of linear spaces and canonical matrix forms. A complete development of the results stated above can be found in Gantmacher [7] . Since the first k columns of W_i form a basis for its range space, it follows that the range space of $[b_i, Ab_i, \ldots, A^{k-1} b_i]$ is equivalent to that of W_i, and

$$\mathrm{rk}\, W_i = \mathrm{rk}[b_i, \ldots, A^{k-1} b_i] = k \tag{2-10}$$

The set of all vectors orthogonal to the range space of W_i (more precisely, orthogonal to every vector in the range space of W_i) also forms a subspace. This subspace is the null space of W_i^T. If x is any vector in this subspace, then

$$W_i^T x = \underline{0} \tag{2-11}$$

The null space of W_i^T will be referred to as the uncontrollable space of b_i. This terminology is motivated by the following observation. Consider a linear scalar function of the state variable given by

$$v_h(t) = h^T x(t) \tag{2-12}$$

where h is a time-invariant n-vector. If h lies in the uncontrollable space of b_i, then the action of the i^{th} effector can have no effect on the dynamic behavior of $v_h(t)$.

The general solution of (2-7) is

$$x(t) = \Phi(t, t_o)\, x(t_o) + \int_{t_o}^{t} \Phi(t, \tau)\, b_i u_i(\tau)\, d\tau \tag{2-13}$$

where $\Phi(t, t_o)$ is the transition matrix defined by

$$\frac{d}{dt} \Phi(t, t_o) = A\, \Phi(t, t_o) \tag{2-14}$$

$$\Phi(t_o, t_o) = I \tag{2-15}$$

(I is the identity matrix.) Since A is time-invariant, $\Phi(t, t_o)$ can be replaced by the matrix exponential

$$\begin{aligned} \Phi(t, t_o) &= e^{A(t-t_o)} \\ &= \sum_{j=0}^{\infty} \frac{A^j (t - t_o)^j}{j\,!} \end{aligned} \tag{2-16}$$

Using this series expansion for $\Phi(t, \tau)$ the integral on the right hand side of (2-13) becomes

$$\int_{t_o}^{t} \Phi(t, \tau)\, b_i u_i(\tau)\, d\tau = \sum_{j=0}^{\infty} A^j b_i \int_{t_o}^{t} \frac{(t-\tau)^j}{j\,!}\, u_i(\tau) d\tau \tag{2-17}$$

The vectors $A^j b_i$ for all j are in the range space of W_i so (2-17) can be expressed as

$$\int_{t_o}^{t} \Phi(t, \tau)\, b_i u_i(\tau)\, d\tau) = W_i g(t) \tag{2-18}$$

where $g(t)$ is some n-vector which depends on $u_i(t)$. ($g(t)$ is not unique if $\text{rk } W_i < n$). Using (2-18), (2-13) becomes

$$x(t) = \Phi(t, t_o)\, x(t_o) + W_i g(t) \tag{2-19}$$

and

$$v_h(t) = h^T x(t) = h^T \Phi(t, t_o)\, x(t_o) + h^T W_i g(t) \tag{2-20}$$

If h is in the null space of W_i^T then $W_i^T h = \underline{0}$ or $h^T W_i = \underline{0}$, and (2-20) reduces to

$$v_h(t) = h^T \Phi(t, t_o)\, x(t_o) \tag{2-21}$$

Clearly $u_i(t)$ has no effect on $v_h(t)$. In this sense the quantity $v_h(t)$ is uncontrollable with respect to the i^{th} effector. These observations concerning the controllable and uncontrollable spaces of b_i describe the capabilities and limitations of individual effectors. They will be used in Section 2.4 to determine the influence of effector failures on system control capabilities and to define more precisely the idea of complementary effectors.

2.3 Partial Observability

The results on observability presented in this section are primarily intended to serve as a basis for evaluating the capabilities of sensors and the effect of their failures on overall system capabilities. Some of the results will be used extensively in Chapter 4 as well.

The system given by (2-1) and (2-2) is said to be observable if given $y(t)$ and $u(t)$ over some time interval $[t_o, t_1]$ it is possible to determine uniquely the starting state $x(t_o)$. Substituting the general solution for $x(t_1)$ into (2-2) yields

$$y(t_1) = C\,\Phi(t_1, t_o)\,x(t_o) + C \int_{t_o}^{t_1} \Phi(t_1, \tau)\,Bu(\tau)\,d\tau \qquad (2\text{-}22)$$

To determine $x(t_o)$ it must be possible to solve the equation

$$C\,\Phi(t_1, t_o)\,x(t_o) = y_o(t_1) \qquad (2\text{-}23)$$

where

$$y_o(t_1) = y(t_1) - C \int_{t_o}^{t_1} \Phi(t_1, \tau)\,Bu(\tau)\,d\tau \qquad (2\text{-}24)$$

is a known quantity. Brockett [4] proves that for a linear time-invariant system $x(t_o)$ can be determined to within an additive constant which lies in the null space of the matrix

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}^T \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

or equivalently, the null space of

$$M = \begin{bmatrix} C \\ C \\ \vdots \\ CA^{n-1} \end{bmatrix} \qquad (2\text{-}25)$$

The system is observable then if and only if the $(m \cdot n) \times n$ matrix M has no null space. This is true if and only if rk M = n. If rk M < n the system is only partially observable.

The range space of M^T will be referred to as the observable space of C. This subspace of the state space determines the ability of the sensors to observe a scalar linear function of the state variables. Consider the scalar

$$v_h(t_o) = h^T x(t_o) \tag{2-26}$$

Given y(t) and u(t) over a time interval $[t_o, t_1]$, $x(t_o)$ can be determined to within an additive constant in the null space of M. Then $x(t_o)$ can be expressed as

$$x(t_o) = x_p + z \tag{2-27}$$

where x_p is a particular solution of (2-23), and z is some unknown vector such that

$$Mz = \underline{0} \tag{2-28}$$

Substituting (2-27) into (2-26) gives

$$v_h(t_o) = h^T x_p + h^T z \tag{2-29}$$

Now $h^T x_p$ is known, but $h^T z$ is, in general, unknown because z is unknown. Therefore $v_h(t_o)$ cannot be determined unless it is known with certainty that $h^T z = \underline{0}$. This will be the case if and only if h is orthogonal to every vector in the null space of M, or equivalently, if h lies in the range space of M^T.

It will become clear in later chapters that in a reorganization scheme sensor outputs are used not only to determine the state of a system, but also to provide information about failures and changes which may have occurred. One part of the reorganization problem is to detect changes in the dynamics of the system described by (2-1), e.g., changes in A or B. The null space of M plays an important part in determining the ability of the sensors to furnish information about such changes. This interpretation of the null space of M will be demonstrated after some basic results are established.

By reasoning similar to that used in Section 2.2 it can be shown that if $\mathrm{rk}\ M = q < n$ the matrix can be truncated after $(m \cdot q)$ rows without altering the null space. That is,

$$\mathrm{rk}\ M = \mathrm{rk} \begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{q-1} \end{bmatrix} = q \qquad (2\text{-}30)$$

and the null space of the truncated matrix is the same as the null space of M. From this fact it is easily established that the null space of M is an invariant subspace with respect to A. Suppose x is in the null space of M. Then

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{q-1} \end{bmatrix} Ax = \begin{bmatrix} CA \\ CA^2 \\ \cdot \\ \cdot \\ \cdot \\ CA^{q} \end{bmatrix} x = \underline{0} \qquad (2\text{-}31)$$

since all the rows of the matrix on the right are included in M (recall $q < n$). If Ax is in the null space of the truncated matrix, it is also in the null space of M. Therefore the null space is invariant with respect to A. A subspace which is invariant with respect to A is also invariant with respect to $\Phi(t, t_o)$ for any t and t_o. This follows from the series expansion for $\Phi(t, t_o)$ given by (2-16).

An invariant subspace with respect to $\Phi(t, t_o)$ is associated with what will be called a free-trajectory subsystem. A free trajectory is a homogeneous (undriven) solution of (2-1) and is given by

$$x(t) = \Phi(t, t_o)\, x(t_o) \tag{2-32}$$

From this equation it is clear that if $x(t_o)$ is in an invariant subspace with respect to $\Phi(t, t_o)$, then the free trajectory x(t) remains in that subspace for all t. Because the trajectory never leaves the subspace, it can be completely described by a reduced state vector whose dimension is the dimension of the subspace. Suppose the subspace has dimension ℓ and the set of vectors $\{w_{I1}, \ldots, w_{I\ell}\}$ is a basis for it. Any x(t) in the subspace can be uniquely expressed as

$$x(t) = w_{I1}\sigma_1(t) + \ldots + w_{I\ell}\sigma_\ell(t) \tag{2-33}$$

for some scalar time functions $\{\sigma_1(t), \ldots, \sigma_\ell(t)\}$. On the other hand, this set of $\sigma_i(t)$ uniquely determines x(t). The ℓ-vector

$$\begin{bmatrix} \sigma_1(t) \\ \cdot \\ \cdot \\ \cdot \\ \sigma_\ell(t) \end{bmatrix}$$

therefore uniquely determines the trajectory $x(t)$ and can be considered the state vector of a subsystem of the original system. The undriven dynamic behavior of this subsystem corresponds to the dynamic behavior of a portion of the complete system given by (2-1).

The null space of M^T, being invariant with respect to $\Phi(t, t_o)$, can be associated with a free-trajectory subsystem. This subsystem is unobservable in several senses. First, for any trajectory in the null space of M

$$y(t) = Cx(t) = \underline{0} \tag{2-34}$$

so $y(t)$ provides no information about the state of the associated subsystem. Moreover, since the dynamic behavior of this system produces no effect on the output $y(t)$, it is clear that any scheme to identify the dynamics of the system from $y(t)$ can never produce any information about that portion of the dynamics associated with the null space of M^T. In light of these observations the null space of M^T will be referred to as the unobservable space of C.

These results are concerned with the capabilities of the complete set of m sensors modeled by (2-2). The same developments can be applied to each row of C to determine the capabilities of each individual sensor.

2.4 Invulnerability to Effector Failures

The material in this section is an attempt to provide some answers to the question of how many effector failures can be tolerated before a system becomes unable to function. Such a question is of interest because one would like to be able to design a self-reorganizing system so that it can cope with the largest possible number of effector failures. There is no unique answer to this question because there are different ways of defining the stage at which a system becomes "unable to function". In this section the concept of controllability will be used to define stages of failure.

2.4.1 Minimum Number of Effectors for Controllability

Consider the system described by (2-1). As in Section 2.2, each component of the control vector will be considered the output of one effector. Each effector is associated with the corresponding column of B. The question to be answered here is, what is the minimum number of effectors necessary to completely control the system? Or in other words, what is the smallest value of r for which there exists an $n \times r$ matrix B such that (A, B) is a controllable pair?

The answer to this question can be obtained from results concerning the invariant polynomials of a square matrix. Extensive results on invariant polynomials can be found in [7]. Only those properties necessary for present purposes will be presented here. Any $n \times n$ matrix A has associated with it a unique set of n invariant polynomials $\{i_1(s), \ldots, i_n(s)\}$ of orders $k_1, \ldots, k_n$ respectively. The polynomials have the following properties:

(1) They are monic, i.e., the coefficient of the highest power in s is unity.

(2) The product of all the invariant polynomials of A yields the characteristic polynomial of A

$$\left| Is - A \right| = i_1(s) \cdot i_2(s) \cdot \ldots \cdot i_n(s) \tag{2-35}$$

Since the characteristic polynomial of A is of order n, it follows that

$$k_1 + \ldots + k_n = n \tag{2-36}$$

(3) Each $i_j(s)$ is evenly divisible by $i_{j+1}(s)$. This implies

$$k_1 \geq k_2 \geq \cdots \geq k_n \tag{2-37}$$

Normally the polynomials become trivial (equal to 1) at some point in the sequence. A typical set might look like

$$\begin{aligned}
i_1(s) &= s^{k_1} + \alpha_{I1k_1} s^{k_1 - 1} + \ldots + \alpha_{I11} \\
i_2(s) &= s^{k_2} + \alpha_{I2k_2} s^{k_2 - 1} + \ldots + \alpha_{I21} \\
&\vdots \\
i_\ell(s) &= s^{k_\ell} + \alpha_{I\ell k_\ell} s^{k_\ell - 1} + \ldots + \alpha_{I\ell 1} \\
i_{\ell+1}(s) &= 1 \\
&\vdots \\
i_n(s) &= 1
\end{aligned} \tag{2-38}$$

where the α_{Iij} are scalars. For this set $k_{\ell+1} = k_{\ell+2} = \ldots = k_n = 0$

and

$$k_1 + \ldots + k_\ell = n \tag{2-39}$$

It will not be proven here, but the only matrices which have all non-trivial invariant polynomials are of the form σI, where σ is a scalar and I is the identity matrix.

The answer to the question posed at the beginning of the section is obtained by counting the number of nontrivial polynomials. Specifically, the minimum number of effectors necessary to make (2-1) a controllable system is equal to the number of nontrivial invariant polynomials of A. To see why this is true it is necessary to investigate the way in which the invariant polynomials are obtained. The first polynomial $i_1(s)$ is the minimal polynomial for the entire state space. This means that for any vector x in the state space

$$i_1(A)x = A^{k_1} x + \alpha_{I1k_1} A^{k_1 - 1} x + \ldots + \alpha_{I11} x = \underline{0} \tag{2-40}$$

This, in fact, implies $i_1(A) = \underline{0}$. Equation (2-40) can be solved for $A^{k_1} x$ in terms of the vectors $\{x, Ax, \ldots, A^{k_1 - 1} x\}$. This implies that

$$\mathrm{rk}\,[\,x, Ax, \ldots, A^{n-1} x\,] = \mathrm{rk}\,[x, \ldots, A^{k_1 - 1} x\,] \leq k_1 \tag{2-41}$$

for any x. Replacing x in this expression by the vector b_i associated with any effector shows that the controllable space of any effector cannot have dimension larger than k_1. In other words, the largest possible subspace which is controllable by a single effector can have dimension no larger than k_1. It can be shown that there always exists a vector for which the equality in (2-41) holds. By taking b_i equal to that vector, the i^{th} effector will have a controllable space of dimension k_1. Denote such a subspace by E_1. From Section 2.2 it is known that the vectors $\{b_i, Ab_i, \ldots, A^{k_1-1} b_i\}$ form a basis for E_1.

The second polynomial $i_2(s)$ is the minimal polynomial for the state space modulo E_1. That is, for any vector x in the state space

$$i_2(A)x = z \tag{2-42}$$

where z is some vector in E_1. This equation can be solved for $A^{k_2} x$ in terms of the vectors $\{x, Ax, \ldots, A^{k_2-1} x\}$ and z. But z can be expressed in terms of the basis vectors $\{b_i, Ab_i, \ldots, A^{k_1-1} b_i\}$ for E_1. Therefore $A^{k_2} x$ can be expressed as a linear combination of the vectors $\{x, Ax, \ldots, A^{k_2-1} x, b_i, Ab_i, \ldots, A^{k_1-1} b_i\}$. This together with (2-41) implies

$$\begin{aligned} &\text{rk}[x, Ax, \ldots, A^{n-1} x, b_i, Ab_i, \ldots, A^{n-1} b_i] \\ &\quad = \text{rk}[x, Ax, \ldots, A^{k_2-1} x, b_i, Ab_i, \ldots, A^{k_1-1} b_i] \\ &\quad \leq k_1 + k_2 \end{aligned} \tag{2-43}$$

for any x. Replacing x by the vector b_j associated with any second effector and reordering the columns in (2-43) yields

$$\text{rk}[(b_i, b_j), A(b_i, b_j), \ldots, A^{n-1}(b_i, b_j)] \leq k_1 + k_2 \tag{2-44}$$

This construction demonstrates that the largest possible subspace which is controllable by two effectors can have dimension no larger than $(k_1 + k_2)$. Again it can be shown it is possible to find a b_j for which equality holds in (2-44). The same reasoning can be applied to $i_3(s)$ and so on. In general, the largest possible subspace which is controllable by r effectors has dimension $(k_1 + \ldots + k_r)$. The entire state space (and the system) is controllable by r effectors if and only if

$$k_1 + \ldots + k_r = n \tag{2-45}$$

Comparing this with (2-35) one can conclude that the minimum value of r for which (2-41) is satisfied is

$$r_{min} = \ell \tag{2-46}$$

Gantmacher [7] discusses several methods for generating the invariant polynomials from which r_{min} can be determined. One method is to reduce the characteristic matrix (Is - A) to a diagonal matrix by elementary row and column operations. Then the invariant polynomials of A appear as the diagonal elements.

A minimal set of vectors $\{b_1, \ldots, b_{r_{min}}\}$ capable of controlling the entire state space is by no means unique -- in fact, there is an infinity of such sets. No systematic procedure for determining all possible minimal sets is presented here. However, one way of selecting at least one minimal set is to transform A to one of the block diagonal standard forms derived by Gantmacher. When this is done it is possible to select a minimal set of b_i by inspection.

2.4.2 Supplementary Effectors

The previous section dealt with the question of controllability of the complete system given by (2-1). In this section attention will be focused on the ability to control a scalar linear function of the state, as given by

$$v_h(t) = h^T x(t) \tag{2-47}$$

A subset of j effectors associated with the vectors $\{b_{i_1}, \ldots, b_{i_j}\}$ will be considered supplementary with respect to control of the scalar $v_h(t)$ if they are each alone capable of controlling $v_h(t)$. Applying the results of Section 2.2, it can be seen that the i^{th} effector is capable of controlling $v_h(t)$ if and only if

$$h^T W_i \neq \underline{0} \tag{2-48}$$

where W_i is defined by (2-8). The subset of vectors $\{b_{i_1}, \ldots, b_{i_j}\}$ (from the full set $\{b_1, \ldots, b_r\}$) which satisfy (2-48) corresponds to the subset of effectors which are supplementary with respect to control of $v_h(t)$. The number of effectors in this subset is a measure of the invulnerability of the quantity $v_h(t)$ to effector failures. $v_h(t)$ will be controllable as long as any one of the effectors in the above subset is functioning. Therefore at least j effector failures (specifically, failure of all effectors in the supplementary subset) are necessary before $v_h(t)$ becomes uncontrollable. One can also associate this degree of invulnerability with the vector h. When investigating the invulnerability of a particular h, a more convenient relation which is equivalent to (2-48) is

$$M_h \, b_i \neq \underline{0} \tag{2-49}$$

where

$$M_h = \begin{bmatrix} h^T \\ h^T A \\ \cdot \\ \cdot \\ \cdot \\ h^T A^{n-1} \end{bmatrix} \tag{2-50}$$

(Note that W_i can be truncated after the k^{th} column, where $k = \text{rk } W_i$. Similarly, M_h can be truncated after the ℓ^{th} row, where $\ell = \text{rk } M_h$.)

An invulnerability degree can be associated with every direction in the state space. The direction with the least degree of invulnerability is in a sense the "weakest link" of the system with regard to controllability. This least degree of invulnerability is this minimum number of effector failures necessary for the system to become not controllable.

2.5 Invulnerability to Sensor Failures

The material in this section is analogous to the observations made in Section 2.4 concerning effector failures. The purpose is to provide some answers to the question of how many sensor failures a system can tolerate and still continue to function. Again the answer depends on how one chooses to define the point at which a system is unable to function. Observability criterion will be used for this purpose in the following sections.

2.5.1 Minimum Number of Sensors for Observability

The results of this section are most easily developed by referring to Section 2.4.1 and recognizing the duality relationship between observability and controllability. Let c_i be the i^{th} row of C. The unobservable space of c_i with respect to A coincides with the uncontrollable space of c_i^T with respect to A^T. Similarly, the observable space of c_i with respect to A coincides with the controllable space of c_i^T with respect to A^T. The invariant polynomials of A and A^T are identical [7]. Therefore the results of Section 2.4.1 show that the largest subspace which is controllable (with respect to A^T) by m effectors has a dimension $(k_1 + \ldots + k_m)$. It follows by duality that the largest subspace which is observable (with respect to A) by m sensors has dimension $(k_1 + \ldots + k_m)$. For a system matrix A with invariant polynomials (2-38), the minimum number of sensors necessary for observability is $m_{min} = \ell$. The minimum number of sensors for observability is equal to the minimum number of effectors for controllability.

2.5.2 Supplementary Sensors

This section presents two viewpoints of supplementation among sensors. The first is based on the ability to observe a scalar linear function of the state. The second is based on the ability to provide information about the subsystem dynamics.

Consider the system (2-1) with sensor outputs given by (2-2). Each component of the output vector y(t) will be considered the output of one sensor. The i^{th} sensor is associated with c_i, the i^{th}

row of C. The observations of Section 2.3 can be applied to each c_i. The scalar function $v_h(t_o)$ given by (2-26) is observable by the i^{th} sensor if and only if h lies in the range space of M_i^T, where

$$M_i = \begin{bmatrix} c_i \\ c_i A \\ \cdot \\ \cdot \\ \cdot \\ c_i A^{n-1} \end{bmatrix} \tag{2-51}$$

If rk $M_i = q_i$, then there are $n - q_i$ independent solutions of the equation

$$M_i z = \underline{0} \tag{2-52}$$

Let $\{z_{i1}, \ldots, z_{i,n-q_i}\}$ be a set of such independent solutions. These vectors form a basis for the null space of M_i. Now h is in the range space of M_i^T if and only if it is orthogonal to every vector in the null space of M_i. This will be the case if $h^T z_{i\ell} = 0$ for $\ell = 1, \ldots, n-q_i$, or equivalently,

$$h^T N_i = \underline{0} \tag{2-53}$$

where

$$N_i = [z_{i1}, \ldots, z_{i,n-q_i}] \tag{2-54}$$

By forming the subset $\{c_{i_1}, \ldots, c_{i_j}\}$ of all rows of C for which (2-53) is satisfied, one obtains the set of sensors which are supplementary with respect to the observation of $v_h(t_o)$. The number of sensors in this set is a measure of the invulnerability of $v_h(t_o)$ with respect to sensor failures. This invulnerability can be associated with the vector h as well. As in the case of effector failures, an (observation) invulnerability can be associated with every direction in the state space. The direction (or directions) with the least degree of invulnerability is the weakest part of the system in terms of observability. This least degree of observation invulnerability is the minimum number of sensor failures necessary for the system to become not observable.

It is also possible to interpret invulnerability in terms of determining subsystem dynamics. As indicated in Section 2.2, an invariant subspace with respect to A can be associated with a free-trajectory subsystem. Suppose the subsystem of interest is associated with a certain ℓ-dimensional invariant subspace defined by the basis vectors $\{w_{I1}, \ldots, w_{I\ell}\}$. Define the $n \times \ell$ matrix

$$X_I = [w_{I1}, \ldots, w_{I\ell}] \tag{2-55}$$

The invariant subspace is the range space of X_I. It can be shown that if $rk(M_i X_I) < \ell$, then the i^{th} sensor can provide information about only a portion of the dynamics of the subsystem associated with the range space of X_I. Assume

$$\ell - rk(M_i X_I) = k > 0 \tag{2-56}$$

Then there are k independent solutions of the equation

$$M_i X_I \beta_I = \underline{0} \tag{2-57}$$

where β_I is an ℓ-vector. Let $\{\beta_{I1}, \ldots, \beta_{Ik}\}$ be a set of such independent solutions. Define an $n \times k$ matrix

$$Z_I = X_I [\beta_{I1}, \ldots, \beta_{Ik}] \tag{2-58}$$

Note that the range space of Z_I consists of all vectors which are both in the null space of M_i and in the range space of X_I. In other words, the range space of Z_I is the intersection of the null space of M_i and the range space of X_I. Since it is the intersection of two invariant subspaces, the range space of Z_I is itself an invariant subspace. A second free-trajectory subsystem can be associated with the range space of Z_I. It is, in fact, a subsystem of the first subsystem because the range space of Z_I is contained in the range space of X_I. The range space of Z_I is also in the null space of M_i, so one may conclude from the results of Section 2.2 that the output of the i^{th} sensor can never yield any information about the dynamics of this second subsystem. In this sense, a portion of the dynamics of the first subsystem is unobservable by the i^{th} sensor. By counting the number of sensors for which $rk(M_i X_i) = \ell$ one can obtain the degree of invulnerability to sensor failures for the subsystem associated with the range space of X_I.

2.6 Summary

This chapter uses the concepts of partial controllability and observability as the basis for some criteria for evaluating the ability of a system to cope with effector and sensor failures. These criteria are offered as possible design goals for the basic system in a self-reorganizing scheme. However, they measure only a potential ability. The actual ability of a system to withstand component failures and other changes depends also on the effectiveness of the self-reorganizing loops whose function is to make advantageous use of the supplementary features built into the basic system. These self-reorganizing loops are the subject of the remaining chapters.

CHAPTER 3

SELF-REORGANIZATION

3.1 General Principles

This chapter outlines some general concepts concerning self-reorganization schemes. Specific areas to which the major results of this research apply are described in more detail. The formulations of the problems considered and the methods of attack are presented as an introduction to the following chapters.

Reorganization of a system is made necessary when a malfunction or change in the system or in the environment causes an unacceptable deterioration in the performance level. (Such an occurrence will be referred to as simply an "event".) The object of the reorganization or restructuring is, of course, to restore the performance to an acceptable level. One is quickly led to the observation that any restructuring decision is based upon information about either the performance of the system or the event which has occurred. Without at least one of these two types of information available, there is no logical basis for selecting a new structure.

Information about the performance of a system might be obtained directly from sensor outputs or it may be obtainable only indirectly by inference from measurable quantities. For example, accessible outputs of the system might be compared to a reference model. The most common types of performance information are performance level and performance gradient with respect to some structural parameters.

Higher derivatives of the performance function are usually too difficult to generate for on-line use. Knowing the performance and performance gradient for a certain system structure amounts to a local knowledge of a performance as a function of structure. Structure-changing algorithms based on such information will be local searching techniques. The local knowledge of the performance surface is used to guide small changes in structure to achieve higher performance levels. Many techniques for locally directed searches have been developed in connection with maximizing (or minimizing) a function of several variables and more recently in connection with finding optimal controls for dynamic systems. Many of the adaptive systems proposed in the literature over the past decade use performance information and locally directed searching techniques [12, 13, 22, 25].

Perhaps the greatest appeal of this approach to reorganization is that it is not necessary to make a detailed analysis of the relationships between performance and structure. The search process takes the place of such analyses, and therefore this approach is most useful in cases where accurate analysis is difficult or impossible in the design stage. Moreover, a substantial amount of imperfect knowledge about the basic system can usually be tolerated when only performance level information is required. As one would expect, a more complete knowledge of the system characteristics is required to generate performance gradient information. If these characteristics are themselves subject to change, it may be necessary to identify them before reliable performance gradient information can be generated. Thus a reorganization scheme based on performance information may also require a certain amount of event information (about system characteristics) as well.

Performance-directed searching methods have several limitations. One limitation is that it is not always possible to determine system performance. This is the case when a performance index is based on inaccessible quantities. For instance, the performance index of an inertial navigation system might be the error magnitude between estimated and true position. Since true position is not known, the performance cannot be determined on-line.

In other cases it may be possible to define a performance measure that is accessible, but which in practice becomes unsatisfactory because it is influenced too much by inaccessible effects. This may happen, for example, when comparison with a reference model is taken as a performance measure for a plant subject to unknown disturbances. If there are significant disturbances acting on the plant but not on the model, the performance measure may be too sensitive to these disturbances to be useful for reorganization.

Performance information measured on-line indicates present or past performance, whereas the information is used to determine structural changes which affect only future performance (because of delays in the restructuring process and in the system itself). This is not a serious problem provided the performance surface (performance as a function of structure) remains relatively stable in time. However, if the performance measure is significantly influenced by time-varying effects other than the restructuring process, then the performance surface may be altered too rapidly for the reorganization process to follow. The resulting performance can be poorer than if no reorganization were attempted.

Another limitation of performance-directed reorganization is concerned with the speed of the reorganization process and related questions of stability. Gradient information usually produces considerably faster convergence in the search process. However, additional delays associated with the use of gradient information can be substantial. The structural reorganization must proceed slowly enough to allow the changes to be properly reflected in the gradient information, otherwise the gradient information will be invalid. This usually means the adjustments must be made slowly with respect to the dynamic response of the basic system. Because of this, excessive searching times may result when major events occur which require large structural changes. In the meantime serious stability problems can arise. In these situations it would appear to be advantageous to try to make large changes initially which put the system structure at least in the general area of the ideal one. This leads to the concept of reorganization based upon event information.

The second basic approach to reorganization is to attempt to determine what event has occurred and to select a new structure to compensate for it. This approach can be viewed in two steps:

(1) Processing the raw data from the system to obtain information about the event which may have occurred.

(2) Using the event information to select a new structure.

The techniques used to accomplish the second step will depend on the type of event information which is generated in the first step.

As noted in Chapter 1 with the example of pressure loss in a hydraulic system, some events can be identified immediately by simple sensory information. Another source of event information is comparison of redundant data. For example, a substantial discrepancy among the outputs of several duplicate sensors might indicate that one (or more) is defective. A "majority rule" decision can be made if there is sufficient redundancy (e.g., if two out of three sensors agree). In the area of digital logic design considerable attention has been devoted to the problem of detecting errors in redundant data [10, 11, 18, 24]. If discrepancies can be traced back to a particular component, this would be an indication of malfunction.

When redundant data is not available, comparison with data from a reliable model might be used to detect discrepancies. In many cases the outputs or inputs of individual components are not accessible. This makes the localization of a failure or change a more difficult problem than simple comparison (unless, as suggested in Chapter 1, components are grouped into easily diagnosable units). Inferences must be made from observable effects on other parts of the system. Model comparison is often used in the identification of dynamic systems from input and output data. Identification of dynamic systems has received substantial attention in connection with adaptive schemes, as mentioned earlier, and also in the off-line design of process control. One technique which has been employed extensively for this purpose is the use of an adaptive model [12, 14, 16, 17, 27]. Parameters of the model are adjusted to minimize some measure of the difference between the system and the model.

If the event information identifies a specific event, then determining a new structure is a matter of establishing a connection or association between the appropriate structure and the event. The association between event and structure could be a direct association or a logical one. The use of standby redundancy and replacement reorganization described in Chapter 1 is a simple example of direct association. Failure of a component is associated directly with the new structure -- replacement of the failed component by a spare. Direct association can also be used with supplementary redundancy. One example is simply a table listing all events and their associated structures. Or a direct association could consist of a fixed functional relationship between event parameters and structural parameters. A logical association would establish a connection between event and structure on-line through the use of logical algorithms. Such an algorithm might be a kind of quick redesign process shortened by prior analysis of the basic properties of the general type of system. Direct association would be faster but less flexible than logical association.

The event information could be in the form of a set of properties or features which categorize events. Of course, if the features are sufficient to identify a specific event, then the restructuring process could be the same as described above. Instead of attempting to identify a specific event, an alternative approach would be to associate each event feature with some appropriate property or feature which the new structure should possess. These associations between event features and structural features could be established as described previously. They might also be established by a learning process. Such a learning process would amount to discovering high correlations between particular event features and structural features. To achieve learning, some feedback must be available which would indicate whether the restructuring has been successful or unsuccessful. If a training period is provided, this information would be supplied by the trainer or teacher. For on-line learning reinforcement some kind of performance information would be necessary.

Another approach to reorganization based on event information is to formulate the problem in a statistical framework. Events can be modeled as statistical events. Then the whole theory of hypothesis testing can be brought to bear on the problem of event idenfication. Once a decision is made about the occurrence of an event the restructuring process can proceed as previously described. Or in some cases, instead of making a yes or no decision about the occurrence of an event, a probability of occurrence conditioned on available information can be used as a basis for restructuring. A new structure could be selected to maximize the expected performance or minimize an expected risk. For example, the confidence in a sensor (i.e., the weight placed on its measurement in arriving at a statistical estimate) could be based on the probability that it has failed. The statistical viewpoint has been taken by Rockwell [21] in obtaining a state estimate of a system in the face of possible sensor malfunctions.

An aid to event identification which has not been considered here is the possibility of performing tests or experiments on a system or its components. Fault-detection experiments are of considerable interest in digital logic design [9,10,11]. The identification of finite-state sequential machines is often based on the construction of test input sequences which take the machine through all its transitions [3,11]. A wide-band input is often used as an aid to identification of a continuous dynamic system. For purposes of self-reorganization one is normally concerned with the problem of identifying failures and changes while the system is functioning. This usually precludes the use of any extensive tests or experiments because test inputs tend to disturb the normal operation of the system. This is not necessarily always the case, however. Sometimes it is possible to apply low-power test signals which do not adversely affect operational performance. Or, during intermittent periods of idleness a component might be isolated and tested.

In the preceding discussion greater attention has been devoted to passive event identification because it is more widely applicable to on-line use. Moreover, techniques designed for passive event identification can be used in active testing as well. The information provided by a passive event identification scheme is often enhanced when judiciously chosen test inputs can be applied to the system.

One advantage offered by reorganization based on event information is the possibility of guiding large discontinuous structural changes in a system. In this way it is possible to achieve quickly a system structure which is relatively close to the ideal one. Implementing this sub-ideal structure will hopefully achieve a sufficiently high temporary performance level to allow additional time for making smaller "fine tuning" adjustments in the structure. A second advantage of being able to make large structural changes is that it is possible to jump over areas of unstable structures. Local adjustment techniques, on the other hand, may have to go around or through unstable areas which lie in the path from the old structure to the new one. Reorganization based on performance and event information should be considered complementary techniques. When both types of information are available, the most successful reorganization scheme will be a combination of the two. Some adaptive systems presently proposed employ performance and event information at different levels in the adaptive hierarchy. For example, the adjustment of a model (based on performance information) to determine system characteristics (event information) which is then used to generate the primary system performance gradient information. From a general viewpoint it would appear that event information is most useful for initial gross restructuring, and performance information best used for subsequent "fine tuning".

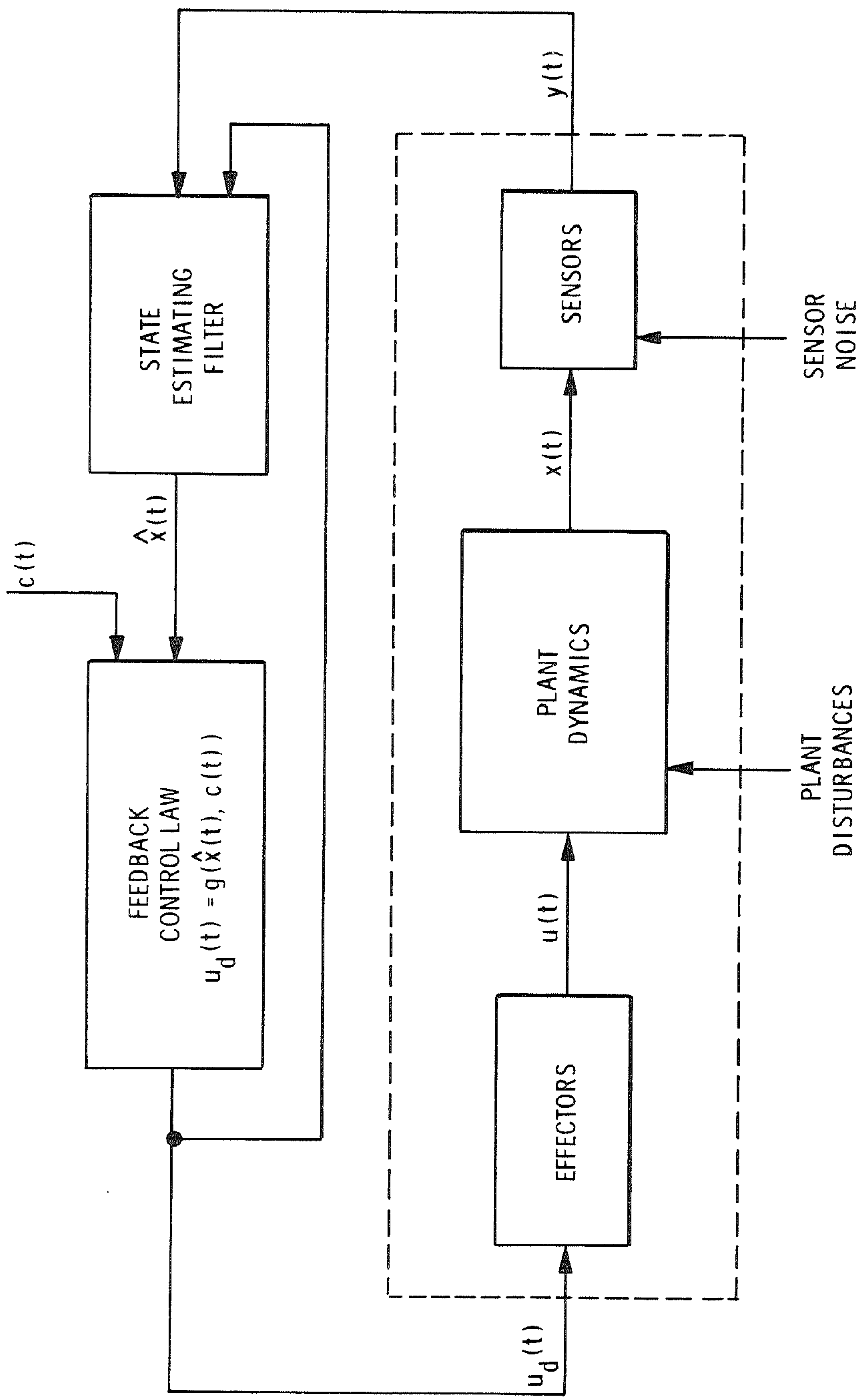
3.2 Method of Approach

The remaining chapters will be concerned with reorganization based on event information. The greatest emphasis will be on obtaining event information from raw system data. Taken together, the results provide a basis for a coherent self-reorganization scheme. However, in so far as is possible, the several areas have been developed independently so they each may be of independent interest.

The basis system configuration is shown in Figure 3-1. The quantities shown are defined as follows:

$x(t)$ -- (n-dimensional) plant state vector.

$u(t)$ -- (r-dimensional) actual control vector. This is the actual control applied to the plant by the effectors. Each component of $u(t)$ corresponds to one effector.



STATE ESTIMATING FILTER
FEEDBACK CONTROL LAW
$u_d(t) = g(\hat{x}(t), c(t))$
$c(t)$
$\hat{x}(t)$
$y(t)$
SENSORS
SENSOR NOISE
$x(t)$
PLANT DYNAMICS
PLANT DISTURBANCES
$u(t)$
EFFECTORS
$u_d(t)$

Figure 3-1.

$y(t)$ -- (m-dimensional) sensor output vector. Each component of $y(t)$ corresponds to one sensor.

$\hat{x}(t)$ -- (n-dimensional) estimated state vector.

$u_d(t)$ -- (r-dimensional) desired control signal.

$c(t)$ -- (r_c-dimensional) command signal. This may be zero for a regulator type control system or nonzero for servomechanism type control.

The plant (enclosed in the dotted line) is defined to include plant dynamics, effectors, and sensors. The following set of equations describe the plant behavior (excluding plant disturbances and sensor noise).

Plant dynamics:	$x(t) = Ax(t) + Bu(t)$	(3-1)
Effectors:	$u(t) = u_d(t)$	(3-2)
Sensors:	$y(t) = Cx(t)$	(3-3)

The matrices A, B, and C are time-invariant and have dimensions $(n \times n)$, $(n \times r)$, and $(m \times n)$ respectively. The significant feature of this plant description is that the effectors and sensors are assumed to be nondynamic. In situations where effectors or sensors have significant dynamics, such dynamics may be included in the linear plant dynamics (3-1) through the use of an enlarged state vector. The simple identity relationship (3-2) assumed for the effectors is taken for convenience. A more general functional relationship such as

$$u(t) = f\big(u_d(t)\big) \tag{3-4}$$

can be brought into the form of (3-2) by defining a new desired control vector

$$u'_d(t) = f\left(u_d(t)\right) \tag{3-5}$$

The feedback loop consists of a state estimating filter and a feedback control law generator. The filter may be designed to minimize some statistical measure of the error between x(t) and $\hat{x}(t)$, such as in a Kalman filter, or it may be designed deterministically so that $\hat{x}(t)$ approaches x(t) asymptotically in the absence of disturbances. The latter is often referred to as an "observer" [15]. This particular configuration for the feedback loop is usually seen in an optimal control formulation. The separation theorem [20] suggests this kind of structure, and it has been heuristically extended with the proposed use of observers [15, 19]. Briefly, the idea is to solve the optimal control problem, assuming the state vector is known, to obtain a state feedback control law. Then since the state vector is not completely known, an estimate of the state (from a Kalman filter or an observer) is used instead to generate the control signal. In these formulations there is no external command signal, c(t). By allowing c(t) to be nonzero, a servomechanism type formulation is possible, and the state feedback control law can be designed to satisfy classical servoanalysis criteria.

For the purpose of this research it will be assumed that all events occur in the plant and restructuring takes place in the feedback loop. A reliable data processing capability is presumed. The data processing equipment may have internal redundancy and self-correcting capabilities of its own in order to achieve reliability. The design of reliable data processing equipment is the subject of considerable (and continuing) research [9, 10, 11, 18, 24], so it will not be belabored here. The following events will be considered:

(1) Effector failure -- a departure from the intended operation of the effectors described by Equation (3-2). A failure in the i^{th} effector is modeled mathematically as

$$u(t) = u_d(t) + \hat{e}_{ri} n(t) \tag{3-6}$$

where $\hat{e}_{ri}$ is a unit r-vector in the i^{th} coordinate direction

$$\hat{e}_{ri} = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \longleftarrow i^{th} \text{ position} \\ \\ \\ \\ \\ \\ \end{matrix} \tag{3-7}$$

and n(t) is an arbitrary scalar time function.

(2) Sensor failure -- a departure from the intended operation of the sensors as described by Equation (3-3). A failure in the i^{th} sensor is modeled as

$$y(t) = Cx(t) + \hat{e}_{mi} n(t) \tag{3-8}$$

where $\hat{e}_{mi}$ is a unit m-vector in the i^{th} coordinate direction and $n(t)$ is an arbitrary scalar time function.

(3) Changes in plant dynamics -- changes in the elements of the matrices A, B, or C.

The problem of detecting and identifying these events is discussed in greater detail in Section 3.3.

The restructuring problem is concerned with altering the feedback control law and the state-estimating filter to compensate for the events described above. This problem is discussed in Section 3.4.

3.3 Detection and Identification Problem

The problem of identifying events from raw system data is considered in two steps -- detection and identification.

3.3.1 Detection and the Detection Filter

Detection refers to the process of obtaining event information based on accessible signals from the plant. The desired control vector $u_d(t)$ and the sensor output vector $y(t)$ are assumed to be accessible signals. Since all events are assumed to occur in the plant, the feedback loop is not considered in the detection process.

A solution to the detection problem is developed in Chapter 4 in the form of a detection filter. The detection filter is a linear filter driven by the accessible signals $u_d(t)$ and $y(t)$. The output of the filter is an "expected" sensor output vector. It represents the sensor outputs which would be obtained if there were no failures, changes, or other disturbances. That is, if there are no disturbances, the filter output will approach the actual sensor output vector asymptotically as the effect of initial condition errors settles out. When disturbances do occur there will be a difference between the expected output from the filter and the actual output from the sensors. This difference or error signal is the source of the desired event information. The detection filter is designed so that when particular events occur the resulting error signal behaves in a manner which is unusual and easily recognizable. Event information is obtained by looking for these unusual error responses.

It happens that in the absence of any disturbances, not only does the filter output approach the sensor output, but the state of the filter approaches the state of the plant. In this sense the detection filter is also a state-estimating filter. In some cases it may even be desirable to allow the detection filter to serve also as a state estimator. However, a filter designed for state estimation will not be a successful detection filter except by mere coincidence. Whereas a state-estimating filter is designed to suppress all errors as much as possible, the detection filter is designed to enhance and make easily recognizable those errors which result from certain events. The filter must be specifically designed to achieve this. The reason a detection filter may also be a successful state estimator is that it can (and should) be designed to suppress errors other than those associated with the events it is designed to detect. Therefore, in the absence of those particular events the errors should be small.

3.3.2 Identification Decisions

The event information obtained from the detection process, although highly correlated with the related event, may not be sufficient to identify a specific event with absolute certainty. Such uncertainty may be the result of noise disturbances, simultaneous multiple events, or events which are simply not distinguishable from each other based on the available data. Identification decisions are concerned with the problem of identifying the most likely event or events in the face of these uncertainties. Chapter 5 discusses some standard techniques for making such decisions.

3.4 Feedback Restructuring

Feedback restructuring is concerned with finding a suitable feedback control law and state-estimating filter to compensate for the events defined in Section 3.2. As was mentioned in Section 3.3.1, it is possible to use the state of a detection filter as a state estimate, eliminating the need for a separate state-estimating filter. In this case restructuring of the filter is taken care of in the solution to the detection problem and need not be considered as a separate restructuring problem. Even if a separate observer is used, the detection filter results of Chapter 4 can be used as the basis for restructuring algorithms for the observer. If a true, statistically optimal Kalman filter is desired, the Riccati equation for it will have to be resolved in whole or in part. If the speed of convergence of the Riccati equation solution is doubtful, the use of a detection filter as a temporary state-estimating filter is suggested.

Chapter 6 deals with the problem of restructuring a linear state feedback law. The main objective will be to achieve closed-loop stability with a minimum of calculation. Several secondary objectives will also be considered, however. Although the original feedback law may have been determined optimally, the time required presently to solve most optimal control problems seems to preclude the use of on-line optimal solutions as a basis for reorganization. The quadratic cost, linear regulator problem, which involves solving a matrix Riccati equation, may be one exception. But a linear feedback law, quickly obtained, could be used to achieve a stable operating condition while the more time-consuming optimal control solution is obtained. Or, a performance-directed search might be used to arrive at the final restructuring.

CHAPTER 4

DETECTION FILTERS

4.1 General Discussion

The background and basic formulation of the detection problem was discussed in Sections 3.2 and 3.3 of the previous chapter. A proposed solution -- the detection filter -- was briefly described in Section 3.3.1. This chapter deals with the design of these filters and the information they produce.

The special case in which the plant state vector is fully measurable is treated separately in the next section. It serves as an introduction to the more general case of a partially measurable state vector.

4.2 Fully Measurable State Vector

The plant being considered is the linear time-invariant system, including effectors and sensors, described by the equations

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{4-1}$$

$$u(t) = u_d(t) \tag{4-2}$$

$$y(t) = Cx(t) \tag{4-3}$$

The quantities in this plant description are defined and discussed in detail in Section 3.2. A fully measurable state vector means that for any time t Equation (4-3) can be solved uniquely for $x(t)$, given $y(t)$.

Equation (4-3) is so invertible only if

$$\mathrm{rk}\, C = n \tag{4-4}$$

This implies that there are at least n independent sensors and $m \geq n$.

The detection filter is a linear time-invariant system driven by the accessible signals $u_d(t)$ and $y(t)$. It is described by

$$\dot{z}(t) = Gz(t) + Dy(t) + B_f u_d(t) \tag{4-5}$$

where $z(t)$ is the n-dimensional state vector of the filter. The matrices G, D, and B_f (of dimension $(n \times n)$, $(n \times m)$, and $(n \times r)$ respectively) are to be chosen to produce the desired event information. The error signal which will be the source of this information is defined as the difference between the plant state and the filter state

$$\epsilon(t) = x(t) - z(t) \tag{4-6}$$

From (4-1) to (4-3) and (4-5)

$$\begin{aligned}
\dot{\epsilon}(t) &= \dot{x}(t) - \dot{z}(t) \\
&= Ax(t) + Bu(t) - Gz(t) - Dy(t) - B_f u_d(t) \\
&= (A - DC)x(t) - Gz(t) + (B - B_f)u(t) + B_f\left(u(t) - u_d(t)\right) \\
&= (A - DC)x(t) - Gz(t) + (B - B_f)u(t)
\end{aligned} \tag{4-7}$$

Now let

$$B_f = B \tag{4-8}$$

$$A - DC = G \tag{4-9}$$

Then the error equation becomes

$$\dot{\epsilon}(t) = G\epsilon(t) \tag{4-10}$$

If G is a stable matrix (i.e., if all its eigenvalues have negative real parts) then

$$\lim_{t \to \infty} \epsilon(t) = \underline{0} \tag{4-11}$$

and z(t) will approach x(t) asymptotically provided there are no disturbances. Satisfaction of (4-8) and (4-9) with a stable G therefore yields a state estimating filter. Equation (4-8), of course, can always be satisfied by choice of B_f. Because of condition (4-4), there always exists a D satisfying (4-9) for any G. If $m = n$, then C^{-1} exists and the solution is unique

$$D = (A - G)\, C^{-1} \tag{4-12}$$

If $m > n$, a (nonunique) solution is

$$D = (A - G)\,(C^T C)^{-1} C^T \tag{4-13}$$

which can be verified by substitution into (4-9). Condition (4-4) guarantees that $(C^T C)^{-1}$ exists.

Having satisfied (4-8) and (4-9) by choice of B_f and D, G can now be selected to produce the additional properties desired of a detection filter. The next three subsections will demonstrate that a judicious choice for G is

$$G = -\sigma_f I \tag{4-14}$$

where I is the $n \times n$ identity matrix and σ_f is a positive scalar. It will be shown that this choice for G results in an error signal whose direction and magnitude are directly and simply related to the event which caused the error.

4.2.1 Effector Failure Information

Assume a failure occurs in the i^{th} effector as modeled in Section 3.2 by

$$u(t) = u_d(t) + \hat{e}_{ri}\, n(t) \qquad (4\text{-}15)$$

where $\hat{e}_{ri}$ is an r-dimensional unit vector in the i^{th} coordinate direction, and n(t) is an arbitrary scalar time function. Replacing (4-2) with (4-15) and assuming (4-8) and (4-9) are satisfied, the error equation becomes

$$\begin{aligned} \dot{\epsilon}(t) &= G\epsilon(t) + B\hat{e}_{ri} n(t) \\ &= G\epsilon(t) + b_i n(t) \qquad (4\text{-}16) \end{aligned}$$

where b_i is the i^{th} column of B. Taking G as in (4-14), the solution of (4-16) is

$$\begin{aligned} \epsilon(t) &= e^{-\sigma_f(t-t_o)}\epsilon(t_o) + \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} b_i n(\tau)\, d\tau \\ &= e^{-\sigma_f(t-t_o)}\epsilon(t_o) + b_i \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} n(\tau)\, d\tau \end{aligned}$$

$$(4\text{-}17)$$

Since σ_f is positive, the initial condition term asymptotically approaches zero so

$$\epsilon(t) \approx b_i \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} n(\tau)\, d\tau \quad \text{for } (t-t_o) >> \frac{1}{\sigma_f} \tag{4-18}$$

Note that

$$\int_{t_o}^{t} e^{-\sigma_f(t-\tau)} n(\tau)\, d\tau$$

is a scalar time function, so that for sufficiently large t, $\epsilon(t)$ maintains a fixed direction in state space — namely the direction of b_i. An error signal which maintains a fixed direction in the state space corresponding to some b_i is therefore indicative of a malfunction in the i^{th} effector.

In the strict sense $\epsilon(t)$ is not an accessible signal because x(t) is not accessible. However, $\epsilon(t)$ can be generated since (4-3) can be solved uniquely for x(t). It is not necessary to solve for x(t) if one defines an output error signal.

$$\epsilon'(t) = C\epsilon(t) = y(t) - Cz(t) \tag{4-19}$$

which is directly accessible. From (4-18)

$$\epsilon'(t) \approx Cb_i \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} n(\tau)\, d\tau \quad \text{for } (t-t_o) >> \frac{1}{\sigma_f} \tag{4-20}$$

so $\epsilon'(t)$ maintains a fixed direction, Cb_i, in the m-dimensional output space. Condition (4-4) ensures that each direction in the n-dimensional state space corresponds to a unique direction in the output space.

Whereas the direction of $\epsilon'(t)$ or $\epsilon(t)$ indicates which effector has failed, the error magnitude contains information about the nature of the failure, specifically information about $n(t)$. The magnitude of $\epsilon'(t)$ or $\epsilon(t)$ is proportional to the output of a first-order linear system (with time constant $\frac{1}{\sigma_f}$) driven by $n(t)$.

4.2.2 Plant Dynamics Information

The detection filter also can produce information about changes in the elements of the matrices A, B, and C. However, there are certain changes equivalent to coordinate transformations which can never be detected from the accessible signals $y(t)$ and $u_d(t)$. Even when detectable, coordinate transformation type changes can be interpreted as changes in initial conditions. This will suggest the use of a standard form for modeling plant dynamics.

Consider a plant whose describing matrices $\{A, B, C\}$ undergo a change amounting to a coordinate transformation of the state space. The new matrices are

$$\overline{A} = T^{-1}AT \tag{4-21}$$

$$\overline{B} = T^{-1}B \tag{4-22}$$

$$\overline{C} = CT \tag{4-23}$$

where T is an $n \times n$ nonsingular matrix. Assume the change occurs at time t_o when the state of the plant is $x(t_o) = x_o$. The output for $t > t_o$ is

$$y(t) = \overline{C}e^{\overline{A}(t-t_o)} x_o + \overline{C} \int_{t_o}^{t} e^{\overline{A}(t-\tau)} \overline{B}u(\tau)\, d\tau \tag{4-24}$$

If the change had not occurred, the output would have been

$$y'(t) = Ce^{A(t-t_o)} x_o + C \int_{t_o}^{t} e^{A(t-\tau)} Bu(\tau)d\tau \tag{4-25}$$

Using (4-21) to (4-23), Equation (4-24) can be expressed in terms of the old matrices

$$y(t) = CTe^{T^{-1}AT(t-t_o)} x_o + CT \int_{t_o}^{t} e^{T^{-1}AT(t-\tau)} T^{-1}Bu(\tau)d\tau$$

$$= CTT^{-1}e^{A(t-t_o)} Tx_o + CT \int_{t_o}^{t} T^{-1} e^{A(t-\tau)} TT^{-1} Bu(\tau)d\tau$$

$$= Ce^{A(t-t_o)} Tx_o + C \int_{t_o}^{t} e^{A(t-\tau)} Bu(\tau)d\tau \tag{4-26}$$

Subtracting (4-25) from (4-26) yields

$$y(t) - y'(t) = Ce^{A(t-t_o)} (Tx_o - x_o) \tag{4-27}$$

If x_o is an eigenvector of T with eigenvalue 1, then $Tx_o = x_o$ and $y(t) = y'(t)$ for all $t > t_o$. In this case the changed plant produces the same output as the old plant would have, so it is impossible to detect the change based on y(t) and $u_d(t)$. If $Tx_o \neq x_o$ there will be a transient difference between the two outputs. In either case the control $u(t)$ causes no output differences.

Comparing (4-25) with (4-26) it is clear that the change given by (4-21) to (4-23) could instead be considered a difference in initial conditions starting at t_o. In the present context of self-reorganization the latter interpretation is preferred. Changes in A, B, or C would initiate a restructuring process, whereas a difference in initial conditions is taken care of automatically by the feedback loop. For this reason all plant descriptions which differ only by a coordinate transformation of the state space will be considered equivalent. The set of all such equivalent descriptions forms an equivalence class.

Any member of an equivalence class can be taken as representative of the entire class. For the purpose of identifying plant dynamics it is convenient to take as the representative member that description which puts the matrix C in the simplest form. In the case where there are exactly n independent sensors, C is $n \times n$ and the most convenient plant description is the one for which C is the identity matrix

$$C = I \tag{4-28}$$

With C as in (4-28) the plant equations are

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{4-29}$$

$$u(t) = u_d(t) \tag{4-30}$$

$$y(t) = x(t) \tag{4-31}$$

In this description all plant dynamics changes appear as changes in the elements of A or B. The use of equivalence properties allows changes in C to be interpreted as changes in A and B while retaining $C = I$. This presumes the change in C does not reduce its rank to less than n. If such a change does occur, condition (4-4) is violated and the state vector is no longer fully measurable. This situation is dealt with in Section 4.3 where the state vector is not assumed to be fully measurable.

Assume A and B change at time t_o by an amount ΔA and ΔB, so the plant dynamics then become

$$\dot{x}(t) = (A + \Delta A)\, x(t) + (B + \Delta B)\, u(t) \tag{4-32}$$

Using (4-30), (4-31), and (4-32) for the plant description and the detection filter as previously developed, the error equation is

$$\begin{aligned}
\dot{\epsilon}(t) &= \dot{x}(t) - \dot{z}(t) \\
&= (A + \Delta A)\, x(t) + (B + \Delta B)\, u(t) - Gz(t) \\
&\qquad - Dy(t) - B_f u_d(t) \\
&= (A - D + \Delta A)\, x(t) - Gz(t) + B\big(u(t) - u_d(t)\big) \\
&\qquad + \Delta B u(t) \\
&= G\epsilon(t) + \Delta A x(t) + \Delta B u(t) \\
&= -\sigma_f\, \epsilon(t) + \Delta\, A x(t) + \Delta B u(t)
\end{aligned} \tag{4-33}$$

The solution of (4-33) is

$$\epsilon(t) = e^{-\sigma_f(t-t_o)} \epsilon(t_o) + \Delta A \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} x(\tau) d\tau$$

$$+ \Delta B \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} u(\tau) d\tau \qquad (4\text{-}34)$$

By virtue of (4-30) and (4-31) this can also be written

$$\epsilon(t) = e^{-\sigma_f(t-t_o)} \epsilon(t_o) + \Delta A \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} y(\tau) d\tau$$

$$+ \Delta B \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} u_d(\tau) d\tau \qquad (4\text{-}35)$$

Note that ΔA and ΔB have been assumed time-invariant in obtaining (4-34) and (4-35). After the initial condition term has died out, the settled-out error is

$$\epsilon(t) = \Delta A \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} y(\tau) d\tau + \Delta B \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} u_d(\tau) d\tau$$

(4-36)

$$\text{for } (t-t_o) >> \frac{1}{\sigma_f}$$

With C = I the accessible output error signal defined by (4-19) is simply

$$\epsilon'(t) = C\epsilon(t) = \epsilon(t) \tag{4-37}$$

The components of the vector-valued time functions

$$\phi(t) = \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} y(\tau)d\tau \tag{4-38}$$

$$\psi(t) = \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} u_d(\tau)d\tau \tag{4-39}$$

can be generated as the outputs of first-order linear systems driven by the components of $y(t)$ and $u_d(t)$. Identifying changes in A and B can now be viewed as the problem of solving

$$\epsilon'(t) = \Delta A\,\phi(t) + \Delta B\psi(t) = [\Delta A, \Delta B] \begin{bmatrix} \phi(t) \\ \psi(t) \end{bmatrix} \tag{4-40}$$

given $\epsilon'(t)$, $\psi(t)$, and $\phi(t)$.

Another useful viewpoint is to consider the error produced by a change in one element of A or B. Let a_{ij} be the ij^{th} element of A. Assume a_{ij} undergoes a change to $a_{ij} + \Delta a_{ij}$ at time t_o. Then

$$\Delta A = \Delta a_{ij}\, \hat{e}_i \hat{e}_j^T \tag{4-41}$$

$$\Delta B = \underline{0} \tag{4-42}$$

where $\hat{e}_i$ and $\hat{e}_j$ are unit n-vectors in the i^{th} and j^{th} coordinate directions respectively. The settled-out error for this situation is

$$\begin{aligned} \epsilon'(t) &= a_{ij}\, \hat{e}_i\, \hat{e}_j^T\, \phi(t) \\ &= a_{ij}\, \hat{e}_i\, \phi_j(t) \end{aligned} \tag{4-43}$$

where $\phi_j(t) = \hat{e}_j^T\, \phi(t)$ is the j^{th} component of $\phi(t)$.

For a change Δb_{ij} in the ij^{th} element of B

$$\Delta A = \underline{0} \tag{4-44}$$

$$\Delta B = \Delta b_{ij}\, \hat{e}_i\, \hat{e}_{rj} \tag{4-45}$$

and the settled-out error is

$$\begin{aligned} \epsilon'(t) &= \Delta b_{ij}\, \hat{e}_i\, \hat{e}_{rj}^T\, \psi(t) \\ &= \Delta b_{ij}\, \hat{e}_i\, \psi_j(t) \end{aligned} \tag{4-46}$$

An error signal in the direction of $\hat{e}_i$ with magnitude proportional to $\phi_j(t)$ is indicative of a change in a_{ij}. An error in the same direction with magnitude proportional to $\psi_j(t)$ indicates a change in b_{ij}. The use of error information to determine ΔA and ΔB, or otherwise model the plant dynamics, is discussed in more detail in Chapter 5.

In case there are more than n sensors $(m > n)$ one can take

$$C = \begin{bmatrix} I \\ C_2 \end{bmatrix} \tag{4-47}$$

where I is $n \times n$ and C_2 is $(m - n) \times n$. This presumes that the first n sensors are independent (i.e., the first n rows of C are independent). If this is not the case, the output vector $y(t)$ can be reordered to make it so. The output relation is

$$y(t) = \begin{bmatrix} x(t) \\ C_2 x(t) \end{bmatrix} \tag{4-48}$$

Partition $y(t)$ into two vectors

$$y(t) = \begin{bmatrix} \underline{y}_1(t) \\ \\ \underline{y}_2(t) \end{bmatrix} \tag{4-49}$$

where $\underline{y}_1(t)$ is n-dimensional and $\underline{y}_2(t)$ is $(m - n)$-dimensional. Then (4-48) is equivalent to

$$\underline{y}_1(t) = x(t) \tag{4-50}$$

$$\underline{y}_2(t) = C_2 x(t) \tag{4-51}$$

The output $\underline{y}_1(t)$ can be used to generate an error signal for ΔA and ΔB in exactly the same manner as for the case $C = I$. Changes in C_2 must now be considered in addition to changes in A and B. $\underline{y}_2(t)$ can be used to produce an error signal for this possibility. Define a second error vector

$$\underline{\epsilon}_2(t) = \underline{y}_2(t) - C_2 x(t) = \underline{y}_2(t) - C_2 \underline{y}_1(t) \tag{4-52}$$

If C_2 changes to $C_2 + \Delta C_2$

$$\underline{y}_2(t) = (C_2 + \Delta C_2)\, x(t) = (C_2 + \Delta C_2)\, \underline{y}_1(t) \tag{4-53}$$

and

$$\underline{\epsilon}_2(t) = \Delta C_2 y_1(t) \tag{4-54}$$

Determining ΔC_2 is then a matter of solving (4-54) for ΔC_2 given $\underline{\epsilon}_2(t)$ and $\underline{y}_1(t)$, both of which are accessible signals.

This development assumes that C does not change in such a way that the first n sensors become dependent. If that happens, the first n rows of C would no longer be linearly independent, and there would be no coordinate transformation which could produce the form of (4-47). This technique for handling the case $m > n$ is appropriate only if there exists n sensors which can be counted upon to remain always independent, thus ensuring the state vector will always be fully measurable by those n sensors. If this is not possible, the techniques of Section 4.3.6 can be used to obtain plant dynamics information.

4.2.3 Sensor Failure Information

It was shown in Section 4.2.1 that an effector failure produces an error signal whose direction is associated with the malfunctioning effector. The situation is similar for sensor failures, except that the information provided by the error direction is not as precise. It will be shown that, in general, the error produced by a sensor failure will lie in a two-dimensional plane.

Assume a failure occurs in the i^{th} sensor as modeled in Section 3.2 by

$$y(t) = Cx(t) + \hat{e}_{mi}\, n(t) \tag{4-55}$$

where $\hat{e}_{mi}$ is a unit m-vector in the i^{th} coordinate direction, and $n(t)$ is an arbitrary scalar time function. Replacing (4-3) with (4-55) in the plant description, and using the same detection filter as before, the error equation is

$$\begin{aligned}
\dot{\epsilon}(t) &= \dot{x}(t) - \dot{z}(t) = Ax(t) + Bu(t) - Gz(t) - Dy(t) - B_f u_d(t) \\
&= (A - DC)\,x(t) - Gz(t) + B\big(u(t) - u_d(t)\big) - D\hat{e}_{mi}\, n(t) \\
&= G\epsilon(t) - D\hat{e}_{mi}\, n(t) \\
&= -\sigma_f \epsilon(t) - D\hat{e}_{mi}\, n(t)
\end{aligned} \tag{4-56}$$

The solution of (4-56) is

$$\epsilon(t) = e^{-\sigma_f(t-t_o)}\,\epsilon(t_o) - D\hat{e}_{mi} \int_{t_o}^{t} e^{-\sigma_f(t-\tau)}\, n(\tau)\, d\tau \tag{4-57}$$

and the settled-out error is

$$\epsilon(t) = -D\hat{e}_{mi} \int_{t_o}^{t} e^{-\sigma_f(t-\tau)}\, n(\tau)\, d\tau \tag{4-58}$$

Note that $\epsilon(t)$ is not an accessible signal, nor can it be generated from accessible signals. Equation (4-55) cannot be solved for $x(t)$ because $n(t)$ is an unknown. However, the output error

$$\epsilon'(t) = y(t) - Cz(t) = C\epsilon(t) + \hat{e}_{mi}\, n(t)$$

$$= - CD\hat{e}_{mi} \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} n(\tau)d\tau + \hat{e}_{mi}\, n(t) \tag{4-59}$$

is accessible. $n(t)$ and $\int_{t_o}^{t} e^{-\sigma_f(t-\tau)} n(\tau)d\tau$ are scalars, so this settled-out error always lies in the plane formed in the output space by the two m-vectors, $CD\hat{e}_{mi}$ and $\hat{e}_{mi}$. In general, $\epsilon'(t)$ will move around in this plane. The only cases in which $\epsilon'(t)$ maintains a fixed direction are

(i) if $CD\hat{e}_{mi} = \alpha\hat{e}_{mi}$ (4-60)

or

(ii) if $n(t)$ satisfies the integral equation

$$n(t) = \alpha \int_{t_o}^{t} e^{-\sigma_f(t-\tau)} n(\tau)d\tau \tag{4-61}$$

where α is an arbitrary scalar constant.

The error plane defined by $CD\hat{e}_{mi}$ and $\hat{e}_{mi}$ is the same for all equivalent plant descriptions. Equivalent descriptions are related by the coordinate transformation equations (4-21) to (4-23). The transformation relation for D is

$$\overline{D} = T^{-1}D \tag{4-62}$$

which may be verified by transforming Equation (4-9) for D. Then

$$\overline{C}\overline{D}\hat{e}_{mi} = CTT^{-1}D\,\hat{e}_{mi} = CD\,\hat{e}_{mi} \tag{4-63}$$

When $m = n$ and C is taken as the identity matrix as in Section 4.2.2, (4-9) and (4-14) can be solved uniquely for D to obtain

$$D = A + \sigma_f I \tag{4-64}$$

Then

$$\hat{e}_{mi} = \hat{e}_i \tag{4-65}$$

and

$$\begin{aligned} CD\hat{e}_{mi} &= D\hat{e}_i = (A + \sigma_f I)\,\hat{e}_i \\ &= a_i + \sigma_f\,\hat{e}_i \end{aligned} \tag{4-66}$$

where a_i is the i^{th} column of A. Then (4-59) can be written

$$\begin{aligned} \epsilon'(t) &= -(a_i + \sigma_f\hat{e}_i)\int_{t_o}^{t} e^{-\sigma_f(t-\tau)}\,n(\tau)d\tau + \hat{e}_i\,n(t) \\ &= -a_i\int_{t_o}^{t} e^{-\sigma_f(t-\tau)}\,n(\tau)d\tau \\ &\quad + \hat{e}_i\left[n(t) - \sigma_f\int_{t_o}^{t} e^{-\sigma_f(t-\tau)}\,n(\tau)d\tau\right] \end{aligned} \tag{4-67}$$

The two-dimensional error plane is uniquely determined by a_i and $\hat{e}_i$. (If a_i happens to lie along the direction of $\hat{e}_i$ then the error plane is degenerate, and the settled-out error will lie in the fixed direction of $\hat{e}_i$.) A settled-out output error which remains confined to a plane formed by a_i and $\hat{e}_i$ is indicative of a failure in the i^{th} sensor.

Each of the m sensors can be associated with an error plane in the output space. An error signal which remains in one of these error planes is indicative of a failure in the associated sensor. Since there are m error planes in the m-dimensional output space, these planes will intersect (unless all m planes are degenerate). However, even when the error planes associated with two different sensors intersect, it is still possible to differentiate between failure of the two sensors, except in the following special cases:

(1) The two error planes are coincident, or in effect, both sensors have the same error plane.

(2) The error signal maintains a fixed direction coincident with the intersection of the two error planes. In order for this to occur, the scalar n(t) representing the sensor failure in (4-55) must satisfy a particular equation of the form of (4-61).

Sections 4.2.1, 4.2.2, and 4.3.3 have described the error signal which the detection filter produces in response to individual effector failures, changes in plant dynamics, and sensor failures. Chapter 5 discusses the problem of processing the error signal to identify the most likely event (or events) in the face of uncertainties resulting from noise disturbances, simultaneous multiple events, or events which are indistinguishable based on error direction alone. The exceptional cases mentioned above are examples of the latter.

4.3 Partially Measurable State Vector

A partially measurable state vector means that

$$\text{rk } C < n \tag{4-68}$$

so (4-3) cannot be solved for $x(t)$. In the previous section it was shown that when the state vector is fully measurable, a single detection filter can produce information about all three types of events -- effector failure, sensor failure, and dynamic changes. When the state vector is only partially measurable, the capabilities of a detection filter are more limited. A single filter, in general, will not be able to produce all the information that the filter in Section 4.2 does. However, the results of this section will show that if the plant is observable, i.e., if (A, C) is an observable pair, any piece of event information found in Section 4.2 can be produced by some detection filter. The limited capacity lies in the fact that it may take a number of different filters to provide all the event information.

In order that the results which follow will be generally applicable to all three types of event information, a detection problem will be defined in formal mathematical terms. The detection filter will still be described by Equations (4-5), (4-8), and (4-9). Throughout Section 4.2 the state error defined by (4-6) always satisfied an equation of the form

$$\dot{\epsilon}(t) = G\epsilon(t) + fv_{\epsilon}(t) \tag{4-69}$$

where f is a time-invariant n-vector and $v_{\epsilon}(t)$ is a scalar. Specifically,

(i) $f = b_i$ and $v_{\epsilon}(t) = n(t)$ for an effector failure,

(ii) $f = \hat{e}_i$ and $v_\epsilon(t) = \Delta a_{ij} x_j(t)$ or

$v_\epsilon(t) = \Delta b_{ij} u_j(t)$ for a dynamics change, and

(iii) $f = -D\hat{e}_{mi}$ and $v_\epsilon(t) = n(t)$ for a sensor failure.

Equation (4-69) describes the state error for what will be considered a "simple" event — one effector failure, one sensor failure, or a change in one element of A or B.

As before $\epsilon(t)$ is not an accessible signal. The accessible error signal is the output error

$$\epsilon'(t) = y(t) - Cz(t) \tag{4-70}$$

For effector failures and dynamic changes, (4-3) is valid and

$$\epsilon'(t) = C\epsilon(t) \tag{4-71}$$

For sensor failures (4-55) replaces (4-3) and

$$\epsilon'(t) = C\epsilon(t) + \hat{e}_{mi} n(t) \tag{4-72}$$

The key feature of the detection filter in Section 4.2 is that the settled-out error $\epsilon(t)$ for a single event maintains a fixed direction in the state space. Of course, this also means that $C\epsilon(t)$ maintains a fixed direction in the output space. This is accomplished by choosing $G = -\sigma_f I$. Under condition (4-68), however, Equation (4-9) no longer has a solution, D, for every G, and in particular may not have a solution for $G = -\sigma_f I$. To make the limitations on G more explicit the state error equation (4-69) can be rewritten as

$$\dot{\epsilon}(t) = (A - DC)\,\epsilon(t) + f v_\epsilon(t) \tag{4-73}$$

by use of (4-9).

The design of detection filters is primarily concerned with being able to specify certain properties of the matrix $(A - DC)$ by choice of D. It is known that if (A, C) is an observable pair, then all n eigenvalues of $(A - DC)$ can be arbitrarily specified by choice of D [24]. The following definition concerning specification of eigenvalues of a matrix will be useful in what follows.

Definition 4.1. The eigenvalues of an $n \times n$ matrix can be specified almost arbitrarily if there exists a set of integers $\{n_1, \ldots, n_\ell\}$ with

$$n_1 + \ldots + n_\ell = n \tag{4-74}$$

such that the eigenvalues can be specified n_i at a time.

For a real matrix this imposes a slight restriction on the specification of complex eigenvalues, because they must appear in complex conjugate pairs. For example, in the case of a real 4×4 matrix $(n = 4)$ with $n_1 = 3$ and $n_2 = 1$, three of the eigenvalues must be specified as a group, then the final one is specified separately. Since complex eigenvalues must occur in conjugate pairs, the group of three eigenvalues can have at most one complex pair with one real eigenvalue. The final eigenvalue specified separately (as a group of one) must be real. The possibility of two complex conjugate pairs of eigenvalues is therefore excluded.

A formalized definition of detectability can now be stated.

Definition 4.2. The event associated with the vector f in (4-73) is detectable (or simply, f is detectable) if there exists a matrix D such that

(1) $C\epsilon(t)$ maintains a fixed direction in the output space (where $\epsilon(t)$ is the settled-out solution of (4-73) with $v_\epsilon(t)$ an arbitrary scalar time function), and

(2) at the same time, all eigenvalues of (A - DC) can be specified almost arbitrarily.

Condition (1) is the distinguishing feature of a detection filter and is the source of the event information. There are several reasons for condition (2). The matrix (A - DC) should at least be stable so that the initial condition term in the solution of (4-73) will die out. Otherwise $C\epsilon(t)$ will not settle out to a fixed direction. But beyond this, it would be desirable to have enough control over the eigenvalues of (A - DC) to be able to influence the time required for $C\epsilon(t)$ to settle out. A second reason for wanting to control the eigenvalues of (A - DC) is that it would then be possible to tailor the dynamics of the system (4-73) to the expected dynamic characteristics of the drive function $v_\epsilon(t)$, thereby enhancing the output error signal. Finally, condition (2) is somewhat easier to deal with mathematically than some alternative possibilities. What can be gained (and lost) by weakening condition (2) will become clear later in this chapter.

The next section deals with the detectability of a simple event. Sections 4.3.2, 4.3.3, and 4.3.4 are concerned with the problem of detecting a number of events with a single filter. The final three sections adapt the general results to the three types of events.

4.3.1 Detection Theorem

The main result of this section is the following theorem.

Theorem 4.1. Every vector in the state space (R^n) is detectable in the sense of Definition 4.2 if and only if (A, C) is an observable pair.

The proof of this theorem is based on a number of intermediate results concerning properties of finite-dimensional linear vector spaces. The following lemma establishes the connection between these vector spaces and condition (1) in the definition of detectability.

Lemma 4.1. Condition (1) of Definition 4.2 is satisfied if and only if

$$\operatorname{rk} C[f, (A - DC) f, \ldots, (A - DC)^{n-1} f] = 1 \tag{4-75}$$

Proof: The settled-out solution of (4-73) is

$$\epsilon(t) = \int_{t_o}^{t} e^{-[A-DC](t-\tau)} f \, v_\epsilon(\tau) \, d\tau \tag{4-76}$$

Applying the remarks of Section 2.2 to the present situation, one may conclude that $\epsilon(t)$ in (4-76) lies in the controllable space of f with respect to (A - DC), or equivalently in the range space of

$$W_f = [f, (A - DC) f, \ldots, (A - DC)^{n-1} f] \tag{4-77}$$

Therefore $\epsilon(t)$ may be expressed in the form of (2-18),

$$\epsilon(t) = W_f g(t) \tag{4-78}$$

for some n-vector g(t) which depends on $v_\epsilon(t)$. Then

$$C\epsilon(t) = CW_f g(t) \tag{4-79}$$

If $\mathrm{rk}\ CW_f = 1$, then the range space of CW_f is one-dimensional and it follows immediately that $C\epsilon(t)$ lies in a fixed direction for any $g(t)$. Therefore (4-75) is sufficient.

By the definition in Section 2.2, all states in the controllable space of f can be driven to zero by some (control) $v_\epsilon(t)$. But a state trajectory for (4-73) can be followed in either direction, so it is also possible to reach every state in the controllable space of f starting from the origin. This means $\epsilon(t)$ can be driven to any state in the range space of W_f. Therefore condition (1) can be guaranteed for arbitrary $v_\epsilon(t)$ only if $\mathrm{rk}\ CW_f = 1$. This establishes necessity and completes the proof.

Finding a D which satisfies (4-75) is the first step in designing a detection filter. The following definition is made for future ease of reference.

Definition 4.3. An $n \times m$ matrix, D, satisfying (4-75) will be referred to as a detector gain for f.

The next lemma introduces a type of vector associated with f which will be important not only in the proof of Theorem 4.1 but also in the actual design of detection filters.

Lemma 4.2. If

(i) (A, C) is an observable pair,

(ii) $\mathrm{rk}\ W_f = k$, and

(iii) $\mathrm{rk}\ CW_f = 1$

where W_f is defined by (4-77), then there exists an n-vector, g, in the controllable space of f (with respect to $[A - DC]$) such that

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{k-2} \end{bmatrix} g = \underline{0} \tag{4-80}$$

and

$$CA^{k-1} g \neq \underline{0} \tag{4-81}$$

Proof: Now

$$C(A - DC) = CA - CDC \tag{4-82}$$

$$\begin{aligned} C(A - DC)^2 &= CA(A - DC) - CDC(A - DC) \\ &= CA^2 - CADC - CDC(A - DC) \end{aligned} \tag{4-83}$$

and, in general,

$$\begin{aligned} C(A - DC)^j = CA^j - CA^{j-1}DC - CA^{j-2}DC(A - DC) - \ldots \\ \ldots - CDC(A - DC)^{j-1} \end{aligned} \tag{4-84}$$

for any j. This sequence of equations is equivalent to the single matrix equation

$$\begin{bmatrix} C \\ C(A - DC) \\ \cdot \\ \cdot \\ \cdot \\ C(A - DC)^j \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^j \end{bmatrix} - \hat{T}_j \begin{bmatrix} C \\ C(A - DC) \\ \cdot \\ \cdot \\ \cdot \\ C(A - DC)^j \end{bmatrix} \tag{4-85}$$

where $\hat{T}_j$ is an $m\cdot(j+1)$ $m\cdot(j+1)$ triangular matrix given by

$$\hat{T}_j = \begin{bmatrix} \underline{0} & \cdots & & & \underline{0} \\ CD & \ddots & & & \vdots \\ CAD & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ CA^{j-1}D & \cdots & CAD & CD & \underline{0} \end{bmatrix} \tag{4-86}$$

From (4-85),

$$[I + \hat{T}_j] \begin{bmatrix} C \\ C(A - DC) \\ \vdots \\ C(A - DC)^j \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^j \end{bmatrix} \tag{4-87}$$

From the form of $\hat{T}_j$ in (4-86) it is clear that $[I + \hat{T}_j]$ is nonsingular. Taking $j = k - 2$, (4-87) implies that (4-80) is satisfied if and only if

$$\begin{bmatrix} C \\ C(A - DC) \\ \vdots \\ C(A - DC)^{k-2} \end{bmatrix} g = \underline{0} \tag{4-88}$$

As was noted in Section 2.2, condition (ii) implies that the range space of the truncated matrix

$$W_{fT} = [f,\ (A - DC)f,\ \ldots,\ (A - DC)^{k-1} f] \tag{4-89}$$

has dimension k and coincides with the controllable space of f. Any vector in this space can be expressed (uniquely) as

$$g = W_{fT}\beta \tag{4-90}$$

where β is a k-vector. Substituting (4-90) into (4-88) yields

$$\begin{bmatrix} C \\ C(A - DC) \\ \cdot \\ \cdot \\ \cdot \\ C(A - DC)^{k-2} \end{bmatrix} W_{fT}\beta = \begin{bmatrix} CW_{fT} \\ C(A - DC)\, W_{fT} \\ \cdot \\ \cdot \\ \cdot \\ C(A - DC)^{k-2} W_{fT} \end{bmatrix} \beta = \underline{0} \tag{4-91}$$

Since β is a k-vector, (4-91) will have a nonzero solution if and only if

$$\mathrm{rk} \begin{bmatrix} CW_{fT} \\ C(A - DC)W_{fT} \\ \cdot \\ \cdot \\ \cdot \\ C(A - DC)^{k-2} W_{fT} \end{bmatrix} \leq k - 1 \tag{4-92}$$

Now

$$\mathrm{rk} \begin{bmatrix} CW_{fT} \\ C(A - DC)W_{fT} \\ \cdot \\ \cdot \\ \cdot \\ C(A - DC)^{k-2} W_{fT} \end{bmatrix} \leq \sum_{j=1}^{k-1} \mathrm{rk}\, [\, C(A - DC)^{j-1}\, W_{fT}] \tag{4-93}$$

Recall that the controllable space of f is an invariant subspace.

Therefore

$$(A - DC) W_{fT} = W_{fT} P \tag{4-94}$$

for some $k \times k$ matrix P. Then

$$(A - DC)^j W_{fT} = W_{fT} P^j \tag{4-95}$$

for any $j \geq 0$. From condition (iii)

$$\text{rk } CW_{fT} = \text{rk } CW_f = 1 \tag{4-96}$$

so

$$\text{rk } C(A - DC)^j W_{fT} = \text{rk } CW_{fT} P^j \leq \text{rk } CW_{fT} = 1 \tag{4-97}$$

Applying (4-97) to (4-93),

$$\text{rk} \begin{bmatrix} CW_{fT} \\ C(A - DC) W_{fT} \\ \cdot \\ \cdot \\ \cdot \\ C(A - DC)^{k-2} W_{fT} \end{bmatrix} \leq \sum_{j=1}^{k-1} 1 = k - 1 \tag{4-98}$$

and therefore (4-91) does have a nonzero solution for β. Then g given by (4-90) is also nonzero and satisfies (4-88) and (4-80).

Relation (4-81) follows from condition (i). Suppose

$$CA^{k-1} g = \underline{0} \tag{4-99}$$

Together with (4-80) and (4-87) this would imply

$$\begin{bmatrix} C \\ C(A - DC) \\ \cdot \\ \cdot \\ \cdot \\ C(A - DC)^{k-1} \end{bmatrix} g = \underline{0} \tag{4-100}$$

or equivalently,

$$C[g, (A - DC) g, \ldots, (A - DC)^{k-1} g] = \underline{0} \tag{4-101}$$

Now g is in the controllable space of f, which is an invariant subspace of dimension k. The cyclic space generated by g therefore can have dimension no larger than k. Then (4-101) would imply

$$C[g, (A - DC) g, \ldots, (A - DC)^{n-1} g] = \underline{0} \tag{4-102}$$

or

$$\begin{bmatrix} C \\ C(A - DC) \\ \cdot \\ \cdot \\ \cdot \\ C(A - DC)^{n-1} \end{bmatrix} g = \underline{0} \tag{4-103}$$

But with (4-87) this would mean

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{n-1} \end{bmatrix} g = \underline{0} \tag{4-104}$$

which contradicts condition (i). One must conclude that (4-99) is not true. This completes the proof.

Relation (4-81) guarantees that the cyclic space generated by g (the controllable space of g with respect to [A - DC]) is of dimension k, and so coincides with the controllable space of f. Note also that (4-80) yields

$$[g, (A - DC)g, \ldots, (A - DC)^{k-1}g] = [g, Ag, \ldots, A^{k-1}g] \tag{4-105}$$

so the set of vectors $\{g, Ag, \ldots, A^{k-1}g\}$ form a basis for the controllable space of f. It should not be construed from (4-105) that the cyclic space generated by g <u>with respect to A</u> also has dimension k. It can be larger. Now f can be expressed as

$$f = \alpha_1 g + \alpha_2 Ag + \ldots + \alpha_k A^{k-1} g \tag{4-106}$$

for some set of scalars $\{\alpha_1, \ldots, \alpha_k\}$. The magnitude of g is not restricted by (4-80) and (4-81). It will be convenient to take the magnitude so that the (nonzero) term in (4-106) with the highest power of A has a coefficient of unity. Premultiplying (4-106) by C and using (4-80) gives

$$Cf = \alpha_k CA^{k-1} g \tag{4-107}$$

If $Cf \neq \underline{0}$ then $\alpha_k \neq 0$ and the magnitude of g is taken so that $\alpha_k = 1$. In general, if for some nonnegative integer μ

$$\left.\begin{aligned} CA^j f &= \underline{0} \text{ for } j = 0, \ldots, \mu - 1 \\ CA^{\mu} f &\neq \underline{0} \end{aligned}\right\} \tag{4-108}$$

then

$$\left.\begin{aligned} \alpha_{k-j} &= 0 \text{ for } j = 0, \ldots, \mu - 1 \\ \alpha_{k-\mu} &\neq 0 \end{aligned}\right\} \qquad (4\text{-}109)$$

and g is taken so that $\alpha_{k-\mu} = 1$. The fact that (A, C) is an observable pair guarantees that (4-108) is true for some $\mu \leq k - 1$. This follows by the same reasoning used to prove (4-81). With the magnitude of g taken as above,

$$f = \alpha_1 g + \ldots + \alpha_{k-1} A^{k-2} g + A^{k-1} g \quad \text{if } Cf \neq \underline{0} \qquad (4\text{-}110)$$

or

$$f = \alpha_1 g + \ldots + \alpha_{k-\mu-1} A^{k-\mu-2} g + A^{k-\mu-1} g$$

$$\text{if (4-108) applies} \qquad (4\text{-}111)$$

<u>Definition 4.4.</u> An n-vector, g, satisfying (4-80), (4-81), and either (4-110) or (4-111) is defined to be a <u>k^{th} order detection generator</u> for f.

This terminology is motivated by the role which detection generators play in the design of detection filters. Specifically, a detection generator for f can be used to generate a detector gain for f. Lemma 4.2 demonstrates that there always exists a detection generator associated with a detector gain. The construction of the detection generator in that lemma is based on knowledge of D, since W_{fT} and in (4-90) depend on D. However, the definition of a detection generator depends only on A, C, and f, so conceptually it is independent of any particular D. The next theorem shows that if a detection generator can be found by some means based only on A, C, and f, then it is possible to write down immediately a solvable equation for D which not only yields a detector gain, but also allows arbitrary specification of k eigenvalues of (A - DC), where k is the order of the detection generator. The construction in Lemma 4.2 is not an appropriate method for finding a detection generator because it is based on knowledge of D. The problem of finding a detection generator will be discussed later.

<u>Theorem 4.2.</u> If the conditions of Lemma 4.2 are satisfied, and the k eigenvalues of [A - DC] associated with the controllable space of f are given by the roots of

$$s^k + p_k s^{k-1} + \dots + p_2 s + p_1 = 0 \qquad (4\text{-}112)$$

where the p_i are scalars and s is a complex variable, then D must be a solution of

$$DCA^{k-1}g = p_1 g + p_2 Ag + \dots + p_k A^{k-1} g + A^k g \qquad (4\text{-}113)$$

where g is a k^{th} order detection generator for f. Conversely, if there exists a k^{th} order detection generator, g, then any solution of (4-113) is a detector gain for f, and k eigenvalues of [A - DC] will be given by the roots of (4-112).

Proof:

Assume the hypothesis for the first part of the theorem. Applying the remarks of Section 2.2 to this situation with (4-112) given implies

$$(A - DC)^k f = - p_1 f - p_2(A - DC)f - \dots - p_k(A - DC)^{k-1} f \tag{4-114}$$

Lemma 4.2 establishes the existence of a k^{th} order detection generator g. Since g as well as f is a generator of the controllable space of f, (4-114) applies to g also

$$(A - DC)^k g = - p_1 g - p_2(A - DC)g - \dots - p_k(A - DC)^{k-1} g \tag{4-115}$$

Using (4-105), (4-115) reduces to

$$(A - DC)A^{k-1} g = A^k g - DCA^{k-1} g = - p_1 g - \dots - p_k A^{k-1} g \tag{4-116}$$

which is equivalent to (4-113). This proves the first part of the theorem.

Assume now there exists a k^{th} order detection generator, g. Let D be any solution of (4-113). Equation (4-115) follows from (4-113) by reversing the development above. Therefore g generates a cyclic space of dimension k with associated eigenvalues given by (4-112). Moreover,

$$\begin{aligned} \text{rk } C[g, (A - DC)g, \dots, (A - DC)^{k-1} g] &= \text{rk } C[g, Ag, \dots, A^{k-1} g] \\ &= \text{rk}[\underline{0}, \dots, \underline{0}, CA^{k-1} g] = 1 \end{aligned} \tag{4-117}$$

so D is a detector gain for g. But f is contained in the controllable space of g by virtue of (4-110) or (4-111). Hence the controllable space of f is contained in that of g, and so D is a detector gain for f as well as g. If the controllable space of f has dimension k, then it coincides with the controllable space of g and has associated eigenvalues given by (4-112). But the fact that g is a k^{th} order detection generator and D satisfies (4-113) does not necessarily mean that the controllable space of f has dimension k. For certain values of the coefficients p_i, it may have dimension less than k. In that case the eigenvalues associated with it are a subset of the k roots of (4-112). In either case, k eigenvalues of $[A - DC]$ are given by (4-112). This completes the proof of Theorem 4.2.

With the use of (4-110) or (4-111), Equation (4-113) may be put in a more convenient form. Premultiplying (4-110) by C yields

$$CA^{k-1}g = Cf \tag{4-118}$$

which gives

$$DCf = +p_1 g + \dots + p_k A^{k-1} g + A^k g \tag{4-119}$$

as the equation for a detector gain when $Cf \neq \underline{0}$. Premultiplying (4-111) by CA^{μ} yields

$$CA^{k-1}g = CA^{\mu} f \tag{4-120}$$

which gives

$$DCA^{\mu} f = +p_1 g + \dots + p_k A^{k-1} g + A^k g \tag{4-121}$$

for the detector gain when (4-108) applies. It is cumbersome and unnecessary to carry along results from both (4-110) and (4-111), since (4-110) can be viewed as a special case of (4-111) with $\mu = 0$. But rather than using the general form, the algebra will be simpler and more readable if (4-110) is used and (4-111) is brought into the form of (4-110). This can be done by premultiplying (4-111) by A^{μ} to get

$$A^{\mu} f = \alpha_1 A^{\mu} g + \dots + \alpha_{k-\mu-1} A^{k-2} g + A^{k-1} g \tag{4-122}$$

All the results which follow from (4-110) can be applied to the general case by replacing f with $A^{\mu}f$ and α_i with $\alpha_{i-\mu}$ for $i = 1, \dots, k$ (defining $\alpha_{i-\mu} = 0$ for $i \leq \mu$).

The solution of (4-119) is developed in the lemma below. Because the results will be used again later, it is presented in a general form.

Lemma 4.3. Let D, S, and Q be matrices of dimension $n \times m$, $m \times \ell$, and $n \times \ell$ respectively. If $\text{rk}\, S = \ell$ then the general solution of the equation

$$DS = Q \tag{4-123}$$

is

$$D = Q(S^T S)^{-1} S^T + D'[I - S(S^T S)^{-1} S^T] \tag{4-124}$$

where D' is an arbitrary $n \times m$ matrix.

Proof:

The general solution of (4-123) can be expressed in the form

$$D = D_p + D_o \tag{4-125}$$

where D_p is a particular solution of (4-123) and D_o is the general solution of the homogeneous equation

$$D_o S = \underline{0} \tag{4-126}$$

Since rk $S = \ell$, $\text{rk}(S^T S) = \ell$ and $(S^T S)^{-1}$ exists. A particular solution of (4-123) is

$$D_p = Q(S^T S)^{-1} S^T \tag{4-127}$$

which can be verified by direct substitution.

It can be shown that the general solution of (4-126) can be expressed in the form

$$D_o = D'[I - S(S^T S)^{-1} S^T] \tag{4-128}$$

where D' is an arbitrary $n \times m$ matrix. Let D_o' be any solution of (4-126). Take $D' = D_o'$. Then

$$D'[I - S(S^T S)^{-1} S^T] = D_o'[I - S(S^T S)^{-1} S^T] = D_o' \tag{4-129}$$

Therefore, all solutions of (4-126) can be expressed in the form of (4-128). On the other hand, $D'[I - S(S^T S)^{-1} S^T]$ is a solution of (4-126) for any D', since

$$D'[I - S(S^T S)^{-1} S^T] S = D'[S - S] = \underline{0} \tag{4-130}$$

Substituting (4-127) and (4-128) into (4-125) gives (4-124) and completes the proof.

Specializing this result to (4-119) gives

$$D = [p_1 g + \dots + p_k A^{k-1} g + A^k g] \; [(Cf)^T Cf]^{-1} (Cf)^T$$
$$+ D' \left[I - Cf[(Cf)^T Cf]^{-1} (Cf)^T \right] \tag{4-131}$$

as the general solution of (4-119). Note that $[(Cf)^T Cf]$ is a nonzero scalar since $Cf \neq \underline{0}$. For D given by (4-131)

$$A - DC = A - [p_1 g + \dots + A^k g] \; [(Cf)^T Cf]^{-1} (Cf)^T C$$
$$- D' \left[I - Cf[(Cf)^T Cf]^{-1} (Cf)^T \right] C$$
$$= A' - D'C' \tag{4-132}$$

where

$$A' = A - [p_1 g + \dots + p_k A^{k-1} g + A^k g] \; [(Cf)^T Cf]^{-1} (Cf)^T C \tag{4-133}$$

and

$$C' = \left[I - Cf[(Cf)^T Cf]^{-1} (Cf)^T \right] C \tag{4-134}$$

A brief summary of what has been accomplished up to this point is probably useful. The $[A - DC]$ given by (4-132) satisfies (4-75) which is equivalent to condition (1) for the detection of f. Condition (2) remains to be dealt with. In the process of finding a detector gain given by (4-131), k eigenvalues of $[A - DC]$ can be specified arbitrarily by selecting the set of coefficients $\{p_1, \dots, p_k\}$ as desired. Condition (2) will be satisfied only if there is enough freedom left in the choice of D to almost arbitrarily specify the remaining (n - k) eigenvalues of $[A - DC]$. The arbitrary matrix D' represents the freedom left in the choice of D after having satisfied

(4-119). Regardless of the choice of D', condition (1) will be satisfied and k eigenvalues of $[A - DC]$ will be given by (4-112). The question which now must be answered is, how many additional eigenvalues of $[A - DC] = [A' - D'C']$ can be specified by free choice of D'? The following lemma answers this question.

Lemma 4.4. If A', C', and D' are real matrices of dimension $n \times n$, $m \times n$, and $n \times m$ respectively, the number of eigenvalues of $[A' - D'C']$ which can be arbitrarily specified by free choice of D' is equal to q', where

$$q' = rk \begin{bmatrix} C' \\ C'A' \\ \cdot \\ \cdot \\ \cdot \\ C'A'^{n-1} \end{bmatrix} \qquad (4\text{-}135)$$

Moreover, for any D' the remaining $(n - q')$ eigenvalues of $[A' - D'C']$ are equal to corresponding eigenvalues of A'.

Proof:

This lemma can be proved using the fact mentioned earlier that all eigenvalues of $[A' - D'C']$ can be arbitrarily specified if and only if (A', C') is an observable pair. Let

$$M' = \begin{bmatrix} C' \\ C'A' \\ \cdot \\ \cdot \\ \cdot \\ C'A'^{n-1} \end{bmatrix} \qquad (4\text{-}136)$$

Since rk $M' = q'$, there are $(n - q')$ independent solutions of

$$M'z = \underline{0} \tag{4-137}$$

Let $\{z_1, \ldots, z_{n-q'}\}$ be a set of such independent solutions and define the $n \times (n - q')$ matrix

$$N' = [z_1, \ldots, z_{n-q'}] \tag{4-138}$$

Then

$$\text{rk } N' = n - q' \tag{4-139}$$

and

$$M'N' = \underline{0} \tag{4-140}$$

The range space of N' coincides with the null space of M'. The results in Section 2.3 show that the null space of M' is an invariant subspace with respect to A'. It follows that the range space of N' is an invariant space and therefore

$$A'N' = N'P'_N \tag{4-141}$$

for some $(n - q') \times (n - q')$ matrix P'_N. Let N'_c be any $n \times q'$ matrix such that the $n \times n$ composite matrix

$$T_{N'} = [N'_c, N'] \tag{4-142}$$

is nonsingular. $T_{N'}$ can be used to define a coordinate transformation

$$\overline{A}' = T^{-1}_{N'} A' T_{N'} \tag{4-143}$$

$$\overline{C}' = C'T_{N'} \tag{4-144}$$

$$\overline{D}' = T^{-1}_{N'} D' \tag{4-145}$$

Then

$$\bar{A}' - \bar{D}'\bar{C}' = T_{N'}^{-1} A' T_{N'} - T_{N'}^{-1} D'C' T_{N'}$$

$$= T_{N'}^{-1} [A' - D'C'] T_{N'} \qquad (4-146)$$

so $[\bar{A}' - \bar{D}'\bar{C}']$ and $[A' - D'C']$ are similar matrices and have identical eigenvalues. Also

$$\bar{M}' = \begin{bmatrix} \bar{C}' \\ \bar{C}'\bar{A}' \\ \cdot \\ \cdot \\ \cdot \\ \bar{C}'\bar{A}'^{n-1} \end{bmatrix} = \begin{bmatrix} C'T_{N'} \\ C'A'T_{N'} \\ \cdot \\ \cdot \\ \cdot \\ C'A'^{n-1}T_{N'} \end{bmatrix} = M'T_{N'}$$

$$(4-147)$$

and since $T_{N'}$ is nonsigular

$$\text{rk } \bar{M}' = \text{rk } [M'T_{N'}] = \text{rk } M' = q' \qquad (4-148)$$

From (4-143)

$$T_{N'} \bar{A}' = A'T_{N'} \qquad (4-149)$$

If $\bar{A}'$ is partitioned into

$$\bar{A}' = \begin{bmatrix} \bar{A}'_{11} & \bar{A}'_{12} \\ \bar{A}'_{21} & \bar{A}'_{22} \end{bmatrix} \qquad (4-150)$$

with block dimensions

$$\bar{A}'_{11} \;-\; q' \times q'$$

$$\bar{A}'_{22} \;-\; (n - q') \times (n - q')$$

$$\bar{A}'_{12} \;-\; q' \times (n - q')$$

$$\bar{A}'_{21} \;-\; (n - q') \times q'$$

then

$$T_{N'}\bar{A}' = [N'_c, N'] \begin{bmatrix} \bar{A}'_{11} & \bar{A}'_{12} \\ \bar{A}'_{21} & \bar{A}'_{22} \end{bmatrix}$$

$$= [(N'_c\bar{A}'_{11} + N'\bar{A}'_{21}), \; (N'_c\bar{A}'_{12} + N'\bar{A}'_{22})] \tag{4-151}$$

Using (4-141)

$$A'T_{N'} = [A'N'_c, A'N'] = [A'N'_c, N'P'_N] \tag{4-152}$$

Substituting (4-151) and (4-152) into (4-149) yields

$$[(N'_c\bar{A}'_{11} + N'\bar{A}'_{21}), (N'_c\bar{A}'_{12} + N'\bar{A}'_{22})] = [A'N'_c, N'P'_N] \tag{4-153}$$

Taking just the last (n - q') columns of this matrix equation

$$N'_c\bar{A}'_{12} + N'\bar{A}'_{22} = N'P'_N \tag{4-154}$$

or

$$[N'_c, N'] \begin{bmatrix} \bar{A}'_{12} \\ \bar{A}'_{22} - P'_N \end{bmatrix} = T_{N'} \begin{bmatrix} \bar{A}'_{12} \\ \bar{A}'_{22} - P'_N \end{bmatrix} = \underline{0} \tag{4-155}$$

But $T_{N'}$ is nonsingular, so this implies

$$\begin{bmatrix} \overline{A}'_{12} \\ \\ \overline{A}'_{22} - P'_N \end{bmatrix} = \underline{0} \tag{4-156}$$

In particular

$$\overline{A}'_{12} = \underline{0} \tag{4-157}$$

so

$$\overline{A}' = \begin{bmatrix} \overline{A}'_{11} & \underline{0} \\ \\ \overline{A}'_{21} & \overline{A}'_{22} \end{bmatrix} \tag{4-158}$$

Partitioning $\overline{C}'$ and $\overline{D}'$ to conform with $\overline{A}'$

$$\overline{C}' = [\overline{C}'_1, \overline{C}'_2] \tag{4-159}$$

$$\overline{D}' = \begin{bmatrix} \overline{D}'_1 \\ \\ \overline{D}'_2 \end{bmatrix} \tag{4-160}$$

Now by (4-140)

$$\begin{aligned} \overline{C}' = C'T_{N'} &= [C'N'_c, C'N'] \\ &= [C'N'_c, \underline{0}] \end{aligned} \tag{4-161}$$

so

$$\overline{C}'_1 = C'N'_c \tag{4-162}$$

$$\overline{C}'_2 = \underline{0} \tag{4-163}$$

Then

$$\bar{A}' - \bar{D}'\bar{C}' = \begin{bmatrix} \bar{A}'_{11} & \underline{0} \\ \bar{A}'_{21} & \bar{A}'_{22} \end{bmatrix} - \begin{bmatrix} \bar{D}'_1\bar{C}'_1 & \underline{0} \\ \bar{D}'_2\bar{C}'_1 & \underline{0} \end{bmatrix}$$

$$= \begin{bmatrix} \bar{A}'_{11} - \bar{D}'_1\bar{C}'_1 & \underline{0} \\ \bar{A}'_{21} - \bar{D}'_2\bar{C}'_1 & \bar{A}'_{22} \end{bmatrix} \tag{4-164}$$

From the block triangular form of (4-164) it is clear that the eigenvalues of $[\bar{A}' - \bar{D}'\bar{C}']$ (and therefore of $[A' - D'C']$) are the combined eigenvalues of $[\bar{A}'_{11} - \bar{D}'_1\bar{C}'_1]$ and $\bar{A}'_{22}$. Now

$$\bar{M}' = \begin{bmatrix} \bar{C}' \\ \bar{C}'\bar{A}' \\ \cdot \\ \cdot \\ \cdot \\ \bar{C}'\bar{A}'^{n-1} \end{bmatrix} = \begin{bmatrix} (\bar{C}'_1, \underline{0}) \\ (\bar{C}'_1\bar{A}'_{11}, \underline{0}) \\ \cdot \\ \cdot \\ \cdot \\ (\bar{C}'_1\bar{A}'^{n-1}_{11}, \underline{0}) \end{bmatrix}$$

$$= \left[\begin{pmatrix} \bar{C}'_1 \\ \bar{C}'_1\bar{A}'_{11} \\ \cdot \\ \cdot \\ \cdot \\ \bar{C}'_1\bar{A}'^{n-1}_{11} \end{pmatrix}, \underline{0} \right] \tag{4-165}$$

so

$$\text{rk} \begin{bmatrix} \bar{C}'_1 \\ \bar{C}'_1\bar{A}'_{11} \\ \cdot \\ \cdot \\ \cdot \\ \bar{C}'_1\bar{A}'^{n-1}_{11} \end{bmatrix} = \text{rk}\, \bar{M}' = q' = \text{rk} \begin{bmatrix} \bar{C}'_1 \\ \bar{C}'_1\bar{A}'_{11} \\ \cdot \\ \cdot \\ \cdot \\ \bar{C}'_1\bar{A}'^{q'-1}_{11} \end{bmatrix} \tag{4-166}$$

Since $\overline{A}'_{11}$ is $q' \times q'$, this implies that $(\overline{A}'_{11}, \overline{C}'_1)$ is an observable pair. Therefore, all q' eigenvalues of the $q' \times q'$ matrix $[\overline{A}'_{11} - \overline{D}'_1\overline{C}'_1]$ can be specified arbitrarily by choice of $\overline{D}'_1$. The remaining $(n - q')$ eigenvalues of $[\overline{A}' - \overline{D}'\overline{C}']$, and thus of $[A' - D'C']$, are the eigenvalues of $\overline{A}'_{22}$ which are not affected by any choice of D'. From (4-158) it can be seen that the $(n - q')$ eigenvalues of $\overline{A}'_{22}$ are eigenvalues of $\overline{A}'$ and thus of A'. This completes the proof of the lemma.

With the result of Lemma 4.4 it is now possible to conclude that the total number of eigenvalues of $A - DC = A' - D'C'$ which can be specified while satisfying (4-119) is $(k + q')$ where q' is given by (4-135) and k is the order of the detection generator in (4-119). Condition (2) of detectability will be satisfied if and only if $k + q' = n$. The next problem is to find under what circumstances (e.g., for which detection generators of what order) is $k + q' = n$. Since A' depends on g, it appears that M' given by (4-136) and $q' = \text{rk}\, M'$ also must depend on g. The following theorem shows that this is not the case. It establishes the very significant fact that the number of additional eigenvalues of $[A - DC]$ which can be specified after satisfying (4-119) does not depend on the particular detection generator g or its order k.

<u>Theorem 4.3.</u> If D is constrained to be a solution of (4-119) (or equivalently (4-113)), then the number of eigenvalues of $[A - DC]$ which can be arbitrarily specified, in addition to those given by (4-112), is equal to

$$\operatorname{rk} \begin{bmatrix} C' \\ C'K \\ \cdot \\ \cdot \\ \cdot \\ C'K^{n-1} \end{bmatrix}$$

where C' is defined by (4-134) and

$$K = A - Af[(Cf)^T Cf]^{-1} (Cf)^T C \tag{4-167}$$

Proof:

By Lemma 4.3 all possible solutions of (4-119) are given by (4-131) with D' arbitrary. The number of additional eigenvalues which can be specified is therefore the number of eigenvalues of $[A' - D'C']$ which can be specified by free choice of D', where A' is defined by (4-133). By Lemma 4.4 this number is q' given by (4-135).

Premultiplying (4-110) by A yields

$$Af = \alpha_1 Ag + \ldots + \alpha_{k-1} A^{k-1} g + A^k g \tag{4-168}$$

Solving this equation for $A^k g$ and substituting the result into (4-133) for A' gives

$$\begin{aligned} A' &= A - [p_1 g + \ldots + p_k A^{k-1} g \\ &\qquad + Af - \alpha_1 Ag - \ldots - \alpha_{k-1} A^{k-1} g] [(Cf)^T Cf]^{-1} (Cf)^T C \\ &= K - z_d [(Cf)^T Cf]^{-1} (Cf)^T C \end{aligned} \tag{4-169}$$

where

$$z_d = p_1 g + (p_2 - \alpha_1) Ag + \ldots + (p_k - \alpha_{k-1}) A^{k-1} g \tag{4-170}$$

By (4-80)

$$C'A^j g = \left[I - Cf[(Cf)^T Cf]^{-1}(Cf)^T \right] CA^j g$$

$$= \underline{0} \quad \text{for} \quad j = 0, \ldots, k-2 \tag{4-171}$$

and with $CA^{k-1}g = Cf$ from (4-118)

$$C'A^{k-1}g = \left[I - Cf[(Cf)^T Cf]^{-1}(Cf)^T \right] CA^{k-1}g$$

$$= \left[I - Cf[(Cf)^T Cf]^{-1}(Cf)^T \right] Cf$$

$$= C'f = Cf - Cf = \underline{0} \tag{4-172}$$

Then

$$C'KA^j g = C'A^{j+1}g - C'Af[(Cf)^T Cf]^{-1}(Cf)^T CA^j g$$

$$= \underline{0} \quad \text{for} \quad j = 0, \ldots, k-2 \tag{4-173}$$

and solving (4-110) for $A^{k-1}g$ gives

$$C'KA^{k-1}g = C'K[f - \alpha_1 g - \ldots - \alpha_{k-1} A^{k-2} g]$$

$$= C'Kf = \underline{0} \tag{4-174}$$

since

$$Kf = Af - Af = \underline{0} \tag{4-175}$$

Assume now

$$C'K^i A^j g = \underline{0} \quad \text{for} \quad j = 0, \ldots, k-1 \tag{4-176}$$

Then

$$C'K^{i+1}A^j g = C'K^i A^{j+1} g - C'K^i Af\,[(Cf)^T Cf]^{-1}(Cf)^T CA^j g$$

$$= C'K^i A^{j+1} g = \underline{0} \quad \text{for} \quad j = 0, \ldots, k-2$$

$$\tag{4-177}$$

and

$$C'K^{i+1}A^{k-1}g = C'K^{i+1}[f - \alpha_1 g - \ldots - \alpha_{k-1} A^{k-2} g]$$

$$= C'K^{i+1}f = \underline{0} \tag{4-178}$$

since $Kf = \underline{0}$. Therefore, by induction, (4-176) is valid for all $i \geq 0$ and $j = 0, \ldots, k-1$. Since z_d in (4-170) is a linear combination of the vectors $\{A^j g;\ j = 0, \ldots, k-1\}$, it follows that

$$C'K^i z_d = \underline{0} \quad \text{for all} \quad i \geq 0 \tag{4-179}$$

Then

$$C'A' = C'K - C'z_d\,[(Cf)^T Cf]^{-1}(Cf)^T C = C'K \tag{4-180}$$

and, in general,

$$C'A'^i = C'K^i \quad \text{for all} \quad i \geq 0 \tag{4-181}$$

Therefore

$$M' = \begin{bmatrix} C' \\ C'A' \\ \cdot \\ \cdot \\ \cdot \\ C'A'^{n-1} \end{bmatrix} = \begin{bmatrix} C' \\ C'K \\ \cdot \\ \cdot \\ \cdot \\ C'K^{n-1} \end{bmatrix}$$

$$\tag{4-182}$$

Substituting this into (4-135) gives the desired result and completes the proof.

Note that K and C' do not depend on g or k. Therefore, M' and $q' = rkM'$ are independent of g and k. This means that regardless of what detection generator is used to solve for a detector gain and regardless of its order, the amount of freedom left in D for specifying additional eigenvalues is always the same. It depends only on A, C, and f. Recall that the number of eigenvalues which can be specified in the process of satisfying (4-119) is equal to the order of the detection generator. It now becomes clear that condition (2) of detectability can be satisfied if and only if it is possible to find a detection generator of order $(n - q')$. Note also that a detection generator can never have order larger than $(n - q')$, because this would imply specification of more than n eigenvalues, which is impossible for an $n \times n$ matrix. This motivates the following definitions.

Definition 4.5. The null space of M' given by (4-182) is defined to be the detection space of f.

Definition 4.6. The dimension of detection space of f is defined to be the detection order of f.

Definition 4.7. A detection generator for f whose order is equal to the detection order of f is defined to be a maximal detection generator (or simply, maximal generator) for f.

Let the detection order of f be denoted by ν. The detection order of f is equal to the dimension of the null space of M', so

$$\nu = n - rk\,M' = n - q' \tag{4-183}$$

where $q' = rk\,M'$ with M' given by (4-182). The detectability of f now depends on being able to find a maximal generator. The next theorem establishes the conditions under which this is possible.

Theorem 4.4. If (A, C) is an observable pair, then every n-vector f has a maximal detection generator and it is unique.

Proof:

For an arbitrary n-vector f, let K, M', and N' be defined by (4-167), (4-182), and (4-138) respectively with rk $M' = q'$. The detection order of f is $\nu = n - q'$. Let

$$g = N'\beta' \tag{4-184}$$

where β' is a ν-vector to be determined. For a maximal generator it is necessary that

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{\nu-2} \end{bmatrix} g = \underline{0} \tag{4-185}$$

Note that K in (4-167) has the same form as $[A - DC]$ with $D = Af[(Cf)^T Cf]^{-1}(Cf)^T$. Therefore, Equation (4-87) can be applied to K to obtain

$$[I + \hat{T}'_{\nu-2}] \begin{bmatrix} C \\ CK \\ \cdot \\ \cdot \\ \cdot \\ CK^{\nu-2} \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{\nu-2} \end{bmatrix} \tag{4-186}$$

where $\hat{T}'_{\nu-2}$ has the form of (4-86) with D replaced by $Af[(Cf)^T Cf]^{-1}(Cf)^T$.

Since $I + \hat{T}'_{\nu-2}$ is nonsingular, (4-185) is equivalent to

$$\begin{bmatrix} C \\ CK \\ \cdot \\ \cdot \\ \cdot \\ CK^{\nu-2} \end{bmatrix} g = \underline{0} \tag{4-187}$$

or with (4-184)

$$\begin{bmatrix} C \\ CK \\ \cdot \\ \cdot \\ \cdot \\ CK^{\nu-2} \end{bmatrix} N'\beta' = \begin{bmatrix} CN' \\ CKN' \\ \cdot \\ \cdot \\ \cdot \\ CK^{\nu-2}N' \end{bmatrix} \beta' = \underline{0} \tag{4-188}$$

This equation will have a nonzero solution if and only if

$$\mathrm{rk} \begin{bmatrix} CN' \\ CKN' \\ \cdot \\ \cdot \\ \cdot \\ CK^{\nu-2}N' \end{bmatrix} \leq \nu - 1 \tag{4-189}$$

Now

$$\begin{bmatrix} CN' \\ CKN' \\ \cdot \\ \cdot \\ \cdot \\ CK^{\nu-2}N' \end{bmatrix} \leq \mathrm{rk}\ CN' + \mathrm{rk}\ CKN' + \ldots + \mathrm{rk}\ CK^{\nu-2}N' \tag{4-190}$$

Since $M'N' = \underline{0}$,

$$C'K^iN' = \underline{0} \text{ for } i = 0, \ldots, n-1 \tag{4-191}$$

Substituting (4-134) into (4-191) gives

$$CK^iN' - Cf[(Cf)^TCf]^{-1}(Cf)^TCK^iN' = \underline{0} \tag{4-192}$$

or

$$CK^iN' = Cf[(Cf)^TCf]^{-1}(Cf)^TCK^iN' \tag{4-193}$$

Then

$$\begin{aligned} rk\left[CK^iN'\right] &= rk\left[Cf[(Cf)^TCf]^{-1}(Cf)^TCK^iN'\right] \\ &\leq rk(Cf) = 1 \text{ for } i = 0, \ldots, n-1 \end{aligned} \tag{4-194}$$

Applying (4-194) to (4-190) yields (4-189) and proves that (4-188) has a nonzero solution. Since $rk\ N' = n - q' = \nu$, g given by (4-184) is also nonzero and satisfies (4-187) and (4-185).

It will now be shown that

$$CA^{\nu-1}g \neq \underline{0} \tag{4-195}$$

First note that with (4-185)

$$Kg = Ag - Af[(Cf)^TCf]^{-1}(Cf)^TCg = Ag \tag{4-196}$$

$$K^2g = KAg = A^2g - Af[(Cf)^TCf]^{-1}(Cf)^TCAg = A^2g \tag{4-197}$$

and, in general,

$$K^ig = A^ig \quad \text{for } i = 0, \ldots, \nu - 1 \tag{4-198}$$

Then (4-195) is equivalent to $CK^{\nu-1}g \neq 0$. From the form of M' in (4-182) it follows (from Section 2.3) that the null space of M' is an (n - q')-dimensional invariant subspace with respect to K. Therefore, g, which is in the null space of M', can generate a cyclic subspace with respect to K of dimension no larger than $(n - q') = \nu$. This means that the range space of $[g, Kg, \ldots, K^{n-1}g]$ coincides with the range space of $[g, Kg, \ldots, K^{\nu-1}g]$. Now if $CK^{\nu-1}g = \underline{0}$, this together with (4-187) gives $C[g, Kg, \ldots, K^{\nu-1}g] = \underline{0}$ which implies $C[g, Kg, \ldots, K^{n-1}g] = \underline{0}$, or

$$\begin{bmatrix} C \\ CK \\ \cdot \\ \cdot \\ \cdot \\ CK^{n-1} \end{bmatrix} g = \underline{0} \tag{4-199}$$

Again applying (4-87) to K with $j = n - 1$, (4-199) would imply

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{n-1} \end{bmatrix} g = \underline{0} \tag{4-200}$$

which would mean (A, C) is not an observable pair, since g is nonzero. But this contradicts the hypothesis, so one must conclude that

$$CK^{\nu-1}g \neq \underline{0} \tag{4-201}$$

which by (4-198) gives (4-195).

Relation (4-201) guarantees that

$$rk[g, Kg, \ldots, K^{\nu-1} g] = \nu \tag{4-202}$$

since by (4-187) $K^{\nu-1}g$ must be independent of the vectors $\{K^i g;\ i = 0, \ldots, \nu - 2\}$ in order to satisfy (4-201). Therefore, the set of vectors $\{K^i g;\ i = 0, \ldots, \nu - 1\}$ form a basis for the null space of M' (or equivalently, the detection space of f). By (4-198) this set of basis vectors is the same as the set $\{A^i g;\ i = 0, \ldots, \nu - 1\}$. Now

$$C'f = Cf - Cf = \underline{0} \tag{4-203}$$

and $C'K^i f = \underline{0}$ for all $i > 0$ because $Kf = \underline{0}$. Therefore,

$$\begin{bmatrix} C' \\ C'K \\ \cdot \\ \cdot \\ \cdot \\ C'^{n-1} \end{bmatrix} f = M'f = \underline{0} \tag{4-204}$$

so f is in its own detection space. Then f can be expressed as a linear combination of the basis vectors $\{A^i g;\ i = 1, \ldots, \nu - 1\}$

$$f = \alpha_1 g + \alpha_2 A g + \ldots + \alpha_\nu A^{\nu-1} g \tag{4-205}$$

It has been shown that g is nonzero and satisfies (4-80), (4-81), and (4-106) with $k = \nu$. By the same argument used previously, the magnitude of g can be taken so that g satisfies (4-110) or (4-111), thus making it a ν^{th} order detection generator for f. By Definition 4.7 this g is a maximal detection generator for f.

For completeness, some clarifying remarks should be made concerning the general case described by (4-108). As mentioned earlier, this case is obtained by replacing f by $A^\mu f$. Equation (4-204) then becomes

$$M'A^\mu f = \underline{0} \tag{4-206}$$

which shows only that $A^{\mu}f$ is in the detection space of f. However, it can be shown that f is in this space as well. By the same development used to obtain (4-198) from (4-185), it follows from (4-108) that

$$K^i f = A^i f \quad \text{for} \quad i = 0, \ldots, \mu \tag{4-207}$$

Substituting this back into (4-108) yields

$$\begin{bmatrix} C \\ CK \\ \cdot \\ \cdot \\ \cdot \\ CK^{\mu} \end{bmatrix} f = \underline{0} \tag{4-208}$$

which in turn gives

$$\begin{bmatrix} C' \\ C'K \\ \cdot \\ \cdot \\ \cdot \\ C'K^{\mu} \end{bmatrix} f = \underline{0} \tag{4-209}$$

Substituting (4-207) into (4-206) yields

$$M'A^{\mu}f = M'K^{\mu}f = \begin{bmatrix} C'K^{\mu} \\ \cdot \\ \cdot \\ \cdot \\ C'K^{n-1+\mu} \end{bmatrix} f = \underline{0} \tag{4-210}$$

Combining (4-209) and (4-210) gives

$$\begin{bmatrix} C' \\ C'K \\ \cdot \\ \cdot \\ \cdot \\ C'K^{n-1} \end{bmatrix} f = M'f = \underline{0} \qquad (4\text{-}211)$$

and proves that f is in the detection space. Equation (4-205) therefore is valid for the general case.

The observability condition guarantees that g is unique. Suppose g_1 and g_2 are both maximal generators for f. Let $\Delta g = g_1 - g_2$. Then

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{\nu-2} \end{bmatrix} \Delta g = \begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{\nu-2} \end{bmatrix} (g_1 - g_2) = \underline{0} \qquad (4\text{-}212)$$

by (4-185). But

$$CA^{\nu-1}\Delta g = CA^{\nu-1}g_1 - CA^{\nu-1}g_2 = Cf - Cf = \underline{0} \qquad (4\text{-}213)$$

If $\Delta g \neq \underline{0}$ (4-212) and (4-213) would imply (A, C) is not observable by the same argument used to show $CA^{\nu-1}g \neq \underline{0}$. Therefore, $\Delta g = \underline{0}$ and

$$g_1 = g_2 \qquad (4\text{-}214)$$

which establishes uniqueness of g. This completes the proof of Theorem 4.4.

Theorem 4.1 follows quite simply from Theorems 4.2, 4.3, and 4.4. By Theorem 4.4 observability of the pair (A, C) is sufficient to guarantee existence of the maximal generator, which by

Theorems 4.2 and 4.3 makes it possible to satisfy both conditions (1) and (2) of detectability. Moreover, the observability of (A, C) is necessary in order to satisfy condition (2). This follows from Lemma 4.4.

The following observations are made to reemphasize several important points and to highlight some additional facts which are of interest.

1) For a given observable pair (A, C) each n-vector f has one and only one detection space, detection order, and maximal generator. Moreover, if A is replaced by $A'' = [A - D''C]$ for arbitrary D'' (with appropriate dimension), the detection space, the detection order, and maximal generator for f remain invariant. This property can be of considerable value in determining the detection order and maximal generator of a vector. As will be seen later, when A and C have a certain standard form, it is a simple matter to choose a D'' which produces an A'' with all elements zero or one, thus making computations much simpler.

It should be noted also that the developments in this section remain valid under a coordinate transformation of the state space. Therefore, the detection order of f is invariant under a coordinate transformation. The detection space and maximal generator transform in the same way as f.

2) Theorem 4.2 states that in order to be a detector gain D must be a solution of (4-113) for some detection generator. By constraining D to be a solution (4-113), $(n - q') = \nu$ eigenvalues of $[A - DC]$ are completely fixed. Of these, k eigenvalues can be arbitrarily specified by choice of the coefficients $\{p_i; \ i = 1, \ldots, k\}$ in (4-113).

If a nonmaximal detection generator is used (i.e., $k < \nu$) then $(\nu - k) = (n - q' - k)$ eigenvalues are fixed without the control of the designer. In any case, the remaining q' eigenvalues can be specified arbitrarily by choice of D' in the general solution of (4-113).

3) All detection generators for f (of all orders up to the maximal) lie in the detection space of f. This follows from the fact, established in the proof of Theorem 4.3, that $C'K^iA^jg = \underline{0}$ for all $i \geq 0$ and $j = 0, \ldots, k - 1$ where g is a k^{th} order detection generator for f. In fact, this shows that all the vectors $\{A^jg;\ j = 0, \ldots, k - 1\}$ are in the detection space of f. By the same reasoning used to obtain (4-198) from (4-185), it can be shown that any k^{th} order detection generator satisfies (4-198).

It should also be noted that every n-vector contained in the detection space of f has the same detection order and detection space as f. Suppose f has detection order ν and g is its maximal generator. Clearly, g satisfies (4-80) and (4-81) with $k = \nu$. Let f_2 be any other vector in the detection space of f. Since the set of vectors $\{A^jg;\ j = 0, \ldots, \nu - 1\}$ span the detection space, f_2 can be expressed as a linear combination of these vectors. Then, with the possible exception of magnitude, g satisfies the requirements to be a ν^{th} order detection generator for f_2. This implies the detection order of f_2 is greater than or equal to ν. Also, by the remarks in the preceding paragraph, the vectors $\{A^jg;\ j = 0, \ldots, \nu - 1\}$ all lie in the detection space of f_2. Since these vectors span the detection space of f, one may conclude that the detection space of f is contained in the detection space of f_2.

This means f is contained in the detection space of f_2. But by the above argument (with the roles of f and f_2 reversed) this implies the detection order of f is greater than or equal to the detection order of f_2, and the detection space of f contains the detection space of f_2. Therefore, one must conclude that f and f_2 have the same detection order, and their detection spaces coincide.

4) Although observability of (A, C) is necessary to satisfy condition (2) of detectability, it is not necessary for condition (1). A detector gain can always be found provided f does not lie in the unobservable space of C. This can be shown by employing a coordinate transformation similar to that used in the proof of Lemma 4.4, which transforms A and C into the forms

$$\overline{A} = T^{-1}AT = \begin{bmatrix} \overline{A}_{11} & \underline{0} \\ \overline{A}_{21} & \overline{A}_{22} \end{bmatrix} \tag{4-215}$$

$$\overline{C} = CT = [\overline{C}_1, \underline{0}] \tag{4-216}$$

where $(\overline{A}_{11}, \overline{C}_1)$ is an observable pair. Partitioning $\overline{f}$ and $\overline{D}$ to conform with $\overline{A}$ and $\overline{C}$

$$\overline{f} = T^{-1}f = \begin{bmatrix} \overline{f}_1 \\ \overline{f}_2 \end{bmatrix} \tag{4-217}$$

$$\overline{D} = T^{-1}D = \begin{bmatrix} \overline{D}_1 \\ \overline{D}_2 \end{bmatrix} \tag{4-218}$$

it is easily shown from the form of $\overline{A}$ and $\overline{C}$ that

$$C[f, \ldots, (A - DC)^{n-1} f] = \overline{C}[\overline{f}, \ldots, (\overline{A} - \overline{D}\overline{C})^{n-1} \overline{f}$$

$$= \overline{C}_1 [\overline{f}_1, \ldots, (\overline{A}_{11} - \overline{D}_1 \overline{C}_1)^{n-1} \overline{f}_1] \tag{4-219}$$

Theorem 4.1 can be applied to the observable pair $(\overline{A}_{11}, \overline{C}_1)$ to show there exists a $\overline{D}_1$, and thus a D, which satisfies condition (1). If f lies in the unobservable space of C, then the settled-out output error is zero for any D. Lemma 4.4 shows that if (A, C) is not observable then there will be a number of eigenvalues of [A - DC] which will be equal to those of A and which cannot be changed by any D (specifically, the eigenvalues of $\overline{A}_{22}$ in (4-215)). Nothing can be gained by accepting a weaker control over the eigenvalues which can be changed. Therefore, the observability condition can be relaxed only if one is willing to give up all control over a certain number of eigenvalues of A.

5) It was suggested previously that it would be desirable to tailor the detection filter dynamics to the dynamic characteristics of the drive $v_\epsilon(t)$. It is of interest therefore to determine the resulting error dynamics when D is a detector gain. The Laplace transform is a convenient tool for studying the settled-out output error. Consider

$$E'(s) = \mathcal{L}\{C\epsilon(t)\} = C\mathcal{L}\{\epsilon(t)\} = C[Is - (A - DC)]^{-1} f \, V_\epsilon(s) \tag{4-220}$$

where

$$V_\epsilon(s) = \mathcal{L}\{v_\epsilon(t)\} \tag{4-221}$$

Let D be a solution of (4-113). The transfer from $V_\epsilon(s)$ to $E'(s)$ is invariant under a coordinate transformation of the state space. Define a coordinate transformation by the $n \times n$ matrix

$$T_f = [T_{f1}, T_{f2}] \tag{4-222}$$

with

$$T_{f1} = [g, Ag, \ldots, A^{k-1}g] = [g, (A - DC)g, \ldots, (A - DC^{k-1}g] \tag{4-223}$$

where g is a k^{th} order detection generator for f and T_{f2} is any $n \times (n - k)$ matrix which makes T_f nonsingular. Let

$$\begin{aligned} \widehat{G} &= T_f^{-1}(A - DC)T_f \\ \widehat{C} &= CT_f \\ \widehat{f} &= T_f^{-1} f \end{aligned} \tag{4-224}$$

Now,

$$(A - DC)T_f = [(A - DC)T_{f1}, (A - DC)T_{f2}] \tag{4-225}$$

From (4-115)

$$(A - DC)T_{f1} = [(A - DC)g, \ldots, (A - DC)^k g] = T_{f1}\widehat{G}_{11} \tag{4-226}$$

where

$$\widehat{G}_{11} = \begin{bmatrix} 0 & 0 & & 0 & -p_1 \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & -p_k \end{bmatrix} \tag{4-227}$$

so

$$(A - DC)T_f = T_f \begin{bmatrix} \hat{G}_{11} & \hat{G}_{12} \\ \underline{0} & \hat{G}_{22} \end{bmatrix} \tag{4-228}$$

where

$$(A - DC)T_{f2} = T_{f1}\hat{G}_{12} + T_{f2}\hat{G}_{22} \tag{4-229}$$

Premultiplying (4-228) by T_f^{-1} yields

$$T_f^{-1}(A - DC)T_f = \hat{G} = \begin{bmatrix} \hat{G}_{11} & \hat{G}_{12} \\ \underline{0} & \hat{G}_{22} \end{bmatrix} \tag{4-230}$$

Also

$$\hat{C} = [CT_{f1}, CT_{f2}] \tag{4-231}$$

with

$$\begin{aligned} CT_{f1} &= [\underline{0}, \ldots, \underline{0}, Cf] \\ &= Cf[0, \ldots, 0, 1] \end{aligned} \tag{4-232}$$

by (4-80) and (4-118) for $Cf \neq \underline{0}$. Finally, using (4-110)

$$f = T_f\hat{f} = [T_{f1}, T_{f2}]\begin{bmatrix} \hat{f}_1 \\ \\ \underline{0} \end{bmatrix} = T_{f1}\hat{f}_1 \tag{4-233}$$

with

$$\hat{f}_1 = \begin{bmatrix} \alpha_1 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{k-1} \\ 1 \end{bmatrix} \tag{4-234}$$

so

$$\hat{f} = \begin{bmatrix} \hat{f}_1 \\ \underline{0} \end{bmatrix} \qquad (4\text{-}235)$$

Now

$$\begin{aligned} C[Is - (A - DC)]^{-1}f &= \hat{C}[Is - \hat{G}]^{-1}f \\ &= [CT_{f1}, CT_{f2}] \begin{bmatrix} (Is - \hat{G}_{11}) & \hat{G}_{12} \\ \underline{0} & (Is - \hat{G}_{22}) \end{bmatrix}^{-1} \begin{bmatrix} \hat{f}_1 \\ \underline{0} \end{bmatrix} \\ &= CT_{f1}(Is - \hat{G}_{11})^{-1}\hat{f}_1 \\ &= Cf\{[0, \ldots, 0, 1](Is - \hat{G}_{11})^{-1}\hat{f}_1\} \\ &= Cf \frac{(s^{k-1} + \alpha_{k-1} s^{k-2} + \ldots + \alpha_1)}{(s^k + p_k s^{k-1} + \ldots + p_1)} \end{aligned} \qquad (4\text{-}236)$$

So

$$E'(s) = Cf\, H(s)\, V_\epsilon(s) \qquad (4\text{-}237)$$

where

$$H(s) = \frac{s^{k-1} + \alpha_{k-1} s^{k-2} + \ldots + \alpha_1}{s^k + p_k s^{k-1} + \ldots + p_1} \qquad (4\text{-}238)$$

for $Cf \neq \underline{0}$. For the general case of (4-108),

$$D'(s) = CA^{\mu}f\, H(s)\, V_\epsilon(s) \qquad (4\text{-}239)$$

with

$$H(s) = \frac{s^{k-\mu-1} + \alpha_{k-\mu-1} s^{k-\mu-2} + \ldots + \alpha_1}{s^k + p_k s^{k-1} + \ldots + p_1} \qquad (4\text{-}240)$$

The direction of $E'(s)$ is, of course, fixed and given by Cf or $CA^{\mu}f$. The magnitude of $E'(s)$ can be considered the output of a k-dimensional single-input, single-output linear system with dynamics given by (4-238) or (4-240). The significant fact to note here is that whereas the denominator of H(s) -- the poles of the system -- are under the complete control of the designer, the numerator -- the zeroes of the system -- cannot be altered by any D. Once a detection generator is found, (4-113) can be solved to obtain a detector gain without knowing the coefficients α_i in (4-110) or (4-111). However, if time allows it may be desirable to find these coefficients and determine where the zeroes of the system lie before deciding where to put the poles.

6) The construction used in Theorem 4.4 to show the existence of the maximal generator is a feasible method for finding the maximal generator for f, because all the quantities used in that construction depend only on A, C, and f. Note C' and K are defined in terms of A, C, and f only. The matrix N' is constructed from M' which in turn can be defined in terms of C' and K by (4-182). Therefore, M' and N' also can be constructed from A, C, and f. Appendix A describes an algorithm for finding the maximal generator of a vector. The algorithm is based on the construction in Theorem 4.4, but is somewhat more direct.

The results of this section show that if (A, C) is observable, any n-vector f in the state space has a unique maximal detection generator, which can be constructed from A, C, and f only. It has not been proven, in general, that f has detection generators of orders less than the maximal. Lemma 4.2 proved only that a k^{th} order detection generator must exist if a detector gain D exists which satisfies the conditions of the lemma. It was noted previously that the construction used in that lemma is not an appropriate method for finding a detection generator, because prior knowledge of D is assumed. It is easily verified, however, that f is a unique first order detection generator for itself. This suggests a tentative speculation that f has a unique detection generator of every order from one up to the maximal.

7) There is a duality relationship between these results on detection and the design of linear state feedback control, which is concerned with the properties of the matrix (A + BL) with A and B given and L to be selected. The dual significance of the results in this section and later sections in this chapter are discussed in Chapter 6.

The results of this section deal only with the detection of a single event. One of the appealing features of the detection filter for the case of a fully measurable state vector was that a single filter could provide all types of event information. As noted at the beginning of Section 4.3, this will not be possible, in general, when the state vector is only partially measurable. The next three sections consider the problem of detecting a number of events with a single filter.

Before proceeding to the next section, a simple example will serve to illustrate some of the preceding remarks.

Example E1:

Suppose

$$A = \begin{bmatrix} 0 & 3 & 4 \\ 1 & 2 & 3 \\ 0 & 2 & 5 \end{bmatrix} \tag{E1-1}$$

$$C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{E1-2}$$

$$f = \begin{bmatrix} -3 \\ 1 \\ 0 \end{bmatrix} \tag{E1-3}$$

Note the (A, C) is an observable pair. As noted in remark 1), the maximal generator, detection order, and detection space of f remain unchanged if A is replaced by $A'' = A - D''C$ for any D''. It is convenient to take

$$D'' = \begin{bmatrix} 3 & 4 \\ 2 & 3 \\ 2 & 5 \end{bmatrix} \tag{E1-4}$$

since this yields the simple form

$$A'' = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{E1-5}$$

Now

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

and from the definition of C'

$$C' = C - Cf\,[\,(Cf)^T Cf\,]^{-1}(Cf)^T C$$

$$= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} [\,0 \quad 1 \quad 0\,]$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{E1-6}$$

Using A'' to form K

$$K = A'' - A''f\,[\,(Cf)^T Cf\,]^{-1}(Cf)^T C$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 \\ -3 \\ 0 \end{bmatrix} [\,0 \quad 1 \quad 0\,]$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{E1-7}$$

Then

$$M' = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{E1-8}$$

and

$$\text{rk } M' = q' = 1 \tag{E1-9}$$

The detection order of f is

$$\nu = n - q' = 3 - 1 = 2 \tag{E1-10}$$

Consider the three-dimensional state space shown in Figure 4-1. Note that

$$C_1^T = \hat{e}_2 \tag{E1-11}$$

$$C_2^T = \hat{e}_3 \tag{E1-12}$$

so the output vector

$$y(t) = Cx(t) \tag{E1-13}$$

is simply the projection of the state vector $x(t)$ on the $(\hat{e}_2 - \hat{e}_3)$- plane.

From M' it can be seen that the detection space of f (the null space of M') is the $(\hat{e}_1 - \hat{e}_2)$- plane. The maximal generator of f must be in this plane and in addition satisfy

$$Cg = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} g = \underline{0} \tag{E1-14}$$

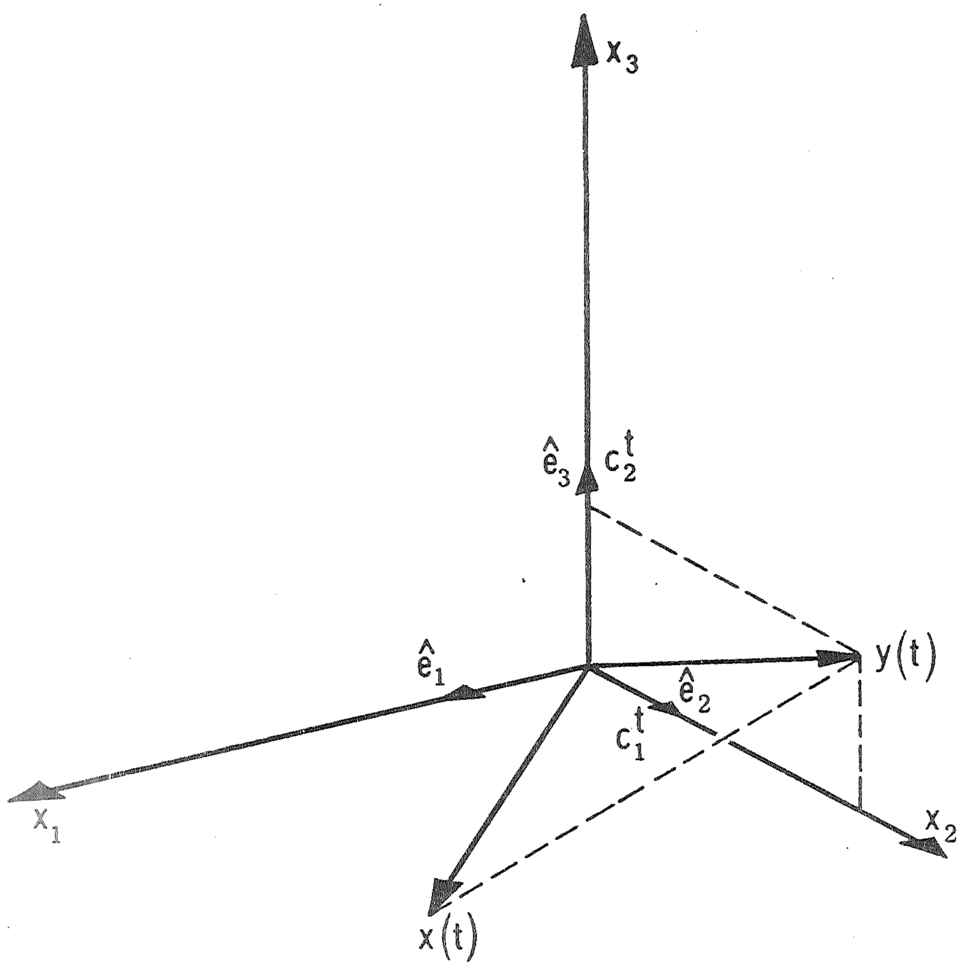


Figure 4-1.

and

$$CA''g = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} g = Cf = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{(E1-15)}$$

These two equations imply that

$$g = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \hat{e}_1 \qquad \text{(E1-16)}$$

Now

$$A''g \;=\; Kg \;=\; Ag \;=\; \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \;=\; \hat{e}_2 \qquad \text{(E1-17)}$$

Note that g and Ag span the detection space of f, as illustrated in Figure 4-2, and

$$f \;=\; -3g + Ag \qquad \text{(E1-18)}$$

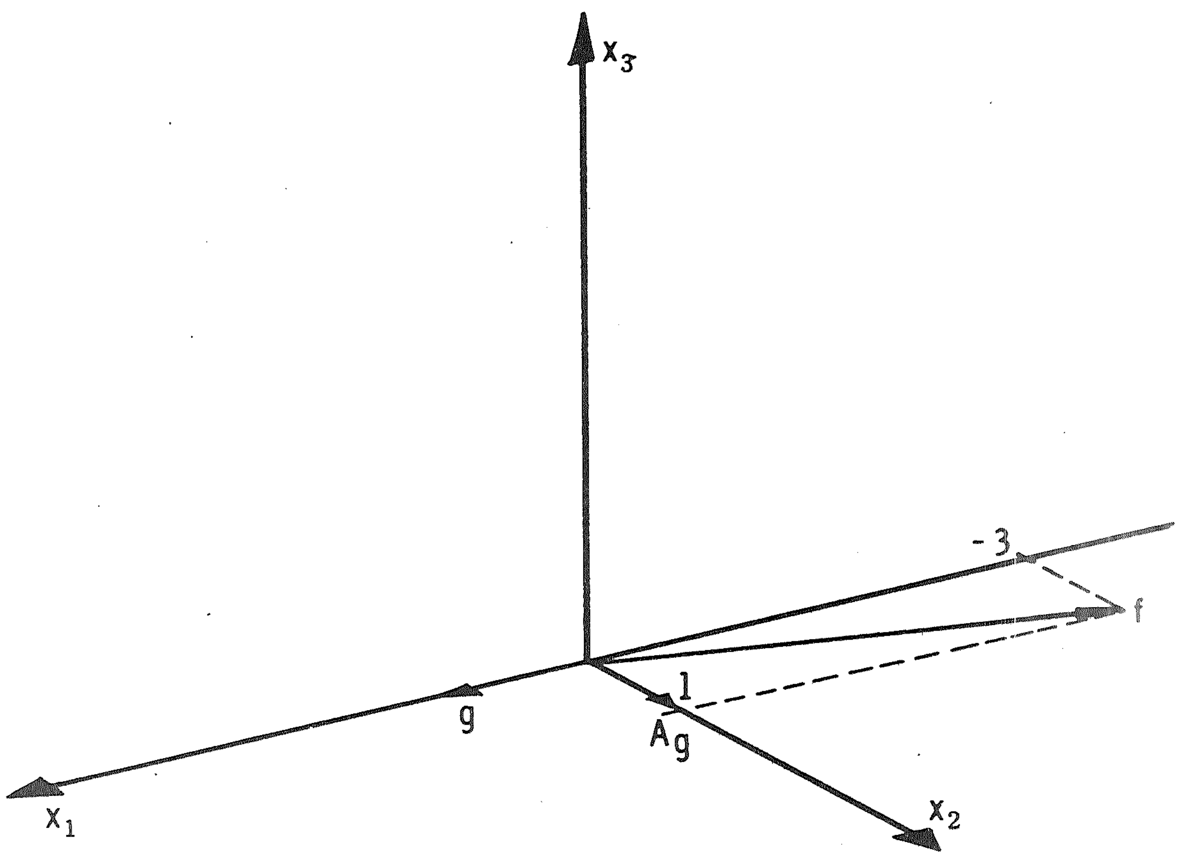


Figure 4-2.

The 3×2 matrix

$$D = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{31} & d_{32} \end{bmatrix} \tag{E1-19}$$

will be a detector gain for f if it satisfies

$$DCf = D\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} d_{11} \\ d_{21} \\ d_{31} \end{bmatrix} = p_1 g + p_2 A g + A^2 g \tag{E1-20}$$

for arbitrary p_1 and p_2. From remark 5) and (E1-18) it is known that if D satisfies (E1-20) the output error transfer function will be

$$\frac{E'(s)}{V_\epsilon(s)} = Cf\ H(s) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{(s - 3)}{(s^2 + p_2 s + p_1)} \tag{E1-21}$$

If poles of H(s) are desired at $s = -2$ and $s = -3$, for example, then

$$(s + 2)\ (s + 3) = s^2 + 5s + 6 \tag{E1-22}$$

and

$$p_1 = 6 \tag{E1-23}$$

$$p_2 = 5 \tag{E1-24}$$

The transfer is then

$$\frac{E'(s)}{V_\epsilon(s)} = Cf\ H(s) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{s - 3}{(s + 2)\ (s + 3)} \tag{E1-25}$$

To produce this transfer the first column of D must satisfy

$$\begin{bmatrix} d_{11} \\ d_{21} \\ d_{31} \end{bmatrix} = 6\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 5\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 2 \end{bmatrix} \tag{E1-26}$$

Then

$$D = \begin{bmatrix} 9 & d_{12} \\ 7 & d_{22} \\ 2 & d_{32} \end{bmatrix} \tag{E1-27}$$

and

$$A - DC = \begin{bmatrix} 0 & -6 & 4 - d_{12} \\ 1 & -5 & 3 - d_{22} \\ 0 & 0 & 5 - d_{32} \end{bmatrix} \tag{E1-28}$$

Note that after constraining D to satisfy (E1-20), the entire second column of D is still arbitrary. The same result is obtained if Lemma 4.3 is used to obtain a solution of (E1-20). In that case

$$D = \begin{bmatrix} 9 \\ 7 \\ 2 \end{bmatrix} (1)^{-1} [1 \quad 0] + D' \left[I - \begin{bmatrix} 1 \\ 0 \end{bmatrix} (1)^{-1} [1 \quad 0] \right]$$

$$= \begin{bmatrix} 9 & 0 \\ 7 & 0 \\ 2 & 0 \end{bmatrix} + D' \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \tag{E1-29}$$

and

$$A - DC = \begin{bmatrix} 0 & -6 & 4 \\ 1 & -5 & 3 \\ 0 & 0 & 5 \end{bmatrix} - D' \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & -6 & 4 - d'_{12} \\ 1 & -5 & 3 - d'_{22} \\ 0 & 0 & 5 - d'_{32} \end{bmatrix} \qquad \text{(E1-30)}$$

where

$$D' = \begin{bmatrix} d'_{11} & d'_{12} \\ d'_{21} & d'_{22} \\ d'_{31} & d'_{32} \end{bmatrix} \qquad \text{(E1-31)}$$

is arbitrary.

Two eigenvalues of (A - DC) are $s = -2$ and $s = -3$ by virtue of the choice of p_1 and p_2. From the block diagonal form of (A - DC) in (E1-28) it is easily seen that the third eigenvalue is $(5 - d_{32})$, so it can be arbitrarily specified by choice of d_{32}. Therefore, all three eigenvalues of (A - DC) can be specified.

This will not be the case if a nonmaximal detection generator is used to find a detector gain. Note that f is a first order detection generator for itself. Hence, a detector gain for f can be found by solving

$$DCf = p_1 f + Af \qquad \text{(E1-32)}$$

This will yield an error transfer of

$$\frac{E'(s)}{V_\epsilon(s)} = Cf\,H(s) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{1}{s + p_1} \qquad \text{(E1-33)}$$

Equation (E1-32) yields

$$\begin{bmatrix} d_{11} \\ d_{21} \\ d_{31} \end{bmatrix} = p_1 \begin{bmatrix} -3 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} -3p_1 + 3 \\ p_1 - 1 \\ 2 \end{bmatrix} \qquad \text{(E1-34)}$$

Then

$$D = \begin{bmatrix} -3p_1 + 3 & d_{12} \\ p_1 - 1 & d_{22} \\ 2 & d_{32} \end{bmatrix} \qquad \text{(E1-35)}$$

and

$$A - DC = \begin{bmatrix} 0 & 3p_1 & 4 - d_{12} \\ 1 & -p_1+3 & 3 - d_{22} \\ 0 & 0 & 5 - d_{32} \end{bmatrix} \qquad \text{(E1-36)}$$

The eigenvalues of (A - DC) are given by the roots of

$$\begin{aligned} \left| Is - (A - DC) \right| &= \left(s^2 + (p_1 - 3)s - 3p_1\right) (s - 5 + d_{32}) \\ &= (s + p_1)(s - 3)(s - 5 + d_{32}) \\ &= 0 \end{aligned} \qquad \text{(E1-37)}$$

Two eigenvalues of (A - DC) are $s = -p_1$ and $s = 5 - d_{32}$, and can be arbitrarily specified by choice of p_1 and d_{32}. However, the third eigenvalue of (A - DC) is always $s = 3$. This eigenvalue is automatically determined when D is constrained to satisfy (E1-32), and it cannot be altered by any choice of p_1, d_{12}, d_{22}, or d_{32}. This is an example of the uncontrolled eigenvalues which result when a nonmaximal detection generator is used to solve for a detector gain, as noted in remark 2). In this example the uncontrolled eigenvalue produces an unstable filter, but this is not necessarily always the case. For some other f the uncontrolled eigenvalue may yield a stable filter. However, to maintain control over all eigenvalues of (A - DC), the maximal generator must be used in determining D.

Consider again the matrix (A - DC) in (E1-28) obtained with the use of the maximal generator. Even after specifying the third eigenvalue of (A - DC) by choice of d_{32}, there is still freedom left in the choice of d_{12} and d_{22}. One might ask if this freedom can be used to make D a detector gain for a second vector, f_2, as well as for f. In this case the answer to that question is yes. First, assume f_2 lies in the detection space of f -- the $(\hat{e}_1 - \hat{e}_2)$ plane. Then

$$f_2 = \alpha_{21} g + \alpha_{22} A g \qquad \text{(E1-38)}$$

for some scalars α_{21} and α_{22}, and it is easily shown that a detector gain for f, determined with the use of the maximal generator g, is a detector gain for f_2 as well. However, the output error direction cannot distinguish between events associated with f and f_2, because

$$Cf_2 = \alpha_{22}\, Cf \qquad \text{(E1-39)}$$

so the output error direction is the same for both f and f_2. Some possible methods for distinguishing such events are discussed later. As a matter of interest, the error transfer function for f_2 in the detection space of f is

$$\frac{E'(s)}{V_\epsilon(s)} = Cf_2 \frac{s + \frac{\alpha_{21}}{\alpha_{22}}}{(s+2)(s+3)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{\alpha_{22}s + \alpha_{21}}{(s+2)(s+3)} \tag{E1-40}$$

When f_2 lies in the detection space of f, the freedom in the choice of d_{12} and d_{22} is not necessary to obtain a detector gain for both f and f_2.

Now suppose f_2 does not lie in the $(\hat{e}_1 - \hat{e}_2)$ plane. Suppose, for example,

$$f_2 = \begin{bmatrix} 1 \\ -1/2 \\ 1/2 \end{bmatrix} \tag{E1-41}$$

It will be found for this example that the detection order of f_2 (and, in fact, of any vector not in the $(\hat{e}_1 - \hat{e}_2)$ plane) is $\nu_2 = 1$. This means that the maximal generator for f_2 is f_2. To be a detector gain for f_2, D must satisfy

$$DCf_2 = D \begin{bmatrix} -1/2 \\ \\ 1/2 \end{bmatrix} = p_{21}f_2 + Af_2 = \begin{bmatrix} p_{21} + 1/2 \\ -1/2\, p_{21} + 3/2 \\ -1/2\, p_{21} + 3/2 \end{bmatrix} \tag{E1-42}$$

for arbitrary p_{21}. The third eigenvalue of (A - DC) will then be $s = -p_{21}$. Let

$$p_{21} = 4 \tag{E1-43}$$

Then (E1-42) yields

$$-\frac{1}{2}\begin{bmatrix} d_{11} \\ d_{21} \\ d_{31} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} d_{12} \\ d_{22} \\ d_{32} \end{bmatrix} = \begin{bmatrix} 9/2 \\ -1/2 \\ 7/2 \end{bmatrix} \tag{E1-44}$$

Substituting (E1-26) into this equation yields

$$\begin{bmatrix} d_{12} \\ d_{22} \\ d_{32} \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 2 \end{bmatrix} + \begin{bmatrix} 9 \\ -1 \\ 7 \end{bmatrix} = \begin{bmatrix} 18 \\ 6 \\ 9 \end{bmatrix} \tag{E1-45}$$

Then

$$D = \begin{bmatrix} 9 & 18 \\ 7 & 6 \\ 2 & 9 \end{bmatrix} \tag{E1-46}$$

and

$$G = A - DC = \begin{bmatrix} 0 & -6 & -14 \\ 1 & -5 & -3 \\ 0 & 0 & -4 \end{bmatrix} \tag{E1-47}$$

It is easily verified that for f

$$E'(s) = C[Is - G]^{-1} f V_\epsilon(s)$$

$$= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{s - 3}{(s + 2)(s + 3)} V_\epsilon(s) \qquad \text{(E1-48)}$$

and for f_2

$$E'(s) = C[Is - G]^{-1} f_2 V_\epsilon(s)$$

$$= \begin{bmatrix} -1/2 \\ 1/2 \end{bmatrix} \frac{1}{s + 4} V_\epsilon(s) \qquad \text{(E1-49)}$$

so D given by (E1-46) is a detector gain for both f and f_2. The settled-out output error produced by the event associated with f always lies in the direction $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ in the output space. The event associated with f_2 produces a settled-out output error lying in the direction $\begin{bmatrix} -1/2 \\ 1/2 \end{bmatrix}$.

In addition to making D a detector gain for f_2, it was possible to specify all three eigenvalues of (A - DC). Unfortunately, this is not always possible. Consider what happens when D is constrained to be a detector gain for f_2 given by (E1-41) and f_1 given by

$$f_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \text{(E1-50)}$$

The detection order of f_1 is $\nu_1 = 1$, and f_1 is the maximal generator. A detector gain for f_1 must satisfy

$$DCf_1 = D \begin{bmatrix} 0 \\ 1 \end{bmatrix} = p_{11} f_1 + Af_1 = \begin{bmatrix} 4 \\ 3 \\ p_{11} + 5 \end{bmatrix} \qquad \text{(E1-51)}$$

for arbitrary p_{11}. This together with Equation (E1-42) for a detector gain for f_2 gives

$$D \begin{bmatrix} 0 & -1/2 \\ 1 & 1/2 \end{bmatrix} = \begin{bmatrix} 4 & p_{21} + 1/2 \\ 3 & -1/2\, p_{21} + 3/2 \\ p_{11} + 5 & 1/2\, p_{21} + 3/2 \end{bmatrix} \tag{E1-52}$$

which has the unique solution

$$D = \begin{bmatrix} -2p_{21} + 3 & 4 \\ p_{21} & 3 \\ -p_{21} + p_{11} + 2 & p_{11} + 5 \end{bmatrix} \tag{E1-53}$$

Then

$$A - DC = \begin{bmatrix} 0 & 2p_{21} & 0 \\ 1 & -p_{21} + 2 & 0 \\ 0 & p_{21} - p_{11} & -p_{11} \end{bmatrix} \tag{E1-54}$$

The D given by (E1-53) is a detector gain for both f_1 and f_2. The eigenvalues of (A - DC) are given by the roots of

$$\begin{aligned} \left| Is - (A - DC) \right| &= \left(s^2 + (p_{21} - 2)s - 2p_{21}\right) (s + p_{11}) \\ &= (s - 2)(s + p_{21})(s + p_{11}) \\ &= 0 \end{aligned} \tag{E1-54}$$

Two eigenvalues of (A - DC) can be specified by choice of p_{11} and p_{21}. However, the third eigenvalue is always $s = 2$ regardless of the choice for p_{11} and p_{21}. This eigenvalue is automatically determined when D is constrained to be a detector gain for both f_1 and f_2. In this example the uncontrolled eigenvalue produces an unstable filter. This implies that it is not possible to detect both f_1 and f_2 with a single filter. It is necessary to use two separate filters -- one for f_1 and another for f_2.

The uncontrolled eigenvalues do not always cause instability. If, for example, instead of (E1-41) f_2 is

$$f_2 = \begin{bmatrix} 1 \\ 1/2 \\ 1/2 \end{bmatrix} \tag{E1-55}$$

then the detector gain for f_1 and f_2 is

$$D = \begin{bmatrix} 2p_{21} + 3 & 4 \\ p_{21} + 4 & 3 \\ p_{21} - p_{11} + 2 & p_{11} + 5 \end{bmatrix} \tag{E1-56}$$

and

$$A - DC = \begin{bmatrix} 0 & -2p_{21} & 0 \\ 1 & -p_{21} - 2 & 0 \\ 0 & -p_{21} + p_{11} & -p_{11} \end{bmatrix} \tag{E1-57}$$

In this case the uncontrolled eigenvalue of (A - DC) is $s = -2$. If a pole at $s = -2$ yields an acceptable settling time for the filter, then f_1 and f_2 can be detected by a single filter with a detector gain given by (E1-56). The next three sections investigate the problem of detecting a number of events with a single filter.

4.3.2 Mutual Detectability

Consider a set of r n-vectors $\{f_1, \ldots, f_r\}$ associated with a set of r events. The problem considered here is, given such a set, to determine if it is possible to detect all vectors in the set with a single detection filter.

Definition 4.8. The vectors $\{f_1, \ldots, f_r\}$ are defined to be mutually detectable if there exists a D which satisfies the conditions of Definition 4.2 for all the f_i, $i = 1, \ldots, r$.

An important special case of this problem is encountered when the vectors are "output separable" as defined below.

Definition 4.9. The vectors $\{f_1, \ldots, f_r\}$ are defined to be output separable if

$$\text{rk } CF = r \tag{4-241}$$

where F is an $n \times r$ matrix given by

$$F = [A^{\mu_1} f_1, A^{\mu_2} f_2, \ldots, A^{\mu_r} f_r] \tag{4-242}$$

with μ_i for each i defined by

$$\left.\begin{array}{lll} CA^j f_i = \underline{0} & ; & j = 1, \ldots, \mu_i - 1 \\ CA^{\mu_i} f_i \neq \underline{0} & & \end{array}\right\} \tag{4-243}$$

Note that (4-241) implies $r \leq m$ where C is $m \times n$. This definition is motivated by the following observation. Suppose two vectors f_1 and f_2 are not output separable. Then rk CF = 1 and $CA^{\mu_1} f_1$ and $CA^{\mu_2} f_2$ lie in the same direction in the output space. This means that even if a D can be found which is a detector gain for both f_1 and f_2 the output error for both events will lie in the same direction. Thus the output error direction will not separate these two events. More will be said about nonseparable vectors at the end of this section.

The next theorem provides a test for mutual detectability of output separable vectors. Before stating the theorem, some preliminary results and definitions are necessary. By Theorem 4.2 a detector gain for the vectors $\{f_1, \ldots, f_r\}$ must satisfy a set of r equations of the form

$$DCA^{k_i-1} g_i = p_{i1}g_i + \ldots + p_{ik_i} A^{k_i-1} g_i + A^{k_i} g_i$$

$$\text{for } i = 1, \ldots, r \tag{4-244}$$

where g_i is a k_i^{th} order detection generator for f_i. Using the form of (4-121) this set of equations can be written as a single matrix equation

$$DCF = Q_d \tag{4-245}$$

where F is defined by (4-242) and

$$Q_d = [w_{d1}, \ldots, w_{dr}] \tag{4-246}$$

with

$$w_{di} = p_{i1}g_i + \ldots + p_{ik_i} A^{k_i-1} g_i + A^{k_i} g_i$$

$$\text{for } i = 1, \ldots, r \tag{4-247}$$

When the f_i are mutually separable (4-245) always has a solution. If D is a solution of (4-244) each g_i generates a cyclic subspace of dimension

k_i with respect to $(A - DC)$. The eigenvalues associated with each of these invariant subspaces can be specified, k_i at a time, by choice of the coefficients $\{p_{ij};\ j = 1, \ldots, k_i;\ i = 1, \ldots, r\}$. The fact that the eigenvalues for each invariant subspace can be specified independently of the remaining subspaces implies that these subspaces are all nonintersecting. This is verified independently by the following lemma.

<u>Lemma 4.5.</u> Let $\{f_1, \ldots, f_r\}$ be a set of output separable vectors. If, for each i, g_i is a k_i^{th} order detection generator for f_i, then the $(k_i + \ldots + k_r)$ vectors $\{g_1, \ldots, A^{k_1 - 1} g_1, g_2, \ldots, A^{k_r - 1} g_r\}$ are all linearly independent.

Proof:

Suppose the above vectors are linearly dependent. Then for some set of scalars $\{\sigma_{ij};\ j = 1, \ldots, k_i;\ i = 1, \ldots, r\}$, not all zero,

$$\sum_{i=1}^{r} \sum_{j=1}^{k_i} \sigma_{ij} A^{j-1} g_i = \underline{0} \tag{4-248}$$

Premultiplying this equation by C and using the properties of a detection generator gives

$$\sum_{i=1}^{r} \sigma_{ik_i} CA^{k_i - 1} g_i = \sum_{i=1}^{r} \sigma_{ik_i} CA^{i} f_i = \underline{0} \tag{4-249}$$

But the vectors $\{CA^{\mu_1} f_1, \ldots, CA^{\mu_r} f_r\}$ are linearly independent because the f_i are output separable. Therefore, (4-249) implies

$$\sigma_{ik_i} = 0 \quad \text{for } i = 1, \ldots, r \qquad (4\text{-}250)$$

Premultiplying (4-248) by CA and using (4-250)

$$\sum_{i=1}^{r} \sigma_{i,k_i-1} \; CA^{k_i-1} g_i = \sum_{i=1}^{r} \sigma_{i,k_i-1} \; CA^{\mu_i} f_i = \underline{0} \qquad (4\text{-}251)$$

which implies

$$\sigma_{i,k_i-1} = 0 \quad \text{for } i = 1, \ldots, r \qquad (4\text{-}252)$$

This procedure can be continued until all the σ_{ij} are shown to be zero. It must therefore be concluded that the vectors $\{g_1, \ldots, A^{k_1-1} g_1, g_2, \ldots, A^{k_r-1} g_r\}$ are all linearly independent. This proves the lemma.

Lemma 4.3 gives the general solution of (4-245) as

$$D = Q_d[(CF)^T CF]^{-1}(CF)^T + D'\left[I - CF[(CF)^T CF]^{-1}(CF)^T\right] \qquad (4\text{-}253)$$

When this D is put into (A - DC) the result is $A - DC = A' - D'C'$, where

$$A' = A - Q_d[(CF)^T CF]^{-1}(CF)^T C \qquad (4\text{-}254)$$

and

$$C' = \left[I - CF[(CF)^T CF]^{-1}(CF^T\right] C \qquad (4\text{-}255)$$

Equation (4-111) for each f_i can be used to obtain an expression for A' corresponding to (4-169),

$$A' = K - Z_d\,[(CF)^T CF]^{-1}(CF)^T C \tag{4-256}$$

where

$$K = A - AF\,[(CF)^T CF]^{-1}(CF)^T C \tag{4-257}$$

and

$$Z_d = [z_{d1}, \ldots, z_{dr}]$$

with

$$z_{di} = w_{di} - A^{\mu_i+1} f_i \tag{4-258}$$

The expression analogous to (4-122) for each $A^{\mu_i} f_i$ is

$$A^{\mu_i} f_i = \alpha_{i1} A^{\mu_i} g_i + \ldots + \alpha_{i,k_i-\mu_i-1} A^{k_i-2} g_i + A^{k_i-1} \tag{4-259}$$

Premultiplying this equation by A and substituting into (4-258) yields

$$z_{di} = p_{i1} g_i + \ldots + (p_{i,\mu_i+2} - \alpha_{i1}) A^{\mu_i+1} g_i + \ldots$$
$$\ldots + (p_{i,k_i} - \alpha_{i,k_i-\mu_i-1}) A^{k_i-1} g_i \tag{4-260}$$

By the same development used to obtain (4-182) it can be shown that

$$\begin{bmatrix} C' \\ C'A' \\ \cdot \\ \cdot \\ \cdot \\ C'A'^{n-1} \end{bmatrix} = \begin{bmatrix} C' \\ C'K \\ \cdot \\ \cdot \\ \cdot \\ C'K^{n-1} \end{bmatrix} = M' \tag{4-261}$$

The following definition is a generalization of the definition of the detection order for a single vector.

Definition 4.10. The dimension of the null space of M', $(n - \text{rk } M')$, is defined to be the group detection order of the set $\{f_1, \ldots, f_r\}$.

A necessary and sufficient condition for mutual detectability can now be presented.

Theorem 4.5. The output separable vectors $\{f_1, \ldots, f_r\}$ are mutually detectable if and only if the sum of the individual detection orders of the f_i is equal to the group detection order.

Proof:

Let M', K, and C' be defined by (4-261), (4-257), and (4-255). The group detection order of $\{f_1, \ldots, f_r\}$ is $(n - q')$ where $q' = \text{rk } M'$. Let ν_i be the detection order of f_i. If the maximal generator for each f_i is used in Equation (4-245) for D then $(\nu_1 + \ldots + \nu_r)$ eigenvalues of $(A - DC)$ can be specified, ν_i at a time, by the selection of the coefficients p_{ij}. An additional q' eigenvalues can be arbitrarily specified by the choice of D' in (4-253). The total number of eigenvalues which can be almost arbitrarily specified is therefore $(q' + \nu_1 + \ldots + \nu_r)$. This is the maximum number of eigenvalues which can be specified while constraining D to be a detector gain for all the f_i. Condition (2) of detectability will be satisfied if and only if $q' + \nu_1 + \ldots + \nu_r = n$, or

$$\nu_1 + \ldots + \nu_r = n - q' \tag{4-262}$$

This completes the proof.

When $(q' + \nu_1 + \ldots + \nu_r) < n$, there are $n - (q' + \nu_1 + \ldots$ $\ldots + \nu_r)$ eigenvalues over which the designer has no control after D is constrained to be a solution of (4-245). It will be shown in Section 4.3.4 that these uncontrolled eigenvalues depend only on A, C, and F. They do not depend on the coefficients p_{ij} or D' in (4-253). Therefore, it is not possible to gain even partial control over these eigenvalues by relaxing control over the other $(q' + \nu_1 + \ldots + \nu_r)$ eigenvalues. As in the case of a single event, nothing is gained by relaxing condition (2) unless one is willing to accept the uncontrolled eigenvalues which result when $(q' + \nu_1 + \ldots + \nu_r) < n$. This may be desirable if the uncontrolled eigenvalues are such that they do not adversely affect the dynamic behavior of the detection filter. Identifying the uncontrolled eigenvalues in the case of nonmutually detectable vectors is discussed in Section 4.3.4 .

Example E1 at the end of the previous section illustrates the above remarks. Each pair of event vectors considered in that example has a group detection order of three, because in every case the C' defined by (4-255) is a zero matrix, which means that M' given by (4-261) is also a zero matrix with rank zero. For the first pair of vectors $\{f_, f_2\}$ given by (E1-3) and (E1-41), the sum of the individual detection orders is $\nu + \nu_2 = 2 + 1 = 3$, which is equal to the group detection order. As shown in the example, all eigenvalues of (A - DC) can be specified while constraining D to be a detector gain for both f and f_2. For the second pair $\{f_1, f_2\}$ given by (E1-50) and (E1-41), the sum of the individual detection orders is only $\nu_1 + \nu_2 = 1 + 1 = 2$, and as (E1-54) verifies, one eigenvalue of (A - DC) is automatically fixed at s = 2 when D is constrained to be a detector gain for both f_1 and f_2. For the third pair $\{f_1, f_2\}$ given by (E1-50) and (E1-55), the sum of the individual detection orders is again only two. But in this case the uncontrolled eigenvalue is $s = -2$, so it is possible to obtain a stable detection filter which detects both f_1 and f_2 in spite of the fact that these two vectors are not mutually detectable.

Results on the mutual detectability of nonseparable vectors are incomplete, but a few useful facts are available. If a number of vectors have identical detection spaces, then a detection filter for one will be a detection filter for all. Since the error signal for all the vectors will lie in the same direction in the output space, the output error direction will not distinguish between the events associated with these vectors. However, the error magnitude may provide additional distinguishing information. This special case of nonseparable vectors is important in the detection of dynamic changes and is discussed in more detail in Section 4.3.6. For the general case of nonseparable vectors Equation (4-245) for D may or may not have a solution. A necessary condition for it to be a consistent matrix equation is that

$$\mathrm{rk}\, Q_d \leq \mathrm{rk}\, CF < r \tag{4-263}$$

Each column w_{di} in Q_d is in a subspace spanned by the vectors $\{A^j g_i;\ j = 0, 1, \ldots, k_i\}$. This subspace contains the detection space of f_i and can be one dimension larger because of the presence of $A^{k_i} g_i$. Condition (4-263) implies that these subspaces cannot all be independent because if they were, rk Q_d would be equal to r. Since Q_d depends on the coefficients p_{ij}, it appears that (4-263) imposes some restrictions on these coefficients. It is not clear at this point what restrictions, if any, this places on the specification of eigenvalues.

It is possible to show that if (A, C) is observable, then D cannot be a detector gain for two nonseparable vectors unless their detection spaces coincide. Let f_1 and f_2 be nonseparable. Assume for simplicity that $Cf_1 \neq \underline{0}$. (The same development is valid for the general case given by (4-243).) Since f_1 and f_2 are nonseparable

$$\mathrm{rk}\,\{C[f_1, f_2]\} = \mathrm{rk}\,[Cf_1, Cf_2] = 1 \tag{4-264}$$

This implies that the m-vectors Cf_1 and Cf_2 have the same direction. Suppose D is a detector gain for both f_1 and f_2. Then by Lemma 4.1

$$\mathrm{rk}\,(CW_{f1}) = 1 \tag{4-265}$$

$$\mathrm{rk}\,(CW_{f2}) = 1 \tag{4-266}$$

where

$$W_{f1} = [f_1, (A - DC)f_1, \ldots, (A - DC)^{n-1} f_1] \tag{4-267}$$

$$W_{f2} = [f_2, (A - DC)f_2, \ldots, (A - DC)^{n-1} f_2] \tag{4-268}$$

By (4-265) the range space of CW_{f1} is one-dimensional and, in fact, coincides with the direction of Cf_1 (Cf_1 is the first column of CW_{f1}). Similarly, the range space of CW_{f2} is one-dimensional and coincides with the direction of Cf_2. Since Cf_1 and Cf_2 have the same direction, the range spaces of CW_{f1} and CW_{f2} must coincide. Therefore

$$\mathrm{rk}\,[CW_{f1}, CW_{f2}] = \mathrm{rk}\,\{C[W_{f1}, W_{f2}]\} = 1 \tag{4-269}$$

Define

$$k_{12} = \text{rk}\,\{W_{f1}, W_{f2}] \tag{4-270}$$

Now form an $n \times k_{12}$ matrix, W_{f12}, whose columns consist of k_{12} independent columns from $[W_{f1}, W_{f2}]$. Then the range space of W_{f12} coincides with the range space of $[W_{f1}, W_{f2}]$. In particular, f_1 and f_2 are both in the range space of W_{f12}. By virtue of (4-269)

$$\text{rk}\, CW_{f12} = 1 \tag{4-271}$$

The development of Lemma 4.2 can be applied to W_{f12} to construct an n-vector g such that

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{k_{12}-2} \end{bmatrix} g = \underline{0} \tag{4-272}$$

$$CA^{k_{12}-1} g \neq \underline{0} \tag{4-273}$$

The set of vectors $\{A^j g;\ j = 0, \ldots, k_{12}-1\}$ span the range space of W_{f12}. Therefore, both f_1 and f_2 can be expressed as linear combinations of these vectors. This means that g, with an appropriate adjustment in magnitude, can be made a k_{12}^{th} order detection generator for either f_1 or f_2. Let g be a detection generator for f_1. By remark 3) at the end of Section 4.3.1, the vectors $\{A^j g;\ j = 0, \ldots, k_{12}-1\}$ are contained in the detection space of f_1. Then f_2 must be contained in the detection space of f_1. Again by remark 3) this implies the detection space of f_1 and f_2 coincide. This result does not generalize to sets of more than two nonseparable vectors.

Theorem 4.5 offers only a pass-fail type of test for mutual detectability. If the vectors in a given set are found to be not all mutually detectable, there is no way to discover which vectors are mutually detectable except by repeated application of Theorem 4.5 to all subsets of vectors in the original set. It would be desirable to have a systematic way of forming subsets of vectors which are mutually detectable. The next section is addressed to this problem.

4.3.3 Constructing Sets of Mutually Detectable Vectors

This section deals with the following problem. Given a set of output separable vectors $\{f_1, \ldots, f_r\}$ which are not all mutually detectable, determine which vectors can be removed from the set to leave a subset whose members are all mutually detectable. Each f_i has a detection space of dimension ν_i, the detection order of f_i. It will be shown that each of these detection spaces is an invariant subspace with respect to K given by (4-257) and is contained in the null space of M' given by (4-261). Since the f_i are output separable, Lemma 4.5 guarantees that the detection spaces are all nonintersecting. Together they make up a subspace of dimension $(\nu_1 + \ldots + \nu_r)$ contained in the $(n - q')$-dimensional null space of $M'(q' = \text{rk } M')$. When the f_i are not all mutually detectable $(n - q') > (\nu_1 + \ldots + \nu_r)$ and it is possible to define an "excess" subspace of dimension

$$k_e = n - q' - \nu_1 - \ldots - \nu_r \tag{4-274}$$

which is contained in the null space of M' and does not intersect any of the detection spaces. The precise definition of this space will be presented shortly. Its special properties and relationship to the detection spaces are of central concern in the investigation of the problem stated above.

First it will be verified that the detection space for each f_i is an invariant subspace with respect to K and is in the null space of M'. Let g_i be the maximal generator for f_i. Then

$$C'A^j g_i = \left[I - CF[(CF)^T CF]^{-1}(CF)^T\right] CA^j g_i = \underline{0} \tag{4-275}$$

for j = 0, 1, ..., ν_i-2 and by (4-120)

$$\begin{aligned} C'A^{\nu_i-1} g_i &= \left[I - CF[(CF)^T CF]^{-1}(CF)^T\right] CA^{\nu_i-1} g_i \\ &= \left[I - CF[(CF)^T CF]^{-1}(CF^T\right] CA^{\mu_i} f_i \\ &= CA^{\mu_i} f_i - CA^{\mu_i} f_i = \underline{0} \end{aligned} \tag{4-276}$$

Similarly

$$C'KA^j g_i = C'A^{j+1} g_i - C'AF[(CF)^T CF]^{-1}(CF)^T CA^j g_i = \underline{0} \tag{4-277}$$

for j = 0, 1, ..., ν_i-2 and with (4-259)

$$\begin{aligned} C'KA^{\nu_i-1} &= C'K[A^{\mu_i} f_i - \alpha_{i1} A^{\mu_i} g_i - \dots - \alpha_{i,\nu_i-\mu_i-1} A^{\nu_i-2} g_i] \\ &= C'KA^{\mu_i} f_i = \underline{0} \end{aligned} \tag{4-278}$$

since $KA^{\mu_i} f_i = \underline{0}$. This development can be repeated any number of times to show that

$$C'K^{\ell}A^{j}g_i = \underline{0} \tag{4-279}$$

for $j = 0, 1, \ldots, \nu_i-1$ and all integers $\ell \geq 0$. Then

$$M'A^{j}g_i = \underline{0} \text{ for } j = 0, \ldots, \nu_i-1 \text{ and } i = 1, \ldots, r \tag{4-280}$$

which shows that the basis vectors $\{A^j g_i;\ j = 0, \ldots, \nu_i-1\}$ for each detection space all lie in the null space of M'. From (4-80) and the form of K in (4-257) is follows that

$$K^{j}g_i = A^{j}g_i \text{ for } j = 0, \ldots, \nu_i-1 \tag{4-281}$$

so $\{K^j g_i;\ j = 0, \ldots, \nu_i-1\}$ form a basis for the detection space of f_i. Substituting (4-281) into (4-159)

$$A^{\mu_i}f_i = \alpha_{i1}K^{\mu_i}g_i + \ldots + \alpha_{i,\nu_i-\mu_i-1}K^{\nu_i-2}g_i + K^{\nu_i-1}g_i \tag{4-282}$$

Premultiplying this equation by K and recalling $KA^{\mu_i}f_i = \underline{0}$ yields

$$\alpha_{i1}K^{\mu_i+1}g_i + \ldots + \alpha_{i,\nu_i-\mu_i-1}K^{\nu_i-1}g_i + K^{\nu_i}g_i = \underline{0} \tag{4-283}$$

which shows that g_i generates an ν_i-dimensional cyclic subspace with respect to K. For each i define an $n \times \nu_i$ matrix

$$W_{gi} = [g_i, Ag_i, \ldots, A^{\nu_i-1}g_i] = [g_i, Kg_i, \ldots, K^{\nu_i-1}g_i] \tag{4-284}$$

Then using (4-283)

$$KW_{gi} = [Kg_i, \ldots, K^{\nu_i} g_i] = W_{gi} P_{\alpha i} \qquad (4\text{-}285)$$

where $P_{\alpha i}$ is an $\nu_i \times \nu_i$ matrix of the form

$$P_{\alpha i} = \begin{bmatrix} 0 & 0 & & 0 & 0 \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & \ddots & \vdots & -\alpha_{i1} \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & -\alpha_{i,\nu_i-\mu_i-1} \end{bmatrix} \qquad (4\text{-}286)$$

Now let the set of n-vectors $\{z_{ei}, \ldots, z_{ek_e}\}$ be a basis for the excess subspace mentioned earlier. These vectors are linearly independent of each other and of the basis vectors for the detection space of the f_i. The complete set of vectors $\{g_1, \ldots, A^{\nu_1-1} g_1, \ldots, \ldots, A^{\nu_r-1} g_r, z_{e1}, \ldots, z_{ek_e}\}$ forms a basis for the null space of M'. Define the $n \times k_e$ matrix

$$Z_e = [z_{e1}, \ldots, z_{ek_e}] \qquad (4\text{-}287)$$

Since the z_{ei} are in the null space of M'

$$M'Z_e = \underline{0} \qquad (4\text{-}288)$$

and

$$C'K^{j-1} Z_e = \underline{0} \text{ for all } j \geq 1 \qquad (4\text{-}289)$$

With (4-255) this gives

$$CK^{j-1} Z_e = CF[(CF)^T CF]^{-1}(CF)^T CK^{j-1} Z_e$$

$$= \sum_{i=1}^{r} CA^{\mu_i} f_i \gamma_{ij} \tag{4-290}$$

where the γ_{ij} are $1 \times k_e$ row vectors and

$$\begin{bmatrix} \gamma_{ij} \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{rj} \end{bmatrix} = [(CF)^T CF]^{-1}(CF)^T CK^{j-1} Z_e \tag{4-291}$$

The basis vectors $\{z_{e1}, \ldots, z_{ek_e}\}$ are to be chosen so that

$$\gamma_{ij} = \underline{0}, \text{ for } j = 1, \ldots, \nu_i \tag{4-292}$$

It must now be demonstrated that this is, in fact, possible. Let $\{z_1', \ldots, z_{k_e}'\}$ be any set of independent vectors which together with the set $\{A^j_{g_i}; j = 0, \ldots, \nu_i - 1; i = 1, \ldots, r\}$ form a basis for the null space of M'. Define

$$Z' = [z_1', \ldots, z_{k_e}'] \tag{4-293}$$

An equation analogous to (4-290) can be written for Z'

$$CK^{j-1} Z' = \sum_{i=1}^{r} CA^{\mu_i} f_i \gamma_{ij}' \tag{4-294}$$

where

$$\begin{bmatrix} \gamma'_{ij} \\ \cdot \\ \cdot \\ \cdot \\ \gamma'_{rj} \end{bmatrix} = [(CF)^T CF]^{-1}(CF)^T CK^{j-1} Z' \tag{4-295}$$

Let

$$Z_e = Z' + \sum_{i=1}^{r} W_{gi} J_i \tag{4-296}$$

where the J_i are $\nu_i \times k_e$ matrices chosen so that

$$\hat{u}_i P_{\alpha i}^{j-1} J_i = -\gamma'_{ij} \text{ for } j = 1, \ldots, \quad i \tag{4-297}$$

with $P\alpha_i$ defined by (4-286) and $\hat{u}_i$ a $1 \times \nu_i$ unit row vector

$$\hat{u}_i = [0, \ldots, 0, 1] \tag{4-298}$$

The set of equations (4-297) defines J_i uniquely as can be seen when they are combined into a single matrix equation

$$\begin{bmatrix} \hat{u}_i \\ \hat{u}_i P_{\alpha i} \\ \cdot \\ \cdot \\ \cdot \\ \hat{u}_i P_{\alpha i}^{\nu_i - 1} \end{bmatrix} J_i = -\begin{bmatrix} \gamma'_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \gamma'_{i\nu_i} \end{bmatrix} \tag{4-299}$$

The $\nu_i \times \nu_i$ matrix on the left has the triangular form

$$\begin{bmatrix} \hat{u}_i \\ \hat{u}_i P_{\alpha i} \\ \vdots \\ \hat{u}_i P_{\alpha i}^{\nu_i - 1} \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & \iddots & \sigma \\ 0 & \iddots & & \vdots \\ 1 & \sigma & \cdots & \sigma \end{bmatrix} \tag{4-300}$$

and is clearly nonsingular (σ denotes possible nonzero elements). With J_i so defined

$$CK^{j-1} Z_e = CK^{j-1} Z' + \sum_{i=1}^{r} CK^{j-1} W_{gi} J_i$$

$$= \sum_{i=1}^{r} CA^{\mu_i} f_i \gamma'_{ij} + \sum_{i=1}^{r} CK^{j-1} W_{gi} J_i \tag{4-301}$$

Noting that

$$K^{j-1} W_{gi} = W_{gi} P_{\alpha i}^{j-1} \tag{4-302}$$

by repeated application of (4-285) and also

$$CW_{gi} = [\underline{0}, \ldots, \underline{0}, CA^{\mu_i} f_i] = CA^{\mu_i} f_i \hat{u}_i \tag{4-303}$$

Equation (4-301) then becomes

$$CK^{j-1} Z_e = \sum_{i=1}^{r} CA^{\mu_i} f_i \gamma'_{ij} + \sum_{i=1}^{r} CW_{gi} P_{\alpha i}^{j-1} J_i$$

$$= \sum_{i=1}^{r} CA^{\mu_i} f_i [\gamma'_{ij} + \hat{u}_i P_{\alpha i}^{j-1} J_i] \tag{4-304}$$

Comparing this with (4-290) one may conclude that

$$\gamma_{ij} = \gamma'_{ij} + \hat{u}_i P_{\alpha i}^{j-1} J_i \tag{4-305}$$

since the $CA^{\mu_i} f_i$ are linearly independent. Then (4-292) follows directly from (4-297).

Equation (4-285) shows that the range space of each W_{gi} is an invariant space with respect to K. The range space of Z_e is not an invariant space itself, but is at least contained in the null space of M', which is an invariant space. Therefore KZ_e is also in the null space of M' and can be expressed as a linear combination of Z_e and the W_{gi}, since the combined range spaces of these matrices coincide with the null space of M'. So

$$KZ_e = Z_e \Lambda + \sum_{i=1}^{r} W_{gi} \Gamma_i \tag{4-306}$$

for some $k_e \times k_e$ matrix Λ and some $\nu_i \times k_e$ matrices Γ_i. Then

$$CK^j Z_e = CK^{j-1} Z_e \Lambda + \sum_{i=1}^{r} CK^{j-1} W_{gi} \Gamma_i$$

$$= \sum_{i=1}^{r} CA^{\mu_i} f_i \gamma_{ij} \Lambda + \sum_{i=1}^{r} CW_{gi} P_{\alpha i}^{j-1} \Gamma_i$$

$$= \sum_{i=1}^{r} CA^{\mu_i} f_i [\gamma_{ij} \Lambda + \hat{u}_i P_{\alpha i}^{j-1} \Gamma_i] \tag{4-307}$$

Comparing this expression with (4-290) with (j - 1) replaced by j, one may conclude that

$$\gamma_{i,j+1} = \gamma_{ij} + \hat{u}_i P_{\alpha i}^{j-1} \Gamma_i \tag{4-308}$$

This along with (4-292) implies that

$$\hat{u}_i P_{\alpha i}^{j-1} \Gamma_i = \begin{cases} \underline{0} & ; \quad j=1, \ldots, \nu_i - 1 \\ \gamma_{i,\nu_i+1} & ; \quad j = \nu_i \end{cases} \tag{4-309}$$

The row vectors γ_{i,ν_i+1} will be referred to frequently in what follows, so it will be convenient to introduce a simpler notation for them

$$\theta_i = \gamma_{i,\nu_i+1} \tag{4-310}$$

Writing (4-309) in matrix form,

$$\begin{bmatrix} \hat{u}_i \\ \hat{u}_i P_{\alpha i} \\ \cdot \\ \cdot \\ \cdot \\ \hat{u}_i P_{\alpha i}^{\nu_i - 1} \end{bmatrix} \Gamma_i = \begin{bmatrix} \underline{0} \\ \cdot \\ \cdot \\ \cdot \\ \underline{0} \\ \theta_i \end{bmatrix} \tag{4-311}$$

Noting the triangular form of (4-309) this equation is easily solved for Γ_i to yield

$$\Gamma_i = \begin{bmatrix} \theta_i \\ \underline{0} \\ \cdot \\ \cdot \\ \cdot \\ \underline{0} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \theta_i \tag{4-312}$$

Then (4-306) reduces to

$$KZ_e = Z_e\Lambda + \sum_{i=1}^{r} g_i \theta_i \tag{4-313}$$

The $k_e \times k_e$ matrix Λ and the $1 \times k_e$ row vector θ_i associated with each f_i is sufficient to determine which vectors in the set can be removed to leave all the remaining vectors mutually detectable. The following theorem is the basis for that determination.

<u>Theorem 4.6.</u> Let Z_e, Λ , and the θ_i for each f_i be defined as above. Assume ℓ vectors $\{f_{i_1}, \ldots, f_{i_\ell}\}$ are removed from the original set of r, $\{f_1, \ldots, f_r\}$. Then for the remaining $(r - \ell)$ vectors the new excess subspace has dimension

$$k = k_e - rk \begin{bmatrix} \Theta \\ \Theta\Lambda \\ \vdots \\ \Theta\Lambda^{k_e-1} \end{bmatrix} \tag{4-314}$$

where Θ is an $\ell \times k_e$ matrix whose rows are the θ_i corresponding to the f_i which were removed

$$\Theta = \begin{bmatrix} \theta_{i_1} \\ \vdots \\ \theta_{i_\ell} \end{bmatrix} \tag{4-315}$$

Furthermore, a basis for the new excess subspace is formed by the set of vectors

$$y_i = Z_e \beta_{ei} \quad \text{for } i = 1, \ldots, k \tag{4-316}$$

where the set of k_e-vectors $\{\beta_{e1}, \ldots, \beta_{ek}\}$ is any basis for the null space of

$$\begin{bmatrix} \Theta \\ \Theta\Lambda \\ \cdot \\ \cdot \\ \cdot \\ \Theta\Lambda^{k_e-1} \end{bmatrix}$$

The following corollary demonstrates the effect of removing a single vector from the original set.

Corollary 4.6.1. If f_i is removed from the set of vectors, then the dimension of the excess subspace will be reduced by an amount equal to

$$\text{rk} \begin{bmatrix} \theta_i \\ \theta_i\Lambda \\ \cdot \\ \cdot \\ \cdot \\ \theta_i\Lambda^{k_e-1} \end{bmatrix}$$

Proof:

Simply take $\Theta = \theta_i$ in Theorem 4.6. The next corollary provides an answer to the problem stated at the beginning of this section.

Corollary 4.6.2. The vectors remaining after the removal of ℓ vectors $\{f_{i_1}, \ldots, f_{i_\ell}\}$ are mutually detectable if and only if (Λ, Θ) is an observable pair.

Proof:

The remaining vectors are mutually detectable if and only if the new excess subspace has zero dimension. By Theorem 4.6 this will be the case if and only if

$$\mathrm{rk}\begin{bmatrix} \Theta \\ \Theta\Lambda \\ \vdots \\ \Theta\Lambda^{k_e-1} \end{bmatrix} = k_e \tag{4-317}$$

which is the condition for (Λ, Θ) to be an observable pair.

Proof of Theorem 4.6:

For convenience of notation, assume that the first ℓ vectors are removed from the original set to leave $\{f_{\ell+1}, \ldots, f_r\}$. Define

$$F_2 = [A^{\mu_{\ell+1}} f_{\ell+1}, \ldots, A^{\mu_r} f_r] \tag{4-318}$$

$$K_2 = A - AF_2[(CF_2)^T CF_2]^{-1}(CF_2)^T C \tag{4-319}$$

$$C_2' = \left[I - CF_2[(CF_2)^T CF_2]^{-1}(CF_2)^T\right] C \tag{4-320}$$

$$M_2' = \begin{bmatrix} C_2' \\ C_2'K_2 \\ \vdots \\ C_2'K_2^{n-1} \end{bmatrix} \tag{4-321}$$

which are analogous to F, K, C', and M' for the original set. The detection spaces of $\{f_{\ell+1}, \ldots, f_r\}$ are contained in the null space of M_2'. These vectors are mutually detectable if and only if the dimension of this null space is exactly $(\nu_{\ell+1} + \ldots + \nu_r)$. Suppose its dimension is larger than this. Then there will exist some n-vector z in the null space of M_2' which is independent of the detection spaces. Any vector in the null space of M_2' is also in the null space of M'. Moreover, $M_2' z = \underline{0}$ if and only if

$$\begin{bmatrix} C_2' \\ C_2'K \\ \vdots \\ C_2'K^{n-1} \end{bmatrix} z = \underline{0} \tag{4-322}$$

These two facts follow from the lemma below.

<u>Lemma 4.6.</u> If $C_2' z = \underline{0}$ for some n-vector z then

$$Kz = K_2 z \tag{4-323}$$

and

$$C'z = \underline{0} \tag{4-324}$$

Proof:

Now

$$C_1'z = Cz - CF_2[(CF_2)^T CF_2]^{-1}(CF_2)^T Cz = \underline{0} \tag{4-325}$$

so

$$Cz = CF_2 \xi_2 \tag{4-326}$$

where

$$\xi_2 = [(CF_2)^T CF_2]^{-1}(CF_2)^T Cz \tag{4-327}$$

From the definitions of F and F_2

$$F_2\xi_2 = F\xi \tag{4-328}$$

where ξ is defined as

$$\xi = \begin{bmatrix} \underline{0} \\ \\ \xi_2 \end{bmatrix} \tag{4-329}$$

Then

$$Cz = CF\xi \tag{4-330}$$

and thus

$$\begin{aligned} C'z &= Cz - CF[(CF)^T CF]^{-1}(CF)^T Cz \\ &= Cz - CF[(CF)^T CF]^{-1}(CF)^T CF\xi \\ &= Cz - CF\xi = \underline{0} \end{aligned} \tag{4-331}$$

Also

$$\begin{aligned} K_2 z &= Az - AF_2[(CF_2)^T CF_2]^{-1}(CF_2)^T Cz \\ &= Az - AF_2\xi_2 = Az - AF\xi \\ &= Az - AF[(CF)^T CF]^{-1}(CF)^T CF\xi \\ &= Az - AF[(CF)^T CF]^{-1}(CF)^T Cz = Kz \end{aligned} \tag{4-332}$$

which completes the proof.

Successive application of (4-323) to $K_2^j z$ and $K^j z$ with $j = 0, 1, \ldots, n-1$ yields (4-322). The fact that z is in the null space of M' follows from (4-322) and (4-324). It is therefore possible to express z as

$$z = Z_e \beta_e + \sum_{i=1}^{r} W_{gi} \beta_i \tag{4-333}$$

for some k_e-vector β_e and ν_i-vectors β_i. With (4-290) and (4-302)

$$\begin{aligned} CK^{j-1} z &= CK^{j-1} Z_e \beta_e + \sum_{i=1}^{r} CK^{j-1} W_{gi} \beta_i \\ &= \sum_{i=1}^{r} CA^{\mu_i} f_i \gamma_{ij} \beta_e + \sum_{i=1}^{r} CA^{\mu_i} f_i \hat{u}_i P_{\alpha i}^{j-1} \beta_i \end{aligned} \tag{4-334}$$

Premultiplying by $I - CF_2[(CF_2)^T CF_2]^{-1}(CF_2)^T$

$$C_2' K^{j-1} z = \sum_{i=1}^{r} C_2' A^{\mu_i} f_i [\gamma_{ij} \beta_e + \hat{u}_i P_{\alpha i}^{j-1} \beta_i] \tag{4-335}$$

Because the f_i are output separable, the vectors $\{CA^{\mu_i} f_i;\ i = 1, \ldots, r\}$ are linearly independent. (These vectors make up the columns of CF which has rank r.) Equation (4-326) shows that if $C_2' A^{\mu_i} f_i = \underline{0}$ then $CA^{\mu_i} f_i$ can be expressed as a linear combination of the columns of CF_2, or in other words a linear combination of the vectors $\{CA^{\mu_i} f_i;\ i = \ell + 1, \ldots, r\}$. But the vectors $\{CA^{\mu_i} f_i;\ i = 1, \ldots, \ell\}$ are independent of the vectors $\{CA^{\mu_i} f_i ;\ i = \ell+1, \ldots, r\}$, so $C_2' A^{\mu_i} f_i \neq \underline{0}$ for $i = 1, \ldots, \ell$. Consequently,

$$C_2' A^{\mu_i} f_i = CA^{\mu_i} f_i - CF_2 [(CF_2)^T CF_2]^{-1} (CF_2)^T CA^{\mu_i} f_i$$

$$\left\{ \begin{array}{l} = \underline{0} \text{ if } i = \ell + 1, \ldots, r \\ \neq \underline{0} \text{ if } i = 1, \ldots, \ell \end{array} \right\} \tag{4-336}$$

Then (4-335) reduces to

$$C_2' K^{j-1} z = \sum_{i=1}^{\ell} C_2' A^{\mu_i} f_i [\gamma_{ij} \beta_e + \hat{u}_i P_{\alpha i}^{j-1} \beta_i] \tag{4-337}$$

But from (4-332) $C_2' K^{j-1} z = \underline{0}$ for all $j \geq 1$. Therefore

$$\gamma_{ij} \beta_e + \hat{u}_i P_{\alpha i}^{j-1} \beta_i = \underline{0} \quad \text{for } i = 1, \ldots, \ell \tag{4-338}$$

and for all $j \geq 1$, since the $C_2' A^{\mu_i} f_i$ are independent. By (4-292) this reduces to

$$\hat{u}_i P_{\alpha i}^{j-1} \beta_i = \underline{0} \quad \text{for } j = 1, \ldots, \nu_i \tag{4-339}$$

or

$$\begin{bmatrix} \hat{u}_i \\ \hat{u}_i P_{\alpha i} \\ \vdots \\ \hat{u}_i P_{\alpha i}^{\nu_i - 1} \end{bmatrix} \beta_i = \underline{0} \quad \text{for } i = 1, \ldots, \ell \tag{4-340}$$

which implies

$$\beta_i = \underline{0} \quad \text{for } i = 1, \ldots, \ell \tag{4-341}$$

by virtue of (4-300). Then (4-338) becomes

$$\gamma_{ij}\beta_e = 0 \quad \text{for } i = 1, \ldots, \ell \text{ and } j \geq \nu_i + 1 \tag{4-342}$$

Define

$$S_j = \begin{bmatrix} \gamma_{1,\nu_1+j} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{\ell,\nu_\ell+j} \end{bmatrix} \tag{4-343}$$

Then (4-342) can be written

$$S_j\beta_e = \underline{0} \quad \text{for } j \geq 1 \tag{4-344}$$

Now from (4-308) and (4-312)

$$\gamma_{i,j+1} = \gamma_{ij}\Lambda + \alpha'_{ij}\,\theta_i \tag{4-345}$$

where

$$\alpha'_{ij} = \hat{u}_i\, P_{\alpha i}^{j-1} \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \tag{4-346}$$

Repeated application of (4-345) starting with $j = \nu_i + 1$ and $\gamma_{i,\nu_i+1} = \theta_i$ yields the general expression

$$\gamma_{i,\nu_i+j} = \theta_i\Lambda^{j-1} + \alpha'_{i,\nu_i+1}\,\theta_i\Lambda^{j+2} + \dots + \alpha'_{i,\nu_i+j-1}\,\theta_i \tag{4-347}$$

for all $j \geq 1$. Using this expression in the definition of S_j

$$S_j = S_1\Lambda^{j-1} + Q_1 S_1\Lambda^{j-2} + \dots + Q_{j-1}\,S_1 \tag{4-348}$$

where

$$Q_j = \begin{bmatrix} \alpha'_{1,\nu_1+j} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \alpha'_{\ell,\nu_\ell+j} \end{bmatrix} \tag{4-349}$$

Noting that $S_1 = \Theta$

$$\begin{bmatrix} S_1 \\ \vdots \\ S_{k_e} \end{bmatrix} = \hat{T}_Q \begin{bmatrix} S_1 \\ S_1\Lambda \\ \vdots \\ S_1\Lambda^{k_e-1} \end{bmatrix} = \hat{T}_Q = \begin{bmatrix} \Theta \\ \Theta\Lambda \\ \vdots \\ \Theta\Lambda^{k_e-1} \end{bmatrix} \tag{4-350}$$

where

$$\hat{T}_Q = \begin{bmatrix} I & \underline{0} & \cdots & \underline{0} \\ Q_1 & \ddots & \ddots & \vdots \\ Q_2 & \ddots & \ddots & \underline{0} \\ \vdots & \ddots & \ddots & \\ Q_{k_e-1} & \cdots\ Q_2 & Q_1 & I \end{bmatrix} \tag{4-351}$$

Now (4-344) for $j = 1, \ldots, k_e$ can be written

$$\begin{bmatrix} S_1 \\ \cdot \\ \cdot \\ \cdot \\ S_{k_e} \end{bmatrix} \beta_e = \underline{0} \tag{4-352}$$

which is equivalent to

$$\begin{bmatrix} \Theta \\ \Theta\Lambda \\ \cdot \\ \cdot \\ \cdot \\ \Theta\Lambda^{k_e-1} \end{bmatrix} \beta_e = \underline{0} \tag{4-353}$$

since T_Q is nonsingular. If (4-353) is satisfied, then $S_j\beta_e = \underline{0}$ for all $j \geq 1$ because

$$\text{rk} \begin{bmatrix} S_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ S_j \end{bmatrix} = \text{rk} \begin{bmatrix} \Theta \\ \Theta\,\Lambda \\ \cdot \\ \cdot \\ \cdot \\ \Theta\Lambda^{j-1} \end{bmatrix} \leq k_e \tag{4-354}$$

for any $j \geq 1$. With (4-341), Equation (4-333) reduces to

$$z = Z_e\beta_e + \sum_{i=\ell+1}^{r} W_{gi}\beta_i = [\, Z_e, W_{g,\ell+1}, \ldots, W_{gr}] \begin{bmatrix} \beta_e \\ \beta_{\ell+1} \\ \cdot \\ \cdot \\ \cdot \\ \beta_r \end{bmatrix} \tag{4-355}$$

where β_e must satisfy (4-353). The only constraint placed on z in arriving at (4-355) was that it lie in the null space of M_2'. All vectors in this null space can therefore be expressed in the form of (4-355). Since the $\mathrm{rk}[Z_e, W_{g,\ell+1}, \ldots, W_{gr}] = (k_e + \nu_{\ell+1} + \ldots + \nu_r)$, the dimension of the null space of M_2' is simply the number of independent $(k_e + \nu_{\ell+1} + \ldots + \nu_r)$-vectors of the form

$$\begin{bmatrix} \beta_e \\ \beta_{\ell+1} \\ \cdot \\ \cdot \\ \cdot \\ \beta_r \end{bmatrix}$$

where β_e must satisfy (4-353). The β_i ($i = \ell + 1, \ldots, r$) are unconstrained so there are at least $(\nu_{\ell+1} + \ldots + \nu_r)$ such vectors. This was expected because the detection spaces of $\{f_{\ell+1}, \ldots, f_r\}$ are known to lie in the null space of M_2'. The number of additional independent vectors in the null space is the number of independent solutions of (4-353). This number is

$$k = k_e - \mathrm{rk} \begin{bmatrix} \Theta \\ \Theta\Lambda \\ \cdot \\ \cdot \\ \cdot \\ \Theta\Lambda^{k_e-1} \end{bmatrix}$$

This, then, is the dimension of the excess subspace for $\{f_{\ell+1}, \ldots, f_r\}$. Let $\{\beta_{e1}, \ldots, \beta_{ek}\}$ be k independent solutions of (4-353). Define

$$Y_e = [\beta_{e1}, \ldots, \beta_{ek}] \tag{4-356}$$

and

$$Z'_e = Z_e Y_e = [z'_{e1}, \ldots, z'_{ek}] \tag{4-357}$$

with

$$z'_{ei} = Z_e \beta_{ei} \quad \text{for } i = 1, \ldots, k \tag{4-358}$$

The columns of Z'_e are in the form of (4-355) (with the $\beta_i = \underline{0}$) and are therefore in the null space of M'_2. Then by Lemma 4.6

$$K_2 Z'_e = KZ'_e = KZ_e Y_e = Z_e \Lambda Y_e + \sum_{i=1}^{r} W_{gi} \Gamma_i Y_e \tag{4-359}$$

Now the range space of Y_e is an invariant subspace with respect to Λ because it coincides with the null space of

$$\begin{bmatrix} \Theta \\ \Theta\Lambda \\ \cdot \\ \cdot \\ \cdot \\ \Theta\Lambda^{k_e - 1} \end{bmatrix}$$

Thus

$$\Lambda Y_e = Y_e \Lambda' \tag{4-360}$$

for some $k \times k$ matrix Λ'. Note also that for $i = 1, \ldots, \ell$

$$\Gamma_i Y_e = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \theta_i Y_e = \underline{0} \tag{4-361}$$

For $i = \ell + 1, \ldots, r$

$$\Gamma_i Y_e = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \theta_i Y_e = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \theta_i' \tag{4-362}$$

where

$$\theta_i' = \theta_i Y_e \tag{4-363}$$

Substituting (4-360), (4-361) and (4-362) into (4-359) gives

$$K_2 Z_e' = Z_e Y_e \Lambda' + \sum_{i=\ell+1}^{r} W_{gi} \Gamma_i Y_e$$

$$= Z_e' \Lambda' + \sum_{i=\ell+1}^{r} g_i \theta_i' \tag{4-364}$$

This equation is analogous to (4-313).

The columns of Z_e' form a basis for the new excess subspace for $\{f_{\ell+1}, \ldots, f_r\}$. To see this, first note that the columns of Z_e' are indeed independent of the detection spaces of $\{f_{\ell+1}, \ldots, f_r\}$, since by (4-357) the range space of Z_e' is contained in the range space of Z_e, which by construction is independent of all the detection spaces. It was noted earlier that the columns of Z_e' are in the null space of M_2' and therefore $KZ_e' = K_2 Z_e'$ by Lemma 4.6. Since the null space of M_2' is invariant with respect to K_2, the range space of $K_2^{j-1} Z_e'$ is also in null space of M_2' for all $j \geq 1$, and

$$K_2^{j-1} Z_e' = K^{j-1} Z_e' \tag{4-365}$$

Then

$$CK_2^{j-1} Z_e' = CK^{j-1} Z_e' = CK^{j-1} Z_e Y_e \tag{4-366}$$

Substituting (4-290) into this equation yields

$$CK_2^{j-1} Z_e' = \sum_{i=1}^{r} CA^{\mu_i} f_i \gamma_{ij} Y_e \tag{4-367}$$

Now the columns of Y_e satisfy (4-353) which is equivalent to (4-342). Also $\gamma_{ij} = \underline{0}$ for $j = 1, \ldots, \nu_i$ so one may conclude that

$$\gamma_{ij} Y_e = \underline{0} \text{ for } i = 1, \ldots, \ell$$

$$\text{and for all } j \geq 1 \tag{4-368}$$

Then (4-367) reduces to

$$CK_2^{j-1} Z_e' = \sum_{i=\ell+1}^{r} CA^{\mu_i} f_i \gamma_{ij} Y_e \tag{4-369}$$

The row vectors $(\gamma_{ij} Y_e)$ for $i = \ell+1, \ldots, r$ play the same role as the γ_{ij} for the original excess subspace. From (4-292)

$$\gamma_{ij} Y_e = \underline{0} \quad \text{for } j = 1, \ldots, \nu_i \tag{4-370}$$

so Z_e' satisfies the condition analogous to (4-292) used to define the excess subspace. This completes the proof of Theorem 4.6.

Appendix B describes an algorithm for generating a basis for the excess subspace, plus the Λ matrix and the row vectors θ_i. Corollary 4.6.2 reduces the problem of constructing a subset of mutually detectable vectors to the problem of finding a subset of the row vectors θ_i which form a Θ such that (Λ, Θ) is an observable pair. At first glance this may seem to be only a pass-fail type test such as provided by Theorem 4.5. However, Λ and the θ_i can provide additional information to guide the choice of which vectors to remove from the original set. Corollary 4.6.1, for example, can be used to identify those vectors whose removal would achieve the greatest reduction in the size of the excess subspace. More information can be obtained from a systematic analysis of Λ and the θ_i as will be seen in the next section. In addition to providing a way of analyzing the problem of detecting a set of vectors with a single filter, Theorem 4.6 has achieved a potentially significant reduction in the dimensionality of the problem. Mutual detectability as originally formulated in Section 4.3.3 deals with an n-dimensional vector space. Theorem 4.6 reduces the problem to considerations in a vector space of dimension k_e, which one might reasonably expect to be significantly smaller than n (recall $k_e = n-q-\nu_1 - \dots - \nu_r$).

4.3.4 Detection of Nonmutually Detectable Vectors with a Single Filter

By definition, a set of vectors which are mutually detectable can be detected with a single filter while retaining control over all the eigenvalues of (A - DC). If one encounters a set of vectors which are not all mutually detectable, the results of the previous section can be used to break up this set into a group of two or more subsets, each of which is made up of only mutually detectable vectors. One detection filter can then be designed for each subset. If this is done, one need consider only the problem of designing a detection filter for mutually detectable vectors. However, if one allows the possibility of using a single filter for nonmutually detectable vectors, it may be possible to reduce the number of detection filters, since a potentially greater number of vectors could be assigned to each filter.

This section investigates the problem of using a single detection filter for a set of output separable but nonmutually detectable vectors. The results of the last two sections show that when this is attempted the resulting (A - DC) matrix will have k_e eigenvalues fixed without the control of the designer, where k_e is the dimension of the excess subspace for the set of vectors. To decide if detection of the set with a single filter is feasible, one must be able to identify these uncontrolled eigenvalues to see if the filter will have satisfactory dynamics. It will be shown in this section that these eigenvalues are indeed uncontrollable -- that they depend only on A, C, and F and are not influenced by the designer's choice of the remaining $(n - k_e)$ eigenvalues of (A - DC). Further, they will be shown to be equal to the eigenvalues of the $k_e \times k_e$ matrix Λ introduced in the previous section. From θ_i it will be possible to determine which of the uncontrolled eigenvalues are eliminated by removing the corresponding f_i from the original set. With this information the designer can eliminate specific undesirable eigenvalues by removing certain f_i from the set.

Suppose D is chosen to be a detector gain for the set of output separable vectors $\{f_1, \ldots, f_r\}$. Define an $n \times n$ coordinate transformation matrix

$$T_F = [W_{g1}, \ldots, W_{gr}, Z_e, T_{F2}] \qquad (4\text{-}371)$$

where Z_e and the W_{gi} are defined as in the last section, and T_{F2} is any $n \times q'$ matrix such that T_F is nonsingular ($q' + k_e + \nu_1 + \ldots + \nu_r = n$). Let

$$\overline{G} = T_F^{-1} (A - DC) T_F \qquad (4\text{-}372)$$

Now by (4-115)

$$(A - DC) W_{gi} = [(A - DC)g_i, \ldots, (A - DC)^{\nu_i} g_i] = W_{gi} P_i \qquad (4\text{-}373)$$

where

$$P_i = \begin{bmatrix} 0 & 0 & 0 & -p_{i1} \\ 1 & 0 & \vdots & \vdots \\ 0 & 1 & \vdots & \vdots \\ \vdots & \ddots & 0 & \vdots \\ 0 & 0 & 1 & -p_{i\nu_i} \end{bmatrix} \qquad (4\text{-}374)$$

From (4-290) and (4-292) with j = 1

$$CZ_e = \underline{0} \qquad (4\text{-}375)$$

(Note that ν_i, the detection order of f_i, is always greater than zero because the null space of M' defined by (4-182) contains f_i and therefore has dimension greater than or equal to one.) Then

$$(A - DC)Z_e = AZ_e = KZ_e = Z_e\Lambda + \sum_{i=1}^{r} W_{gi}\Gamma_i \tag{4-376}$$

With (4-373) and (4-376)

$$(A-DC)T_F = [W_{g1}P_1, \ldots, W_{gr}P_r, (Z_e\Lambda + \sum_{i=1}^{r} W_{gi}\Gamma_i), (A-DC)T_{F2}]$$

$$= T_F \begin{bmatrix} P_1 & \underline{0} & \Gamma_1 & \overline{G}_{1,r+2} \\ \underline{0} & \ddots & \vdots & \vdots \\ \vdots & \underline{0} & \vdots & \vdots \\ \vdots & P_r & \Gamma_r & \vdots \\ \vdots & \underline{0} & \Lambda & \vdots \\ \underline{0} & \underline{0} & \underline{0} & \overline{G}_{r+2,r+2} \end{bmatrix} \tag{4-377}$$

where the $\overline{G}_{i,r+2}$ are defined by

$$(A-DC)T_{F2} = \sum_{i=1}^{r} W_{gi}\overline{G}_{i,r+2} + \Lambda\,\overline{G}_{r+1,r+2} + T_{F2}\overline{G}_{r+2,r+2} \tag{4-378}$$

Premultiplying (4-377) by T_F^{-1} and comparing the result with (4-372) yields

$$\overline{G} = \begin{bmatrix} P_1 & & & \underline{0} & \Gamma_1 & \overline{G}_{1,r+2} \\ \underline{0} & \ddots & & \vdots & \vdots & \vdots \\ \vdots & & \ddots & \underline{0} & \vdots & \vdots \\ \vdots & & & P_r & \Gamma_r & \vdots \\ \vdots & & & \underline{0} & \Lambda & \vdots \\ \underline{0} & & & \underline{0} & \underline{0} & \overline{G}_{r+2,r+2} \end{bmatrix} \qquad (4\text{-}379)$$

Since $\overline{G}$ and (A - DC) are similar, they have identical eigenvalues. From the block diagonal form of $\overline{G}$ one can conclude that the eigenvalues of (A - DC) are equal to the combined eigenvalues of Λ, the P_i, and $\overline{G}_{r+2,r+2}$. Recall from Section 4.3.3 that the construction of Λ depends only on A, C, and the detection spaces of the f_i. It does not depend on the coefficients p_{ij} which appear in the P_i. Therefore, the k_e eigenvalues of Λ, which are equal to k_e eigenvalues of (A - DC), are independent of the eigenvalues of the P_i. The eigenvalues of $\overline{G}_{r+2,r+2}$ are determined by the choice of D' in (4-253). By Lemma 4.4, D' does not influence the eigenvalues of the P_i or Λ. This shows that the eigenvalues of Λ are, in fact, the uncontrolled eigenvalues which result when D is constrained to be a detector gain for the set of output separable, nonmutually detectable vectors.

Consider θ_i, as defined in Section 4.3.3, which is associated with one vector, f_i, in the set $\{f_1, \ldots, f_r\}$. If that vector is removed from the set, the new excess subspace will have dimension

$$k = k_e - rk \begin{bmatrix} \theta_i \\ \theta_i \Lambda \\ \vdots \\ \theta_i \Lambda^{k_e - 1} \end{bmatrix} \qquad (4\text{-}380)$$

Equation (4-350) means that

$$\mathrm{rk}\begin{bmatrix} \theta_i \\ \theta_i\Lambda \\ \cdot \\ \cdot \\ \cdot \\ \theta_i\Lambda^{k_e-k-1} \end{bmatrix} = k_e - k \tag{4-381}$$

and

$$\theta_i\Lambda^{k_e-k} = -\alpha_{e1}\theta_i - \alpha_{e2}\theta_i\Lambda - \dots - \alpha_{e,k_e-k}\,\theta_i\Lambda^{k_e-k-1} \tag{4-382}$$

for some set of scalars $\{\alpha_{e1}, \dots, \alpha_{e,k_e-k}\}$. Moreover, $(k_e - k)$ eigenvalues of Λ are given by the roots of the equation

$$s^{k_e-k} + \alpha_{e,k_e-k}\, s^{k_e-k-1} + \dots + \alpha_{e2}\, s + \alpha_{e1} = 0 \tag{4-383}$$

It will be shown that these $(k_e - k)$ eigenvalues are exactly the ones which are eliminated when f_i is removed from the original set.

Removal of f_i results in a new excess subspace of dimension k. The matrix Λ is replaced by the $k \times k$ matrix Λ' satisfying (4-360). By the development at the first of this section it is known that the remaining uncontrolled eigenvalues are the eigenvalues of Λ'. Now define a $k_e \times k_e$ coordinate transformation matrix

$$T_Y = [Y_e, T_{Y2}] \tag{4-384}$$

with Y_e given by (4-356) and T_{Y2} any $k_e \times k$ matrix which makes T_Y nonsingular.

Let

$$\overline{\Lambda} = T_Y^{-1} \Lambda T_Y \tag{4-385}$$

and

$$\theta_i = \theta_i T_Y \tag{4-386}$$

By (4-360)

$$\begin{aligned} \Lambda T_Y &= [\Lambda Y_e, \Lambda T_{Y2}]. \\ &= [Y_e \Lambda', \Lambda T_{Y2}] \\ &= [Y_e, T_{Y2}] \begin{bmatrix} \Lambda' & \overline{\Lambda}_{12} \\ \underline{0} & \overline{\Lambda}_{22} \end{bmatrix}. \end{aligned} \tag{4-387}$$

where

$$\Lambda T_{Y2} = Y_e \overline{\Lambda}_{12} + T_{Y2} \overline{\Lambda}_{22} \tag{4-388}$$

Premultiplying (4-387) by T_Y^{-1} and comparing the result with (4-385 gives

$$\overline{\Lambda} = \begin{bmatrix} \Lambda' & \overline{\Lambda}_{12} \\ \underline{0} & \overline{\Lambda}_{22} \end{bmatrix} \tag{4-389}$$

The eigenvalues of $\overline{\Lambda}$, and thus of Λ, are equal to the combined eigen- of Λ' and $\overline{\Lambda}_{22}$. The eigenvalues of Λ' remain after removal of f_i, so the eigenvalues which are eliminated are the eigenvalues of $\overline{\Lambda}_{22}$. It must now be shown that these eigenvalues are given by (4-383).

By the definition of Y_e and (4-353), $\theta_i \Lambda^j Y_e = \underline{0}$ for all $j \geq 0$, so

$$\bar{\theta}_i = \theta_i T_Y = [\theta_i Y_e, \theta_i T_{Y2}] = [\underline{0}, \bar{\theta}_{i2}] \tag{4-390}$$

where

$$\bar{\theta}_{i2} = \theta_i T_{Y2} \tag{4-391}$$

and also

$$\bar{\theta}_i \bar{\Lambda}^j = \theta_i \Lambda^j T_Y = [\underline{0}, \bar{\theta}_{i2} \bar{\Lambda}_{22}^j] \tag{4-392}$$

Then

$$\begin{bmatrix} \bar{\theta}_i \\ \bar{\theta}_i \bar{\Lambda} \\ \vdots \\ \bar{\theta}_i \bar{\Lambda}^{k_e-1} \end{bmatrix} = \begin{bmatrix} \theta_i \\ \theta_i \Lambda \\ \vdots \\ \theta_i \Lambda^{k_e-1} \end{bmatrix} T_Y = \begin{bmatrix} \underline{0}, \begin{pmatrix} \bar{\theta}_{i2} \\ \bar{\theta}_{i2} \bar{\Lambda}_{22} \\ \vdots \\ \bar{\theta}_{i2} \bar{\Lambda}_{22}^{k_e-1} \end{pmatrix} \end{bmatrix} \tag{4-393}$$

Since T_Y is nonsingular, this implies

$$\mathrm{rk} \begin{bmatrix} \bar{\theta}_{i2} \\ \bar{\theta}_{i2} \bar{\Lambda}_{22} \\ \vdots \\ \bar{\theta}_{i2} \bar{\Lambda}_{22}^{k_e-1} \end{bmatrix} = \mathrm{rk} \begin{bmatrix} \theta_i \\ \theta_i \Lambda \\ \vdots \\ \theta_i \Lambda^{k_e-1} \end{bmatrix} = k_e - k \tag{4-394}$$

by (4-380). Postmultiplying (4-382) by T_Y and using (4-392) yields

$$\bar{\theta}_{i2} \bar{\Lambda}_{22}^{k_e-k} = -\alpha_{e1} \bar{\theta}_{i2} - \cdots - \alpha_{e,k_e-k} \bar{\theta}_{i2} \bar{\Lambda}_{22}^{k_e-k-1} \tag{4-395}$$

Equations (4-394) and (4-395) prove that the eigenvalues of $\bar{\Lambda}_{22}$ are given by the roots of (4-383). This establishes the earlier claim that the eigenvalues given by (4-383) are eliminated by removing f_i from the set $\{f_1, \ldots, f_r\}$

From Λ one can determine the uncontrolled eigenvalues. If some of these are found to be undesirable, the θ_i will identify that vector (or vectors) whose removal will eliminate those particular eigenvalues. The following example illustrates the result of this and the previous sections.

Example E2:

Suppose

$$A = \begin{bmatrix} -3 & 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 3 & 0 \\ 1 & 1 & -1 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & -4 & -5 \\ 2 & 0 & 0 & 1 & 2 & 6 \\ 0 & 1 & 4 & 5 & 1 & 3 \end{bmatrix} \tag{E2-1}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{E2-2}$$

and there are four event vectors

$$f_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad f_2 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ -1 \\ 0 \\ 0 \end{bmatrix} \quad f_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -3 \\ 1 \\ 0 \end{bmatrix} \quad f_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{E2-3}$$

Since $Cf_i \neq \underline{0}$ for i = 1, 2, 3, 4 the matrix F defined by (4-232) is

$$F = [f_1, f_2, f_3, f_4] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & -3 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{E2-4}$$

Then

$$CF = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = I \tag{E2-5}$$

Now replace A by the simpler form

$$A'' = A - D''C = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 5 & 0 & 0 \end{bmatrix} \tag{E2-6}$$

which is obtained by taking the first, second, third, and fourth columns of D'' equal to the first, third, fifth, and sixth columns of A respectively. Using A'' to form K yields

$$K = A'' - A''F\,[(CF)^T CF]^{-1}(CF)^T C = A'' - A''FC$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 3 & 0 \\ 0 & 1 & 3 & 5 & 15 & 0 \end{bmatrix} \tag{E2-7}$$

For the full set of event vectors, C' defined by (4-255) becomes

$$C' = C - CF[(CF)^T CF]^{-1}(CF)^T C = C - C = \underline{0} \tag{E2-8}$$

and therefore M' defined by (4-249) is

$$M' = \underline{0} \tag{E2-9}$$

Hence, the group detection order of the set $\{f_1, f_2, f_3, f_4\}$ is six, the dimension of the state space. When the results of Section 4.3.1 are applied to each f_i, it will be found that the detection order is

$$\nu_i = 1 \quad \text{for } i = 1, 2, 3, 4 \tag{E2-10}$$

and each f_i is its own maximal generator. The sum of the individual detection orders is

$$\nu_1 + \nu_2 + \nu_3 + \nu_4 = 4 \tag{E2-11}$$

which means that the vectors $\{f_1, f_2, f_3, f_4\}$ are not mutually detectable and the excess subspace has dimension

$$k_e = 6 - 4 = 2 \tag{E2-12}$$

To determine if it is necessary or desirable to remove one or more vectors from the set, Λ and θ_i will be generated with the algorithm presented in Appendix B. Since $M' = \underline{0}$, the reduction procedure applied to the rows of this matrix produce no reductions. The terminating matrix which results from processing M' is simply the symmetric, positive-definite starting matrix. Let this matrix be the 6×6 identity matrix

$$\Omega_1 = I \tag{E2-13}$$

According to Appendix B the reduction procedure now starts with Ω_1 and is applied to the rows of the matrix $\breve{M}$ defined by (B-2).

Now $\breve{C}$ defined by (B-5) is

$$\breve{C} = [(CF)^T CF]^{-1}(CF)^T C = C \tag{E2-14}$$

Recalling that $\nu_i = 1$ for $i = 1, 2, 3, 4$, the $\breve{M}_1$ defined by (B-3) is simply

$$\breve{M}_1 = \begin{bmatrix} \breve{c}_1 \\ \breve{c}_2 \\ \breve{c}_3 \\ \breve{c}_4 \end{bmatrix} = \breve{C} = C \tag{E2-15}$$

and $\breve{M}_2$ defined by (B-4) is

$$\breve{M}_2 = \begin{bmatrix} \breve{C}K \\ \breve{C}K^2 \end{bmatrix} \tag{E2-16}$$

So

$$\breve{M} = \begin{bmatrix} \breve{M}_1 \\ \breve{M}_2 \end{bmatrix} = \begin{bmatrix} C \\ CK \\ CK^2 \end{bmatrix} \tag{E2-17}$$

The first reduction occurs at the first row in $\breve{M}_1$ (i.e., at $\breve{c}_1 = c_1$)

and

$$w_1 = \Omega_1 c_1^T = c_1^T = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{E2-18}$$

and

$$\Omega_2 = \Omega_1 - \frac{w_1 w_1^T}{c_1 w_1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{E2-19}$$

Reductions also occur at each of the next three rows, c_2, c_3, and c_4. The positive semi-definite matrix which results after these reductions is

$$\Omega_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{E2-20}$$

This completes the reduction process applied to $\breve{M}_1$. The first row of $\breve{M}_2$ is $c_1K = \underline{0}$, so $w_5 = \underline{0}$ and $\Omega_6 = \Omega_5$. No reduction occurs at this row, so c_1 is terminated. The second row of $\breve{M}_2$ is

$$c_2K = [0 \quad 1 \quad -2 \quad 0 \quad 0 \quad 0] \tag{E2-21}$$

Then

$$w_6 = \Omega_6(c_2K)^T = \Omega_5(c_2K)^T = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{E2-22}$$

and

$$\Omega_7 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{E2-23}$$

The third row of $\breve{M}_2$ is

$$c_3K = [0 \quad 0 \quad 1 \quad 1 \quad 3 \quad 0] \tag{E2-24}$$

Then

$$w_7 = \Omega_7(c_3K)^T = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \text{(E2-25)}$$

and

$$\Omega_8 = \underline{0} \qquad \text{(E2-26)}$$

so the reduction process is fully terminated. The two final nonzero auxiliary vectors needed to generate Λ and the θ_i are

$$\breve{w}_{f2} = w_6 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \breve{w}_{f3} = w_7 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \text{(E2-27)}$$

These two vectors occurred at rows $\breve{c}_2K = c_2K$ and $\breve{c}_3K = c_3K$ in $\breve{M}_2$. Therefore

$$\nu_2 + k_{e2} - 1 = 1 \qquad \text{(E2-28)}$$

and

$$k_{e2} = 2 - \nu_2 = 2 - 1 = 1 \qquad \text{(E2-29)}$$

also

$$k_{e3} = 2 - \nu_3 = 2 - 1 = 1 \qquad \text{(E2-30)}$$

$$k_{e1} = 1 - \nu_1 = 0 \qquad \text{(E2-31)}$$

$$k_{e4} = 1 - \nu_4 = 0 \qquad \text{(E2-32)}$$

Then from (B-28)

$$Z_e = [\breve{w}_{f2}, \breve{w}_{f3}] = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{(E2-33)}$$

From (B-32)

$$\theta_1 = \breve{c}_1 K^{\nu_1} Z_e = c_1 K Z_e = \underline{0} \qquad \text{(E2-34)}$$

since $c_1 K = \underline{0}$. Similarly

$$\theta_2 = c_2 K^{\nu_2} Z_e = c_2 K Z_e$$

$$= [0 \quad 1 \quad -2 \quad 0 \quad 0 \quad 0] \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = [1 \quad 0] \qquad \text{(E2-35)}$$

$$\theta_3 = c_3 K^{\nu_3} Z_e = c_3 K Z_e$$

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

(E2-36)

$$\theta_4 = c_4 K^{\nu_4} Z_e = c_4 K Z_e$$

$$= \begin{bmatrix} 0 & 1 & 3 & 5 & 15 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 5 \end{bmatrix}$$

(E2-37)

From (B-36)

$$Z_e \Lambda = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \Lambda = K Z_e - \sum_{i=1}^{4} \theta_i f_i$$

$$= \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 5 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 2 & 0 \\ 1 & 0 \\ -1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & -3 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 5 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 \\ -2 & 0 \\ 0 & 0 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{E2-38}$$

The first, third, fifth, and sixth rows of this vector equation are identically zero and may be discarded. The second and fourth rows yield

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \Lambda = \Lambda = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix} \tag{E2-39}$$

Note that the eigenvalues of Λ are $s = -2$ and $s = 3$. These are the uncontrolled eigenvalues which $(A - DC)$ will have if D is constrained to be a detector gain for all four vectors $\{f_1, f_2, f_3, f_4\}$. This Λ and the θ_i given by (E2-34) to (E2-37) yield the following conclusions:

1) Since $\theta_1 = \underline{0}$, removing f_1 from the set of event vectors will not reduce the excess subspace.

2) Since

$$\theta_2 \Lambda = [-2 \quad 0] = -2\theta_2 \tag{E2-40}$$

$$\text{rk} \begin{bmatrix} \theta_2 \\ \theta_2 \Lambda \end{bmatrix} = \text{rk} \begin{bmatrix} 1 & 0 \\ -2 & 0 \end{bmatrix} = 1 \tag{E2-41}$$

This means that removal of f_2 from the set will reduce the excess subspace by one dimension. The eigenvalue $s = -2$ will be eliminated and the uncontrolled eigenvalue which will remain for the set $\{f_1, f_3, f_4\}$ is $s = 3$.

3) Since

$$\text{rk} \begin{bmatrix} \theta_3 \\ \theta_3 \Lambda \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 3 \end{bmatrix} = 2 \tag{E2-42}$$

the removal of f_3 from the set will eliminate the excess subspace entirely. Therefore, the vectors $\{f_1, f_2, f_4\}$ are mutually detectable.

4) Since

$$\theta_4 \Lambda = [3 \quad 15] = 3\theta_4 \tag{E2-43}$$

$$\text{rk} \begin{bmatrix} \theta_4 \\ \theta_4 \Lambda \end{bmatrix} = \begin{bmatrix} 1 & 5 \\ 3 & 15 \end{bmatrix} = 1 \tag{E2-44}$$

This means that removal of f_4 from the set will reduce the excess subspace by one dimension. The eigenvalue s = 3 will be eliminated, and the uncontrolled eigenvalue which will remain for the set $\{f_1, f_2, f_3\}$ is s = -2.

5) From Corollary 4.6.2 it may be concluded that the following (nontrivial) subsets of vectors are mutually detectable:

(a) $\{f_1, f_2, f_4\}$

(b) Any subset of (a)

(c) $\{f_1, f_3\}$

Detection of all four event vectors requires a minimum of two detection filters. The set $\{f_1, f_2, f_3, f_4\}$ can be subdivided into two subsets of mutually detectable vectors. All the vectors in each such subset can be detected by one detection filter. The possible subdivisions are:

(i) $\{f_1, f_2, f_4\}$; $\{f_3\}$

(ii) $\{f_1, f_3\}$; $\{f_2, f_4\}$

Although the vectors $\{f_1, f_2. f_3\}$ are not mutually detectable, they can all be detected by a single stable detection filter, since the uncontrolled eigenvalue is s = -2. If this eigenvalue is acceptable, two additional subdivisions are possible:

(iii) $\{f_1, f_2, f_3\}$; $\{f_4\}$

(iv) $\{f_1, f_4\}$; $\{f_2, f_3\}$

In case (iii) the detection filter for $\{f_1, f_2, f_3\}$ will have the uncontrolled eigenvalue s = -2. In case (iv) the detection filter for $\{f_2, f_3\}$ will have the uncontrolled eigenvalue s = -2.

4.3.5 Effector Failure Information

The results of the previous four sections can be applied directly to the design of filters which detect effector failures. For the system described by (4-1) to (4-3), failure of the i^{th} effector is associated with b_i, the i^{th} column of B. This b_i replaces the f_i in the previous sections as the vector associated with a particular event. The design of the detection filter proceeds as follows:

1) For each column vector b_i in $B = [b_1, \ldots, b_r]$ determine the maximal generator with the algorithm of Appendix A. If two or more b_i have the same detection space, then only one of those vectors need be considered in the remaining steps. Any detection filter for one such vector will be a detection filter for all vectors having the same detection space.

2) Form F as defined by (4-242) with f_i replaced by b_i. If rk CF = r, the b_i are output separable. If rk CF < r, subdivide the b_i into two or more subsets so that each subset consists of output separable vectors.

3) Generate the θ_i and Λ for each of the subsets from step 2) using the algorithm of Appendix B. If a Λ exists (i.e., has nonzero dimension), identify the eigenvalues and decide if they are satisfactory. If not, use the results of Section 4.3.4 to subdivide that set further so that the undesirable eigenvalues are eliminated.

4) A detector gain for each subset of vectors from step 3) can be found by solving an equation of the form of (4-245) with the p_{ij} selected to give the desired eigenvalues. If the subset has fewer vectors than rk C, then the remaining eigenvalues of (A - DC)

are specified by choice of D'. Appendix A presents a comvenient method for doing this. The resulting detection filter has a state equation

$$\dot{z}(t) = (A - DC)\,z(t) + Bu_d(t) + Dy(t) \tag{4-396}$$

Suppose a failure as modeled by (4-15) occurs in the i^{th} effector. The detection filter for that effector will produce a settled-out output error of

$$\epsilon'(t) = C\epsilon(t) = CA^{\mu_i} b_i \int_{t_o}^{t} h_{bi}(t - \tau)\, n(\tau)\, d\tau \tag{4-397}$$

where μ_i is defined by condition (4-243) for b_i and

$$h_{bi}(t) = \mathcal{L}^{-1}\{H_{bi}(s)\} \tag{4-398}$$

with $H_{bi}(s)$ given by (4-240) for $f = b_i$. This result follows from remark 5) at the end of Section 4.3.1. The failure can then be identified by the fixed direction $(CA^{\mu_i} b_i)$ of the error signal.

If there are other detection filters, they will also produce error signals, but these errors will not lie in a fixed direction for arbitrary n(t) as the error given by (4-397) does. Note the qualification, "for arbitrary n(t)". For any filter there always exists a specific n(t) which can make the error lie in a fixed direction. An example which works for all stable filters is n(t) = constant. Even with the qualification there is still one possible exception to the above statement. A detection filter gain D designed for another set of vectors could, by coincidence, happen to be a detector gain for the b_i in (4-397). In that case this filter also would produce a fixed direction output error. However, no confusion should result, because to interpret the error signal from a detection filter one compares its direction with those directions for which the filter was designed. Even though the signal from another filter by coincidence lies in a fixed direction, that direction will not match any direction for which the filter was designed. This fact is assured by the following observation. If there is a detection filter designed for another vector b_j for which $CA^{\mu_j} b_j$ has the same direction as $CA^{\mu_i} b_i$, but b_i and b_j have different detection spaces, then the remarks at the end of Section 4.3.2 guarantee that the gain D for this second filter (for b_j) cannot be a detector gain for b_i. Therefore, the error signal from this filter (resulting from a failure of the i^{th} effector) will not lie in a fixed direction for arbitrary n(t).

If b_i and b_j have the same detection space, they would be assigned to the same detection filter by the procedure suggested in step 1). As mentioned in Section 4.3.2, events associated with such vectors cannot be differentiated on the basis of error direction alone. Error magnitude may provide additional information if something is known about the dynamic characteristics of such failures. If, for example, the n(t) for different events is expected to have different frequency spectra, then the frequency spectrum of the error magnitude may identify the most likely event. Chapter 5 discusses the problem of identifying effector failures from detection filter error signals when those signals are corrupted by errors caused by other simultaneous events or noise disturbances.

4.3.6 Plant Dynamics Information

For reasons discussed in Section 4.2.2, it will be convenient to model the plant dynamics given by (4-1) to (4-3) in a form for which all dynamics changes appear as changes in A or B, leaving C in fixed and simple form. Additional considerations will suggest a standard form for A as well. For the resulting plant description it will be especially simple to design a detection filter to detect dynamics changes. The detector gain can, in fact, be determined by inspection and the algorithms of Appendices A and B will be unnecessary for this situation.

The error equation for a change in the ij^{th} element of A is obtained as in the development of (4-33) using (4-41) and (4-42).

$$\dot{\epsilon}(t) = (A - DC)\,\epsilon(t) + \Delta a_{ij}\,\hat{e}_i\,x_j(t) \tag{4-399}$$

The detection filter for this event should be designed to detect the vector $\hat{e}_i$, in which case the settled-out output error is

$$\epsilon'(t) = C\epsilon(t) = \Delta a_{ij}\,CA^{\mu_i}\,\hat{e}_i \int_{t_o}^{t} h_i(t - \tau)\,x_j(\tau)\,d\tau \tag{4-400}$$

where μ_i is defined by condition (4-243) for $\hat{e}_i$ and

$$h_i(t) = \mathcal{L}^{-1}\{H_i(s)\} \tag{4-401}$$

with $H_i(s)$ given by (4-240) for $f = \hat{e}_i$. Note that the direction of the output error in (4-400) is the same for all j. A knowledge of the error magnitude factor

$$\phi_{ij}(t) = \int_{t_o}^{t} h_i(t - \tau)\, x_j(\tau)\, d\tau \tag{4-402}$$

is necessary to be able to decide which element in the i^{th} row of A has undergone a change. When the state vector is fully measurable, $x_j(t)$ can be determined directly from the sensor outputs (assuming noiseless sensors) as shown by (4-31). When the state vector is partially measurable only a part of it is so available. The remainder of the state vector must be reconstructed by a state-estimating filter. But when the model of the plant is inaccurate, as it will be if A or B undergo changes as assumed here, the state estimate will be unreliable even if there are no noise disturbances in the plant or sensors. This suggests the use of a standard form for A in which all the elements subject to change appear only in those columns, j, for which the corresponding state component, $x_j(t)$, can be determined directly from the sensor outputs. Such a standard form is

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mm} \end{bmatrix} \quad (n \times n) \tag{4-403}$$

where

$$A_{ii} = \begin{bmatrix} 0 & 0 & & 0 & a_{ii1} \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & a_{iin_i} \end{bmatrix} \quad (n_i \times n_i) \tag{4-404}$$

and for $i \neq j$

$$A_{ij} = \begin{bmatrix} 0 & \cdots & 0 & a_{ij1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & a_{ijn_i} \end{bmatrix} \quad (n_i \times n_j) \tag{4-405}$$

with

$$n_1 + \ldots + n_m = n \tag{4-406}$$

and

$$C = \begin{bmatrix} \hat{e}^T_{s_1} \\ \vdots \\ \hat{e}^T_{s_m} \end{bmatrix} \quad (m \times n) \tag{4-407}$$

where

$$s_i = n_1 + \ldots + n_i \tag{4-408}$$

(Note $s_1 = n_1$ and $s_m = n$.) The form of (4-407) implies that $\text{rk}\, C = m$. This point will be mentioned later. The process of producing this standard form for A and C is also discussed later in this section.

With A and C in the above form, plant dynamics changes appear as changes in the scalars $\{a_{ij\ell};\ i,j = 1, \ldots, m;\ \ell = 1, \ldots, n_i\}$ and the elements of B. The results of the previous sections can be applied to A and C given by (4-403) to (4-408) to design a detection filter for all $\hat{e}_i$, $i = 1, \ldots, n$. In this situation the maximal generators for the $\hat{e}_i$ have a simple form, and the equation for the detector gain can be solved by inspection. When the steps for designing a detection filter given in Section 4.3.5 are followed, the results below are easily established.

1) Taking advantage of the fact that A can be replaced by $A'' = A - D''C$ for arbitrary D'', as mentioned at the end of Section 4.3.1, let

$$D'' = \begin{bmatrix} d''_{11} & \cdots & d''_{1m} \\ \vdots & & \vdots \\ d''_{m1} & \cdots & d''_{mm} \end{bmatrix} \tag{4-409}$$

where

$$d''_{ij} = \begin{bmatrix} a_{ij1} \\ \vdots \\ a_{ijn_i} \end{bmatrix} \tag{4-410}$$

Then

$$A'' = \begin{bmatrix} A''_{11} & \cdots & A''_{1m} \\ \vdots & & \vdots \\ A''_{m1} & \cdots & A''_{mm} \end{bmatrix} \tag{4-411}$$

with

$$A''_{ii} = \begin{bmatrix} 0 & 0 & & 0 & 0 \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & 0 \end{bmatrix} \tag{4-412}$$

and

$$A''_{ij} = \underline{0} \qquad \text{for } i \neq j \tag{4-413}$$

2) The detection order of $\hat{e}_i$ is n_{j+1} where $s_j < i \leq s_{j+1}$ (s_j given by (4-408)), and its maximal generator is $\hat{e}_{s_j+1}$. (For $0 < i \leq n_1$ the detection order of $\hat{e}_i$ is n_1 and the maximal generator is $\hat{e}_1$.) This means that all $\hat{e}_i$ for which $s_j < i \leq s_{j+1}$ have the same maximal generator and detection space. By the remark in step 1) of Section 4.3.5, only one of these $\hat{e}_i$ need be considered. Then let $\hat{e}_{s_{j+1}}$ be retained as the representative of all $\hat{e}_i$ for $s_j < i \leq s_{j+1}$. The set of vectors remaining is then $\{\hat{e}_{s_1}, \ldots, \hat{e}_{s_m}\}$.

3) All vectors in the set $\{\hat{e}_{s_1}, \ldots, \hat{e}_{s_m}\}$ are output separable and mutually detectable. The F for this set is

$$F = [\hat{e}_{s_1}, \ldots, \hat{e}_{s_m}] = C^T \tag{4-414}$$

so

$$CF = CC^T = I \tag{4-415}$$

Then Equation (4-245) for D can be solved by inspection

$$\text{DCF} = D = \begin{bmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & & \vdots \\ d_{m1} & \cdots & d_{mm} \end{bmatrix} \tag{4-416}$$

where

$$d_{ii} = \begin{bmatrix} a_{ii1} + p_{i1} \\ \vdots \\ a_{iin_i} + p_{in_i} \end{bmatrix} \tag{4-417}$$

and $d_{ij} = d''_{ij}$ given by (4-410) for $i \neq j$. Then

$$A - DC = \begin{bmatrix} P_1 & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \ddots & & \vdots \\ \vdots & & \ddots & \underline{0} \\ \underline{0} & \cdots & \underline{0} & P_m \end{bmatrix} \tag{4-418}$$

with

$$P_i = \begin{bmatrix} 0 & 0 & & 0 & -p_{i1} \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & -p_{in_i} \end{bmatrix} \tag{4-419}$$

This filter is a detection filter for all the coordinate directions $\hat{e}_i$, $i = 1, \ldots, n$. A change $\Delta a_{ij\ell}$ in one element $a_{ij\ell}$ of A given by (4-403) to (4-405) produces a settled-out error of

$$\epsilon'(t) = C\epsilon(t) = \Delta a_{ij\ell}\, \hat{e}_{mi} \int_{t_o}^{t} h_{i\ell}(t-\tau)\, x_{s_j}(\tau)\, d\tau \qquad (4\text{-}420)$$

where

$$h_{i\ell}(t) = \mathcal{L}^{-1}\left\{ \frac{s^{\ell-1}}{s^{n_i} + p_{in_i} s^{n_i - 1} + \ldots + p_{i2}\, s + p_{i1}} \right\} \qquad (4\text{-}421)$$

$x_{s_j}(t)$ is the s_j^{th} component of the state vector $x(t)$, and $\hat{e}_{mi}$ is a unit m-vector in the i^{th} coordinate direction. From the form of C in (4-407)

$$x_{s_j}(t) = y_j(t) \qquad (4\text{-}422)$$

where $y_j(t)$ is the j^{th} component of the sensor output vector. Then (4-420) can be written

$$\epsilon'(t) = \Delta a_{ij\ell}\, \phi_{ij\ell}(t)\, \hat{e}_{mi} \qquad (4\text{-}423)$$

where

$$\phi_{ij\ell}(t) = \int_{t_o}^{t} h_{i\ell}(t-\tau)\, y_j(\tau)\, d\tau \qquad (4\text{-}424)$$

The p_{ij} in (4-421) are at the discretion of the designer and are known. Since $y_j(t)$ is an accessible signal, the scalar function $\phi_{ij\ell}(t)$ can be generated on-line from sensor output without knowledge of the plant dynamics. For consistency of notation the B matrix can be partitioned to conform with A

$$B = \begin{bmatrix} b_{11} & \cdot \; \cdot \; \cdot \; \cdot & b_{1r} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ b_{m1} & \cdot \; \cdot \; \cdot \; \cdot & b_{mr} \end{bmatrix} (n \times r) \qquad (4\text{-}425)$$

where

$$b_{ij} = \begin{bmatrix} b_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ b_{ijn_i} \end{bmatrix} (n_i \times 1) \qquad (4\text{-}426)$$

A change $b_{ij\ell}$ in $b_{ij\ell}$ produces a settled-out error signal of

$$\epsilon'(t) = C\epsilon(t) = \Delta b_{ij\ell} \, \psi_{ij\ell}(t) \, \hat{e}_{mi} \qquad (4\text{-}427)$$

with

$$\psi_{ij\ell}(t) = \int_{t_o}^{t} h_{i\ell}(t - \tau) \, u_{dj}(\tau) \, d\tau \qquad (4\text{-}428)$$

where $u_{dj}(t)$ is the j^{th} component of $u_d(t)$ and $h_{i\ell}(t)$ is given by (4-421). As in the case of $\phi_{ij\ell}(t)$, $\psi_{ij\ell}(t)$ can be generated on-line from accessible signals $(u_d(t))$ without knowledge of the plant dynamics.

It has been shown that (4-403) to (4-408) are especially convenient forms for A and C. In Section 4.2.2 it was demonstrated that all plant descriptions which are related by a state space coordinate transformation can be considered equivalent. Unfortunately it is not always possible, in general, to put A and C into the form of (4-403) to (4-408) by a coordinate transformation. However, it can be shown that these standard forms can always be obtained by augmenting (enlarging) the state space. Appendix C presents a way of constructing a coordinate transformation which puts A and C into the form of (4-403) to (4-408) except that the off-diagonal blocks of A in general have the form

$$A_{ji} = \begin{bmatrix} 0 & \cdots & 0 & a_{ji1} \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & a_{jin_i} \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix} \quad (n_j \times n_i) \tag{4-429}$$

and

$$A_{ij} = \begin{bmatrix} 0 & \cdots & 0 & 0 & a_{ij1} \\ \vdots & & \vdots & \vdots & \vdots \\ \vdots & & \vdots & 0 & \vdots \\ 0 & \cdots & 0 & a^*_{ij} & a_{ijn_i} \end{bmatrix} \quad (n_i \times n_j) \tag{4-430}$$

where $n_j > n_i$. If $n_j = n_i$ $(i \neq j)$ then A_{ij} and A_{ji} have the form of (4-405). The appearance of the nonzero element a^*_{ij} in (4-430) violates the form of (4-405). For a general A and C, (4-430) is as close as one can get to the form of (4-405) by a coordinate transformation which does not change the dimension of the state space. To explain the appearance of the elements a^*_{ij} and determine how they may be eliminated (made zero) by enlarging the state space, it will be convenient to introduce the concept of output decoupling.

Definition 4.11. The matrix pair (A, C) is defined to be output decouplable if A and C can be put into the forms of (4-403) to (4-408) by a state space coordinate transformation.

This terminology is motivated by the fact that with proper choice of D the observable spaces of the c_i (i^{th} row of C) with respect to (A - DC) can all be made nonintersecting (which is, in a sense, output decoupled). The (A - DC) given by (4-418) is an example. Note that this definition implies that an output decouplable pair is also observable and rk C = m. The definition could be generalized to include nonobservable pairs, but that is unnecessary for purposes of plant dynamics identification. This point is discussed later.

Definition 4.12. Consider the pair (A, C), and let c_i be the i^{th} row of C. The output decoupling order (or simply, decoupling order) of c_i is defined to be the largest integer value of j such that

$$\operatorname{rk}\begin{bmatrix} M_{T,j-1} \\ c_i A^{j-1} \end{bmatrix} = \operatorname{rk} M_{T,j-1} + 1 \tag{4-431}$$

where

$$M_{T,j-1} = \begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{j-2} \end{bmatrix} \tag{4-432}$$

(For j = 1, M_{T0} is taken as the zero matrix.)

An equivalent definition is the smallest positive integer value of j such that

$$\mathrm{rk}\begin{bmatrix} M_{Tj} \\ c_i A^j \end{bmatrix} = \mathrm{rk}\, M_{Tj} \tag{4-433}$$

It can be shown from (4-431) or (4-433) that decoupling order is invariant with respect to coordinate transformations of the state space. Note that for A and C in (4-403) to (4-408) the decoupling order of each c_i is n_i and $n_1 + \ldots + n_m = n$. From the algorithm used to obtain the form of (4-429) and (4-430) it can be verified that the decoupling order of each c_i is greater than or equal to n_i, and the equality holds if and only if $a^*_{ij} = 0$ for all $j = 1, \ldots, m$. These observations establish the following theorem.

Theorem 4.7. The pair (A, C), with A of dimension $n \times n$ and C of dimension $m \times n$, is output decouplable if and only if $q_1 + \ldots + q_m = n$ where q_i is the decoupling order of c_i, the i^{th} row of C. If this is the case, then $n_i = q_i$ for the standard forms (4-403) to (4-408).

Output decoupling order has an interesting and useful relationship to detection order which is stated in the following theorem.

Theorem 4.8. If f is any n-vector for which $c_i f \neq \underline{0}$ (or $c_i A^{\mu} f \neq \underline{0}$ in the case of (4-108)), then the detection order of f cannot exceed the decoupling order of c_i.

Proof:

Let ν be the detection order of f. Then f has a maximal generator g which satisfies

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{\nu-2} \\ CA^{\nu-1} \end{bmatrix} g = \begin{bmatrix} \underline{0} \\ \underline{0} \\ \cdot \\ \cdot \\ \cdot \\ \underline{0} \\ Cf \end{bmatrix} \tag{4-434}$$

But $c_i A^{\nu-1} g = c_i f \neq \underline{0}$, which means that $c_i A^{\nu-1}$ must be independent of the rows of

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{\nu-2} \end{bmatrix}$$

This implies that (4-431) is satisfied for $j = \nu$. Therefore, ν must be less than or equal to the decoupling order of c_i, since that is the largest integer satisfying (4-431). This completes the proof.

It is easy to show that there always exists a vector which has a detection order equal to the decoupling order of c_i. If q_i is the decoupling order of c_i, condition (4-431) implies that there must exist some vector f such that $M_{T, q_i - 1} f = \underline{0}$ and $c_i A^{q_i - 1} f \neq \underline{0}$. The detection order of this f must be at least q_i because f is a q_i^{th} order detection generator for itself. On the other hand, Theorem 4.8 shows that the detection order of f cannot exceed q_i. The only consistent conclusion is that the detection order of f is equal to q_i. The fact that such an f exists shows that decoupling order has the same invariance properties as detection order. Specifically, decoupling order is invariant with respect to replacement of A by $(A - D''C)$ for any D''.

The possibility of obtaining an output decouplable pair by augmenting the state space will now be investigated. A plan description given by (4-1) to (4-3) is represented by the matrix triplet (A, B, C). Referring back to Equations (4-24) and (4-25), from which the notion of equivalent plant descriptions was developed, it can be seen that the property which makes two descriptions, (A, B, C) and $(\tilde{A}, \tilde{B}, \tilde{C})$, equivalent is that

$$C e^{A(t - t_o)} B = \tilde{C} e^{\tilde{A}(t - t_o)} B \tag{4-435}$$

for all t. When this condition is satisfied, both (A, B, C) and $(\tilde{A}, \tilde{B}, \tilde{C})$ have the same dynamic transfer from $u_d(t)$ to $y(t)$, i.e., starting from zero initial conditions, $u_d(t)$ elicits the same output $y(t)$ from both descriptions. In Section 4.2.2 only coordinate transformations were considered, for which A and $\tilde{A}$ have the same dimensions. However, (4-435) can also be satisfied for A and $\tilde{A}$ of different dimensions. Using the terminology of Brockett [4], a representation (A, B, C) of the plant dynamics with the smallest possible state space dimension (i.e., smallest n where A is $n \times n$) will be referred to as a minimal representation. Any equivalent representation $(\tilde{A}, \tilde{B}, \tilde{C})$ (i.e., satisfying (4-435)) having a larger state is considered nonminimal. Brockett shows that a minimal representation is both controllable and observable. If $(\tilde{A}, \tilde{B}, \tilde{C})$ is nonminimal it can be controllable or observable, but not both.

It must now be shown that it is possible to obtain a decouplable representation of the plant by allowing augmentations which preserve the equivalence property (4-435). The following theorem places a lower bound on the dimension of the state space which is necessary for an equivalent, decouplable representation.

Theorem 4.9. If (A, B, C) is a minimal representation and $(\tilde{A}, \tilde{B}, \tilde{C})$ is any other equivalent representation, then the decoupling order of the i^{th} row of $\tilde{C}$ cannot be less than the decoupling order of the i^{th} row of C.

Proof:

Both matrix exponentials in (4-435) can be expanded in an infinite series of the form (2-16). Since (4-435) must be satisfied for all t, the series expansions must be equal term by term. Equation (4-435) is therefore equivalent to

$$CA^jB = \tilde{C}\tilde{A}^j\tilde{B} \quad \text{for all } j \geq 0 \tag{4-436}$$

This implies that

$$\begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^j \end{bmatrix} [B,\ AB,\ \ldots,\ A^{n-1}B] = \begin{bmatrix} \tilde{C} \\ \tilde{C}\tilde{A} \\ \cdot \\ \cdot \\ \cdot \\ \tilde{C}\tilde{A}^j \end{bmatrix} [\tilde{B},\ \tilde{A}\tilde{B},\ \ldots,\ \tilde{A}^{n-1}\tilde{B}] \tag{4-437}$$

for all $j \geq 0$.

Define

$$W = [B, AB, \ldots, A^{n-1}B] \tag{4-438}$$

$$\tilde{W} = [\tilde{B}, \tilde{A}\tilde{B}, \ldots, \tilde{A}^{n-1}\tilde{B}] \tag{4-439}$$

$$M_{Tj} = \begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{j-1} \end{bmatrix} \tag{4-440}$$

$$\tilde{M}_{Tj} = \begin{bmatrix} \tilde{C} \\ \tilde{C}\tilde{A} \\ \cdot \\ \cdot \\ \cdot \\ \tilde{C}\tilde{A}^{j-1} \end{bmatrix} \tag{4-441}$$

Let c_i be the i^{th} row of C, and $\tilde{c}_i$ the i^{th} row of $\tilde{C}$. Also let q_i be the decoupling order of c_i. Suppose the decoupling order of $\tilde{c}_i$ is less than q_i. Then (4-433) implies that $\tilde{c}_i\tilde{A}^{q_i-1}$ can be expressed as a linear combination of the rows of $\tilde{M}_{T,q_i-1}$, that is

$$\tilde{c}_i\tilde{A}^{q_i-1} = \tilde{\gamma}\tilde{M}_{T,q_i-1} \tag{4-442}$$

for some $1 \times m \cdot (q_i - 1)$ row vector $\tilde{\gamma}$. Now (4-437) implies that

$$c_iA^{q_i-1} W = \tilde{c}_i\tilde{A}^{q_i-1} \tilde{W} \tag{4-443}$$

Since (A, B, C) is minimal, (A, B) is a controllable pair and rk W = n. Therefore, (4-443) can be solved uniquely for $c_iA^{q_i-1}$

$$c_i A^{q_i - 1} = \tilde{c}_i \tilde{A}^{q_i - 1} \tilde{W} W^T [W W^T]^{-1} \quad (4\text{-}444)$$

and similarly (4-437) with $j = q_i - 2$ yields

$$M_{T, q_i - 1} = \tilde{M}_{T, q_i - 1} \tilde{W} W^T [W W^T]^{-1} \quad (4\text{-}445)$$

Substituting (4-442) and (4-445) into (4-444) gives

$$\begin{aligned} c_i A^{q_i - 1} &= \tilde{\gamma} \tilde{M}_{T, q_i - 1} \tilde{W} W^T [W W^T]^{-1} \\ &= \tilde{\gamma} M_{T, q_i - 1} \end{aligned} \quad (4\text{-}446)$$

But this contradicts the fact that the decoupling order of c_i is q_i. Therefore, the decoupling order of $\tilde{c}_i$ cannot be less than q_i. This completes the proof.

By this theorem the decoupling order of any row of C cannot be decreased when the state space is made larger than the minimal one. Therefore, to obtain a decouplable representation (if the minimal one is not decouplable) the state space must be enlarged to a dimension of at least $(q_i + \ldots + q_m)$, where q_i is the decoupling order of the i^{th} row of C in a minimal representation. Appendix C demonstrates that this lower bound is, in fact, reachable. It presents a way of augmenting a representation to obtain an equivalent decouplable representation with dimension $(q_1 + \ldots + q_m)$.

To reiterate, a plant representation in the form of (4-403) to (4-408) was shown to be desirable for the detection of changes in plant dynamics. The extended development on output decoupling and augmented representations was necessary because it is essential to be aware of the assumptions tacitly made about the plant when it is represented in the form of (4-403) to (4-408). Specifically the assumptions are as follows:

(1) The plant is observable.

(2) The output decoupling order of the i^{th} sensor (i.e., the decoupling order of c_i in the minimal representation) does not exceed n_i.

The first assumption is entirely reasonable when dealing with the identification of plant dynamics from sensor outputs. It was noted in Chapter 2 that the unobservable portion of the dynamics cannot be determined from the output (and input). It does not make sense, then, to model the plant with an unobservable representation when the unobservable portion cannot be identified. The second assumption places a restriction on the kind of dynamics changes which the standard form model can handle. To be specific, the plant dynamics should not change in such a way that the decoupling order of the i^{th} sensor exceeds n_i. If this happens (4-403) to (4-408) cannot be a valid model (i.e., an equivalent representation) of the plant for any values of the elements $a_{ij\ell}$. This means that the less prior knowledge one has about the possible plant dynamics changes, the larger the model will have to be to guarantee a valid representation. Suppose, for example, it is known that the decoupling orders of the sensors will remain fixed at known values (n_i for the i^{th} sensor). Then the plant can be safely modeled by a representation of the form (4-403) to (4-408) with a state space of dimension $(n_1 + \ldots + n_m)$. If the decoupling orders of the sensors do not necessarily remain fixed, but an upper bound $\bar{n}_i$ is known for each sensor, then the plant can be modeled in the form of (4-403) to (4-408) with a state space of dimension $(\bar{n}_1 + \ldots + \bar{n}_m)$. If the dimension of the minimal plant representation is known to be fixed at (or at least does not exceed) n, and it is further known that the sensors all remain independent (i.e., that rk C = m in the minimal representation), then an upper bound on the decoupling order of any sensor is $(n - m + 1)$. In this case the plant can be modeled with a state space dimension of $m \cdot (n - m + 1)$. It is interesting to note that this number attains a maximum value for m near $\frac{n}{2}$ and approaches n as m approaches 1 or n. Finally, if it is known only that for the minimal representation rk C is at least k and the dimension of the state space does not exceed n, then the upper bound on the decoupling order of any sensor is $(n - k + 1)$. In this case a model with an $[m \cdot (n - k + 1)]$ dimensional state space will always be valid.

The standard form of (4-403) to (4-408) can be interpreted in a different way which may have more physical meaning in many cases. The state space description of the plant given by (4-1) to (4-3) is equivalent to a set of m linear, coupled, scalar differential equations relating the output variables $\{y_i(t) \; ; \; i = 1, \ldots, m\}$ to the input variables $\{u_{dj}(t) ; \; j = 1, \ldots, r\}$. In Chapter 5 this set of differential equations is developed for the case in which A and C are in the form of (4-403) to (4-408) (Equations (5-52) to (5-55)). From these equations it can be seen that each row of blocks of A in (4-403) corresponds to one differential equation. For example, the blocks $\{A_{i1}, \ldots, A_{im}\}$ (and the corresponding row of block is B) are associated with a differential equation for the output component $y_i(t)$. This differential equation is of order n_i(where A_{ii} is $n_i \times n_i$). The highest derivative of $y_i(t)$ in this equation is n_i. The significant feature of this equation is that the highest derivative of any other variables ($y_j(t)$ for $j \neq i$ and $u_{d\ell}(t)$ for all $\ell = 1, \ldots, r$) is less than n_i. In other words, the driving terms, involving $u_{d\ell}(t)$ for $\ell = 1, \ldots, r$, and the cross-coupling terms, involving $y_j(t)$ for all $j \neq i$, all have lower order derivatives than the highest order derivative of $y_i(t)$, which is

$$\frac{d^{n_i}}{dt^{n_i}} \; y_i(t)$$

If the plant dynamics can be described by a set of input-output equations having this property, then the state space description can be put into the form of (4-403) to (4-408), and vice versa. The meaning of the general form of (4-430) is that if some $a^*_{ij} \neq \underline{0}$ then there exists a cross-coupling term involving

$$\frac{d^{n_i}}{dt^{n_i}} \; y_j(t)$$

whose order is equal to the highest derivative of $y_i(t)$.

In closing this section, some final observations should be made.

1) Although it was not proven, it will be found that if the form of (4-430) with $a_{ij}^* \neq 0$ is used for a plant model then, in addition to the objections already noted, more than one detection filter may be necessary to detect all of the coordinate directions. This happens because the presence of a nonzero a_{ij}^* makes certain nonseparable coordinate directions have nonidentical detection spaces. This results in uncontrolled eigenvalues which must then be investigated for satisfactory filter dynamics.

2) The form of C in (4-407) implies rk C = m where m is the number of sensors. It may happen that in the minimal representation for the plant rk C < m. Appendix C considers this possibility, and in any case the $\tilde{C}$ in the augmented representation will have full rank m.

3) Because of the form of $h_{i\ell}(t)$ in (4-421), the $\phi_{ij\ell}(t)$ for $\ell = 1, \ldots, n_i$ in (4-424) are the components of the state vector for the n_i-dimensional system

$$\dot{\underline{\phi}}_{ij}(t) = P_i^T \underline{\phi}_{ij}(t) + \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} y_j(t) \qquad (4\text{-}447)$$

with

$$\underline{\phi}_{ij}(t) = \begin{bmatrix} \phi_{ij1}(t) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \phi_{ijn_i}(t) \end{bmatrix} \qquad (4\text{-}448)$$

where P_i is given by (4-419). Similarly the $\psi_{ij\ell}(t)$ in (4-428) are the components of the state vector for

$$\dot{\underline{\psi}}_{ij}(t) = P_i^T \underline{\psi}_{ij}(t) + \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} u_{dj}(t) \tag{4-449}$$

with

$$\underline{\psi}_{ij}(t) = \begin{bmatrix} \psi_{ij_1}(t) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \psi_{ijn_i}(t) \end{bmatrix} \tag{4-450}$$

Chapter 5 discusses several methods for processing the error signals given by (4-423) and (4-427) to determine $\Delta a_{ij\ell}$ and $\Delta b_{ij\ell}$.

4.3.7 Sensor Failure Information

In Section 4.2.3 it was found that the best information a detection filter could provide about the sensor failures was an error signal constrained to a two-dimensional plane. It will be shown in this section that this can also be achieved in the case of a partially measurable state vector.

When the i^{th} sensor of the plant given by (4-1) to (4-3) suffers a failure as described by (4-55) the equation for the state error can be obtained from (4-56)

$$\dot{\epsilon}(t) = (A - DC)\,\epsilon(t) + d_i\, n(t) \tag{4-451}$$

where d_i is the i^{th} column of D.

$$d_i = D\hat{e}_{mi} \qquad (4\text{-}452)$$

The accessible output error is defined by (4-72) as

$$\epsilon'(t) = y(t) - Cz(t) = C\epsilon(t) + \hat{e}_{mi}\, n(t) \qquad (4\text{-}453)$$

Theorem 4.1 is not directly applicable to (4-451) because d_i corresponding to f is not fixed, but depends on the detector gain D which is under the control of the designer. Therefore, some additional results are necessary to show that a detector gain does exist which will constrain the output error to a plane. In previous sections an event has been associated with the drive term of the state error equation; for example, f in Equation (4-73). It is not satisfactory to associate a sensor failure with d_i, however, because this vector can be changed at will and has no inherent relationship to the sensor. For this reason failure of the i^{th} sensor will be associated with c_i, the i^{th} row of C, and detectability of this event will be defined accordingly.

<u>Definition 4.13.</u> The i^{th} row of C, $c_i = \hat{e}_{mi}^T C$, is defined to be <u>sensor detectable</u> if there exists a matrix D such that

(1) $\epsilon'(t)$ is constrained to lie in a two-dimensional plane in the output space, where $\epsilon'(t)$ is given by (4-453) and $\epsilon(t)$ is the settled-out solution of (4-451) with n(t) an arbitrary scalar time function, and

(2) at the same time, all eigenvalues of (A - DC) can be specified almost arbitrarily.

The following theorem provides sufficient conditions for sensor detectability. Its proof will lead to the design procedure for a sensor failure detection filter.

Theorem 4.10. If (A, C) is an observable pair and c_i, the i^{th} row of C, is linearly independent of all the other rows in C, Then c_i is sensor detectable.

Proof:

Let f be an n-vector satisfying

$$Cf = \hat{e}_{mi} \tag{4-454}$$

Note that a necessary and sufficient condition for the existence of such an f is that c_i be linearly independent of all the other rows of C. By Theorem 4.1, f is detectable. Let ν be the detection order of f, and g its maximal generator. First choose D to be a detector gain for f by constraining it to be a solution of (4-113), or equivalently (4-119). Then as shown in Section 4.3.1, $A - DC = A' - D'C'$ where A' and C' are given by (4-133) and (4-134), and D' is arbitrary. With (4-454), Equation (4-119) for D reduces to

$$DCf = D\hat{e}_{mi} = d_i = p_1 g + \ldots + p_\nu A^{\nu-1} g + A^\nu g \tag{4-455}$$

or using (4-168)

$$d_i = z_d + Af \tag{4-456}$$

where z_d is given by (4-170).

The purpose of making D a detector gain for f is that d_i has been fixed, as shown by (4-456). The sensor failure detection filter can now be obtained by making D' a detector gain for d_i. Note carefully, however, that in determining this second detector gain one must start with the matrix pair (A', C') instead of (A, C). In applying the results of Section 4.3.1, A and C must be replaced by A' and C'. The only additional consideration necessary is the fact that (A', C') is not an observable pair, since

$$\mathrm{rk} \begin{bmatrix} C' \\ C'A' \\ \cdot \\ \cdot \\ \cdot \\ C'A'^{n-1} \end{bmatrix} = n - \nu \qquad (4\text{-}457)$$

It was shown at the end of Section 4.3.1 that even for a nonobservable pair a detector gain can be found for any vector which does not lie in the unobservable space. Assume first that d_i does not lie in the unobservable space of C' with respect to A'. Then it is possible to find a D' which is a detector gain for d_i (with respect to (A', C')), and at the same time specify almost arbitrarily $(n - \nu)$ eigenvalues of $A' - D'C' = A - DC$. The remaining ν eigenvalues are associated with the unobservable space of C' (the detection space of f) and have already been specified by constraining D to be a solution of (4-455). Therefore, all the eigenvalues of $(A - DC)$ can be almost arbitrarily specified.

It must now be verified that the output error given by (4-453) will be constrained to lie in a plane. With D' selected to be a detector gain for d_i with respect to (A', C'), it is known that $C'\epsilon(t)$ must lie in a fixed direction, where $\epsilon(t)$ is the settled-out solution of

$$\dot{\epsilon}(t) = (A' - D'C')\,\epsilon(t) + d_i\,n(t)$$

$$= (A - DC)\,\epsilon(t) + d_i\,n(t) \tag{4-458}$$

Let the fixed direction be represented by an m-vector y_d. Then $C'\epsilon(t)$ can be expressed as

$$C'\epsilon(t) = y_d\,\phi_d(t) \tag{4-459}$$

where $\phi_d(t)$ is a scalar function depending on n(t). Now from (4-134)

$$C' = C - Cf[(Cf)^T Cf]^{-1}(Cf)^T C$$

$$= C - \hat{e}_{mi}\,\hat{e}_{mi}^T\,C = C - \hat{e}_{mi}\,c_i \tag{4-460}$$

where $c_i = \hat{e}_{mi}^T C$ is the i^{th} row of C. (Note that C' is simply C with the i^{th} row set to zero.) Then

$$C\epsilon(t) = C'\epsilon(t) + \hat{e}_{mi}\,c_i\,\epsilon(t)$$

$$= y_d\,\phi_d(t) + \hat{e}_{mi}\,c_i\,\epsilon(t) \tag{4-461}$$

and the output error is

$$\epsilon'(t) = C\epsilon(t) + \hat{e}_{mi}\, n(t)$$

$$= y_d \phi_d(t) + \hat{e}_{mi}\Big(n(t) + c_i \epsilon(t)\Big) \qquad (4\text{-}462)$$

Since $\Big(n(t) + c_i \epsilon(t)\Big)$ is a scalar function, it is clear that $\epsilon'(t)$ lies in the two-dimensional plane formed by y_d and $\hat{e}_{mi}$.

In obtaining this result it was assumed that d_i did not lie in the unobservable space of C'. Suppose now that d_i does lie in this space. Then

$$\begin{bmatrix} C' \\ C'A' \\ \cdot \\ \cdot \\ \cdot \\ C'A'^{n-1} \end{bmatrix} d_i = \underline{0} \qquad (4\text{-}463)$$

By (4-182) and Definition 4.5 this means that d_i lies in the detection space of f. This, in turn, means that D satisfying (4-455) is a detector gain for d_i as well as f. In this case the second step of making D' a detector gain for d_i is unnecessary, and one can immediately conclude that $C\epsilon(t)$ lies in a fixed direction. If this direction is represented by y_d, then $\epsilon'(t)$ lies in the two-dimensional plane formed by y_d and $\hat{e}_{mi}$. The choice of D' is unconstrained and can be selected to arbitrarily specify $(n - \nu)$ eigenvalues of (A - DC). As before, the remaining ν eigenvalues are specified by choice of the coefficients in (4-455). This completes the formal proof of the theorem.

This proof shows in a general way how to proceed in designing a detection filter for sensor failures. Some additional material will now be presented which is of significant value in developing practical design procedures for these detection filters. In remark 4) at the end of Section 4.3.1, a coordinate transformation was used to demonstrate how a detector gain could be found for a nonobservable pair. In effect the problem was transformed so that the unobservable part of the state space was eliminated from consideration, and the results of Section 4.3.1 could be applied to a subspace which was observable — specifically the observable pair $(\overline{A}_{11}, \overline{C}_1)$. In practice it is neither necessary nor desirable to actually perform a coordinate transformation to find a detector gain D'. The same result can be achieved with the notion of vector equivalence classes. A complete formal development of this concept can be found in [7]. Only a brief introduction will be given here.

Denote the unobservable space of C' with respect to A' by E. Two vectors x_1 and x_2 in the state space are defined to be <u>equivalent modulo E</u> (denoted $x_1 \equiv x_2 \pmod{E}$) if their difference lies in E. The set of all equivalent vectors forms an equivalence class. The equivalence classes themselves can then be considered members of a new vector space replacing the original state space. Because E is an invariant subspace with respect to A', it can be shown that A' is a linear operator in the vector space of equivalence classes (mod E). Also, C' can be viewed as a linear operator from the space of equivalence classes into the ordinary m-vector output space. All the results of Section 4.3.1 can then be applied to this new state space (with A and C replaced by A' and C'). The end result is that all vector equations in the state space (i.e., vector equations with n rows) remain valid except that "=" is replaced by "$\equiv$ (mod E)". All other equations (for example, (4-80) and (4-91) retain the true equality sign. There is one exception to this rule. An equation in the state space retains the true equality sign if it is derived entirely from equations in which true equality holds. An example is (4-105) which is derived from (4-80).

Let ν' be the detection order of d_i with respect to (A', C') and g' its maximal generator (mod E). In this situation the maximal generator (mod E) is not unique because any vector equivalent to g' is also a maximal generator. The uniqueness assertion of Theorem 4.4 applies to the equivalence class of maximal generators rather than a specific n-vector. The algorithm of Appendix A for finding a maximal generator is applicable to nonobservable pairs, so it can be used to generate a g'. Specific note is made of the nonobservable case in the appendix. The equation for D' corresponding to (4-113) is

$$D'C'A'^{\nu'-1} g' \equiv p_1' g' + p_2' A' g' + \ldots + p_{\nu'}' A'^{\nu'-1} g' + A'^{\nu'} g' \text{ (mod E)} \tag{4-464}$$

This is equivalent to the equation

$$D'C'A'^{\nu'-1} g' = p_1' g' + \ldots + p_{\nu'}' A'^{\nu'-1} g' + A'^{\nu'} g' + z_E \tag{4-465}$$

where z_E is any vector in E. The coefficients p_i' and the vector z_E can be arbitrarily specified by the designer except that z_E must lie in E. A simple choice for z_E is $\underline{0}$.

When a D' satisfying (4-465) is used to form $(A' - D'C') = (A - DC)$, this matrix will have ν eigenvalues given by the roots of

$$s^{\nu} + p_{\nu}s^{\nu-1} + \ldots + p_2 s + p_1 = 0 \qquad (4\text{-}466)$$

and ν' eigenvalues given by the roots of

$$s^{\nu'} + p'_{\nu'}s^{\nu'-1} + \ldots + p'_2 s + p'_1 = 0 \qquad (4\text{-}467)$$

This fact can be verified by introducing the coordinate transformation

$$\overline{G} = T_g^{-1}(A - DC)\,T_g \qquad (4\text{-}468)$$

where

$$T_g = [W_g, W'_g, T_{g2}] \qquad (4\text{-}469)$$

with

$$W_g = [\,g, (A - DC)g, \ldots, (A - DC)^{\nu-1}g] \qquad (4\text{-}470)$$

$$\begin{aligned} W'_g &= [\,g', (A' - D'C')g', \ldots, (A' - D'C')^{\nu'-1}g'] \\ &= [\,g', (A - DC)g', \ldots, (A - DC)^{\nu'-1}g'\,] \end{aligned} \qquad (4\text{-}471)$$

and T_{g2} is any $n \times (n - \nu - \nu')$ matrix which makes T_g nonsingular. From (4-115)

$$(A - DC)\,W_g = W_g P \qquad (4\text{-}472)$$

where

$$P = \begin{bmatrix} 0 & 0 & & 0 & -p_1 \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & -p_\nu \end{bmatrix} \quad (\nu \times \nu) \tag{4-473}$$

The equation with D' corresponding to (4-115) is

$$(A' - D'C')^{\nu'} g' = -p_1' g' - \dots - p_{\nu'}'(A' - D'C')^{\nu'-1} g' + z_E \tag{4-474}$$

where z_E is the same vector appearing in (4-465). Then

$$(A - DC)\, W_g' = (A' - D'C')W_g' = W_g' P' + W \overline{G}_{12} \tag{4-475}$$

where

$$P' = \begin{bmatrix} 0 & 0 & & 0 & -p_1' \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & -p_{\nu'}' \end{bmatrix} \quad (\nu' \times \nu') \tag{4-476}$$

and

$$\overline{G}_{12} = \begin{bmatrix} 0 & \cdots & 0 & \alpha_{E1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \alpha_{E\nu} \end{bmatrix} \quad (\nu \times \nu') \tag{4-477}$$

The scalars $\{\alpha_{E1}, \ldots, \alpha_{E\nu}\}$ are defined by

$$z_E = \alpha_{E1} g + \alpha_{E2} A g + \ldots + \alpha_{E\nu} A^{\nu-1} g \tag{4-478}$$

Any vector in E can be expressed uniquely in this form because the set of vectors $\{g, Ag, \ldots, A^{\nu-1} g\}$ form a basis for E. Using these results the coordinate transformation yields

$$\overline{G} = \begin{bmatrix} P & \overline{G}_{12} & \overline{G}_{13} \\ \underline{0} & P' & \overline{G}_{23} \\ \underline{0} & \underline{0} & \overline{G}_{33} \end{bmatrix} \tag{4-479}$$

where

$$(A - DC)T_{g2} = W_g \overline{G}_{13} + W'_g \overline{G}_{23} + T_{g2} \overline{G}_{33} \tag{4-480}$$

From the block triangular form of $\overline{G}$ it is clear that $(\nu + \nu')$ eiganvalues of $\overline{G}$, and thus $(A - DC)$, are given by (4-466) and (4-467). The remaining $(n - \nu - \nu')$ eigenvalues can be specified by the freedom left in D' after constraining it to satisfy (4-465).

The design procedure suggested by the above material is quite straightforward. First g, the maximal generator of f, is found. The coefficients p_i are selected and together with g, A' and d_i can be formed. Then starting with A', C', and d_i the standard design procedure for an ordinary detection filter can be followed to determine a suitable D' to detect d_i. The only difference is that the designer has some additional free choices to make, such as the vector z_E in (4-465).

By taking advantage of equivalence properties arising from the vector equivalence classes it is possible to introduce a number of simplifications in the procedure described above. To begin with, d_i can be replaced by any vector which is equivalent (mod E). Since z_d in (4-426) is in E, Af is such a vector. Besides being simpler to form, Af does not depend on the coefficients p_i. The matrix A' can also be replaced by any other which is equivalent (mod E). The matrix K given by (4-167) is equivalent to A'. Like Af, it is simpler to form and does not depend on the p_i. To show that K and A' are equivalent (mod E), let x be an arbitrary n-vector, and note from (4-169) that

$$\begin{aligned}(K - A')x &= z_d\,[(Cf)^T Cf]^{-1}(Cf)^T Cx \\ &= z_d(c_i x) \qquad (4\text{-}481)\end{aligned}$$

since $Cf = \hat{e}_{mi}$. But $(c_i x)$ is a scalar so the vector on the right is always in E. Hence

$$(K - A')x \equiv \underline{0} \pmod{E} \qquad (4\text{-}482)$$

for arbitrary x. This implies that $K - A' \equiv \underline{0} \pmod{E}$ or

$$A' \equiv K \pmod{E} \qquad (4\text{-}483)$$

Equation (4-465) can be written in terms of K as

$$D'C'K^{\nu'-1}g' = p_1'g' + \ldots + p_{\nu'}'K^{\nu'-1}g' + K^{\nu'}g' + z_E' \qquad (4\text{-}484)$$

where z_E' is any vector in E.

Replacement of A' and d_i by K and Af, which do not depend on the p_i, also allows certain steps in the design procedure to be performed in a different order. In particular it becomes possible to generate g' during the same sequence of operations in which g is generated. (Previously, g had to be found and the p_i selected before A' and d_i could be formed to generate g'.) Generating g and g' in the same operation is more efficient computationally than the two-step process necessary when g' is found using A' and d_i. The procedure is described in Appendix A.

Returning to (4-459), the vector y_d can now be more precisely identified. If $C'Af \neq \underline{0}$ then

$$y_d = C'd_i = C'Af$$

$$= CAf - \hat{e}_{mi}(c_i Af) \qquad (4\text{-}485)$$

using (4-460). Then the output error $\epsilon'(t)$ given by (4-462) lies in the plane formed by CAf and $\hat{e}_{mi}$. In general, if $C'A'^j Af = C'K^j Af = \underline{0}$ for $j = 0, 1, \ldots, \ell - 1$ and $C'A'^{\ell} Af = C'K^{\ell} Af \neq \underline{0}$, then

$$y_d = C'K^{\ell} Af = CK^{\ell} Af - \hat{e}_{mi}(c_i K^{\ell} Af) \qquad (4\text{-}486)$$

and $\epsilon'(t)$ will lie in the plane formed by $CK^{\ell} Af$ and $\hat{e}_{mi}$. Note that the error plane does not depend on the eigenvalues specified for (A - DC) (i.e., on the p_i or p_j'). A Laplace transform analysis of the complete error dynamics can be performed in a manner similar to that in remark 5) at the end of Section 4.3.1. The coordinate transformation given by (4-469) to (4-471) can be used for this purpose. If this is done it will be found that, in addition to results corresponding to those in remark 5), the error dynamics also depend, in part, on z_E in (4-465) and even on the particular g' used in that same equation (recall g' is not unique). Unfortunately the complete results of the Laplace transform analysis in this case are considerably more complicated than those obtained in remark 5). The significantly greater amount of computation required to obtain and interpret the results reduces their practical usefulness.

Up to this point the design of a filter to detect only a single sensor failure has been considered. With the use of equivalence classes (mod E) the results of Sections 4.3.2, 4.3.3, and 4.3.4 can be applied to the problem of designing a detection filter to detect a number of sensor failures. The steps in design correspond in a general way to those listed in Section 4.3.5 with some additional considerations. Below is a brief description of a straightforward design procedure. It is not necessarily the most efficient computationally.

1) Consider k rows of C, each of which is independent of all other rows in C. For convenience of notation let these be the first k rows $\{c_1, \ldots, c_k\}$. For each c_i determine f_i such that $Cf_i = \hat{e}_{mi}$.

2) Form $F = [f_1, \ldots, f_k]$. By construction in step 1) the f_i are all output separable vectors. Generate the θ_i and as described in Appendix B. If Λ does not exist (has zero dimension), the f_i are mutually detectable. If Λ does exist, identify its eigenvalues and decide if they are satisfactory. If not, apply the results of

Section 4.3.4 to subdivide the set $\{f_1, \ldots, f_k\}$ so that the undesirable eigenvalues are eliminated. If the standard form model of the plant suggested in Section 4.3.6 is used, the f_i will always be mutually detectable. This step can be skipped in that case.

3) Let $\{f_1, \ldots, f_{k_1}\}$ be a set resulting from step 2). Form the vectors $\{Af_1, \ldots, Af_{k_1}\}$ and the matrices A' and C' defined by (4-254) and (4-255) with $F = [f_1, \ldots, f_{k_1}]$. For each vector Af_i one of three possibilities must hold.

(i) Af_i does not lie in the unobservable space of C' with respect to A'.

(ii) Af_i does lie in the unobservable space of C', and any detector gain satisfying (4-245) is also a detector gain for Af_i.

(iii) Af_i lies in the unobservable space of C', but a detector gain satisfying (4-245) is not a detector gain for Af_i.

Case (ii) will result if Af_i lies in the detection space of some f_j. It may also result when Af_i lies in a subspace made up of several detection spaces which have some identical eigenvalues. The chance of this special situation occurring is made more likely by specifying a large number of identical eigenvalues for the detection space of the f_j. In any case, one way to check for the occurrence of case (ii) for any Af_i lying in the unobservable space of C' is to determine if the sequence of vectors $\{CAf_i, CA'Af_i, \ldots, CA'^{n-1}Af_i\}$ all lie in one direction. If they do, case (ii) applies, if not case (iii) applies.

Retain all f_i for which (i) or (ii) holds and remove any others from the set.

4) Let $\{f_1, \ldots, f_{k_2}\}$ be a set resulting from step 3). Define A' and C' by (4-254) and (4-255) with $F = [f_1, \ldots, f_{k_2}]$. The Af_i in category (i) of step 3) must now be checked for mutual detectability with respect to (A', C'). This means essentially repeating step 2) with A, C, and the f_i replaced by A', C', and the Af_i. For any Af_i which produces undesirable eigenvalues, the corresponding f_i is removed from the set $\{f_1, \ldots, f_{k_2}\}$. If some vectors are removed, some Af_i may move from category (ii) to category (i). Then mutual detectability of the Af_i must be rechecked with the new members.

5) Let $\{f_1, \ldots, f_{k_3}\}$ be a set resulting from step 4). A detector gain for the Af_i in category (i) can be found by solving a set of equations for D' of the form of (4-245). The remaining freedom in D', if any, is used to specify the remaining eigenvalues of $(A' - D'C')$. A procedure analogous to that mentioned in step 4) of Section 4.3.5 can be used to do this. The resulting matrix $(A' - D'C') = (A - DC)$ yields a detection filter which will detect the failure of any of the k_3 sensors associated with the vectors $\{f_1, \ldots, f_{k_3}\}$.

It should be emphasized that when the plant is modeled in the standard form suggested in Section 4.3.6, many of these steps are considerably simplified and can often be completed by simple inspection. Chapter 5 discusses the processing of detection filter error signals to diagnose sensor failures.

4.4 Summary

The concept of a detection filter and the motivation for its development was discussed in Chapter 3. Basically it is designed to provide information which will aid in the detection and identification of effector and sensor failures and changes in the linear plant dynamics as described in Chapter 3. The detection filter produces an output estimate which asymptotically approaches the actual output of the sensors when there are no failures, plant changes, or other disturbances. A deviation from the undisturbed condition produces an accessible error signal which is the difference between the actual sensor outputs and the filter estimate of those outputs. The essential feature of a detection filter is that it is designed to respond in a special way to certain failures or changes. Of course any other disturbance may also elicit an error response from the filter, but by knowing and looking for the special responses it is possible to detect and identify the occurrence of a failure or change even though it is obscured by the ambient disturbance level.

When a failure or change occurs which a certain filter has been designed to detect, that filter will produce an output error signal which has a fixed direction (the output error is a vector-valued signal). That fixed direction is identified with a certain failure or plant change. There are two qualifications to this ideal situation. First, several failures or changes may be associated with a single error direction. Often additional information (e.g., dynamic properties of the error magnitude) can help to differentiate among such possibilities. Second, it is not possible, in general, to construct a filter which produces a fixed-direction error in the case of a sensor failure. The best that can be done is to constrain the error to a two-dimensional plane.

When there are a sufficient number of independent sensors to be able to determine instantaneously the state of the plant (assuming perfect measurements), the state vector is considered to be fully measurable. In this case, as is shown in Section 4.2, a single detection filter can provide information about all the events described in Chapter 3 -- effector failures, sensor failures, and changes in plant dynamics. This filter is of the same order (state vector dimension) as the plant. In response to a single failure or change it produces an error signal fixed in direction, with a magnitude equivalent to the response of a first order linear system driven by the magnitude of the failure or change (i.e., the magnitude of the deviation from the normal operating characteristics of the plant). The time constant of this first order response can be arbitrarily specified by the designer, but is the same for all events. Of course it is not necessary to use a single all-purpose filter. In some situations it may be preferable to use several filters and tailor their dynamic characteristics to match the characteristics of different events. It would seem desirable, however, to keep the number of detection filters small.

When the state vector of the plant is not fully measurable, it is not possible to construct a single all-purpose filter which provides information about all events. It is not difficult to show that even in this case it is possible to construct a filter which produces the characteristic fixed-direction error signal in response to one event at least. But there are two other important considerations in the design of a detection filter.

The first is the ability to control certain dynamic properties of the filter while achieving the fixed-direction error characteristic. Not only is it important to be able to avoid undesirable (e.g., unstable) filter dynamics, but also to be able to tailor those dynamics to enhance the response to the events of interest and suppress the response to other disturbances. The results of Section 4.3.1 show how it is possible to obtain the fixed-direction error response for one event and at the same time retain control over the poles of the detection filter. It is found that the error magnitude response is not necessarily that of a first order system as it was in the case of a fully measurable state vector. However, for each event there is a maximum system order for the magnitude response beyond which the fixed-direction property cannot be achieved. This order is defined as the detection order of the event. It is found that the order of the error magnitude response should be made a maximum, i.e., equal to the detection order, if one wishes to remain control over as many poles of lthe filter as possible. The poles associated with the magnitude response can be arbitrarily specified by the designer, but the zeros cannot. It is possible to determine the location of the zeros before specifying the poles, so zeros in the left half of the complex plane can be cancelled with poles if desired.

Because the control of the detection filter poles is included in the problem of detection, the condition of observability of the plant appears in the results. When a plant model is not observable, then a detection filter which considers the full plant will have a certain number of poles equal to those of the plant, and these cannot be controlled by the designer of the filter. In a practical sense observability plays only a superficial role, however. The whole subject of detection here is based on obtaining information from only accessible signals. As noted in Chapter 2, when a plant is not observable, the unobservable portion has no effect on the accessible signals. That portion then is "unknowable" with respect to accessible signals, so for the purpose of detection it does not make sense to model the plant dynamics with an unobservable representation.

The second important consideration in the design of a detection filter is to make the filter as versatile as possible, i.e., able to provide information about as many events as possible. This problem is the subject of Sections 4.3.2, 4.3.3, and 4.3.4. It is found that in constructing a filter to detect a number of events it is not always possible to retain control over all the poles of the filter. Section 4.3.3 shows how to determine which events can be detected by the same filter while still retaining control over all poles. Section 4.3.4 takes a broader view and allows the possibility of uncontrolled poles in the filter. It demonstrates how to identify such poles and how undesirable poles can be eliminated by removing certain events from the set of events which the filter is required to detect.

The final three sections in the chapter specialize the previous general results to the three types of events described in Chapter 3. Section 4.3.5 deals with the detection of effector failures. A brief step-by-step design procedure is presented, and the error response of the resulting filter is discussed. Section 4.3.6 considers the use of detection filters to determine changes in plant dynamics. It describes a standard form model for the plant which simplifies the design process and makes it possible to produce information about all changes in plant dynamics. This model may have a larger state vector dimension than the minimum dimension necessary to represent the plant when the dynamics are completely determined. The enlarged state vector reflects the uncertainty introduced by the possibility of changes in the plant dynamics. Section 4.3.7 deals with the most complex problem in detection filter design -- the detection of sensor failures. It is shown that the error response to a sensor failure can be restricted to a two-dimensional plane if that sensor output is modeled as being independent of the other outputs driving the filter. In the standard form suggested in Section 4.3.6, every sensor output is modeled as independent of all the others. If in the minimal plant representation some sensors are dependent and are so modeled, then a more direct way of detecting a failure is by a simple comparison of outputs. This point is illustrated in Section 4.2.3. The detection-filter method of detecting sensor failures complements the direct-comparison method. The direct-comparison method can be used only if the sensor is dependent on other sensors, whereas for the detection-filter method the sensor is assumed to be independent of the other sensors.

A detection filter for any type of event is of course based on a model of the plant dynamics. One detection filter, at least, will have the responsibility for detecting and identifying changes in these dynamics — in effect forming a new plant model. Having obtained a new plant model, all the other detection filters must be rechecked and adjusted, if necessary, to fit the new model. Therefore, it is important to the overall reorganization scheme to have efficient filter design algorithms which can be carried out by on-line computers. For this reason reference is made throughout Chapter 4 to Appendices A and B which describe algorithms for obtaining the various vector and matrix qnaitities necessary in the filter design process. These algorithms are developed for a general linear plant description. When a standard form plant model is used, a number of significant simplifications result.

CHAPTER 5

IDENTIFICATION DECISIONS

5.1 General Discussion

This chapter investigates the problem of identifying events from the error signals produced by detection filters. The detection filter is designed to produce a fixed-direction error in response to certain events. Ideally the identification problem is a simple matter of noticing the fixed-direction error and associating it with a specific event. The actual identification problem is more difficult than this for two reasons. The first is that the detection filter may be responding to other disturbances besides the specific event producing a fixed-direction error. When these extraneous errors are added to the fixed-direction error the result is an error signal not fixed in direction. The total error must be processed somehow to recover the fixed-direction signal from the extraneous errors. Noise disturbance in the sensor outputs or entering through the plant dynamics is one source of extraneous errors. A second source is the occurrence of multiple events which must be detected by different filters. For example, changes in plant dynamics will cause extraneous errors in the output of a filter designed to detect effector failures.

The second complicating factor in the identification problem is the case of nonseparable events which cannot be distinguished on the basis of error direction alone. The most important example of this arises in the detection of changes in plant dynamics. As was seen in

Sections 4.2.2 and 4.3.6, error direction alone is not sufficient to determine which elements of A or B have changed. Error magnitude information is also necessary. The identification of plant dynamics is treated as a special case in the next section. The identification of effector and sensor failures is investigated in the final section.

5.2 Plant Dynamics Identification

This section discusses the problem of determining changes in plant dynamics from the error signal produced by a detection filter. The problem will be considered first in a formal mathematical framework. This will show, in theory, what information the error signal can and cannot provide about plant dynamics. Such results will establish the limitations on what can be expected from any dynamics identification scheme based on detection filters. Section 5.2.2 compares the detection-filter method of dynamics identification to some other methods.

5.2.1 Conditions for Identifiability

This section investigates the conditions under which the plant dynamics can (and cannot) be uniquely determined from the information provided by a detection filter, assuming perfect knowledge of the input and output vectors of the plant.

It will be assumed that the plant is modeled by

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{5-1}$$

$$u(t) = u_d(t) \tag{5-2}$$

$$y(t) = Cx(t) \tag{5-3}$$

with A and C in the standard form suggested in Section 4.3.6

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mm} \end{bmatrix} \quad (n \times n) \tag{5-4}$$

$$A_{ii} = \begin{bmatrix} 0 & 0 & & 0 & a_{ii1} \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & a_{iin_i} \end{bmatrix} \quad (n_i \times n_i) \tag{5-5}$$

$$A_{ij} = \begin{bmatrix} 0 & \cdots & 0 & a_{ij1} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & a_{ijn_i} \end{bmatrix} \quad (n_i \times n_j) \tag{5-6}$$

$$A_{ji} = \begin{bmatrix} 0 & \cdots & 0 & a_{ji1} \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & a_{jin_i} \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix} \quad (n_j \times n_i) \tag{5-7}$$

where $n_j \geq n_i$, and

$$C = \begin{bmatrix} \hat{e}_{s_1}^T \\ \vdots \\ \hat{e}_{s_m}^T \end{bmatrix} \quad (m \times n) \tag{5-8}$$

where

$$s_i = n_1 + \ldots + n_i \tag{5-9}$$

and

$$n_1 + \ldots + n_m = n \tag{5-10}$$

The matrix B is partitioned to conform to the blocks of A as in Section 4.3.6

$$B = \begin{bmatrix} b_{11} & \cdot & \cdot & \cdot & \cdot & \cdot & b_{1r} \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ b_{m1} & \cdot & \cdot & \cdot & \cdot & \cdot & b_{rr} \end{bmatrix} \quad (n \times r) \tag{5-11}$$

$$b_{ij} = \begin{bmatrix} b_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ b_{ijn_i} \end{bmatrix} \quad (n_i \times 1) \tag{5-12}$$

The error response to changes in individual elements of A and B is given by (4-423) and (4-427) respectively. Adding together the effects of all allowable changes in A and B yields a total settled-out output error of

$$\begin{aligned} \epsilon'(t) = & \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{\ell=1}^{\bar{n}_{ij}} \Delta a_{ij\ell}\, \phi_{ij\ell}(t)\, \hat{e}_{mi} \\ & + \sum_{i=1}^{m} \sum_{j=1}^{r} \sum_{\ell=1}^{n_i} \Delta b_{ij\ell}\, \psi_{ij\ell}(t)\, \hat{e}_{mi} \end{aligned} \tag{5-13}$$

with $\phi_{ij\ell}(t)$ and $\psi_{ij\ell}(t)$ given by (4-424) and (4-428). In the first term on the right side of (5-13) the summation on ℓ has the upper limit of

$$\bar{n}_{ij} = \min\{n_i, n_j\} \tag{5-14}$$

instead of simply n_i because nonsquare blocks of A have the form of (5-7) in which $a_{ij\ell}$ is identically zero for all $\ell > \bar{n}_{ij}$. This results from the algorithm of Appendix C for obtaining the standard form. The i^{th} component of $\epsilon'(t)$ is

$$\epsilon'_i(t) = \sum_{j=1}^{m} \sum_{\ell=1}^{\bar{n}_{ij}} \Delta a_{ij\ell}\, \phi_{ij\ell}(t) + \sum_{j=1}^{r} \sum_{\ell=1}^{n_i} \Delta b_{ij\ell}\, \psi_{ij\ell}(t) \tag{5-15}$$

Define the following vectors:

$$\pi_{ij} = \begin{bmatrix} \Delta a_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ \Delta a_{ijn_{ij}} \end{bmatrix} \quad \text{for } j = 1, \ldots, m \tag{5-16}$$

$$\pi_{i,m+j} = \begin{bmatrix} \Delta b_{ij1} \\ \cdot \\ \cdot \\ \cdot \\ \Delta b_{ijn_i} \end{bmatrix} \quad \text{for } j = 1, \ldots, r \tag{5-17}$$

$$\xi_{ij}(t) = \begin{bmatrix} \phi_{ij1}(t) \\ \cdot \\ \cdot \\ \cdot \\ \phi_{ij\bar{n}_{ij}}(t) \end{bmatrix} \quad \text{for } j = 1, \ldots, m \tag{5-18}$$

$$\xi_{i,m+j}(t) = \begin{bmatrix} \psi_{ij1}(t) \\ \cdot \\ \cdot \\ \cdot \\ \psi_{ijn_i}(t) \end{bmatrix} \quad \text{for } j = 1, \ldots, r \tag{5-19}$$

and with these vectors form the composite vectors

$$\pi_i = \begin{bmatrix} \pi_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \pi_{i,m+r} \end{bmatrix} \tag{5-20}$$

$$\xi_i(t) = \begin{bmatrix} \xi_{i1}(t) \\ \cdot \\ \cdot \\ \cdot \\ \xi_{i,m+r}(t) \end{bmatrix} \tag{5-21}$$

Now (5-15) can be written as

$$\epsilon'_i(t) = \xi_i^T(t)\ \pi_i \tag{5-22}$$

The basic problem in identification of plant dynamics is to solve (5-22) for π_i, given $\epsilon'_i(t)$ and $\xi_i(t)$. The question of interest here is to determine under what circumstances this is theoretically possible. Equation (5-22) can be viewed as a linear mapping from Euclidean space into the vector space of continuous scalar functions over some time interval $t_1 \leq t \leq t_2$. From the theory of linear mappings (Section 12 in [4]) it is known that π_i in (5-22) can be determined to within an additive constant vector which lies in the null space of $\xi_i^T(t)$. The null space of $\xi_i^T(t)$ is the set of all vectors π_o for which $\xi_i^T(t)\ \pi_o$ is identically zero on the interval $[t_1, t_2]$. A time-invariant vector equation can be obtained from (5-22) by multiplying by $\xi_i(t)$ and integrating over $[t_1, t_2]$. This yields

$$\xi_{\epsilon i} = M_i(t_1, t_2)\ \pi_i \tag{5-23}$$

where

$$\xi_{\epsilon i} = \int_{t_1}^{t_2} \xi_i(t)\, \epsilon'_i(t)\, dt \tag{5-24}$$

and

$$M_i(t_1, t_2) = \int_{t_1}^{t_2} \xi_i(t)\, \xi_i^T(t)\, dt \tag{5-25}$$

Any π_i which satisfies (5-23) also satisfies (5-22) and conversely. The null space of $\xi_i^T(t)$ over $[t_1, t_2]$ coincides with the null space of $M_i(t_1, t_2)$. This result is proven by Brockett (Lemma 1, Section 14 in [4]). It is clear from (5-23) that π_i can be determined uniquely if and only if $M_i(t_1, t_2)$ is nonsingular. If $M_i(t_1, t_2)$ is singular then for any π_i which lies in the null space, $\epsilon'_i(t)$ will be zero over the interval $[t_1, t_2]$. This means that all standard form models whose parameters have a vector difference π_o lying in the null space of $M_i(t_1, t_2)$ can reproduce exactly the output of the i^{th} sensor over the interval $[t_1, t_2]$. All such models adequately explain the dynamic behavior of the plant over $[t_1, t_2]$ as measured by the i^{th} sensor. Without additional information there is no basis for choosing among these models. In other words, any π_i which satisfies (5-23) will yield a model which can duplicate the plant behavior over $[t_1, t_2]$ as seen by the i^{th} sensor. Of course the main purpose of having a plant model is to be able to predict future plant behavior. It is of interest, therefore, to determine the conditions under which differences between plant and model are indeterminant and to investigate the nature of those differences. For this purpose it is necessary to determine what circumstances produce a singular $M_i(t_1, t_2)$. Suppose $M_i(t_1, t_2)$ is singular and π_o is a nonzero vector in the null space. Since, as noted above, the null spaces of $M_i(t_1, t_2)$ and $\xi_i^T(t)$ coincide

$$\xi_i^T(t)\,\pi_o \quad = \quad 0 \qquad \text{for all } t_1 \leq t \leq t_2 \tag{5-26}$$

Partition π_o into $(m + r)$ vectors conforming to $\xi_i(t)$.

$$\pi_o \quad = \quad \begin{bmatrix} \pi_{o1} \\ \cdot \\ \cdot \\ \cdot \\ \pi_{o,m+r} \end{bmatrix} \tag{5-27}$$

with

$$\pi_{oj} \quad = \quad \begin{bmatrix} \pi_{oj1} \\ \cdot \\ \cdot \\ \cdot \\ \pi_{oj\bar{n}_{ij}} \end{bmatrix} \text{for } j = 1, \ldots, m \tag{5-28}$$

and

$$\pi_{oj} \quad = \quad \begin{bmatrix} \pi_{oj1} \\ \cdot \\ \cdot \\ \cdot \\ \pi_{ojn_i'} \end{bmatrix} \quad \text{for } j = m, \ldots, m + r \tag{5-29}$$

where the $\pi_{oj\ell}$ are scalars. From the definition of $\xi_i(t)$ (5-26) can be written as

$$\sum_{j=1}^{m} \sum_{\ell=1}^{\bar{n}_{ij}} \pi_{oj\ell}\,\phi_{ij\ell}(t) + \sum_{j=1}^{r} \sum_{\ell=1}^{n_i} \pi_{o,m+j,\ell}\,\psi_{ij\ell}(t) \quad = \quad 0 \tag{5-30}$$

This equation is equivalent to a linear differential equation for $\phi_{ij1}(t)$ and $\psi_{ij1}(t)$. To see this, note from the definition of $h_{i\ell}(t)$ (4-421) used in forming $\phi_{ij\ell}(t)$ and $\psi_{ij\ell}(t)$ it follows that

$$\phi_{ij\ell}(t) = \frac{d^{\ell-1}}{dt^{\ell-1}} \phi_{ij1}(t) \quad \text{for } \ell = 1, \ldots, n_i \tag{5-31}$$

and

$$\psi_{ij\ell}(t) = \frac{d^{\ell-1}}{dt^{\ell-1}} \psi_{ij1}(t) \quad \text{for } \ell = 1, \ldots, n_i \tag{5-32}$$

Then (5-30) becomes

$$\sum_{j=1}^{m} \sum_{\ell=1}^{\bar{n}_{ij}} \pi_{oj\ell} \frac{d^{\ell-1}}{dt^{\ell-1}} \phi_{ij1}(t) + \sum_{j=1}^{r} \sum_{\ell=1}^{n_i} \pi_{o,m+j,\ell} \frac{d^{\ell-1}}{dt^{\ell-1}} \psi_{ij1}(t) = 0 \tag{5-33}$$

To simplify notation define

$$\lambda^{\ell} = \frac{d^{\ell}}{dt^{\ell}} \tag{5-34}$$

Then (5-33) can be written

$$\sum_{j=1}^{m} \eta_j(\lambda) \phi_{ij1}(t) + \sum_{j=1}^{r} \rho_j(\lambda) \psi_{ij1}(t) = 0 \tag{5-35}$$

where

$$\eta_j(\lambda) = \sum_{\ell=1}^{\bar{n}_{ij}} \pi_{oj\ell} \lambda^{\ell-1} \tag{5-36}$$

and

$$\rho_j(\lambda) = \sum_{\ell=1}^{n_i} \pi_{o,m+j,\ell} \lambda^{\ell-1} \tag{5-37}$$

By their definitions $\phi_{ij1}(t)$ and $\psi_{ij1}(t)$ are related to $y_j(t)$ and $u_{dj}(t)$ through the differential equations

$$\mu_i(\lambda)\, \phi_{ij1}(t) = y_j(t) \tag{5-38}$$

$$\mu_i(\lambda)\, \psi_{ij1}(t) = u_{dj}(t) \tag{5-39}$$

where

$$\mu_i(\lambda) = \lambda^{n_i} + \sum_{\ell=1}^{n_i} p_{i\ell} \lambda^{\ell-1} \tag{5-40}$$

The $p_{i\ell}$ which appear in (5-40) are the same as those appearing in (4-419). These are the coefficients chosen by the designer to specify the poles of the detection filter. Applying the differential operator $\mu_i(\lambda)$ to (5-35) gives

$$\sum_{j=1}^{m} \mu_i(\lambda)\, \eta_j(\lambda)\, \phi_{ij1}(t) + \sum_{j=1}^{r} \mu_i(\lambda)\, \rho_j(\lambda)\, \psi_{ij1}(t) = 0 \tag{5-41}$$

Interchanging the order of the differential operators and using (5-38) and (5-39) yields

$$\sum_{j=1}^{m} \eta_j(\lambda)\, y_j(t) + \sum_{j=1}^{r} \rho_j(\lambda)\, u_{dj}(t) = 0 \tag{5-42}$$

This shows that (5-42) is a necessary condition for $M_i(t_1, t_2)$ to be singular. It can also be shown that it is sufficient. Suppose (5-30) is satisfied for any $\eta_j(\lambda)$ and $\rho_j(\lambda)$ having the form of (5-36) and (5-37) with arbitrary coefficients $\pi_{oj\ell}$ (not all zero). Substituting (5-38) and (5-39) into (5-42) and interchanging the order of the differential operators gives (5-41). Defining

$$q(t) = \sum_{j=1}^{m} \eta_j(\lambda)\ \phi_{ij1}(t) + \sum_{j=1}^{r} \rho_j(\lambda)\ \psi_{ij1}(t) \tag{5-43}$$

equation (5-41) can be written as

$$\mu_i(\lambda)\ q(t) = 0 \tag{5-44}$$

Recall that for the error signal given by (5-15) it was assumed that the initial condition effects in the detection filter had settled out. The n_i roots of $\mu_i(s) = 0$ are poles of the detection filter. This means that the initial condition effects of any solution of (5-44) have the same settling times as those of the detection filter. If t is large enough so that the filter has settled out, then the solution of (5-44) will have settled out also. Since (5-44) is undriven, the settled-out solution is

$$q(t) = 0 \tag{5-45}$$

which gives (5-35) be definition of q(t). The development from (5-26) to (5-35) is equally valid in reverse so (5-35) implies (5-26) which in turn implies $M_i(t_1, t_2)$ is singular. This shows that condition (5-42) is both necessary and sufficient for $M_i(t_1, t_2)$ to be singular.

To see clearly what condition (5-42) means, it must be interpreted in terms of the dynamic behavior of the plant. This condition is a differential equation relating the control signal $u_d(t)$ and the sensor output vector $y(t)$. These quantities are of course already related by Equations (5-1) to (5-3) describing the plant dynamics. These relationships must be clearly delineated before (5-42) can be properly interpreted. Consider the plant representation (5-1) to (5-3) with A and C in the form of (5-4) to (5-10). Partition the state vector $x(t)$ into m n_i-vectors to conform with the partitioning of A,

$$x(t) = \begin{bmatrix} \underline{x}_1(t) \\ \cdot \\ \cdot \\ \cdot \\ \underline{x}_m(t) \end{bmatrix} \tag{5-46}$$

with

$$\underline{x}_k(t) = \begin{bmatrix} x_{k1}(t) \\ \cdot \\ \cdot \\ \cdot \\ x_{kn_k}(t) \end{bmatrix} \tag{5-47}$$

Then

$$y_k(t) = x_{kn_k}(t) \tag{5-48}$$

and

$$\begin{aligned} \dot{\underline{x}}_k(t) &= A_{kk}\underline{x}_k(t) + \sum_{\substack{j=1 \\ j\neq k}}^{m} A_{kj}\,\underline{x}_j(t) + \sum_{k=1}^{m} b_{kj}\,u_{dj}(t) \\ &= A_{kk}\,\underline{x}_k(t) + \sum_{\substack{j=1 \\ j\neq k}}^{m} \underline{a}_{kj}\,y_j(t) + \sum_{k=1}^{m} b_{kj}\,u_{dj}(t) \end{aligned} \tag{5-49}$$

where

$$\underline{a}_{kj} = \begin{bmatrix} a_{kj1} \\ \cdot \\ \cdot \\ \cdot \\ a_{kjn_k} \end{bmatrix} \tag{5-50}$$

and

$$a_{kj\ell} = 0 \quad \text{if } \ell > n_j \tag{5-51}$$

Equations (5-48) and (5-49) are equivalent to the scalar differential equation

$$\sum_{j=1}^{m} \nu_{kj}(\lambda)\, y_j(t) + \sum_{j=1}^{r} \gamma_{kj}(\lambda)\, u_{dj}(t) = 0 \tag{5-52}$$

where

$$\nu_{kk}(\lambda) = -\lambda^{n_k} + \sum_{\ell=1}^{n_k} a_{kk\ell}\, \lambda^{\ell-1} \tag{5-53}$$

$$\nu_{kj}(\lambda) = \sum_{\ell=1}^{\overline{n}_{kj}} a_{kj\ell}\, \lambda^{\ell-1} \quad \text{for } j \neq k \tag{5-54}$$

$$\gamma_{kj}(\lambda) = \sum_{\ell-1}^{n_k} b_{kj\ell}\, \lambda^{\ell-1} \tag{5-55}$$

Note that $\nu_{kk}(s)$ is the characteristic polynomial of A_{kk} and always has order n_k. The order of $\nu_{kj}(s)$ $(j \neq k)$ is less than or equal to $(\overline{n}_{kj} - 1)$ and $\gamma_{kj}(s)$ has order no larger than $(n_k - 1)$. Equation (5-52) for any k does not satisfy condition (5-42) because $\nu_{kk}(\lambda)$ has order n_k whereas the operator $n_k(\lambda)$ associated with $y_k(t)$ in (5-42) must have order no larger than $(n_k - 1)$ as can be seen from (5-36). This means that $M_i(t_1, t_2)$ for all i will be nonsingular as long as the dynamic behavior of $u_d(t)$ and $y(t)$ cannot be described by any equations of lower order than those given by (5-52) for $k = 1, \ldots, m$. In other words, the plant should exhibit the full dynamic properties attributed to it by the representation (5-1) to (5-3).

It is possible to associate the singularities of $M_i(t_1, t_2)$ with several specific situations. A nonminimal model may yield a singular $M_i(t_1, t_2)$. It was noted in Section 4.3.6 that a nonminimal representation cannot be both controllable and observable. The standard form model is constructed to be observable, so if it is nonminimal it must be noncontrollable. When a representation is not minimal it is possible to reduce the dimension of the state space to obtain a representation which is minimal and which has the same dynamic relationship between input and output. In effect the uncontrollable part of the system is discarded to obtain the minimal representation. The reduced representation yields a set of differential equations relating $y(t)$ and $u_d(t)$ to replace those given by (5-52) for $k = 1, \ldots, m$. One or more of these equations will be of lower order than (5-52) for some k since the state vector has been made smaller. Any such equation will fit the form of (5-42) suggesting that some $M_i(t_1, t_2)$ can be singular if the nonminimal model is used. This may or may not be the case depending on the initial conditions. The reason a nonminimal representation can be reduced is because the uncontrollable portion of the dynamics is never excited by the input. As far as the relationship between input and output is concerned, this portion of the dynamics can be ignored. However, this does not mean that the effect of the uncontrollable portion is never seen in the output. Because the model is observable, the full effect of the uncontrollable portion can be evident in the output provided the initial conditions are such that the uncontrollable modes are excited by transients. In this case the reduced minimal representation will not be adequate to explain all the dynamics appearing in the output. The lower order equations suggested by the reduced representation will not be valid and $M_i(t_1, t_2)$ will not be singular. The lower order equations are valid only if the initial conditions for the uncontrollable modes are zero or their effect has settled out by the time t_1.

There are two reasons why the model may be nonminimal. As noted in Section 4.3.6, it may be necessary to enlarge the state space in order to achieve the standard form of (5-4) to (5-10). If this is done the model will be nonminimal. The method described in Appendix C for enlarging the state space demonstrates the arbitrary nature of the added portion of the augmented model. Because of this an augmented model is not unique, and this nonuniqueness is reflected in the singularity of certain $M_i(t_1, t_2)$ (implying the solution of (5-23) is not unique). Singularities in $M_i(t_1, t_2)$ which result from an augmented model present no theoretical problem because any solution of (5-23) will yield a plant representation which correctly models the plant behavior. The multiple solutions of (5-23) simply correspond to the arbitrary portion of the augmented model, which is not related to any dynamics in the actual plant.

A second reason for a nonminimal representation is that the actual plant may be nonminimal. This could be the result of effector failures, sensor failures, or dynamics changes which have caused the plant to become unobservable or uncontrollable. In this case a portion of the actual plant dynamics may be unidentifiable. An unobservable plant may result from sensor failures or changes in dynamics. In this case $M_i(t_1, t_2)$ for some i will be singular. As shown in Chapter 2, the unobservable portion of the plant dynamics will never appear in the output. This means that the plant behavior as seen by the detection filter can be fully explained by a reduced state vector which results when the unobservable portion of the plant is ignored. This implies the relationship between input and output satisfies a differential equation of lower order than those derived from the original state vector, which is the same size as the state vector of the model. This means condition (5-42) is satisfied, and therefore some $M_i(t_1, t_2)$ will be singular.

An uncontrollable plant may result from effector failures or changes in dynamics. In this case $M_i(t_1, t_2)$ may be singular or nonsingular. The uncontrollable modes of the plant dynamics will be seen in the output if and only if they are excited by the initial conditions. If some uncontrollable modes of the plant are not excited by the initial conditions, then some $M_i(t_1, t_2)$ will be singular. If the uncontrollable portion of the plant is fully excited by initial conditions, and the controllable portion is fully excited either by the inputs or initial conditions or both, then $M_i(t_1, t_2)$ will be nonsingular. Of course, initial condition transients can identify uncontrollable modes of the plant only if their settling times are significantly longer than the settling time of the detection filter. Otherwise the transients will settle to zero in the time allowed for the filter to settle out.

Even for a minimal plant and model $M_i(t_1, t_2)$ may be singular if there is external low-order coupling between y(t) and $u_d(t)$

or between components of $u_d(t)$. External low-order coupling means dynamic coupling of the form given by (5-42) caused by effects external to the plant. The most obvious example of external coupling is a feedback loop. If $y(t)$ and $u_d(t)$ are related through feedback by a low-order relation in the form of (5-42), then some $M_i(t_1, t_2)$ will be singular.

Coupling between components of $u_d(t)$ may also cause a singular $M_i(t_1, t_2)$. It can be shown that for a minimal representation some $M_i(t_1, t_2)$ will be singular only if there exists a set of polynomials $\{\chi_j(s), \ j = 1, \ldots, r\}$ (not all identically zero) each with order no larger than $(n + \bar{n} - 1)$, such that

$$\sum_{j=1}^{r} \chi_j(\lambda)\, u_{dj}(t) = 0 \tag{5-56}$$

where n is the state dimension of the minimal representation and

$$\bar{n} = \max\{n_1, \ldots, n_m\} \tag{5-57}$$

Define the matrices of polynomials

$$N(s) = \begin{bmatrix} \nu_{11}(s) & \cdots & \nu_{1m}(s) \\ \vdots & & \vdots \\ \nu_{m1}(s) & \cdots & \nu_{mm}(s) \end{bmatrix} \tag{5-58}$$

$$\Gamma(s) = \begin{bmatrix} \gamma_{11}(s) & \cdots & \gamma_{1r}(s) \\ \vdots & & \vdots \\ \gamma_{m1}(s) & \cdots & \gamma_{rr}(s) \end{bmatrix} \tag{5-59}$$

Then the equations (5-52) for $j = 1, \ldots, m$ can all be written in one vector equation

$$N(\lambda)\, y(t) + \Gamma(\lambda)\, u_d(t) = \underline{0} \tag{5-60}$$

Let $\overline{N}(s)$ be the matrix of cofactors of $N(s)$ having the property

$$\overline{N}(s)\, N(s) = \left| N(s) \right| I = \nu_o(s)\, I \tag{5-61}$$

($\nu_o(s)$ is the characteristic polynomial of A.) Applying the operator $\overline{N}(\lambda)$ to (5-60) yields

$$\begin{aligned} &\overline{N}(\lambda)\, N(\lambda)\, y(t) + \overline{N}(\lambda)\, \Gamma(\lambda)\, u_d(t) \\ &= \nu_o(\lambda)\, y(t) + \overline{N}(\lambda)\, \Gamma(\lambda)\, u_d(t) = \underline{0} \end{aligned} \tag{5-62}$$

Assume $y(t)$ and $u_d(t)$ also satisfy (5-42) for some $\eta_j(\lambda)$ and $\rho_j(\lambda)$. Define the vectors of polynomials

$$\eta(s) = \begin{bmatrix} \eta_1(s) \\ \cdot \\ \cdot \\ \cdot \\ \eta_m(s) \end{bmatrix} \tag{5-63}$$

$$\rho(s) = \begin{bmatrix} \rho_1(s) \\ \cdot \\ \cdot \\ \cdot \\ \rho_r(s) \end{bmatrix} \tag{5-64}$$

Then (5-42) can be written

$$\eta^T(\lambda)\, y(t) + \rho^T(\lambda)\, u_d(t) = 0 \tag{5-65}$$

Applying the operator $\nu_o(\lambda)$ to this equation and using (5-62) yields

$$\nu_o(\lambda)\,\eta^T(\lambda)\,y(t) + \nu_o(\lambda)\,\rho^T(\lambda)\,u_d(t)$$

$$= \eta^T(\lambda)\,\nu_o(\lambda)\,y(t) + \rho^T(\lambda)\,\nu_o(\lambda)\,u_d(t)$$

$$= [-\eta^T(\lambda)\,\overline{N}(\lambda)\,\Gamma(\lambda) + \rho^T(\lambda)\,\nu_o(\lambda)]\,u_d(t) = \underline{0} \tag{5-66}$$

Now $\nu_o(s)$ has order n because it is the characteristic polynomial of A which is $n \times n$. By (5-37) the highest order polynomial in $\rho(s)$ can have order no larger than $\overline{n} - 1$ where $\overline{n} = \max\{n_1, \ldots, n_m\}$. Then the polynomial elements of $\rho^T(s)\,\nu_o(s)$ are of order no larger than $(n + \overline{n} - 1)$. The matrix $\overline{N}(s)\,\Gamma(s)$ has no polynomial element with order larger than $(n - 1)$. This can be shown from (5-62). Taking the Laplace transform of both (5-62) and (5-1) to (5-3) and equating the transfer functions from $\mathcal{L}\{u_d(t)\}$ to $\mathcal{L}\{y(t)\}$ yields

$$C[Is - A]^{-1}B = \frac{\overline{N}(s)\,\Gamma(s)}{\nu_o(s)} \tag{5-67}$$

The elements of $C[Is - A]^{-1}B$ are rational polynomials each with a larger order denominator than numerator. The same must be true of $\overline{N}(s)\,\Gamma(s)/\nu_o(s)$. Hence, no polynomial element of $\overline{N}(s)\,\Gamma(s)$ can have order greater than $(n - 1)$, since $\nu_o(s)$ has order n. From (5-36) it is clear that $(\overline{n} - 1)$ is the highest order polynomial allowable in $\eta(s)$. The complete differential operator $[-\eta^T(\lambda)\,\overline{N}(\lambda)\,\Gamma(\lambda) + \rho^T(\lambda)\,\nu_o(\lambda)]$ has order no larger than $(n + \overline{n} - 1)$, and therefore (5-66) has the form of (5-56) where

$$[\chi_1(s), \ldots, \chi_r(s)] = \chi^T(s) = -\eta^T(s)\,\bar{N}(s)\,\Gamma(s) + \rho^T(s)\,\nu_o(s) \tag{5-68}$$

The case where $[-\eta^T(s)\,\bar{N}(s)\,\Gamma(s) + \rho^T(s)\,\nu_o(s)]$ is identically zero corresponds to a nonminimal representation. If $[-\eta^T(s)\,\bar{N}(s)\,\Gamma(s) + \rho^T(s)\,\nu_o(s)] = 0$ for all complex values of s, then

$$[-\eta^T(\lambda)\,\bar{N}(\lambda)\,\Gamma(\lambda) + \rho^T(\lambda)\ {}_o(\lambda)]\, u_d(t) = \underline{0} \tag{5-69}$$

for any $u_d(t)$. From (5-62) this implies

$$\nu_o(\lambda)\,[\eta^T(\lambda)\,y(t) + \rho^T(\lambda)\,u_d(t)] = \underline{0} \tag{5-70}$$

for any $u_d(t)$. This means that, ignoring initial condition transients, (5-65) is a valid relationship between y(t) and $u_d(t)$, <u>for any $u_d(t)$</u>, which in turn implies that a reduction is possible in the order of Equations (5-52).

The above development shows that if (5-42) is satisfied for some $\eta_j(\lambda)$ and $\rho_j(\lambda)$ then (5-56) is satisfied for the $\chi_j(\lambda)$ given by (5-68). Since (5-42) is a necessary and sufficient condition for $M_i(t_1, t_2)$ to be singular, one may conclude that (5-56) is a necessary condition for some $M_i(t_1, t_2)$ to be singular if the model is minimal. It is not a sufficient condition in general. The negation of (5-56) is a sufficient condition for all $M_i(t_1, t_2)$ to be nonsingular. That is, if (5-56) is not satisfied for any $\chi_j(\lambda)$ (not all zero) with order less than or equal to $(n + \bar{n} - 1)$, then all $M_i(t_1, t_2)$ for $i = 1, \ldots, m$ will be nonsingular, provided the model is minimal.

In the case of a scalar-input plant ($r = 1$), these remarks can be made more tangible if the result is interpreted in the case where the input $u_d(t)$ is assumed to be a periodic signal made up of a number of discrete frequency components. For $r = 1$, (5-56) reduces to

$$\chi(\lambda)\ u_d(t) = 0 \qquad (5\text{-}71)$$

where $u_d(t)$ is a scalar and $\chi(s)$ is a polynomial of order no larger than $(n + \bar{n} - 1)$. If $u_d(t)$ is a periodic signal with discrete frequency components, (5-71) will be satisfied for some $\chi(\lambda)$ only if the number of distinct frequency components in $u_d(t)$ is less than or equal to $(n + \bar{n} - 1)/2$. Therefore, a sufficient condition for all $M_i(t_1, t_2)$ to be nonsingular (and the minimal representation to be completely identifiable) is that $u_d(t)$ have at least $(n + \bar{n})/2$ different frequency components. For a scalar-output plant $\bar{n} = n$, so one may conclude that the minimal representation of a scalar-input, scalar-output plant can be completely identified if the periodic input has at least n distinct frequency components. This agrees with a similar statement made by Young [27].

The results of this section have been derived for the general case where all $a_{ij\ell}$ in A given by (5-4) to (5-7) and all $b_{ij\ell}$ in B given by (5-11) and (5-12) are subject to change. If in a particular situation only a limited number of $a_{ij\ell}$ and $b_{ij\ell}$ are subject to change, then it is necessary to identify only those particular elements. In that case π_i in (5-20) should include only those elements subject to change, and $\xi_i(t)$ should be shortened accordingly. Conditions for identifiability of a limited number of elements can be derived in the same way as shown here for the general case. For a limited number of changeable elements sufficient conditions for identifiability should be less restrictive.

5.2.2 On-Line Identification Methods

The previous section demonstrated how the plant dynamics can be determined after observing the detection filter error signal over a finite period of time. The method used in that analytical development involved generating the vector $\xi_{\epsilon i}$ and the matrix $M_i(t_1, t_2)$ for a given time interval $[t_1, t_2]$, then solving (5-23) for the difference between plant parameters and model parameters. This may be a feasible method for determining changes in plant dynamics on-line, provided there is sufficient time and computing capacity to solve the equation (5-23). The actual dimension of the vector equation (5-23) depends on the number of changeable parameters in A and B, since, as noted in the previous section, only changeable parameters need to be considered in the identification process. Determining the plant parameters by analytical solution of (5-23) would be most effective in situations where the number of changeable parameters is small and the changes are expected to occur in sudden jumps (as might be expected in the event of a failure).

In situations where the number of changeable parameters is large, and the changes are nearly continuous and slowly time-varying, a more suitable method for on-line identification is a reference model approach. There are several reference model identification methods which have received considerable attention in the literature. The detection filter can also be used in a reference model approach. In the remainder of this section certain properties of the detection filter method will be compared to the properties of some other reference model techniques.

The basic philosophy of reference model identification is to adjust certain parameters in the model to null or minimize some measure of the error between plant and model. Two basic distinguishing features of a reference model identification scheme are the error signal and the parameter adjustment process. The goal of the parameter adjustment process is simply to null or minimize some measure of the error signal. Many algorithms for parameter optimization can be used to obtain a parameter adjustment law which attempts to minimize the error measure. Gradient or "steepest descent" methods are the most common example [12, 14, 25]. Such gradient adjustment laws may be discrete [12] or continuous [12, 14]. In some cases a recursive solution of a linear least squares problem may be used to update parameter estimates at discrete points in time [27]. Another method for determining a parameter adjustment law is based on Liapunov functions [17]. Most of these techniques can also be used with the detection filter error signal. There is a substantial body of literature on the theory and use of such methods of parameter adjustment and their application to reference model identification, so they will not be analyzed further here. It will be instructive, however, to compare some important properties of the error signal from a detection filter with those produced by other reference model methods.

Most reference model identification schemes are variants on one of two basic methods. The first method is often referred to as the response error method [14], or sometimes the "closed" method by Russian authors [16]. The basic philosophy of this method is to apply to the model the same observed input that is acting on the plant, and to observe the difference between the plant output vector (as measured by the sensors) and the model output vector. This output or response error vector is taken as the error between plant and model. The second basic method is usually referred to as the equation error method [14, 27], or the "open" method [16]. The basic philosophy of this method is to substitute the observed input and output vectors of the plant into an equation describing the estimated plant behavior (the equation is the model in this case). If the equation accurately describes the plant behavior (and there are no unobservable disturbances), then the observed input and output vectors should satisfy the equation. If they do not, the discrepancy is taken as the error between plant and model. The model equation is chosen so that the error signal is an algebraic function of the parameters. This means that the error signal at any instant in time depends on the parameter values only at that same instant. This is not the case for the response error. In general the response error depends on past values of the parameters as well. This is an important distinction between these two basic methods.

An important variant on the equation error method is the generalized equation error method [14, 27]. One of the difficulties of the equation error method is that substitution of the observed input and output vectors into the model equation often involves performing operations (e.g., pure time derivatives) which are undesirable with regard to noise suppression. This problem is avoided by the generalized equation error method. The equation describing the plant is replaced by a generalized equation which involves no pure time derivatives of the input and output vectors. Satisfaction of the generalized equation implies satisfaction of the original model equation.

The use of a detection filter for plant identification is a variant on the response error method. The error signal produced by a detection filter is a kind of response error -- the observed difference between the plant and model outputs when the same observed input is applied to both plant and model. The distinction between the detection filter method and the response error method is that the error signal from the detection filter is fed back into the model. This interpretation can be seen from the state equation for the detection filter

$$\begin{aligned} \dot{z}(t) &= (A - DC)\, z(t) + Dy(t) + Bu_d(t) \\ &= Az(t) + Bu_d(t) + D\big(y(t) - Cz(t)\big) \\ &= Az(t) + Bu_d(t) + D\epsilon'(t) \end{aligned} \tag{5-72}$$

where

$$\epsilon'(t) = y(t) - Cz(t) \tag{5-73}$$

is the observed or accessible error signal. Equation (5-72) represents a model of the plant with the error feedback term $D\epsilon'(t)$, as illustrated in Figure 5-1. If the detector gain D is made zero, then the error feedback would be eliminated and the result would be a true response error configuration. The effects of the error feedback on the identification process will become apparent as the detection filter method is compared with the other methods.

One advantage of the equation error method and its variants is a result of the fact that the error signal is an algebraic function of the parameters. Because of this fact, the effect of parameter

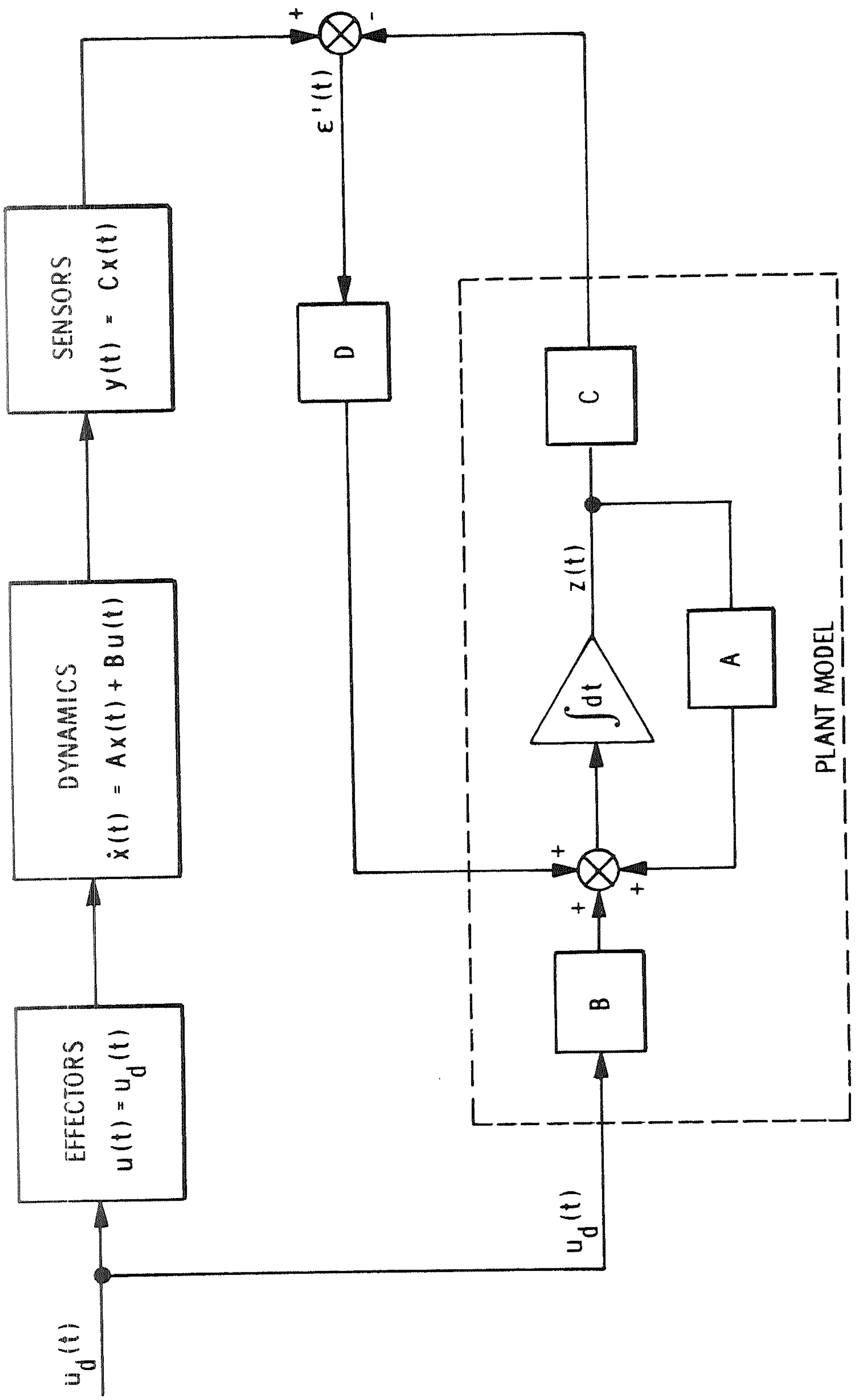

SENSORS
y(t) = Cx(t)
DYNAMICS
ẋ(t) = Ax(t) + Bu(t)
EFFECTORS
u(t) = u_d(t)
u_d(t)
ε'(t)
D
C
z(t)
∫dt
A
B
PLANT MODEL

Figure 5-1.

changes is reflected immediately in the error signal. This means that parameter adjustments can be made rapidly without destroying the validity of the error signal. On the other hand, the response error signal does not, in general, reflect accurately the effect of parameter changes instantaneously. If parameter changes in the model are made too rapidly without waiting for the effect to appear in the response error, the meaning of the response error becomes doubtful and stability problems may arise [12, 17].

The parameter adjustment law often involves partial derivatives of the error signal with respect to the parameters. In this case the above remarks can be made more specific. For the equation error signal the partial derivatives with respect to the parameters are true instantaneous partial derivatives (i.e., holding time constant). For the response error method such an interpretation is not appropriate because the error signal depends on past values of the parameters. The partial derivative of the response error with respect to a parameter is usually interpreted as a sensitivity function [12, 13, 22]. It is the relative change in the error trajectory over some finite time interval which would result if the parameter were subjected to an infinitesimal time-invariant change over that same time interval. This means that the parameters should be time-invariant during the time interval in which the partial derivatives (sensitivity functions) are being generated. This condition will be satisfied if the parameter adjustments are made at discrete points in time and the partial derivatives are generated in the intervening intervals. If the parameter adjustments are made continuously, they should be made slowly enough so that the parameters appear to be approximately time invariant compared to the response time of the model (which is comparable to the response time of the plant). The equation error method and its variants have no such theoretical limitation on speed of parameter adjustment.

As a variant on the response error method, the detection filter method also has a limitation on the speed of parameter adjustment. Specifically, the parameters should be adjusted slowly enough so that they appear approximately time invariant compared to the response time of the detdction filter. This limitation is much less restrictive than for the response error method. For the response error method the response time of the model is approximately the same as that of the plant (assuming the identification process is successful) and is determined by the eigenvalues of the matrix A in the plant representation (5-1). But the response time of the detection filter is determined by the eigenvalues of $G = (A - DC)$ which, as shown in Chapter 4, can be arbitrarily specified if the model is observable. This means that the response time of the detection filter can be made arbitrarily fast consistent with other practical considerations such as gain magnitudes and noise suppression. Therefore, the speed of parameter adjustment is not limited by the response time of the plant as in the case of the response error method.

These remarks can be made more specific by referring to Equation (5-22) with π_i and $\xi_i(t)$ defined by Equations (5-16) to (5-21). Recall the vector π_i represents the difference between model parameters and plant parameters. A similar equation obtains for the equation error method. (Equation (5-22) represents just one component of the vector error signal. Since all the references mentioned in this section deal only with identification of a scalar-input, scalar-output system, the remarks which follow will be specifically directed to that case. Then the error signal is a scalar as in (5-22). The remarks, however, generalize to the case of a multiple-input, multiple-output system.) For the equation error method, the equation corresponding to (5-22) is valid even if π_i is time varying. Recall for the detection filter method π_i was assumed time-invariant in obtaining (5-22). In a practical sense, then, the parameter adjustments should be made slowly enough so that (5-22) is a valid approximation for the observed error signal. Then the detection filter error signal will have approximately the same interpretation as in the case of the equation error method.

Although the equation error method has no theoretical limitation on the speed of parameter adjustment, it has been shown experimentally that increasing the speed of parameter adjustment beyond a certain level does not necessarily increase (and may decrease) the speed of convergence of the identification process [14]. It was noted above that the error signal for the equation error method can be expressed in a form similar to (5-22). Ideally the parameter adjustment process is intended to converge to the point $\pi_i = \underline{0}$ which, in the absence of sensor noise and plant disturbances, will null the error signal. However, at any time t, any vector π_i which is orthogonal to $\xi_i(t)$ will yield an error signal which is instantaneously zero. The set of all such π_i orthogonal to $\xi_i(t)$ at time t form a hyperplane of dimension $(n_\pi - 1)$ where n_π is the dimension of π_i. The hyperplane moves with time (but always contains the origin, $\pi_i = \underline{0}$) since $\xi_i(t)$ is a time-varying vector. Now if the parameter adjustments are made rapidly enough, the vector π_i could follow approximately the movement of the time-varying hyperplane. This means that π_i could remain near the moving hyperplane, thus producing an approximately nulled error signal without being close to the desired convergence point $\pi_i = \underline{0}$. Such behavior would retard the convergence of the identification process. It is only when π_i is unable to keep up with the motion of the hyperplane that it is forced to converge toward the origin as desired. Lion [14] has demonstrated that the speed of convergence can be substantially increased with the use of multiple generalized equations. Each generalized equation produces an error signal expressible in the form of (5-22). By introducing n_π independent generalized equations (where n_π is the dimension of π_i), n_π independent error equations in the form of (5-22) are obtained. In theory, this implies π_i can be solved for instantaneously (n_π equations, n_π unknowns). In practice, it means that π_i is forced to converge toward the origin regardless of how fast parameter adjustments are made, because there is no nonzero π_i which can null all n_π error signals simultaneously. Of course, this improved convergence is purchased at the expense of substantially increased complexity. Each independent generalized equation requires the equivalent of a plant model.

Assuming parameter adjustments are made slowly enough so that (5-22) is valid, the above remarks can be applied to the detection filter method also. Multiple detection filters, each with different dynamics, can be used to achieve the same effect that Lion has obtained with the use of multiple generalized equations. A similar increase in complexity is the price of the improved convergence. The speed of parameter adjustment is still limited by the response time of the detection filters.

It has been noted in the literature that for the equation error method, disturbances in the observed plant output vector (i.e., sensor noise) will produce an asymptotic bias in the estimate of the plant parameters [16,27]. The magnitude of the bias depends on the signal-to-noise ratio [27]. In the true response error method the estimate of the plant parameters is not biased if, as is often the case, the output of the model and the sensor noise are uncorrelated. In the case of a detection filter, the output error signal is fed back into the model through the detector gain. Hence, the output of the model will be correlated with the sensor noise, producing a bias in the parameter estimates. However, in this case the size of the bias depends on the detector gain as well as the signal-to-noise ratio. To see this, note that if the detector gain is reduced to zero the bias is reduced to zero, because the detection filter becomes simply a response error model.

Norkin [16] has suggested that the equation error method with its faster parameter adjustment potential would be more desirable for initial gross parameter estimates, and the slower but unbiased response error method would be more suitable for final fine tuning. This philosophy would be relatively easy to implement with a detection filter. Adjustment of the detector gain can produce a smooth transition from a fast detection filter with properties similar to the equation error method (i.e., fast parameter adjustment, biased by noise) to the response error method (with detector gain zero).

The purpose of this section has been to compare the potential of the detection filter method of identifying plant dynamics to other related methods. Various techniques for adjusting the model parameters were mentioned briefly. They have not been discussed in detail here because extensive literature already exists in this area. Representative parameter adjustment schemes may be found in [12, 14, 17, 22, 25, 27] as previously noted.

5.3 Identification of Effector and Sensor Failures by Correlation

This section discusses the problem of identifying the occurrence of effector or sensor failures in the presence of other disturbing influences. Consider a detection filter designed to detect the failure of any one of a set of r effectors associated with the vectors $\{b_1, \ldots, b_r\}$. Failure of the i^{th} effector of this set will produce a fixed-direction error signal from the detection filter as given by (4-397). If no other disturbances are acting on the plant or sensors, this error signal is easily identified with the failure of the i^{th} effector. As noted in Section 5.1, the fixed-direction error signal may be obscured by other disturbances such as sensor or plant noise, uncertainties in plant dynamics, and failures the filter is not designed to detect. These extraneous errors in general will not have a fixed direction in the output space. If the fixed-direction error signal makes up a significant portion of the total error, one would expect the error vector to be biased toward that direction. One way of identifying such a directional bias is to form a correlation matrix

$$R(t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \epsilon'(t)\, \epsilon'(t)^T \, dt \tag{5-73}$$

over some time interval $[t_1, t_2]$. $R(t_1, t_2)$ is an $m \times m$ positive semi-definite matrix. It is helpful to associate this matrix with an ellipsoid in m-dimensional Euclidean space. The ellipsoid is defined as the set of all m-vectors y such that $y^T \eta \leq 1$ for any m-vector η satisfying $\eta^T R \eta = 1$. This defines an ellipsoid centered at the origin and having principal axes along the eigenvector directions of $R(t_1, t_2)$ with length equal to twice the corresponding eigenvalues. If $R(t_1, t_2)$ is singular, the ellipsoid will be degenerate, i.e., one or more principal axes will have zero length. When $\epsilon'(t)$ maintains an exact fixed direction over $[t_1, t_2]$, the ellipsoid consists of a single straight line. If other disturbances are present, the additional error signals will fill out the ellipsoid by producing nonzero principal axes in other directions. Because the fixed-direction error signal has all its power concentrated in a single vector direction the ellipsoid will tend to be cigar-shaped with a dominant principal axis in that direction. A scheme for identifying effector failures in the presence of other disturbances is to look for a dominant axis ellipsoid with the major axis near a direction associated with an effector failure (i.e., the direction of $\epsilon'(t)$ in (4-397)). Since the failure directions are known, it is not necessary to analyze completely the shape of the error correlation ellipsoid. It is sufficient to simply check for a dominant axis in one of the known directions. If the failure directions are linearly independent, one way of doing this is to transform the error signal to a coordinate frame where the effector failure directions are along orthogonal coordinate axes. Then a particular effector failure would be indicated by a single large diagonal element in $R(t_1, t_2)$ relative to all the other elements.

The correlation matrix can be used in a similar way to identify sensor failures. The error response to a sensor failure is restricted to a two-dimensional plane. In this case one would expect an error ellipsoid having two dominant axes, i.e., having the shape of a pancake.

CHAPTER 6

FEEDBACK RESTRUCTURING

6.1 General Discussion

After an event as described in Section 3.2 is detected and identified, the next problem is to restructure the system to compensate for it. For the system configuration shown in Figure 3-1, the restructuring takes place in the feedback loop. The plant, which includes effectors, sensors, and plant dynamics, is assumed to be inaccessible for restructuring. This means that effectors and sensors are considered nonrepairable. When the decision is made that an effector or sensor has failed, two courses of action are possible. One is to continue to use the failed component with some appropriate compensation for its irregular behavior. The second possibility is to remove the component from further use and restructure the feedback loop to function without it. The first course of action in general requires more precise information or some a priori assumptions about the nature of the failure in order to determine the appropriate compensation. In the latter course of action knowledge of the exact nature of the failure is not necessary. It is only necessary to identify the failed component. This chapter will be concerned with the second "surgical" restructuring method. Failed effectors and sensors are removed from service and restructuring compensates for the reduction in active components. Some attention has been given to the nonsurgical method. Chien [5] has used this approach in dealing with failures in redundant gyro arrays.

When a malfunctioning gyro has been detected (by a sophisticated method of comparison of redundant information), it is removed from service temporarily while the malfunction is investigated further to determine if it is possible to compensate for it (e.g., a biased sensor output can be compensated for if the bias can be determined). If compensation is possible, the gyro is returned to service after the appropriate compensation has been implemented.

The feedback loop consists of two basic functional parts -- the state-estimating filter and the state feedback law generator. If these two parts are designed independently of each other (Section 3.2 describes the separation philosophy), the restructuring problem for each part may also be considered independently. This leads to some simplification because some events may require restructuring of only one part of the feedback loop. Another part of a self-reorganizing system which may require restructuring is the detection filters. It is of interest to note the types of restructuring required by each type of event.

1) An effector failure requires restructuring of the feedback law only.

2) A sensor failure requires restructuring of both the detection and state-estimating filters. It may or may not require changes in the feedback law depending on the changes made in the plant model. The only necessary change in the plant model is to delete from the C matrix the row corresponding to the failed sensor. In this case only the detection and state-estimating filters need be restructured for the reduced number of sensor outputs. If the plant model is to be kept in the standard form of Section 4.3.6, a coordinate transformation of the

state space will be necessary in addition to deletion of the appropriate row of the C matrix. In this case the same transformation must be applied to the feedback law.

3) Changes in plant dynamics may, in general, require restructuring of the detection filters, state-estimating filter, and the feedback law. The detection filter which identifies the plant dynamics is automatically adjusted in the process of identification, so it does not require any further restructuring. The extent of restructuring necessary in the other detection filters (for effector and sensor failures) and the state-estimating filter depends on where the changes in plant dynamics appear. For purposes of the following discussion, detection filters for sensor failures and detection filters for effector failures are referred to separately because they have different restructuring requirements. In fact, one filter may detect both sensor and effector failures, in which case the restructuring requirements include both those necessary for sensor failure detection and those for effector failure detection. Changes in the B matrix of the plant state equation

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{6-1}$$

require simple adjustments in the state-estimating filter and detection filters for sensor failures. For these filters it is only necessary to adjust the filter state equation

$$\dot{z}(t) = (A - DC)\,z(t) + Dy(t) + Bu_d(t) \tag{6-2}$$

by replacing the old B matrix with the new one. Detection filters for effector failures may require more extensive restructuring because detection orders, detection generators, and the mutual detectability of the columns of the new B matrix may be different. When the change in plant dynamics occurs in the A matrix, the restructuring will be simplest if A and C are in the standard form suggested in previous chapters (e.g., Equations (5-4) to (5-10)). In this case the changed matrix $(A + \Delta A)$ can be expressed as

$$A + \Delta A = A + \Delta A C^T C \tag{6-3}$$

because the changes in A occur only in the last column of each block of A in (5-4). Note from (6-3) that $(A + \Delta A)$ has the form $(A - D''C)$ with $D'' = -\Delta A C^T$. As noted in Chapter 4, detection filter properties (detection orders, detection generators, mutual detectability, etc.) are invariant with respect to replacement of A by $(A - D''C)$ for any D''. Furthermore, if the event vectors (e.g., the columns of B for effector failure detection) are unchanged, it is necessary to make only a simple adjustment in the detector gain D to keep $G = (A - DC)$ unchanged. Specifically, the adjustment ΔD in D is taken to be

$$\Delta D = \Delta A C^T \tag{6-4}$$

Then

$$\begin{aligned}(A + \Delta A) - (D + \Delta D)C &= (A + \Delta A C^T C) - (D + \Delta A C^T)C \\ &= A - DC = G\end{aligned} \tag{6-5}$$

With this adjustment in the detector gain, G remains unchanged and the detection filter detects the same event vectors that it did before the change in A. Therefore, if the columns of B do not change, the adjustment given by (6-4) is all that is necessary for the effector failure detection filters. Detection filters designed to detect sensor failures may require more extensive restructuring. Recall that in order to detect a failure in the i^{th} sensor, a filter must detect d_i, the i^{th} column of the detector gain matrix D. If this column is changed by the adjustment ΔD given by (6-4), then the filter may no longer detect the i^{th} column of the new detector gain matrix. In this case the filter must be partially redesigned so that the filter does detect the new d_i. If the state-estimating filter has the same Kalman-type configuration as a detection filter (Figure 5.1), then the simplest and fastest way to compensate for changes in A is to adjust the feedback gain D by the amount given in (6-4). This adjustment will keep the poles of the filter unchanged and thus guarantee stability, at least. Of course, this may not be the best filter for noise suppression. If the adjustment given by (6-4) increases the feedback gains, then the effect of sensor noise on the state estimate will be increased. If the original filter was statistically optimum (Kalman filter), the adjusted filter will not be optimum in general. If a new Kalman filter is desired, then a Riccati equation must be solved in whole or in part to obtain the new feedback gains. But whatever kind of restructuring is used in the state-estimating filter, the adjustment in the feedback gain matrix given by (6-4) is a quick, simple way to ensure filter stability. It can at least be used as a temporary measure until more sophisticated restructuring can be implemented where necessary.

Beyond the simple adjustments discussed above, the redesign or restructuring of detection filters is a matter of implementing the theory in Chapter 4 with the algorithms presented in Appendices A, B, and C. It has been noted previously that a detection filter can also serve as a state-estimating filter. (In Chapter 4 it was shown that the state of a detection filter approaches the state of the plant asymptotically in the absence of disturbances.) If the state estimate for feedback control is taken from one or more detection filters, then the problem of restructuring a state-estimating filter is taken care of automatically in the restructuring of the detection filter. Even if a separate state-estimating filter is used, detection filter theory can be applied to the restructuring of a state-estimating filter in order to specify its pole locations. As noted above, if a true Kalman filter is desired, it will be necessary to resolve a Riccati equation. If this is attempted, detection filter design algorithms can be used for intermediate restructuring (pole assignment) to serve until a new optimal solution is obtained. For these reasons restructuring of a state-estimating filter will not be considered separately.

The remainder of this chapter will be devoted to restructuring of feedback control law for the primary purpose of maintaining stability of the closed loop system. For reasons stated in the next section, the feedback control to be considered is a linear time-invariant state feedback law of the form

$$u_d(t) = L\hat{x}(t) + L_c c(t) \tag{6-6}$$

where L and L_c are time-invariant matrices of dimension $r \times n$ and $r \times r_c$ respectively. Section 6.2 discusses the linear state feedback control problem and shows how the detection filter theory in Chapter 4 can be applied in dual form to produce some interesting designs for linear state feedback control. Section 6.3 discusses several algorithms for generating a linear state feedback control law. Two of these algorithms implement the feedback designs in Section 6.2.

6.2 Detection Results Applied to State Feedback Control

When a change or failure occurs in a system, the primary immediate concern is usually to achieve stability as quickly as possible. The central focus of the remainder of this chapter will be the restructuring of the feedback law to achieve closed-loop stability for the system shown in Figure 3-1. The linear time-invariant state feedback law given by (6-6) is particularly suited for this purpose. It is one of the more widely used feedback laws. The optimal solution to the infinite interval regulator problem is such a feedback law (without the command input $c(t)$). In addition, this law yields a linear time-invariant closed-loop system whose stability properties are well defined and can be determined analytically. Even if the original and final restructured feedback laws are not of the form of (6-6), the linear constant form can still serve as a temporary law to maintain stability while a more sophisticated law is derived. Therefore, (6-6) is a reasonable starting point for the development of restructuring methods.

It will be assumed that the detection filters have identified the plant dynamics, and any failed effectors or sensors have been detected and removed from service. The information at hand is an up-to-date description of the plant

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{6-7}$$

$$u(t) = u_d(t) \tag{6-8}$$

$$y(t) = Cx(t) \tag{6-9}$$

The restructuring problem to be considered here is to develop methods for selecting the L matrix in (6-6) so that the closed-loop system

$$\dot{x}(t) = Ax(t) + BL\hat{x}(t) + BL_c\, c(t) \tag{6-10}$$

is at least stable (if that is possible).

If the state-estimating filter dynamics are given by

$$\dot{\hat{x}}(t) = (A - DC)\,\hat{x}(t) + Bu_d(t) + Dy(t) \tag{6-11}$$

then the state error

$$\epsilon(t) = x(t) - \hat{x}(t) \tag{6-12}$$

obeys the equation

$$\dot{\epsilon}(t) = (A - DC)\,\epsilon(t) \tag{6-13}$$

Then (6-6) can be written as

$$u_d(t) = Lx(t) - L\epsilon(t) + L_c\, c(t) \tag{6-14}$$

and the complete closed-loop system dynamics are given by

$$\begin{bmatrix} \dot{x}(t) \\ \dot{\epsilon}(t) \end{bmatrix} = \begin{bmatrix} (A + BL) & -BL \\ \underline{0} & (A - DC) \end{bmatrix} \begin{bmatrix} x(t) \\ \epsilon(t) \end{bmatrix} + \begin{bmatrix} BL_c \\ \underline{0} \end{bmatrix} c(t) \tag{6-15}$$

The poles of the complete closed-loop system are given by the eigenvalues of the matrices $(A + BL)$ and $(A - DC)$. The eigenvalues of $(A - DC)$ are the poles of the state-estimating filter. Restructuring of the state-estimating filter was discussed in the previous section. Assuming this restructuring is successful, the eigenvalues of $(A - DC)$ are known to be stable, so the stability of the closed-loop system depends on the eigenvalues of $(A + BL)$. Furthermore, in the absence of disturbances the state error satisfying (6-13) will settle to zero, and the closed-loop system dynamics reduce to

$$\dot{x}(t) = (A + BL)\,x(t) + BL_c\,c(t) \tag{6-16}$$

The restructuring problem may now be simplified to the problem of choosing L so that the system given by (6-16) is stable. Note that L_c does not affect the stability of the closed-loop system, so it is of secondary concern in the restructuring problem. Of course, L_c does not affect the dynamic response of the system to the command $c(t)$. One way of selecting L_c is discussed in Section 6.2.1.

The problem of selecting L to control the dynamics of (6-16) is related by duality to certain aspects of detection filter design. The problem of choosing L to obtain stable eigenvalues for $(A + BL)$ is the dual to the problem of choosing L^T to obtain stable eigenvalues for $(A + BL)^T = (A^T + L^T B^T)$. Selecting L^T to specify eigenvalues of $(A^T + L^T B^T)$ is one of the considerations in detection filter design. In the notation of Chapter 4, A^T, B^T, and L^T correspond to A, C, and $-D$. Since $(A + BL)$ and $(A + BL)^T$ have identical eigenvalues, some results of Chapter 4 are immediately applicable to the feedback restructuring problem. From Lemma 4.4 it can be concluded that by choice of L it is possible to specify arbitrarily exactly κ eigenvalues of (A + BL) where

$$\kappa = \mathrm{rk}[B, AB, \ldots, A^{n-1}B] \tag{6-17}$$

If $\kappa < n$ (A is $n \times n$), the remaining $(n - \kappa)$ eigenvalues of (A + BL) are always equal to corresponding eigenvalues of A and are not influenced by any choice of L. The methods developed in Chapter 4 for finding a detector gain can be applied in their dual form to the problem of selecting L to specify eigenvalues of (A + BL). The design of detection filters involves more than just stability and specification of eigenvalues. The special properties of detection filters and the concept of sensor decoupling in Chapter 4 have interesting dual interpretations in the context of linear feedback control. For the reader's information these interpretations are discussed in Sections 6.2.1 and 6.2.2. It should be repeated, however, that the first objective in feedback restructuring is to generate as quickly as possible a feedback matrix L which ensures stability of the closed-loop system. Hence, the subject of primary concern is the computation involved in the algorithms for generating L. As will be seen in Section 6.3, algorithms based on detection filter theory usually require more computation than algorithms which are concerned solely with ensuring stable closed-loop poles.

6.2.1 Construction of Scalar-Input, Scalar-Output Subsystems by State Feedback

In dual form the basic results for detection filter design in Section 4.3.1 show how it is possible through state feedback to obtain scalar-input control over a scalar output of the plant. It is easiest to explain scalar-input control in terms of Laplace transforms. Let

$$v_h(t) = h^T x(t) \tag{6-18}$$

be the scalar output of interest, where h is a time-invariant n-vector. The simplified closed-loop system dynamics given by (6-16) may rewritten as

$$\dot{x}(t) = (A + BL)\, x(t) + Bu_{dc}(t) \tag{6-19}$$

where

$$u_{dc}(t) = L_c\, c(t) \tag{6-20}$$

is that portion of the desired control signal which is due to the command signal c(t). Then the transfer from control signal to output in Laplace transforms is

$$V_h(s) = h^T [Is - (A + BL)]^{-1} BU_{dc}(s) \tag{6-21}$$

where

$$V_h(s) = \mathcal{L}\{v_h(t)\} \tag{6-22}$$

$$U_{dc}(s) = \mathcal{L}\{u_{dc}(t)\} \tag{6-23}$$

The right side of (6-21) can be expanded to yield

$$V_h(s) = \sum_{i=1}^{r} F_i(s)\, U_{dci}(s) \tag{6-24}$$

where $U_{dci}(s)$ is the i^{th} component of $U_{dc}(s)$ and

$$F_i(s) = h^T[Is - (A + BL)]^{-1} b_i \tag{6-25}$$

with b_i the i^{th} column of B. $F_i(s)$ is a scalar rational function of s representing the closed-loop transfer from the i^{th} component of the control vector to the output. In general, the $F_i(s)$ are different and the complete control vector must be known in order to determine the output. Suppose, however, the $F_i(s)$ differ by only a constant, e.g.,

$$F_i(s) = F(s)\, \eta_i\, U_{dci}(s)$$

$$= F(s)\, \eta^T\, U_{dc}(s) \tag{6-27}$$

where

$$\eta = \begin{bmatrix} \eta_1 \\ \cdot \\ \cdot \\ \cdot \\ \eta_r \end{bmatrix} \tag{6-28}$$

In this case the output $v_h(t)$ does not depend on the full control vector $u_{dc}(t)$, but only on the linear scalar function $\eta^T\, u_{dc}(t)$. This situation will be referred to as scalar control of $v_h(t)$.

Comparison of (6-24) and (6-27) makes it clear that scalar control yields a simpler input-output transfer function. In effect a multiple-input, scalar-output relationship is reduced to a scalar-input, scalar-output relationship. Furthermore, the fact that $v_h(t)$ depends only on a scalar, linear combination of the components of $u_{dc}(t)$ implies that there is freedom left in $u_{dc}(t)$ to perform additional control actions without disturbing $v_h(t)$. For example, suppose L_c is selected so that all its columns except the first are orthogonal to η. Let η and the first column of L_c have an inner product of one. Then

$$\eta^T L_c = [1, 0, \ldots, 0] \tag{6-29}$$

and

$$\eta^T u_{dc}(t) = \eta^T L_c c(t) = [1, 0, \ldots, 0]\, c(t) = c_1(t) \tag{6-30}$$

where $c_1(t)$ is the first component of $c(t)$. This result shows that $v_h(t)$ responds only to the first component of $c(t)$. It is not influenced by any other component of the command signal. Since η and the columns of L_c are r-vectors, L_c can have as many as $(r - 1)$ independent columns which are orthogonal to η. Suppose $c(t)$ is an r-vector $(r_c = r)$ and L_c is chosen to satisfy (6-30) with all columns of L_c independent (L_c is $r \times r$). Then the command components $\{c_2(t), \ldots, c_r(t)\}$ can produce $(r - 1)$ independent control actions, none of which affect the output $v_h(t)$.

The scalar control property is the dual to the fixed-direction error property of a detection filter. The results of Chapter 4 show that for any controllable output of the form of (6-18) (i.e., for any h not lying in the uncontrollable space of B) it is always possible to find an L which achieves scalar control. The dual of Theorem 4.1 shows that in addition to obtaining scalar control, all the eigenvalues of $(A + BL)$ can be almost arbitrarily specified if (A, B) is a controllable pair. If (A, B) is not controllable then κ eigenvalues can be specified where κ is given by (6-17). This follows from remark 4) in Section 4.3.1. In other words scalar control can be achieved while still maintaining control over the maximum number of eigenvalues of $(A + BL)$. This result is most easily verified by considering the transpose of the transfer function in (6-21)

$$\left[h^T[Is - (A + BL)]^{-1}B \right]^T = B^T[Is - (A^T + L^TB^T)]^{-1} h \tag{6-31}$$

Let A^T, B^T, L^T, and h correspond to A, C, $-D$, and f of Section 4.3.1. Let ν be the detection order of h with respect to (A^T, B^T) and let g be its maximal generator. If L satisfies the equation

$$- L^TB^T[A^T]^{\nu-1} g = p_1 g + \ldots + p_{\ -1}[A^T]^{\nu-1} g + \{A^T]^{\nu} g \tag{6-32}$$

and $B^Th \neq \underline{0}$, then

$$B^T[Is - (A^T + L^TB^T)]^{-1} h = B^ThF(s) \tag{6-33}$$

with

$$F(s) = \frac{s^{\nu-1} + \alpha_{\nu-1} s^{\nu-2} + \ldots + \alpha_1}{s^{\nu} + p_{\nu}s^{\nu-1} + \ldots + p_1} \tag{6-34}$$

where the α_i are determined by the relation

$$h = \alpha_1 g + \ldots + \alpha_{\nu-1}[A^T]^{\nu-2} g + [A^T]^{\nu-1} g \tag{6-35}$$

and the p_i are arbitrary. Transposing (6-33), it is clear that for L satisfying (6-32), (6-21) reduces to the form of (6-27) with

$$\eta^T = h^T B \tag{6-36}$$

In general if

$$\left.\begin{aligned} B^T[A^T]^j\, h &= \underline{0} \qquad j = 1, \ldots, \mu - 1 \\ B^T[A^T]^{\mu}\, h &\neq \underline{0} \end{aligned}\right\} \tag{6-37}$$

then

$$h = \alpha_1 g + \ldots + \alpha_{\nu-\mu-1}[A^T]^{\nu-\mu-2} g + [A^T]^{\nu-\mu-1} g \tag{6-38}$$

$$F(s) = \frac{s^{\nu-\mu-1} + \alpha_{\nu-\mu-1} s^{\nu-\mu-2} + \ldots + \alpha_1}{s + p\, s^{\nu-1} + \ldots + p_1} \tag{6-39}$$

$$h^T[Is - (A + BL)]^{-1} B = h^T A^{\mu} B \;\; F(s) \tag{6-40}$$

and

$$\eta^T = h^T A^{\mu} B \tag{6-41}$$

If (6-37) is not satisfied for any μ, then h lies in the uncontrollable space of B, and $v_h(t)$ is not controllable regardless of L.

The results of Sections 4.3.2, 4.3.3, and 4.3.4 are applicable to feedback design for control of multiple outputs. Consider an ℓ-dimensional output vector

$$v_H(t) = Hx(t) \tag{6-42}$$

where H is an $\ell \times n$ time-invariant matrix

$$H = \begin{bmatrix} h_1^T \\ \cdot \\ \cdot \\ \cdot \\ h_\ell^T \end{bmatrix} \tag{6-43}$$

If $\ell \leq r$ and the set of vectors $\{h_1, \ldots, h_\ell\}$ are output separable with respect to (A^T, B^T) (Definition 4.9), it is possible to find a feedback gain L which produces a closed loop transfer of the form

$$V_H(s) = \begin{bmatrix} F_{11}(s) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & F_{\ell\ell}(s) \end{bmatrix} H' \, BU_{dc}(s) \tag{6-44}$$

where

$$H' = \begin{bmatrix} h_1^T A^{\mu_1} \\ \cdot \\ \cdot \\ \cdot \\ h_\ell^T A^{\mu_\ell} \end{bmatrix} \tag{6-45}$$

with the μ_i defined by condition (6-37) for each h_i. The $F_{ii}(s)$ are scalar rational functions of S of the form of (6-39). If the h_i are mutually detectable with respect to (A^T, B^T) κ eigenvalues of $(A + BL)$ can be specified almost arbitrarily. If the h_i are not mutually detectable, control over certain eigenvalues will be lost in achieving (6-44). Such uncontrolled eigenvalues can be identified as described in Section 4.3.4.

Now let $c(t)$ be an ℓ-vector and choose L_c to be a solution of

$$H' BL_c = I \tag{6-46}$$

This equation always has a solution for L_c because the h_i are output separable, which implies $\mathrm{rk}\,[H'^T B] = \ell$. One solution is

$$L_c = B^T H'^T [H' BB^T H'^T]^{-1} \tag{6-47}$$

The inverse exists because $\mathrm{rk}[H'B] = \ell$. With this L_c and the Laplace transform of (6-20), (6-44) becomes

$$V_H(s) = \begin{bmatrix} F_{11}(s) & \cdots & 0 & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & F_{\ell\ell}(s) \end{bmatrix} C(s) \tag{6-48}$$

where

$$C(s) = \mathcal{L}\{c(t)\} \tag{6-49}$$

Or in component form

$$V_{Hi}(s) = F_{ii}(s)\, C_i(s) \tag{6-50}$$

This means that each component of $v_H(t)$ is controlled exclusively by the corresponding component of $c(t)$. A multiple-input, multiple-output system has thus been reduced to a set of scalar-input, scalar-output subsystems.

6.2.2 Effector Decoupling

The concept of output decoupling introduced in Section 4.3.6 has a dual interpretation which leads to the idea of effector decoupling. The results on output decoupling are presented in this section in their dual form and interpreted in the context of linear state feedback control. The main reason for discussing this material is to call the reader's attention to the interesting dual interpretations of previous results. A second reason is that the algorithm for generating effector decoupling feedback control is somewhat simpler and more generally applicable than the algorithm necessary to implement the scalar-input, scalar-output control described in Section 6.2.1, as will be seen in Section 6.3.

Loosely speaking, effector decoupling means that individual effectors control independent parts of the system. The following two definitions formalize the concept of effector decoupling.

Definition 6.1. The system described by (6-19) is defined to be effector decoupled if the controllable space of each b_i (the i^{th} column of B) does not intersect the controllable space of any other column.

Definition 6.2. The matrix pair (A, B) is defined to be effector decouplable if there exists some feedback gain matrix L such that the closed-loop system (6-19) is effector decoupled.

The dynamic behavior of an effector decoupled system is best illustrated by transforming the state space to a special coordinate frame. The transformation can be generated by using the dual form of the algorithm of Appendix C. The same result is obtained if the algorithm as given is applied to the transposed matrix pair $((A + BL)^T, B^T)$. The transformed matrices have the form

$$\overline{(A + BL)} = T^{-1}(A + BL)T = \begin{bmatrix} P_1^T & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \ddots & & \vdots \\ \vdots & & \ddots & \underline{0} \\ \underline{0} & \cdots & \underline{0} & P_r^T \end{bmatrix} (n \times n) \tag{6-51}$$

with

$$P_i^T = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & \cdots & \cdots & 0 & 1 \\ -p_1 & \cdots & \cdots & \cdots & -p_{k_i} \end{bmatrix} (\kappa_i \times \kappa_i) \tag{6-52}$$

and

$$\overline{B} = T^{-1}B = \begin{bmatrix} \overline{b}_{11} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \ddots & & \vdots \\ \vdots & & \ddots & \underline{0} \\ \underline{0} & \cdots & \underline{0} & \overline{b}_{rr} \end{bmatrix} (n \times r) \tag{6-53}$$

with

$$\overline{b}_{ii} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} (\kappa_i \times 1) \tag{6-54}$$

where κ_i is the dimension of the controllable space of b_i with respect to $(A + BL)$. The block diagonal form of (6-51) is a result of the fact that the controllable spaces for the b_i are all nonintersecting. If the transformed state vector is partitioned to conform with the blocks in (6-51)

$$\bar{x}(t) = T^{-1}x(t) = \begin{bmatrix} \bar{x}_1(t) \\ \cdot \\ \cdot \\ \cdot \\ \bar{x}_r(t) \end{bmatrix} \tag{6-55}$$

then the equation for each decoupled subsystem is

$$\dot{\bar{x}}_i(t) = P_i^T \bar{x}_i(t) + \bar{b}_{ii} u_{dci}(t) \tag{6-56}$$

The form of (6-51) assumes that $(A + BL, B)$ (or equivalently (A, B)) is a controllable pair. If (A, B) is not controllable, the controllable portion of the system can be isolated by applying the dual form of the transformation used in Lemma 4.4. Then the above transformation can be applied to the controllable portion. The general form in this case is

$$\overline{(A + BL)} = \begin{bmatrix} P_1^T & \underline{0} & \cdots & \underline{0} & R_1 \\ \underline{0} & \ddots & & \vdots & \vdots \\ \vdots & & \ddots & \underline{0} & \vdots \\ \vdots & & & P_r^T & \vdots \\ \underline{0} & \cdots & \cdots & \underline{0} & R_{r+1} \end{bmatrix} \tag{6-57}$$

$$\overline{B} = \begin{bmatrix} \overline{b}_{11} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \ddots & & \vdots \\ \vdots & & \ddots & \underline{0} \\ \underline{0} & \cdots & \underline{0} & \overline{b}_{rr} \\ \underline{0} & \cdots & \cdots & \underline{0} \end{bmatrix} \tag{6-58}$$

with $\overline{b}_{ii}$ given by (6-54) and P_i^T by (6-52). The R_i are associated with the uncontrollable portion of the system.

The results of Section 4.3.6 concerning output decouplable systems can be applied in their dual forms to the study of effector decoupling. The following definition is the dual of Definition 4.12 for output decoupling order.

Definition 6.3. The effector decoupling order of b_i, the i^{th} column of B, is defined to be the smallest positive integer value of j such that

$$\mathrm{rk}[B, AB, \ldots, A^{j-1}B, A^j b_i] = \mathrm{rk}[B, AB, \ldots, A^{j-1} B] \tag{6-59}$$

It is clear that decoupling order is invariant under coordinate transformations, since the ranks of the matrices in (6-59) are so invariant. It was noted in Section 4.3.6 that output decoupling order is invariant under replacement of A by $(A + D''C)$ for any D''. In the present context this means that effector decoupling order is invariant under replacement of A by $(A + BL)$ for any L. Note that for the decoupled system given by (6-57) to (6-58) the effector decoupling order of the i^{th} column of $\overline{B}$ is κ_i and

$$\kappa_1 + \ldots + \kappa_r = \kappa \tag{6-60}$$

where

$$\kappa = \mathrm{rk}\,[\,\overline{B},\ \overline{(A+BL)}\,\overline{B},\ \ldots,\ \overline{(A+BL)}^{n-1}\,\overline{B}\,] \tag{6-61}$$

By invariance under coordinate transformations the effector decoupling order of b_i must likewise be κ_i. Further,

$$\begin{aligned}
\kappa &= \mathrm{rk}[\,\overline{B},\ \overline{(A+BL)}\,\overline{B},\ \ldots,\ \overline{(A+BL)}^{n-1}\,\overline{B}\,] \\
&= \mathrm{rk}[\,B,\ (A+BL)B,\ \ldots,\ (A+BL)^{n-1}\,B] \\
&= \mathrm{rk}[\,B,\ AB,\ \ldots,\ A^{n-1}\,B]
\end{aligned} \tag{6-62}$$

This is true for any L and follows from the dual of (4-87).

Now if (A, B) is effector decouplable, there exists some L which produces a decoupled closed-loop system. Since condition (6-60) holds for the decoupled system, it must hold for the pair (A, B) as well by virtue of the invariance properties of the κ_i and κ. This means that a necessary condition for (A, B) to be effector decouplable is that the sum of the decoupling orders of all the b_i must be equal to the dimension of the controllable space of B. This condition can be shown to be sufficient by transforming to a standard form. If the above condition holds, the dual of the transformation in Appendix C will transform A and B into the form

$$\overline{A} = T^{-1}AT = \begin{bmatrix} \overline{A}_{11} & \cdots & \overline{A}_{1r} & R_1 \\ \vdots & & \vdots & \vdots \\ \overline{A}_{r1} & \cdots & \overline{A}_{rr} & \vdots \\ \underline{0} & \cdots & \underline{0} & R_{r+1} \end{bmatrix} (n \times n) \tag{6-63}$$

with

$$\bar{A}_{ii} = \begin{bmatrix} 0 & 1 & & & \underline{0} \\ \underline{0} & & \ddots & & \\ & & & \ddots & 1 \\ a_{ii1} & \cdots & \cdots & \cdots & a_{ii\kappa_i} \end{bmatrix} (\kappa_i \times \kappa_i)$$

(6-64)

$$\bar{A}_{ij} = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ a_{ij1} & \cdots & a_{ij\kappa_i} & 0 & \cdots & 0 \end{bmatrix} (\kappa_i \times \kappa_j)$$

(6-65)

$$\bar{A}_{ji} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ a_{ji1} & \cdots & a_{ji\kappa_i} \end{bmatrix} (\kappa_j \times \kappa_i)$$

(6-66)

where $\kappa_j \geq \kappa_i$ and $\bar{B}$ is given by (6-58). It is easy to see now that the decoupled form (6-57) can be obtained from (6-63) to (6-66) by choosing

$$\bar{L} = \begin{bmatrix} \bar{l}_{11} & \cdots & \bar{l}_{1r} & \bar{l}_{1,r+1} \\ \vdots & & \vdots & \vdots \\ \bar{l}_{r1} & \cdots & \bar{l}_{rr} & \bar{l}_{r,r+1} \end{bmatrix} \qquad (6\text{-}67)$$

with

$$\bar{l}_{ii} = [-p_{i1} - a_{ii1}, \ldots, -p_{i\kappa_i} - a_{ii\kappa_i}] \qquad (6\text{-}68)$$

$$\bar{l}_{ij} = [-a_{ij1}, \ldots, -a_{ij\kappa_i}, 0, \ldots, 0] \qquad (6\text{-}69)$$

$$\bar{l}_{ji} = [-a_{ji1}, \ldots, -a_{ji\kappa_i}] \qquad (6\text{-}70)$$

for i, $j = 1, \ldots, r$. The $\bar{l}_{i,r+1}$ for $i = 1, \ldots, r$ are arbitrary. These observations establish the following theorem.

Theorem 6.1. The matrix pair (A, B) is effector decouplable if and only if

$$\kappa_1 + \ldots + \kappa_r = \kappa \qquad (6\text{-}71)$$

where κ_i is the effector decoupling order of b_i and κ (given by (6-56)) is the dimension of the controllable space of B.

This theorem is the dual of Theorem 4.7 with a slight generalization to include noncontrollable systems.

In Section 4.3.6 it was shown that a system representation could be enlarged to obtain a decouplable form. Such enlargement is not appropriate here. It was noted in Section 4.3.6 that the added portion of the representation would not be controllable. In this situation the added portion of the system would not be observable. But obtaining an effector decoupled system depends on state feedback. State feedback in turn depends on knowing the state of the system. Nothing is gained by enlarging the representation, because there will be no information available about the state of the added portion of the representation, which is unobservable.

The transformation of Appendix C was convenient for establishing Theorem 6.1, but in practice it is not necessary to apply this transformation to find a feedback gain which produces a decoupled system. In the next section various algorithms for generating feedback gains will be discussed. Among them is an algorithm which produces a decoupled closed-loop system when the open-loop system is decouplable. If the system is not decouplable, the algorithm will achieve decoupling for as many effectors as possible.

6.3 Algorithms for Generating State Feedback Gains

This section discusses three algorithms for generating constant linear state feedback gains. They all have the capability for achieving the primary goal stated at the beginning of this chapter, namely closed-loop stability for the controllable portion of the system. More specifically, any of the algorithms can be used to specify almost arbitrarily all of the closed-loop poles of the controllable portion of the system. The algorithms differ in two respects. First, the computational requirements for implementing them are different. Second, the closed-loop systems they produce will have different structural characteristics, i.e., the structure of subsystems and dynamic coupling among them.

The first algorithm is simply the dual of the method developed in Chapter 4 (and Appendix A) for generating a detector gain. The structure of the resulting closed-loop system is described in Section 6.2.1. Central attention is focused on a set of outputs as given by (6-42). In the closed-loop system each component of the output is independently controlled by a scalar input. The amount of computation involved in implementing this algorithm is evident from the step-by-step outline in Section 4.3.5. As will be seen later, it appears that this algorithm is the most time-consuming of the three if separability and mutual detectability must be investigated.

The second algorithm produces the effector decoupling described in Section 6.2.2. If the system is decouplable the resulting closed-loop system will be fully effector decoupled. This algorithm is based on the same orthogonal reduction procedure used in Appendix A. The general procedure and its properties are fully described in Appendix A. Only a brief review specialized to the present situation will be presented here. Basically the null space of a symmetric positive semi-definite matrix is sequentially enlarged to contain the vectors from an ordered set. In this case the columns of the matrix

$$W = [B, AB, \ldots, A^{n-1} B] \tag{6-72}$$

taken from left to right form the ordered set of vectors. The procedure begins with any $n \times n$ symmetric positive definite matrix Ω_{11} (the identity matrix is a simple choice). An auxiliary vector is defined by

$$w_{11} = \Omega_{11} b_1 \tag{6-73}$$

This vector is nonzero because b_1 is nonzero and Ω_{11} is positive definite. A new symmetric positive semi-definite matrix which contains b_1 in its null space is defined by

$$\Omega_{21} = \Omega_{11} - \frac{w_{11} w_{11}^T}{w_{11}^T b_1} \tag{6-74}$$

The next auxiliary vector is

$$w_{21} = \Omega_{21} b_2 \tag{6-75}$$

and if $w_{21} \neq \underline{0}$ a new symmetric positive semi-definite matrix which contains both b_1 and b_2 in its null space is

$$\Omega_{31} = \Omega_{21} - \frac{w_{21} w_{21}^T}{w_{21}^T b_2} \tag{6-76}$$

If $w_{21} = \underline{0}$, b_2 is already in the null space of Ω_{21} and

$$\Omega_{31} = \Omega_{21} \tag{6-77}$$

For notational convenience the matrices and auxiliary vectors are double subscripted for easy association with the columns of W. The first subscript refers to the column of B, and the second subscript refers to the power of A (plus one). For example, Ω_{ij} and w_{ij} are associated with the vector $A^{j-1} b_i$. A general iteration in the reduction procedure is as follows:

1) With Ω_{ij} from the previous iteration form the auxiliary vector

$$w_{ij} = \Omega_{ij} A^{j-1} b_i \tag{6-78}$$

2) Define the new matrix by

(i)

$$\Omega_{i+1,j} = \begin{cases} \Omega_{ij} - \dfrac{w_{ij} w_{ij}^T}{w_{ij}^T A^{j-1} b_i} & \text{if } w_{ij} \neq \underline{0} \\ \\ \Omega_{ij} & \text{if } w_{ij} = \underline{0} \end{cases} \tag{6-79}$$

for $i < r$ (r is the number of columns in B)

or

(ii) if $i = r$

$$\Omega_{1,j+1} = \begin{cases} \Omega_{rj} - \dfrac{w_{rj} w_{rj}^T}{w_{rj}^T A^{j-1} b_i} & \text{if } w_{rj} \neq \underline{0} \\ \\ \Omega_{rj} & \text{if } w_{rj} = \underline{0} \end{cases} \tag{6-80}$$

and return to step 1).

Using the Schwarz inequality and induction it can be shown that every Ω_{ij} is positive semi-definite if the initial matrix is at least positive semi-definite. In this case the initial matrix was taken to be positive definite. The positive semi-definiteness of Ω_{ij} ensures that $w_{ij}^T A^{j-1} b_i = 0$ if and only if $w_{ij} = \underline{0}$.

The orthogonal reduction process terminates when all independent columns of W have been considered. The termination point is signaled in one of two ways. The process is obviously terminated if at some point $\Omega_{ij} = \underline{0}$. This means that n independent vectors have been processed. Since W is $n \times (r \cdot n)$, there can be no more than n independent columns. When the process terminates with a zero matrix it implies that (A, B) is a controllable pair. The process can terminate on a nonzero matrix if it becomes clear that there are no additional independent vectors in the remaining columns of W. The cyclic property of the columns of W make it possible to identify such a termination point. For example, if at some point $w_{jk} = \underline{0}$, this means that $A^{k-1} b_j$ is linearly dependent on the preceding columns in W. But then $A^i b_j$ for all $i \geq k - 1$ will also be dependent on the preceding solumns in W, and as a result $w_{ji} = \underline{0}$ for all $i \geq k - 1$. Since Ω_{ji} remains unchanged if $w_{ji} = \underline{0}$, it is not necessary to even consider the vectors $A^i b_j$ for $i \geq k - 1$. In short, if $w_{kj} = \underline{0}$, the reduction process terminates for b_j and all remaining columns in W generated by b_j can be deleted from the ordered set. When the process has so terminated for every column of B, it is completely terminated. If at this point $\Omega_{ij} \neq \underline{0}$, then (A, B) is not a controllable pair. The range space of the final Ω_{ij} is the uncontrollable space of B with respect to A. By counting the number of actual reductions (the number of times $w_{ij} \neq \underline{0}$) one obtains the dimension of the controllable space of B (rk W).

The last nonzero auxiliary vector for each column of B has properties similar to the detection generator of Chapter 4. These vectors can be used to generate the equation for the feedback gain matrix. Let k_j be the integer for which

$$w_{jk_j} \neq \underline{0} \tag{6-81}$$

and

$$w_{j,k_j+1} = \underline{0} \tag{6-82}$$

To simplify notation define

$$g_j = w_{jk_j} \tag{6-83}$$

It should be noted here that if b_j is linearly dependent on the other columns of B, then $w_{j1} = \underline{0}$ and there will be no g_j for that b_j. In this case (A, B) cannot be decouplable because those columns of B which are dependent will always have intersecting controllable spaces regardless of the feedback. This algorithm can still be used to generate a feedback gain. To avoid unnecessary complication this case will be discussed separately later. Until then it will be assumed that all the columns of B are independent so that

$$\text{rk } B = r \tag{6-84}$$

and there is a nonzero g_j for every b_j.

From the reduction procedure it is known that g_j is orthogonal to all the columns of W preceding $A^{k_j-1} b_j$. Specifically

$$g_j^T A^{i-1} B = \underline{0} \qquad i = 1, \ldots, k_j - 1 \tag{6-85}$$

and if $j > 1$

$$g_j^T A^{k_j-1} b_\ell = 0 \qquad \ell = 1, \ldots, j - 1 \tag{6-86}$$

As noted earlier, the positive semi-definiteness of Ω_{jk_j} and (6-81) ensures that

$$g_j^T A^{k_j-1} b_j \neq 0 \tag{6-87}$$

This fact along with (6-85) shows that the vectors $\{g_j, A^T g_j, \ldots$ $\ldots, [A^T]^{k_j - 1} g_j\}$ are all linearly independent. Furthermore, it is easily seen from (6-85) that

$$g_j^T (A + BL)^{i-1} = g_j^T A^{i-1} \qquad i = 1, \ldots, k_j \tag{6-88}$$

and

$$\begin{aligned} g_j^T (A + BL)^{k_j} &= g_j^T A^{k_j - 1} (A + BL) \\ &= g_j^T A^{k_j} + g_j^T A^{k_j - 1} BL \end{aligned} \tag{6-89}$$

for any L. Suppose L is chosen to satisfy the equation

$$g_j^T A^{k_j - 1} BL = - p_{j1} g_j^T - \ldots - p_{jk_j} g_j^T A^{k_j - 1} - g_j^T A^{k_j} \tag{6-90}$$

for some scalars p_{ji}. Then

$$\begin{aligned} g_j^T (A + BL)^{k_j} &= - p_{j1} g_j^T - \ldots - p_{jk_j} g_j^T A^{k_j - 1} \\ &= - p_{j1} g_j^T - \ldots - p_{jk_j} g_j^T (A + BL)^{k_j - 1} \end{aligned} \tag{6-91}$$

from which it can be seen that k_j eigenvalues of (A + BL) are given by the roots of

$$s^{k_j} + s^{k_j - 1} p_{jk_j} + \ldots + p_{j1} = 0 \tag{6-92}$$

It is possible to specify $(k_1 + \ldots + k_r)$ eigenvalues (k_j at a time) by choosing L to satisfy r equations of the form of (6-90). Combining these equations into a single matrix equation yields

$$\begin{bmatrix} g_1^T A^{k_1-1} B \\ \cdot \\ \cdot \\ \cdot \\ g_r^T A^{k_r-1} B \end{bmatrix} L = \begin{bmatrix} z_{f1}^T \\ \cdot \\ \cdot \\ \cdot \\ z_{fr}^T \end{bmatrix} \tag{6-93}$$

where

$$z_{fj}^T = - p_{j1} g_j^T - \ldots - p_{jk_j} g_j^T A^{k_j-1} - g_j^T A^{k_j} \tag{6-94}$$

From (6-81) it can be verified that the matrix premultiplying L in (6-93) has the triangular form

$$\begin{bmatrix} g_1^T A^{k_1-1} B \\ \cdot \\ \cdot \\ \cdot \\ g_r^T A^{k_r-1} B \end{bmatrix} = \begin{bmatrix} g_1^T A^{k_1-1} b_1 & \cdots & g_1^T A^{k_1-1} b_r \\ \underline{0} & \ddots & \vdots \\ \underline{0} \; \cdots \; \underline{0} & & g_r^T A^{k_r-1} b_r \end{bmatrix} \tag{6-95}$$

By virtue of (6-87) the main diagonal elements in this matrix are all nonzero, so the matrix is always nonsingular. This proves that (6-93) always has a unique solution. The diagonal form of (6-95) makes it possible to solve (6-93) most easily by starting with the bottom row and working up. Now

$$k_1 + \ldots + k_r = \kappa = \text{rk } W \tag{6-96}$$

so the number of eigenvalues which can be specified by this method is the maximum possible number.

It will be shown shortly that the closed-loop system with feedback gain L given by the solution of (6-93) will be completely effector decoupled if and only if $g_j A^{k_j - 1} b_\ell = 0$ for all j and ℓ such that $j \neq \ell$. But even if $g_j A^{k_j - 1} b_\ell \neq 0$ for some $\ell \neq j$ this does not necessarily mean that the system cannot be decoupled. In some cases it is possible to modify g_j and form a new g_j' which has the same orthogonality properties as g_j in (6-85) and (6-86) and in addition satisfies $g_j'^T A^{k_j - 1} b_\ell = 0$ for all $\ell \neq j$. By making this modification in g_j where possible, one ensures that the L given by (6-93) achieves as much effector decoupling as possible. Specifically, g_j can (and should) be modified if the following two conditions hold:

(i) $\quad g_j^T A^{k_j - 1} b_\ell \neq 0 \quad$ for some $\ell > j$

and

(ii) $\quad k_\ell \geq k_j$

Let $\{b_{\ell_1}, \ldots, b_{\ell_\rho}\}$ be the set of all vectors for which (i) and (ii) hold. Define a new vector

$$g_j'^T = g_j^T + \sum_{i=1}^{\rho} \eta_i' \, g_{\ell_i}^T A^{(k_{\ell_i} - k_j)} \tag{6-97}$$

The scalars η_i' are the components of the ρ-vector

$$\eta' = \begin{bmatrix} \eta_1' \\ \cdot \\ \cdot \\ \cdot \\ \eta_\rho' \end{bmatrix} \tag{6-98}$$

which satisfies the matrix equation

$$\eta'^T \begin{bmatrix} g_{\ell_1}^T A^{k_{\ell_1}-1} \\ \vdots \\ g_{\ell_\rho}^T A^{k_{\ell_\rho}-1} \end{bmatrix} [b_{\ell_1}, \ldots, b_{\ell_\rho}] = - g_j^T A^{k_j-1} [b_{\ell_1}, \ldots, b_{\ell_\rho}] \tag{6-99}$$

This equation always has a unique solution because the product matrix postmultiplying η'^T has the same triangular form as (6-95). The g_j' defined by (6-97) is used in place of g_j in (6-93). Note that g_j' has the same orthogonality properties as g_j in (6-85) and (6-86) and in addition even for $\ell > j$

$$g_j'^T A^{k_j-1} b_\ell = 0 \quad \text{if } k_\ell \geq k_j \tag{6-100}$$

From these properties it can be shown that the algorithm will produce an effector decoupled closed-loop system when (A, B) is decouplable. From (6-85) and (6-87) it is clear that $A^{k_j-1} b_j$ is independent of the columns of $[B, AB, \ldots, A^{k_j-2} B]$. This shows that the decoupling order of b_j is greater than or equal to k_j. If κ_j is the decoupling order of b_j, then

$$k_j \leq \kappa_j \tag{6-101}$$

and

$$k_1 + \ldots + k_r \leq \kappa_1 + \ldots + \kappa_r \tag{6-102}$$

From (6-96) and Theorem 6.1 it may be concluded that (A, B) is decouplable if and only if equality holds in (6-102). But by (6-101) equality holds in (6-102) if and only if

$$k_j = \kappa_j \qquad j = 1, \ldots, r \tag{6-103}$$

Hence, (A, B) is decouplable if and only if (6-103) holds. If (6-103) holds then

$$g_j'^T A^{k_j - 1} b_\ell = 0 \qquad \text{for all } \ell \neq j \tag{6-104}$$

by the following reasoning. In view of (6-86) and (6-100) the only b_ℓ for which (6-104) could be violated is if $\ell > j$ and $k_j > k_\ell$. But if $g_j'^T A^{k_j - 1} b_\ell \neq 0$, then $\kappa_\ell \geq k_j$ by the same reasoning used to establish $\kappa_j \geq k_j$. This would imply $\kappa_\ell > k_\ell$ which contradicts (6-103). Therefore (6-103) implies (6-104), and one may conclude that (A, B) is decouplable only if (6-104) holds.

If (A, B) is decouplable, the closed-loop system with L given by (6-93) (with g_j replaced by g_j' where appropriate) can now be shown to be effector decoupled by introducing a transformation defined by

$$T_d = \begin{bmatrix} g_1'^T \\ \vdots \\ g_1'^T A^{k_1 - 1} \\ g_2'^T \\ \vdots \\ g_r'^T A^{k_r - 1} \\ T_{d2} \end{bmatrix} \tag{6-105}$$

where T_{d2} is an $(n - \kappa) \times n$ matrix chosen so that the columns of T_{d2}^T form a basis for the uncontrollable space of B. When this transformation is applied to (A + BL) and B, the resulting forms are

$$\overline{(A + BL)} = T(A + BL)T^{-1} = \begin{bmatrix} P_1^T & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \ddots & & \vdots \\ \vdots & & P_r^T & \underline{0} \\ \underline{0} & \cdots & \underline{0} & R \end{bmatrix} \tag{6-106}$$

$$\overline{B} = TB = \begin{bmatrix} \overline{b}_{11} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \ddots & & \vdots \\ \vdots & & \ddots & \underline{0} \\ \underline{0} & \cdots & \underline{0} & \overline{b}_{rr} \end{bmatrix} \tag{6-107}$$

where

$$P_i^T = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & \cdots & & 0 & 1 \\ -p_{i1} & \cdots & & \cdots & -p_{ik_i} \end{bmatrix} (k_i \times k_i) \tag{6-108}$$

and

$$\overline{b}_{ii} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ g_i'^T A^{k_i - 1} b_i \end{bmatrix} (k_i \times 1) \tag{6-109}$$

(recall $g_i'^T A^{k_i - 1} b_i \neq 0$) and R is an $(n - \kappa) \times (n - \kappa)$ matrix satisfying

$$T_{d2}(A + BL) = T_{d2}A = RT_{d2} \tag{6-110}$$

From the block diagonal forms of $(\overline{A + BL})$ and $\overline{B}$ it can be seen that the algorithm has produced an effector decoupled system.

If (A, B) is not a decouplable pair (6-104) will be violated for some b_ℓ for which $\ell > j$ and $\kappa_\ell \geq k_j > k_\ell$. When the transformation (6-105) is applied in this case, $(\overline{A + BL})$ will have the same form as (6-106) but $\overline{B}$ will have the more general form

$$\overline{B} = \begin{bmatrix} \overline{b}_{11} & \cdots & \overline{b}_{1r} \\ \vdots & & \vdots \\ \overline{b}_{r1} & \cdots & \overline{b}_{rr} \end{bmatrix} \tag{6-111}$$

with

$$\overline{b}_{j\ell} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ g_j^T A^{k_j - 1} b_\ell \end{bmatrix} \quad (k_j \times 1) \tag{6-112}$$

The equation for each subsystem is

$$\dot{\overline{x}}_j(t) = P_j^T \overline{x}_j(t) + \overline{b}_{jj} u_{dcj}(t) + \sum_{\substack{\ell=1 \\ \ell \neq j}}^{r} \overline{b}_{j\ell} u_{dc\ell}(t) \tag{6-113}$$

Note that $\overline{b}_{j\ell}$ can be nonzero only if $\kappa_\ell > k_\ell$. This means if $\kappa_\ell = k_\ell$ then $u_{dc\ell}(t)$ controls only the ℓ^{th} subsystem and has no influence on the other subsystems.

In the case where rk $B < r$, indicating a linear dependence among the columns of B, the algorithm can still be used to generate an L. Suppose rk $B = r' < r$. There will then be only r' nonzero generators g_j and only r' equations such as (6-90) for L. The matrix premultiplying L in (6-93) will no longer be square, but will have dimension $r' \times r$. It can be shown from (6-85) and (6-86) that this matrix always has rank r'. This ensures that the equation for L will always have a solution, but it will not be unique. As mentioned earlier, this situation precludes the possibility of obtaining an effector decoupled system because (A, B) is not decouplable.

It seems certain that this algorithm will require less computation than the first one. It is not necessary to generate the auxiliary matrices corresponding to K and C' of Chapter 4. Nor is it necessary to worry about separability and mutual detectability. The solution of the equation for L is made simpler by the triangular form of (6-95). The modification of the g_j seems to require some additional computation, but this is not certain because the use of the modified generators g'_j introduces additional zeros in off-diagonal elements of the matrix in (6-95). In fact, if the system is decouplable, this matrix will be purely diagonal. It should be mentioned that the most efficient way to modify the g_j is to start with $j = r - 1$ (g_r never needs modification) and work backward replacing g_j with g'_j at each step. In this way one obtains the largest number of off-diagonal zeros in the matrix postmultiplying η'^T in (6-99). It is possible to show that none of the g_j will need modification if the starting matrix for the reduction procedure Ω_{11} is properly chosen. Unfortunately no simple way of finding such a Ω_{11} is yet available.

The third algorithm for generating a feedback gain matrix is concerned only with specifying poles of the closed-loop system, rather than producing any specific kind of subsystem structure (e.g., decoupled effectors). It is of interest for feedback restructuring because it allows the possibility of specifying some poles of the closed-loop system as the algorithm proceeds, rather than having to wait until all the computation is completed as in the previous two algorithms. This feature will be described in more detail later.

The third algorithm is computationally very similar to the decoupling algorithm just presented. The orthogonal reduction procedure is again employed. The columns of W make up the ordered set of vectors except the ordering of the set is different. In this case the ordered set of vectors is $\{b_1, Ab_1, \ldots, A^{n-1} b_1, b_2, \ldots, A^{n-1} b_r\}$. The reduction process proceeds as before with appropriate changes in the condition for termination. After starting with b_1, the first intermediate termination point is reached when $w_{1j} = \underline{0}$ for some j (the double subscripts on w_{ij} and Ω_{ij} have the same significance as previously). All further vectors generated by b_1 may be disregarded and the process continues with b_2. The process is completely terminated when either $\Omega_{ij} = \underline{0}$ or the termination point associated with b_r is reached (i.e., $w_{rj} = \underline{0}$ for some j). The terminating Ω_{ij} has the same significance as in the decoupling algorithm. Feedback generating vectors are again defined as the last nonzero auxiliary vectors associated with the columns of B

$$g_j = w_{jk_j} \neq \underline{0} \qquad (6\text{-}114)$$

$$w_{j,k_j+1} = \underline{0} \qquad (6\text{-}115)$$

In general the k_j here are different from those in the decoupling algorithm, but it is still true that $k_1 + \ldots + k_r = \kappa = \text{rk } W$. The equations for L have the same form as (6-90). It is not necessary to modify the g_j. For this algorithm it is more likely that there will be less than r generating vectors g_j. This will certainly be the case if rk $B < r$. Even when rk $B = r$ there will be fewer than r generating vectors if b_j, for example, is contained in the combined controllable spaces of the previous columns of B (i.e., the controllable space of $[b_1, \ldots, b_{j-1}]$). In this case $w_{j1} = \underline{0}$ and there will be no g_j. As noted previously, the presence of less than r generating vectors simply means the solution of the matrix equation for L is not unique.

Just as for the decoupling algorithm the total number of eigenvalues of (A + BL) which can be specified is the maximum possible number, κ = rk W. The significant feature of this algorithm which makes it worthy of mention is that it is possible to specify some eigenvalues of (A + BL) before the orthogonal reduction procedure is completed and without fear of introducing unwanted eigenvalues. To clarify this statement some background information is necessary. At any point in the reduction procedure for either of the last two algorithms it is possible to use the auxiliary vectors to immediately write down an equation for L which will specify a certain number of eigenvalues of (A + BL). For example, at any point in the decoupling algorithm one has at hand the last nonzero auxiliary vectors for each b_j, say $w_{jk'_j} \neq 0$. These vectors have the same orthogonality properties as the g_j in (6-85) and (6-86) except that k_j is replaced by k'_j. By using the $w_{jk'_j}$ in the same way that the g_j were used it is possible to write down an equation for L corresponding to (6-93). The solution of this equation will yield a matrix (A + BL) in which $(k'_1 + \ldots + k'_r)$ eigenvalues can be specified by choice of the coefficients p_{ij}. The problem with this premature specification of eigenvalues is that when the reduction procedure is not complete $(k'_1 + \ldots + k'_r) < \kappa$, and the number of eigenvalues so specified is less than κ. The $(n - \kappa)$ eigenvalues of A associated with the uncontrollable space of B cannot be altered by the feedback. But this still leaves $\kappa - (k'_1 + \ldots + k'_r)$ eigenvalues of (A + BL) which are determined by the feedback and yet are not explicitly specified by the p_{ij}. There is no simple way of ensuring these uncontrolled eigenvalues will be stable.

Using the third algorithm it is possible to specify a number of eigenvalues at each intermediate termination point without introducing uncontrolled eigenvalues. Suppose the first intermediate termination point has been reached, so g_1 is known. Now introduce feedback in just the first control component so the closed-loop system matrix is $(A + b_1\underline{l}_1)$ where $\underline{l}_1$ is an $(1 \times n)$ row vector given by

$$\underline{l}_1 = \frac{1}{(g_1^T A^{k_1-1} b_1)} [- p_{11} g_1^T - \ldots - p_{1k_1} g_1^T A^{k_1-1} - g_1^T A^{k_1}] \tag{6-116}$$

Then

$$g_1^T (A + b_1\underline{l}_1)^{k_1} = g_1^T A^{k_1} + g_1^T A^{k_1-1} b_1 \underline{l}_1$$

$$= \; - p_{11} g_1^T - \dots - p_{1k_1} g_1^T A^{k_1 - 1}$$

$$= \; - p_{11} g_1^T - \dots - p_{1k_1} g_1^T (A + b_1 \underline{l}_1)^{k_1 - 1} \tag{6-117}$$

which shows that k_1 eigenvalues of $(A + b_1 \underline{l}_1)$ are given by the roots of

$$s^{k_1} + p_{1k_1} s^{k_1 - 1} + \dots + p_{11} = 0 \tag{6-118}$$

But

$$k_1 = rk\,[b_1, \dots, A^{n-1} b_1] \tag{6-119}$$

and this is the maximum number of eigenvalues which can be influenced by feedback in only the first control component. All the remaining $(n - k_1)$ eigenvalues of $(A + b_1 \underline{l}_1)$ must be the same as those of A. Therefore no uncontrolled eigenvalues have been introduced. When the second intermediate termination point is reached, feedback can be allowed in the first two control components and the number of eigenvalues which can be specified is

$$k_1 + k_2 = rk[(b_1, b_2), A(b_1, b_2), \dots, A^{n-1}(b_1, b_2)] \tag{6-120}$$

Again no uncontrolled eigenvalues are introduced because all remaining eigenvalues remain unchanged. The process can be repeated at each intermediate termination point. The intermediate specification of eigenvalues may be valuable in situations where instabilities in the open-loop system threaten to exceed acceptable bounds before the orthogonal reduction process can be completed. By specifying eigenvalues at intermediate termination points unstable eigenvalues may be eliminated. Of course, the algorithm need not start with b_1. The columns of B can be arranged in any order. If an unstable eigenvalue of A is known or suspected to be associated with the controllable space of a particular column of B, then the algorithm should begin with that column.

The structure of the closed-loop system produced by the third algorithm is more obvious if a transformation of the form of (6-105) is applied. When this is done, $(\overline{A + BL}) = T(A + BL)T^{-1}$ has the same form as (6-106), but $\overline{B}$ has the form

$$\overline{B} = TB = \begin{bmatrix} \overline{b}_{11} & \cdots & \cdots & \overline{b}_{1r} \\ \underline{0} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \underline{0} & \cdots & \underline{0} & \overline{b}_{rr} \end{bmatrix} \qquad (6\text{-}121)$$

where

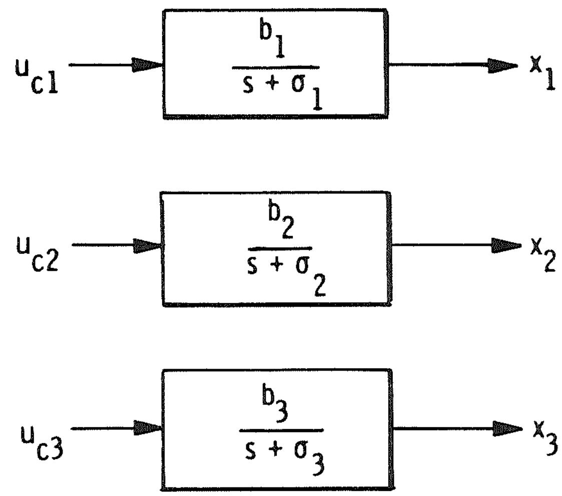
$$\overline{b}_{ii} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ g_i^T A^{k_i - 1} b_i \end{bmatrix} \quad (k_i \times 1) \qquad (6\text{-}122)$$

The off-diagonal vectors $\overline{b}_{ij}$ have no simple form, in general. If there are fewer than r generating vectors, say $r' < r$, then there are only r' of the P_i blocks in $(\overline{A + BL})$ and $\overline{B}$ has the form

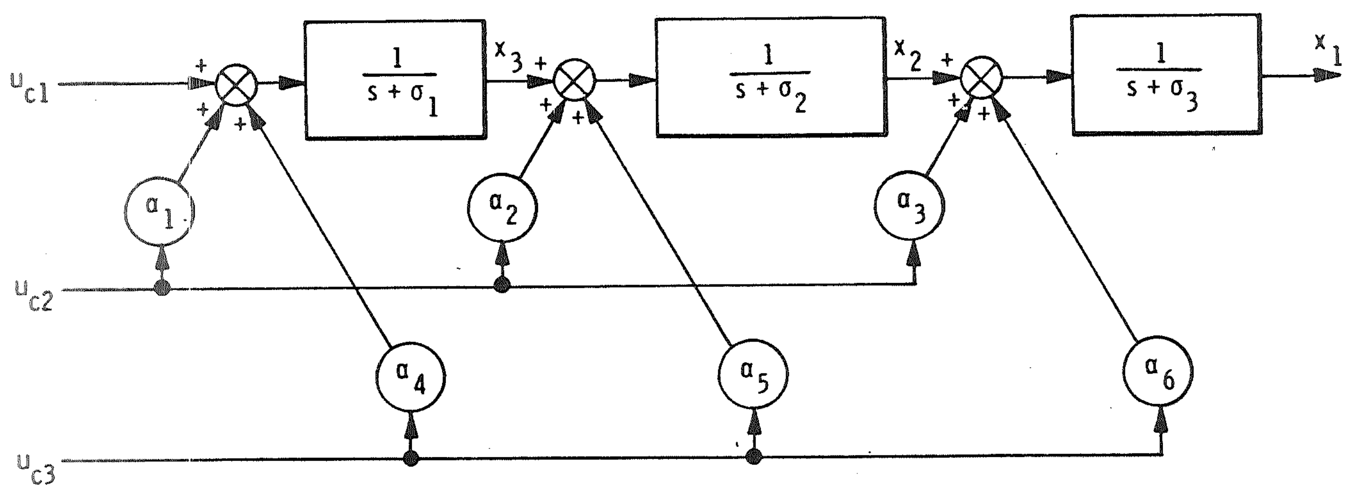
$$\overline{B} = \begin{bmatrix} \overline{b}_{11} & \cdots & \cdots & \cdots & \overline{b}_{1r} \\ \underline{0} & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \underline{0} & \cdots & \underline{0} & \overline{b}_{rr'} \cdots & \overline{b}_{rr} \end{bmatrix} \tag{6-123}$$

Although this algorithm is not designed to yield any specific subsystem structure, it is possible to make some general remarks about the type of structure it tends to produce. For this algorithm the columns of W are reordered so that all vectors generated by b_1 are considered first, and so on. This tends to make the dimensions of the earlier (lower indexed) P_i^T smaller. On the other hand, the decoupling algorithm tends to make the P_i roughly equal in size. In terms of system structure this means that the decoupling algorithm tends to produce a parallel type of structure, whereas the third algorithm leads to a cascade-type structure. As a simple illustration of this consider a third order system controlled by three independent control inputs, each of which can control the system acting alone. Suppose the three closed-loop poles are specified to lie on the negative real axis at $-\sigma_1$, $-\sigma_2$, and $-\sigma_3$. The decoupling algorithm would produce a system of three independent first order subsystems as shown in Figure 6-1(a). The third algorithm would produce the cascade-type structure shown in Figure 6-1(b).

It would appear that the cascade algorithm compares favorably with other pole assignment algorithms discussed in the literature. It is certainly computationally simpler than the straightforward approach of determining the characteristic polynomial of (A + BL) by expanding



(a) Decoupling Algorithm



(b) Cascade Algorithm

Figure 6-1.

the determinant $|Is - (A + BL)|$, setting the coefficients equal to some desired values, then solving the set of n nonlinear equations for the $n \cdot r$ elements of L. It is also simpler than algorithms based on transformations which produce certain canonical matrix forms (such as suggested in [23]). Although it is not necessary to actually perform a complete state space transformation in such an algorithm, it is necessary to compute certain parameters appearing in the canonical form of A, and then transform the feedback gain matrix back to the original coordinate frame.

There is another pole assignment algorithm discussed in the literature (referred to as the spectral algorithm in [23]), which may be useful for feedback restructuring. It is based on the Jordan form of the A matrix (the system matrix for normal mode state variables). This algorithm allows assignment of a small number of closed-loop poles (in some cases a single pole) while leaving the remaining poles of the system undisturbed. Hence, the algorithm can be applied recursively, specifying a small number of closed-loop poles at step. As noted previously in introducing the cascade algorithm, this would seem to be a desirable feature for an on-line restructuring process. The spectral algorithm has some computational disadvantages, however. In order to specify a certain number of closed-loop poles, one must first determine an equal number of open-loop poles (eigenvalues of the A matrix) plus the corresponding eigenvectors of A. In general, determining eigenvalues of A will require solving the characteristic equation for A, which is an n^{th} order polynomial equation. None of the algorithms discussed previously in this section require knowledge of any eigenvalues of A.

Another disadvantage of the spectral algorithm is that it must be modified if A has repeated eigenvalues. This suggests that for a general A matrix it will be necessary to determine all the eigenvalues in order to check for repeated eigenvalues before the correct algorithm can be implemented. This requirement would increase the computation time necessary before specifying the first group of closed-loop poles, thus reducing the speed advantage offered by recursive specification of poles. The cascade algorithm is applicable to a general A matrix, and it is not necessary to have information about repeated eigenvalues or other structural properties in order to implement it.

Because of the necessity for computing eigenvectors of A^T, the computation required in the spectral algorithm increases significantly when specifying a large number of poles. (Simon and Mitter [23] claim the increase is exponential.) Therefore, the cascade algorithm seems better suited to specifying a large number of poles. It would appear, however, that if A happens to be in a form in which some eigenvalues can be readily identified, then the spectral algorithm would probably be the fastest way of changing those particular eigenvalues. The spectral technique would be especially valuable if some way could be found to identify quickly any unstable poles in the existing system, since it would provide a way of concentrating the feedback restructuring efforts on stabilizing those unstable modes.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

The purpose of this research was to develop practical methods of self-reorganization which can give a complex linear dynamic system the ability to restructure itself to compensate for failures in its effectors and sensors and changes in dynamics. The ultimate goal of self-reorganization is to achieve the maximum reliability with the minimum amount of hardware by restructuring the system to make effective use of all hardware available at any given time. The basic approach taken in this research is to identify the failure or change and then restructure the system based on that information. This approach is in contrast to reorganization based on performance information.

Chapter 2 demonstrates how the concepts of controllability and observability may be used to evaluate the potential ability of a linear system to tolerate failures of its effectors and sensors. A lower bound is established for the number of effectors and sensors a linear time-invariant system requires for complete controllability and observability.

Since the reorganization process is based on information about the failures or changes occurring in the system, the greatest attention was devoted to the problem of detecting and identifying such events. The major contribution of this research is the theory and design of detection filters developed in Chapter 4. Detection filters provide a practical way of detecting and identifying effector failures, sensor failures, and dynamic changes in a complex multiple-input, multiple-output linear system. The important features of a detection filter include the following:

1) When a failure or change occurs the detection filter produces a vector error signal whose direction indicates the location of the failure or change, or at least narrows the location down to a small number of possibilities. An effector failure or a change in some parameter in the dynamic equations of the system produces an error signal in a fixed vector direction. This invariant direction indicates which effector is malfunctioning or which parameter has changed. In some cases the invariant direction may be associated with more than one effector or parameter, in which case the location of the failure or change is narrowed down to those effectors or parameters associated with the invariant direction. In this situation the time-varying behavior of the error magnitude often provides enough additional information to identify a particular effector or parameter from the set of possibilities indicated by the invariant error direction. A sensor failure does not produce a fixed-direction error signal, but the error vector is constrained to lie in a two-dimensional invariant plane. This plane identifies the malfunctioning sensor.

2) In the absence of failures or changes in dynamics (or after they have been identified and compensated for) the detection filter produces an estimate of the state of the system. The estimate is asymptotically stable in the sense that in the absence of disturbances the error in the estimate approaches zero asymptotically. The

detection filter may therefore serve also as a state estimating filter.

3) The poles of the detection filter are under the control of the designer. This means the response time of the filter can be made as fast as desired, consistent with other considerations such as noise disturbances and gain magnitudes. It also means that the filter may be designed to enhance the response to failures or changes it is supposed to detect, while suppressing the response to sensor noise and plant disturbances.

4) A detection filter (whose state dimension is equal to that of the system) has the potential to detect a substantial number of different events (failures and changes in dynamics). When a single detection filter is not capable of detecting all possible events, it is merely necessary to use additional filters, each designed to detect a subset of the set of all possible events. Because each filter has the potential to detect a substantial number of events, it should be possible to detect all possible events with a small number of filters. For the special case in which the state vector of the system is fully measurable, a single detection filter can provide information about all possible events -- effector failures, sensor failures, and changes in dynamics. For the more general case of a partially measurable state vector, the number of different failures a detection filter is capable of detecting is, loosely speaking, approximately equal to the number of independent sensors in the system. In particular situations it may be more or less. In any case a single detection filter can provide information about all changes in the dynamics of a linear system.

5) The same basic theory is applicable to designing detection filters for effector failures, sensor failures, and changes in dynamics. For detecting changes in dynamics, the detection filter is especially effective when the possible changes are limited to a small number of parameters. Even when applied to the general problem of identifying or tracking unknown linear system dynamics, detection filter theory yields an identification method which appears comparable to the best tracking model methods now proposed in the literature.

6) The computation required to design detection filters involves mainly the solution of sets of linear algebraic equations. It is not necessary to solve differential equations -- either linear or nonlinear. The computation is substantially less than that required for a Kalman filter, for example, which requires the solution of a Riccati equation.

Chapter 4 develops a substantial body of analytical results on the structure of detection filters. The results have been developed from the viewpoint of actually constructing a detection filter. As a result, some of the algebra may be more extensive than would be necessary if more sophisticated methods of mathematical analysis were used. However, the constructure viewpoint provided a good basis for the development of the design algorithms presented in Appendices A, B, and C. The material in Chapter 4 should continue to provide a good basis for the future development of even more efficient design algorithms. Some of the more important results of Chapter 4 are listed below.

1) Theorem 4.1 is the basic result of detection filter theory. It guarantees that there always exists some detection filter, with poles arbitrarily specified by the designer, which will detect any single

failure or change in the observable dynamics of a system. The other theorems and lemmas in Section 4.3.1 are intermediate results leading to the proof of Theorem 4.1. However, some of them are important in filter design, and these are mentioned in the next item.

2) Lemma 4.2 establishes the existence of detection generators, the vectors which play a central role in the actual design of detection filters. Theorem 4.2 introduces the basic linear algebraic equation for the error feedback gain matrix which gives a detection filter the invariant direction property. The results of Theorems 4.3 and 4.4 show how it is possible to arbitrarily specify all the poles of the detection filter while achieving the invariant direction property. In addition, the proof of Theorem 4.4 shows how to actually determine the maximal detection generator, which allows full specification of the poles of the filter. The algorithm in Appendix A is based on the construction used in that proof.

3) Theorem 4.5 establishes the conditions under which it is possible for a single detection filter to detect a number of different events while allowing the poles of the filter to be arbitrarily specified. (Such events are defined to be mutually detectable.)

4) Theorem 4.6 establishes a method for dividing the set of all possible events into subsets of mutually detectable events. All the events in each subset can then be detected by one detection filter. Often events which are not mutually detectable can still be detected with a single filter by allowing certain poles of the filter to be fixed by the design process rather than specified by the designer. Theorem 4.6 provides the basis for identifying these unspecified poles and regrouping

sets of events so that any undesirable poles are eliminated. This material is developed in Section 4.3.4.

When detection filter theory is interpreted in its dual form the results yield design techniques for determining linear state feedback laws for linear time-invariant systems. It is well known that if a linear time-invariant system is controllable, then a linear state feedback law can always be found which produces closed-loop poles in any desired location in the complex plane (complex poles must appear in complex conjugate pairs). The techniques introduced in this research not only provide for specification of the closed-loop poles of the system, but also can produce several interesting types of subsystem structure such as scalar-input, scalar-output decoupled subsystems or effector decoupled subsystems. Chapter 6 presents the algorithms for implementing these feedback control designs. Also presented is a third algorithm which is concerned only with fast specification of closed-loop poles. These algorithms form the basis for restructuring of the feedback control loop to compensate for failures and changes in the system. The computation involved in implementing these techniques seems sufficiently simple to make their use feasible for on-line restructuring. The results may also be of interest for off-line feedback design.

7.2 Recommendations for Further Study

The next logical step for further research is to substantiate the theoretical analysis of detection filters and test the feasibility of the feedback restructuring algorithms through computer simulation. It would also be most valuable to design detection filter reorganization systems for some example systems to demonstrate computational feasibility and performance in the presence of realistic disturbances. Areas for further analytical studies include the following:

1) The concepts introduced in Chapter 2 merely evaluate the supplementary redundancy of a system after it is constructed. It should be possible to develop these concepts to aid in the actual design of supplementary redundant systems.

2) It would be useful to obtain more general results on the detection of nonseparable events as defined in Chapter 4. Such results could lead to methods for substantially increasing the number of different events a single filter is capable of detecting. For the general case of a partially measurable state vector the number of simple events (e.g., one effector failure) detectable by a single filter is, with present design methods, roughly the same as the number of independent sensors. Recall that for the case of the fully measurable state vector a single filter could detect all the events being considered -- effector failures, sensor failures, and changes in dynamics -- potentially a much larger number of events than the number of independent sensors. It seems reasonable to speculate that as the number of independent sensors increases, it should be possible to construct a detection filter capable of detecting substantially more events than the number of independent sensors.

3) The algorithms in Appendices A and B for implementing the design of detection filters are not intended to be the last word in computational efficiency. It seems reasonable to expect that they can be

improved upon in this respect. The extensive analytical results in Chapter 4 should be useful in developing new methods of implementing the theory of detection filters. Rapid computational algorithms will also be valuable for the design of linear state feedback laws for time-invariant linear systems.

4) Chapter 5 discusses some simple methods for processing the detection filter error information to identify the most likely event (or events) in the face of uncertainties caused by noise disturbances or simultaneous multiple events. It should be possible to develop more sophisticated methods for processing the detection filter information. For example, if statistical information is available on noise disturbances or on the occurrence of events, then this information might be used to develop decision rules which are statistically optimum in some sense.

5) This research has been primarily directed toward designing reorganization methods for an existing dynamic system. A related area which seems lucrative for further research is the design of the basic system (e.g., placement of effectors and sensors) to make failures easy to identify. The material in Chapter 4 should provide a good basis for such research.

APPENDIX A

ALGORITHM FOR DETERMINING THE MAXIMAL GENERATOR

Determination of the maximal generator for a vector f is divided into two basic steps:

I. Finding the null space of M defined by (4-182), i.e., all independent solutions of

$$M'w = \underline{0} \tag{A-1}$$

II. Finding a vector g in the null space of M' satisfying

$$\begin{bmatrix} C \\ \cdot \\ \cdot \\ \cdot \\ CA^{\nu-2} \end{bmatrix} g = \underline{0} \tag{A-2}$$

$$CA^{\nu-1} g = CA^{\mu} f \tag{A-3}$$

where ν is the detection order of f and μ is defined by condition (4-108). Note the similiarity of these two steps. They both involve finding vectors lying in the null space of a given matrix. The following algorithm, referred to as the orthogonal reduction procedure, is a general method for solving such a problem.

Consider an $n' \times n$ matrix

$$V = \begin{bmatrix} v_1^T \\ \cdot \\ \cdot \\ \cdot \\ v_n^T \end{bmatrix} \qquad \text{(A-4)}$$

where the v_i are arbitrary n-vectors. The orthogonal reduction procedure is an iterative process which generates an $n \times n$ positive semi-definite matrix whose range space coincides with the null space of V. In each iteration a row of V is tested to determine if it is orthogonal to the range space of the symmetric matrix. If not, the range space of the matrix is reduced so that this is the case. The procedure begins with any symmetric positive-definite matrix Ω_1. An auxiliary n-vector is defined by

$$w_1 = \Omega_1 v_1 \qquad \text{(A-5)}$$

If v_1 is nonzero w_1 will be nonzero, since Ω_1 is positive definite. Furthermore, $w_1^T v_1$ will be nonzero. A new symmetric positive semi-definite matrix is defined by

$$\Omega_2 = \Omega_1 - \frac{w_1 w_1^T}{w_1^T v_1} \qquad \text{(A-6)}$$

The procedure continues according to the following general iteration:

(i) With Ω_i from the previous iteration, form the auxiliary vector

$$w_i = \Omega_i v_i \qquad \text{(A-7)}$$

(ii) If $w_i \neq \underline{0}$ set

$$\Omega_{i+1} = \Omega_i - \frac{w_i w_i^T}{w_i^T v_i} \tag{A-8}$$

or if $w_i = \underline{0}$ set

$$\Omega_{i+1} = \Omega_i \tag{A-9}$$

and return to (i)

The algorithm has the following important properties:

1) If Ω_i is positive semi-definite, $w_i^T v_i = 0$ if and only if $w_i = \underline{0}$. This follows from the definition of w_i.

2) If Ω_i is positive semi-definite, so is Ω_{i+1}. This is trivially true if $w_i = \underline{0}$. Assume $w_i \neq \underline{0}$. For any arbitrary n-vector z and any scalar α

$$(z - \alpha v_i)^T \Omega_i (z - \alpha v_i) \geq 0 \tag{A-10}$$

In particular, this must be true for

$$\alpha = \frac{w_i^T z}{w_i^T v_i} \tag{A-11}$$

Expanding (A-10) and substituting (A-11) yields

$$(z - \alpha v_i)^T \Omega_i (z - \alpha v_i) = z^T \Omega_i z - 2\alpha v_i^T \Omega_i z + \alpha^2 v_i^T \Omega_i v_i$$

$$= z^T\Omega_i z - 2\alpha w_i^T z + \alpha^2 w_i^T v_i$$

$$= z^T\Omega_i z - \frac{(w_i^T z)^2}{w_i^T v_i}$$

$$= z^T\Omega_{i+1} z \geq 0 \qquad \text{(A-12)}$$

By induction this shows that all Ω_i are positive semi-definite if the starting matrix Ω_1 is at least positive semi-definite.

3) If $w_i \neq 0$, then

$$\text{rk}\, \Omega_{i+1} = \text{rk}\, \Omega_i - 1 \qquad \text{(A-13)}$$

and the null space of Ω_{i+1} is the subspace formed by v_i and the null space of Ω_i. In Equation (A-12) equality holds (and thus $\Omega_{i+1} z = \underline{0}$) if and only if $(z - \alpha v_i)$ lies in the null space of Ω_i. But this implies z must lie in the subspace formed by v_i and the null space of Ω_i.

4) At any point in the process the range space of Ω_i is made up of all vectors orthogonal to the vectors $\{v_1, \ldots, v_{i-1}\}$. This follows from property 3) and the fact that the starting matrix Ω_1 is positive definite. If Ω_1 is only positive semi-definite, the range space of Ω_i is made up of all vectors from the range space of Ω_1 which are orthogonal to $\{v_1, \ldots, v_{i-1}\}$. When all the rows of V have been processed the final matrix $\Omega_{n'+1}$ has a range space which coincides with the null space of V (for Ω_1 positive definite). The number of reductions made (i.e., the number of times (A-8) is performed) is equal to the rank of V.

5) If Ω_1 is positive definite and $w_i = \underline{0}$, then v_i is linearly dependent on the preceding vectors $\{v_1, \ldots, v_{i-1}\}$. By virtue of property 4) the vectors $\{v_1, \ldots, v_{i-1}\}$ span the null space of Ω_i. Since $w_i = \underline{0}$ implies v_i is in the null space of Ω_i, it must be expressible as a linear combination of the vectors $\{v_1, \ldots, v_{i-1}\}$.

The first step in finding the maximal generator for f can now be accomplished by applying the reduction algorithm to the matrix M' defined by (4-182). The algorithm begins with a symmetric positive definite matrix, such as the identity matrix. The rows of M' correspond to the v_i^T in (A-4). Because of the cyclic manner in which the rows of M' are generated it is not necessary to process all the rows. A row can be skipped if it is known that it is linearly dependent on preceding rows, because the auxiliary vector in that case will be zero. When a particular auxiliary vector is found to be zero, for example,

$$w_i = \Omega_i (c'_j K^\ell)^T = \underline{0} \qquad \text{(A-14)}$$

(where c'_j is the j^{th} row of C') it is then known that $c'_j K^\ell$ is linearly dependent on the preceding rows in M'. But if this is so, then all remaining rows of M' generated by c'_j (i.e., $c'_j K^k$ for all $k > \ell$) will also be dependent on preceding rows of M'. The auxiliary vectors associated with these rows will all be zero, so there is no need to consider them in the reduction procedure. The appearance of the first zero auxiliary vector, as in (A-14), will be referred to as the intermediate termination point for c'_j. The reduction process is completely terminated for M' when the intermediate termination points for all rows of C' have been reached. It is of interest to note that since $\text{rk } C' < \text{rk } C$, there is a linear dependence among the rows of C', and at least one row of C' will be terminated when it is first processed. When the algorithm is completely terminated the final matrix, denoted by Ω_f, will have a range space which coincides with the null space of M'. At that point $q' = \text{rk } M'$ is given by the number of reductions performed.

The second step in finding the maximal generator is accomplished by applying the reduction procedure to the rows of the matrix

$$M_{KT} = \begin{bmatrix} C \\ \cdot \\ \cdot \\ \cdot \\ CK^{n-q'-1} \end{bmatrix} \tag{A-15}$$

starting with the final matrix Ω_f from the first reduction process. The rows of C span a subspace which contains and is exactly one dimension larger than the subspace spanned by the rows of C'. Since the range space of Ω_f is orthogonal to all the rows of C', all rows of C except one will be terminated when first encountered in the reduction process. The process will be completely terminated when the termination point for this one row, say c_j, is reached. The final symmetric matrix at termination will be the zero matrix if (A, C) is an observable pair. The maximal generator is formed from the last nonzero auxiliary vector before termination,

$$w_i = \Omega_i \, (c_j K^{\nu-1})^T \neq \underline{0} \tag{A-16}$$

where $\nu = n - q'$ is the detection order of f. By construction w_i lies in the null space of M' and satisfies

$$\begin{bmatrix} C \\ \cdot \\ \cdot \\ \cdot \\ CK^{\nu-2} \end{bmatrix} w_i = \underline{0} \tag{A-17}$$

and

$$CK^{\nu-1} w_i = CA^{\nu-1} w_i \neq \underline{0} \tag{A-18}$$

These are all the requirements for the maximal generator except the magnitude of w_i must be adjusted to satisfy (A-3). The maximal generator for f is then given by

$$g = \left(\frac{c_j A^{\mu} f}{c_j K^{\nu-1} w_i} \right) w_i \tag{A-19}$$

It should be mentioned that the matrix

$$M_T = \begin{bmatrix} C \\ \cdot \\ \cdot \\ \cdot \\ CA^{n-q'-1} \end{bmatrix} \tag{A-20}$$

can be used in place of M_{KT} for the second reduction process. In fact, any matrix of the form $A'' = A - D''C$ with D'' arbitrary can be used in place of K in (A-15). The matrix K was shown because it is usually simpler than A. As noted in Section 4.3.1, A may be in a form (e.g., the standard form (4-403) to (4-405)) which makes it possible to determine by inspection a D'' which yields an $A'' = A - D''C$ considerably simpler than A. In this case A'' can be used in place of A in finding the maximal generator. This includes using A'' in defining K. When such an A'' is available it can also be used in A-15) in place of K.

If the final symmetric matrix at termination is not the zero matrix, then (A, C) is not an observable pair and the range space of the final matrix is the unobservable space of C. The maximal generator was defined in Chapter 4 only for the case where (A, C) was an observable pair. However, it was noted in remark 4) at the end of Section 4.3.1 that condition (1) of detectability can be achieved for an unobservable pair if f does not lie in the unobservable space of C. For this case the g given by (A-19) can be used in exactly the same way as the maximal generator to achieve condition (1). If $(k - 1)$ is the power of A associated with the last nonzero auxiliary vector, then $(k + q')$ is equal to the dimension of the observable space of C, which in this case is less than n. The $(n - q' - k)$ eigenvalues of A associated with the unobservable space of C cannot be altered and will always appear as eigenvalues of $(A - DC)$.

When using this algorithm to find maximal generators for a set of vectors $\{f_1, \ldots, f_r\}$, the following procedure is suggested:

(i) Starting with a symmetric positive definite matrix, apply the reduction process to M' given by (4-261) with K and C' defined by (4-257) and (4-255) for the full set of f_i.

(ii) For each f_i apply the algorithm as presented, except replace the starting matrix Ω_1 with the final terminating matrix from (i).

This procedure requires fewer total reductions than simply repeating the complete algorithm for each f_i.

The last nonzero auxiliary vectors obtained at the intermediate termination points in the first orthogonal reduction process can be used to specify the q' eigenvalues of (A - DC) = (A' - D'C') which remain unspecified after D is constrained to be a detector gain. It was noted earlier that at least one row of C' will be terminated when first encountered in the reduction process. For this row there will be no nonzero auxiliary vector. Additional rows of C' will also be terminated at first encounter if rk C $<$ m, implying a linear dependence among some rows of C (recall C is $m \times n$). Assume, then, there are ℓ independent rows in C' where $\ell \leq (m - 1)$. Each of these rows will have a final nonzero auxiliary vector. Let $\{c'_{j_1}, \ldots, c'_{j_\ell}\}$ be the first ℓ independent rows of C'. Denote by w_{fi} the final nonzero auxiliary vector associated with c'_{j_i} and assume the termination point occurs at the row $c'_{j_i} K^{q'_i}$. Then

$$c'_{j_i} K^{q'_i - 1} w_{fi} = c'_{j_i} A'^{q'_i - 1} w_{fi} \neq \underline{0} \tag{A-21}$$

and w_{fi} is orthogonal to all preceding rows of M'. Specifically

$$\begin{bmatrix} C' \\ \cdot \\ \cdot \\ \cdot \\ C'K^{q'_i - 2} \end{bmatrix} w_{fi} = \begin{bmatrix} C' \\ \cdot \\ \cdot \\ \cdot \\ C'A'^{q'_i - 2} \end{bmatrix} w_{fi} = \underline{0} \tag{A-22}$$

and

$$c'_{\rho} K^{q'_i - 1} w_{fi} = c'_{\rho} A'^{q'_i - 1} w_{fi} = \underline{0} \quad \text{for all } \rho < j_i \tag{A-23}$$

From (A-21) and (A-22) it can be seen that the w_{fi} have orthogonality properties similar to those in (4-80) and (4-81) for a detection generator. They can therefore be used in like manner to specify eigenvalues of $(A' - D'C')$. By arguments similar to those used for detection generators it can be shown that

$$A'^{\rho} w_{fi} = (A' - D'C')^{\rho} w_{fi} \quad \text{for } \rho = 0, \ldots, q'_i - 1 \tag{A-24}$$

and that these q'_i vectors are linearly independent. Further, (A-23) can be used in a development similar to the proof of Lemma 4.5 to show that the entire set of $(q'_1 + \ldots + q'_{\ell}) = q'$ vectors $\{w_{f1}, \ldots, A'^{q_1 - 1} w_{f1}, w_{f2}, \ldots, A'^{q'_{\ell} - 1} w_{f\ell}\}$ are all linearly independent. Now if D' is chosen to satisfy the equation

$$\begin{aligned} D'C'K^{q'_i - 1} w_{fi} &= D'C'A'^{q'_i - 1} w_{fi} \\ &= p_{i1} w_{fi} + \ldots + p'_{iq'_i} A'^{q'_i - 1} w_{fi} + A'^{q'_i} w_{fi} \end{aligned} \tag{A-25}$$

then

$$\begin{aligned} (A' - D'C')^{q'_i} w_{fi} &= A'^{q'_i} w_{fi} - D'C'A'^{q'_i - 1} w_{fi} \\ &= - p'_{i1} w_{fi} - \ldots - p'_{iq'_i} A'^{q'_i - 1} w_{fi} \end{aligned}$$

$$= -p'_{i1} w_{fi} - \dots - p'_{iq'_i}(A' - D'C')^{q'_i - 1} w_{fi}$$

(A-26)

which shows that q'_i eigenvalues of $(A' - D'C')$ are given by the roots of

$$s^{q'_i} + p'_{iq'_i} s^{q'_i - 1} + \dots + p'_{i1} = 0 \qquad \text{(A-27)}$$

By requiring D' to satisfy equations such as (A-25) for $i = 1, \dots, \ell$ a total of $(q'_1 + \dots + q'_\ell) = q'$ eigenvalues can be specified by choice of the $p'_{i\rho}$. Combining all these equations into a single matrix equation yields

$$D'C'[K^{q'_1 - 1} w_{f1}, \dots, K^{q'_\ell - 1} w_{f\ell}] = [w'_1, \dots, w'_\ell]$$

(A-28)

where

$$w'_i = p'_{i1} w_{fi} + \dots + p'_{iq'_i} A'^{q'_i - 1} w_{fi} + A'^{q'_i} w_{fi}$$

(A-29)

Relation (A-23) ensures that

$$\text{rk}\,\{C'[K^{q'_1 - 1} w_{f1}, \dots, K^{q'_\ell - 1} w_{f\ell}]\} = \ell \qquad \text{(A-30)}$$

and therefore by Lemma 4.3, (A-28) always has a solution.

APPENDIX B

ALGORITHM FOR GENERATING Λ AND THE θ_i FOR NONMUTUALLY DETECTABLE VECTORS

It is assumed that the maximal detection generators for the set of output separable vectors $\{f_1, \ldots, f_r\}$ have been found. The detection order of f_i is ν_i. If these vectors are not mutually detectable the dimension of the excess subspace is

$$k_e = n - q' - (\nu_1 + \ldots + \nu_r) \tag{B-1}$$

where $(n - q')$ is the group detection order of the above set of vectors. The orthogonal reduction procedure described in Appendix A can be used to generate a basis for the excess subspace as defined in Section 4.3.3. The algorithm begins with the terminating matrix which results from step (i) in the procedure suggested in Appendix A for finding the maximal generators for a set of vectors. Specifically, this is the terminating matrix which results when the reduction procedure is applied to M' given by (4-261). Starting with this positive semi-definite matrix the reduction process is applied to the rows of the matrix

$$\breve{M} = \begin{bmatrix} \breve{M}_1 \\ \breve{M}_2 \end{bmatrix} \tag{B-2}$$

where

$$\breve{M}_1 = \begin{bmatrix} \breve{c}_1 \\ \cdot \\ \cdot \\ \cdot \\ \breve{c}_1 K^{\nu_1 - 1} \\ \breve{c}_2 \\ \cdot \\ \cdot \\ \cdot \\ \breve{c}_2 K^{\nu_1 - 1} \\ \cdot \\ \cdot \\ \cdot \\ \breve{c}_r \\ \cdot \\ \cdot \\ \cdot \\ \breve{c}_r K^{\nu_1 - 1} \end{bmatrix} \tag{B-3}$$

and

$$\breve{M}_2 = \begin{bmatrix} \breve{c}_1 K^{\nu_1} \\ \cdot \\ \cdot \\ \cdot \\ \breve{c}_r K^{\nu_r} \\ \breve{c}_1 K^{\nu_1 + 1} \\ \cdot \\ \cdot \\ \cdot \\ \breve{c}_r K^{\nu_r + 1} \\ \cdot \\ \cdot \\ \cdot \\ \breve{c}_1 K^{\nu_1 + k_e - 1} \\ \cdot \\ \cdot \\ \cdot \\ \breve{c}_r K^{\nu_1 + k_e - 1} \end{bmatrix} \tag{B-4}$$

with K given by (4-257). The $\breve{c}_i$ $i = 1, \ldots, r$ are the rows of the $r \times n$ matrix

$$\breve{C} = [(CF)^T CF]^{-1} (CF)^T C \tag{B-5}$$

with F given by (4-242).

It can be shown that the rule presented in Appendix A for identifying intermediate termination points is also valid for this algorithm. The reasoning is somewhat different, however. Let g_i be the maximal generator for f_i. From the properties of a maximal generator it can be verified that

$$\breve{c}_i K^{\rho} g_i = 0 \quad \text{if } \rho < \nu_i - 1 \tag{B-6}$$

$$\breve{c}_i K^{\nu_i - 1} g_i \neq 0 \tag{B-7}$$

$$\breve{c}_j K^{\rho} g_i = 0 \quad \text{for all } \rho \geq 0 \quad \text{if } j \neq i \tag{B-8}$$

These relations can be used in a development similar to the proof of Lemma 4.5 to show that all $(\nu_1 + \ldots + \nu_r)$ rows of $\breve{M}_1$ are linearly independent of each other and all rows of M' as well. This means that

$$\text{rk}\, \breve{M}_1 = \nu_1 + \ldots + \nu_r \tag{B-9}$$

and

$$\text{rk} \begin{bmatrix} \breve{M}' \\ \breve{M}_1 \end{bmatrix} = \text{rk}\, M + \text{rk}\, \breve{M}_1 = q' + \nu_1 + \ldots + \nu_r \tag{B-10}$$

All auxiliary vectors associated with the rows of $\breve{M}_1$ must be nonzero because a zero auxiliary vector implies the associated row is dependent on previous rows. Assume the final nonzero auxiliary vector for c_i occurs at row $\breve{c}_i K^{\nu_i + k_{ei} - 1}$, i.e., the intermediate termination point for $\breve{c}_i$ occurs at row $\breve{c}_i K^{\nu_i + k_{ei}}$. Since no nonzero auxiliary vectors can be associated with rows in $\breve{M}_1$,

$$k_{ei} \geq 0 \quad \text{for all } i = 1, \ldots, r \tag{B-11}$$

If $k_{ei} > 0$, let $\breve{w}_i$ denote the final nonzero auxiliary vector for $\breve{c}_i$. Then

$$\breve{c}_i K^{\nu_i + k_{ei} - 1} \breve{w}_i \neq 0 \tag{B-12}$$

When $k_{ei} > 0$, $\breve{w}_i$ must appear during processing of $\breve{M}_2$. It is orthogonal to all preceding rows in $\breve{M}_2$ as well as all rows of $\breve{M}_1$ and M', so

$$M' \breve{w}_i = \underline{0} \tag{B-13}$$

$$\breve{c}_j K^{\rho} \breve{w}_i = 0 \quad \text{for } \rho = 0, \ldots, \nu_j + k_{ei} - 2 \text{ and all}$$
$$j = 1, \ldots, r \tag{B-14}$$

and

$$\breve{c}_j K^{\nu_j + k_{ei} - 1} \breve{w}_i = 0 \quad \text{if } j < i \tag{B-15}$$

Now consider the set of $(k_{e1} + \ldots + k_{er})$ vectors

$$\{\breve{w}_1, \ldots, K^{k_{e1}-1} \breve{w}_1, \breve{w}_2, \ldots, K^{k_{er}-1} \breve{w}_r\}$$

It is assumed here that all the k_{ei} are greater than zero. If some k_{ei} is zero the corresponding $\breve{w}_i$ does not appear in this set at all. But even if some k_{ei} are zero and the corresponding $\breve{w}_i$ do not appear, there is still $(k_{e1} + \ldots + k_{er})$ vectors in the set. All $\breve{w}_i$ for $i = 1, \ldots, r$ are shown in the set to avoid complicating the notation.

The case where some $k_{ei} = 0$ is discussed later. Relations (B-12) to (B-15) can be used in a development similar to the proof of Lemma 4.5 to show that all vectors in the above set are linearly independent. It can also be shown that they all lie in the null spaces of M' and $\breve{M}_1$. By construction each $\breve{w}_i$ lies in the null space M', and since this subspace is invariant with respect to K, all other vectors in the set must also be contained in the null space of M'. The fact that all the vectors lie in the null space of $\breve{M}_1$ follows from (B-14) and the assumption that $k_{ei} > 0$. The maximum possible number of independent vectors contained in the null of M' and $\breve{M}_1$ is

$$n - \mathrm{rk}\begin{bmatrix} M' \\ \breve{M}_1 \end{bmatrix} = n - q' - (\nu_1 + \ldots + \nu_r) = k_e \tag{B-16}$$

Therefore

$$k_{e1} + \ldots + k_{er} \leq k_e \tag{B-17}$$

It can be shown that if (A, C) is an observable pair, the final terminating matrix for this algorithm is the zero matrix (the case (A, C) not observable will be discussed later). If $\breve{\Omega}$ is the final terminating matrix, it must satisfy

$$M'\breve{\Omega} = \begin{bmatrix} C' \\ \cdot \\ \cdot \\ \cdot \\ C'K^{n-1} \end{bmatrix} \breve{\Omega} = \underline{0} \tag{B-18}$$

and

$$\breve{M}\breve{\Omega} = \begin{bmatrix} \breve{M}_1 \\ \breve{M}_2 \end{bmatrix} \breve{\Omega} = \underline{0} \tag{B-19}$$

which implies

$$\begin{bmatrix} \breve{C} \\ \cdot \\ \cdot \\ \cdot \\ \breve{C}K^{n-1} \end{bmatrix} \breve{\Omega} = \underline{0} \tag{B-20}$$

Observing that

$$C = C' + CF\,\breve{C} \tag{B-21}$$

it may be concluded that (B-18) and (B-20) imply

$$\begin{bmatrix} C \\ \cdot \\ \cdot \\ \cdot \\ CK^{n-1} \end{bmatrix} \breve{\Omega} = \underline{0} \tag{B-22}$$

which also implies

$$\begin{bmatrix} C \\ \cdot \\ \cdot \\ \cdot \\ CA^{n-1} \end{bmatrix} \breve{\Omega} = \underline{0} \tag{B-23}$$

If (A, C) is observable, this implies

$$\breve{\Omega} = \underline{0} \tag{B-24}$$

The positive semi-definite matrix which remains after processing $\breve{M}_1$ has a rank of

$$n - rk \begin{bmatrix} M' \\ \breve{M}_1 \end{bmatrix} = n - q' - (\nu_1 + \ldots + \nu_r) = k_e \tag{B-25}$$

Since each reduction reduces the rank of the positive semi-definite matrix by one, k_e reductions must be performed during the processing of $\breve{M}_2$ in order to produce a final terminating matrix of rank zero (the zero matrix). This means that at least k_e rows of $\breve{M}_2$ must be processed before termination. Excluding the rows $\breve{c}_i K^{\rho}$ for $\rho \geq k_{ei} + \nu_i$ (because termination of $\breve{c}_i$ occurs at $\breve{c}_i K^{\nu_i + k_{ei}}$) the total number of rows of $\breve{M}_2$ processed before termination is $(k_{e1} + \ldots + k_{er})$. Therefore

$$(k_{e1} + \ldots + k_{er}) \geq k_e \tag{B-26}$$

This result together with (B-17) implies that

$$k_{e1} + \ldots + k_{er} = k_e \tag{B-27}$$

and shows that the number of reductions is, in fact, equal to the number of rows processed before termination. This means that a reduction is performed for every row processed before termination. No zero auxiliary vector can occur before termination because that row would not produce a reduction. Hence the termination point for each $\breve{c}_i$ is signaled by the first zero auxiliary vector just as for the algorithm in Appendix A.

By virtue of (B-6) to (B-8) no vector lying in the subspace formed by the vectors

$$\{g_1, \ldots, K^{\nu_1 - 1} g_1, g_2, \ldots, K^{\nu_r - 1} g_r\}$$

can be in the null space of $\breve{M}_1$. On the other hand all vectors in the set

$$\{\breve{w}_1, \ldots, K^{k_{e1}-1}\breve{w}_1, \breve{w}_2, \ldots, K^{k_{er}-1}\breve{w}_r\}$$

are in the null space of $\breve{M}_1$. Therefore, the composite set of vectors

$$\{g_1, \ldots, K^{\nu_r-1} g_r, \breve{w}_1, \ldots, K^{k_{er}-1}\breve{w}_r\}$$

are linearly independent and form a basis for the null space of M'. Define the $n \times k_e$ matrix

$$Z_e = [\breve{w}_1, \ldots, K^{k_{e1}-1}\breve{w}_1, \breve{w}_2, \ldots, K^{k_{er}-1}\breve{w}_r] \tag{B-28}$$

Using (B-5), Equation (4-268) can be written

$$\begin{bmatrix} \gamma_{ij} \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{rj} \end{bmatrix} = \breve{C}K^{j-1} Z_e \tag{B-29}$$

and then

$$\gamma_{ij} = \breve{c}_i K^{j-1} Z_e \tag{B-30}$$

From (B-14) and (B-28) it is clear that

$$\gamma_{ij} = \underline{0} \qquad \text{for} \quad j = 1, \ldots, \nu_i \tag{B-31}$$

and so the vectors $\{\breve{w}_1, \ldots, K^{k_{e1}-1}\breve{w}_1, \breve{w}_2, \ldots, K^{k_{er}-1}\breve{w}_r\}$ form a basis for the excess subspace as described in Section 4.3.3. The θ_i are given by

$$\theta_i = \gamma_{i,\nu_i+1} = c_i K^{\nu_i} Z_e \tag{B-32}$$

From (B-14) it can be seen that the θ_i have the form

$$\theta_i = [\theta_{i1}, \dots, \theta_{ir}] \quad (1 \times k_e) \tag{B-33}$$

where

$$\theta_{ij} = [0, \dots, 0, \breve{c}_i K^{\nu_i + k_{ej} - 1} \breve{w}_j] \quad (1 \times k_{ej}) \tag{B-34}$$

and in view of (B-15)

$$\theta_{ij} = \underline{0} \quad \text{if } j > i \tag{B-35}$$

The Λ matrix can be obtained from the equation

$$KZ_e = Z_e \Lambda + \sum_{i=1}^{r} \theta_i g_i \tag{B-36}$$

Since rk $Z_e = k_e$, this equation can be solved for Λ in the closed form

$$\Lambda = [Z_e^T Z_e]^{-1} \quad Z_e^T [KZ_e - \sum_{i=1}^{r} \theta_i g_i] \tag{B-37}$$

This form is more general than is necessary, however, because from the form of Z_e in (B-28) it can be seen that Λ has the form

$$\Lambda = \begin{bmatrix} \Lambda_{11} & \cdots & \Lambda_{1r} \\ \vdots & & \vdots \\ \Lambda_{r1} & \cdots & \Lambda_{rr} \end{bmatrix} (k_e \times k_e) \tag{B-38}$$

where

$$\Lambda_{ii} = \begin{bmatrix} 0 & 0 & & 0 & \lambda_{ii1} \\ & & & \cdot & \cdot \\ 1 & 0 & & \cdot & \cdot \\ & & & \cdot & \cdot \\ 0 & 1 & & \cdot & \cdot \\ \cdot & \cdot & \ddots & \cdot & \cdot \\ \cdot & \cdot & & 0 & \cdot \\ \cdot & \cdot & & \ddots & \cdot \\ 0 & 0 & & 1 & \lambda_{iik_{ei}} \end{bmatrix} (k_{ei} \times k_{ei})$$

(B-39)

and

$$\Lambda_{ij} = \begin{bmatrix} 0 & \cdots & 0 & \lambda_{ij1} \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ 0 & \cdots & 0 & \lambda_{ijk_{ei}} \end{bmatrix} (k_{ei} \times k_{ej})$$

(B-40)

It is only necessary to solve (B-36) for the last column in each of the blocks of Λ.

It was noted earlier that if some $k_{ej} = 0$, then w_j does not appear in Z_e. In this case Λ will have fewer blocks than shown in (B-38). If there are r' nonzero k_{ei}, then there will be r'^2 blocks in Λ.

It was assumed previously that (A, C) is an observable pair. If (A, C) is not observable, the algorithm is still valid and Λ and the θ_i have the same significance as in the observable case. The only difference is that the final terminating matrix will not be zero, but will have a range space which coincides with the unobservable space of C. The eigenvalues of A (and A - DC)) associated with the unobservable space do not appear in Λ, but since they remain unchanged for any D, they are not of interest in investigating mutual detectability. The eigenvalues of Λ are those eigenvalues of $(A - DC)$ which can be influenced by the choice of D, but which are uncontrolled if D is constrained to be detector gain for all vectors in the set $\{f_1, \ldots, f_{\text{[illegible]}}\}$.

APPENDIX C

STANDARD MATRIX FORM AND DECOUPLABLE REPRESENTATION

In this appendix a transformation matrix which produces the standard form described in Section 4.3.6 is derived. Also, it will be shown how a system representation may be augmented to produce a decouplable representation.

Let the matrices A and C be $n \times n$ and $m \times n$ respectively. Assume that (A, C) is an observable pair and that

$$\text{rk } C = m \tag{C-1}$$

so all rows of C are linearly independent. A set of n independent row ($1 \times n$) vectors is to be generated as follows. Consider each row of the matrix

$$M = \begin{bmatrix} C \\ \cdot \\ \cdot \\ \cdot \\ CA^{n-1} \end{bmatrix} \tag{C-2}$$

starting with the top row and working downward. Retain only those rows which are independent of all preceding rows. Let $\{c_1, \ldots, c_1A^{n_1-1}, c_2, \ldots, c_2A^{n_2-1}, \ldots, c_m, \ldots, c_mA^{n_m-1}\}$ be the set of basis vectors so obtained, where c_i is the i^{th} row of C (the vectors are not shown in the order in which they were obtained). Since

(A, C) is observable, there must be n independent rows, so

$$n_1 + \ldots + n_m = n \tag{C-3}$$

The row $c_i A^{n_i}$ for each i does not appear in the set, so it must be dependent of the preceding rows. Then $c_i A^{n_i}$ can be expressed in terms of those basis vectors which precede it in M

$$c_i A^{n_i} = \sum_{\ell=1}^{m} \sum_{\rho=1}^{n_i} \omega_{i\ell\rho}\, c_\ell A^{\rho-1} + \sum_{\ell-1}^{i-1} \omega^*_{i\ell}\, c_\ell A^{n_i} \tag{C-4}$$

The final summation appears only for $i>1$. The terms $c_\ell A^{\rho-1}$ appear in (C-4) only if they are members of the basis, i.e., only if $\rho \leq n_\ell$. This fact can be recognized without changing the summation limits by requiring that

$$\omega_{i\ell\rho} = 0 \quad \text{if } \rho > n_\ell \tag{C-5}$$

Similarly for $\ell < i$

$$\omega^*_{i\ell} = 0 \quad \text{if } n_i \geq n_\ell \tag{C-6}$$

The second summation in (C-4) is written separately in order to call attention to the significance of the $\omega^*_{i\ell}$. From the way in which the basis vectors were selected it is clear that n_i cannot be larger than the decoupling order of c_i. On the other hand, it can be verified from (C-4) and Equation (4-433) in the alternate definition of decoupling order that if the second term in (C-4) is zero (i.e., $\omega^*_{i\ell} = 0$ for all $\ell < i$) then n_i is at least as large as the decoupling order of c_i. This implies that n_i is equal to the decoupling order of c_i if $\omega^*_{i\ell} = 0$ for all $\ell < i$ (note $\omega^*_{i\ell}$ is defined only for $\ell < i$). If $\omega^*_{i\ell} \neq 0$ for some $\ell < i$, then Equation (4-433) is not satisfied for n_i, implying that n_i is less than the decoupling order of c_i. This shows, incidently, that n_1 is always equal to the decoupling order of c_1 because the second summation does not appear in (C-4) when $i = 1$.

Now define a new set of n independent basis vectors as follows:

$$e_{in_i} = c_i \tag{C-7}$$

$$\begin{aligned} e_{ij} = c_i A^{n_i - j} &- \sum_{\ell=1}^{m} \sum_{\rho=j+1}^{n_i} \omega_{i\ell\rho}\, c_\ell\, A^{\rho - j - 1} \\ &- \sum_{\ell=1}^{i=1} \omega^*_{i\ell}\, c_\ell A^{n_i - j} \end{aligned} \tag{C-8}$$

for $j = 1, \ldots, n_i - 1$ (if $n_i > 1$) and $i = 1, \ldots, m$. Define the transformation matrix

$$T_e = \begin{bmatrix} e_{11} \\ \cdot \\ \cdot \\ \cdot \\ e_{1n_1} \\ e_{21} \\ \cdot \\ \cdot \\ \cdot \\ e_{2n_2} \\ \cdot \\ \cdot \\ \cdot \\ e_{m1} \\ \cdot \\ \cdot \\ \cdot \\ e_{mn_m} \end{bmatrix} \tag{C-9}$$

The transformed matrices are

$$\overline{A} = T_e A T_e^{-1} \tag{C-10}$$

$$\overline{C} = C T_e^{-1} \tag{C-11}$$

To identify the forms of $\overline{A}$ and $\overline{C}$, it is necessary to determine expressions for the basis vectors e_{ij} when post-multiplied by A. Now for $j = 2, \ldots, n_i - 1$

$$e_{i,j-1} = c_i A^{n_i-j+1} - \sum_{\ell=1}^{m} \sum_{\rho=j}^{n_i} \omega_{i\ell\rho}\, c_\ell A^{\rho-j}$$

$$- \sum_{e=1}^{i-1} \omega^*_{i\ell}\, c_\ell A^{n_i-j+1}$$

$$= [\, c_i A^{n_i-j} - \sum_{\ell=1}^{m} \sum_{\rho=j+1}^{n_i} \omega_{i\ell\rho}\, c_\ell A^{\rho-j-1}$$

$$- \sum_{\ell=1}^{i-1} \omega^*_{i\ell}\, c_\ell A^{n_i-j}\,]\, A - \sum_{\ell=1}^{m} \omega_{i\ell j}\, c_\ell$$

$$= e_{ij} A - \sum_{\ell=1}^{m} \omega_{i\ell j}\, c_\ell \tag{C-12}$$

or

$$e_{ij}A = e_{i,j-1} + \sum_{\ell=1}^{m} \omega_{i\ell j}\, c_\ell \quad \text{for } j = 2, \ldots, n_i - 1$$

(C-13)

For j = 1

$$e_{i1}A = [\, c_i A^{n_i - 1} - \sum_{\ell=1}^{m} \sum_{\rho=2}^{n_i} \omega_{i\ell\rho}\, c_\ell A^{\rho-2}$$

$$- \sum_{\ell=1}^{i-1} \omega_{i\ell}^{*}\, c_\ell A^{n_i-1} \;]\; A$$

$$= c_i A^{n_i} - \sum_{\ell=1}^{m} \sum_{\rho=2}^{n_i} \omega_{i\ell\rho}\, c_\ell A^{\rho-1}$$

$$- \sum_{\ell=1}^{i-1} \omega_{i\ell}^{*}\, c_\ell A^{n_i} \tag{C-14}$$

Substituting (C-4) for $c_i A^{n_i}$, all terms cancel except those involving c_ℓ, and the result is

$$e_{i1}\, A = \sum_{\ell=1}^{m} \omega_{i\ell 1}\, c_\ell = \sum_{\ell=1}^{m} \omega_{i\ell 1}\, e_{\ell n_\ell} \tag{C-15}$$

Finally

$$e_{i,n_i-1} = c_i A - \sum_{\ell=1}^{m} \omega_{i\ell n_i} c_\ell - \sum_{\ell=1}^{i-1} \omega_{i\ell}^{*} c_\ell A$$

$$= e_{in_i} A - \sum_{\ell=1}^{m} \omega_{i\ell n_i} e_{\ell n_\ell} - \sum_{\ell=1}^{i-1} \omega_{i\ell}^{*} e_{\ell n_\ell} A \tag{C-16}$$

or

$$e_{in_i} A - \sum_{\ell=1}^{i-1} \omega_{i\ell}^{*} e_{\ell n_\ell} A = e_{i,n_i-1} + \sum_{\ell=1}^{m} \omega_{i\ell n_i} e_{\ell n_\ell} \tag{C-17}$$

Combining all such equations for i = 1, . . . , m into a single matrix equation yields

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\omega_{21}^{*} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ -\omega_{m1}^{*} & \cdots & -\omega_{m,m-1}^{*} & 1 \end{bmatrix} \begin{bmatrix} e_{in_1} A \\ \vdots \\ e_{mn_m} A \end{bmatrix} = \begin{bmatrix} e_{1,n_1-1} \\ \vdots \\ e_{m,n_m-1} \end{bmatrix}$$

$$+ \begin{bmatrix} \omega_{11n_1} & \cdots & \omega_{1mn_1} \\ \vdots & & \vdots \\ \omega_{m1n_m} & \cdots & \omega_{mmn_m} \end{bmatrix} \begin{bmatrix} e_{1n_1} \\ \vdots \\ e_{mn_m} \end{bmatrix} \tag{C-18}$$

The triangular matrix on the left is clearly nonsingular and its inverse also has a triangular form

$$
\begin{bmatrix} 1 & 0 & \cdots & & 0 \\ -\omega^*_{21} & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ -\omega^*_{m_1} & \cdots & & -\omega^*_{m,m-1} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ a^*_{21} & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ a^*_{m_1} & \cdots & & a^*_{m,m-1} & 1 \end{bmatrix}
\tag{C-19}
$$

Then (C-18) yields

$$
\begin{bmatrix} e_{1n_1}A \\ \vdots \\ e_{mn_m}A \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ a^*_{21} & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ a^*_{m1} & \cdots & & a^*_{m,m-1} & 1 \end{bmatrix} \begin{bmatrix} e_{1,n_1-1} \\ \vdots \\ e_{m,n_m-1} \end{bmatrix}
$$

$$
\begin{bmatrix} a_{11n_1} & \cdots & a_{1mn_1} \\ \vdots & & \vdots \\ a_{m1n_m} & \cdots & a_{mmn_m} \end{bmatrix} \begin{bmatrix} e_{1n_1} \\ \vdots \\ e_{mn_m} \end{bmatrix}
\tag{C-20}
$$

where

$$
\begin{bmatrix} a_{11n_1} & \cdots & a_{1mn_1} \\ \vdots & & \vdots \\ a_{m1n_m} & \cdots & a_{mmn_m} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ a^*_{21} & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ a^*_{m1} & \cdots & & a^*_{m,m-1} & 1 \end{bmatrix} \begin{bmatrix} \omega_{11n_1} & \cdots & \omega_{1mn_1} \\ \vdots & & \vdots \\ \omega_{m1n_m} & \cdots & \omega_{mmn_m} \end{bmatrix}
\tag{C-21}
$$

The i^{th} row of (C-20) is

$$e_{in_i}A = e_{i,n_i-1} + \sum_{\ell=1}^{i=1} a^*_{i\ell}\, e_{\ell,n_\ell-1} + \sum_{\ell=1}^{m} a_{i\ell n_i}\, e_{\ell n_\ell}$$

(C-22)

Post-multiplying (C-10) by T_e yields

$$\bar{A}\,T_e = T_eA = \begin{bmatrix} e_{11}A \\ \vdots \\ e_{1n_1}A \\ e_{21}A \\ \vdots \\ e_{mn_m}A \end{bmatrix} \qquad \text{(C-23)}$$

From Equations (C-13), (C-15), and (C-22) the form of $\bar{A}$ is seen

$$\bar{A} = \begin{bmatrix} \bar{A}_{11} & \cdots & \bar{A}_{1m} \\ \vdots & & \vdots \\ \bar{A}_{m1} & \cdots & \bar{A}_{mm} \end{bmatrix} \quad (n \times n)$$

(C-24)

with

$$\bar{A}_{ii} = \begin{bmatrix} 0 & 0 & & 0 & a_{ii1} \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & a_{iin_i} \end{bmatrix} \quad (n_i \times n_i)$$

(C-25)

$$\overline{A}_{ij} = \begin{bmatrix} 0 & \cdots & 0 & 0 & a_{ij1} \\ \vdots & & \vdots & \vdots & \vdots \\ \vdots & & \vdots & 0 & \vdots \\ 0 & \cdots & 0 & a^*_{ij} & a_{ijn_i} \end{bmatrix} \quad (n_i \times n_j)$$

(C-26)

$$\overline{A}_{ji} = \begin{bmatrix} 0 & \cdots & 0 & a_{ji1} \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & a_{jin_i} \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix} \quad (n_j \times n_i) \qquad \text{(C-27)}$$

where $n_j \geq n_i$. The $a_{ij\ell}$ are defined as follows:

$$a_{ij\ell} = \omega_{ij\ell} \qquad i, j = 1, \ldots, m \qquad \ell = 1, \ldots, n_i - 1$$

The elements a_{ijn_i} and a^*_{ij} are given by (C-21) and (C-19) respectively. Post-multiplying (C-11) by T_e yields

$$\overline{C}\, T_e = C \qquad \text{(C-28)}$$

and from (C-7) it is easily seen that

$$\overline{C} = \begin{bmatrix} \overline{c}_1 & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \underline{0} \\ \underline{0} & \cdots & \underline{0} & \overline{c}_m \end{bmatrix} \quad (m \times n) \qquad \text{(C-29)}$$

with

$$\bar{c}_i = [0 \ . \ . \ . \ 0 \quad 1] \qquad (1 \times n_i) \tag{C-30}$$

The final zeros in the last column of $\bar{A}_{ji}$ appear when $n_j > n_i$ and are a result of (C-5). From the form of the defining equation (C-19) for the $a^*_{i\ell}$ it can be verified that the conditions on $\omega^*_{i\ell}$ given in (C-6) apply to the $a^*_{i\ell}$ as well, i.e.,

$$a^*_{i\ell} = 0 \quad \text{if } n_i \geq n_\ell \tag{C-31}$$

It is for this reason that there is no a^*_{ji} in $\bar{A}_{ji}$ given by (C-27). Also from (C-19)

$$a^*_{i\ell} = 0 \qquad \text{for } i \leq \ell \tag{C-32}$$

If the $a^*_{i\ell}$ are zero for all $\ell < i$, then n_i is equal to the decoupling order of c_i. If all the $a^*_{i\ell}$ are zero then (A, C) is a decouplable pair, and $\bar{A}$ and $\bar{C}$ have the standard form presented in Section 4.3.6.

It will now be demonstrated how a system representation may be augmented to achieve a decouplable representation. Let (A, B, C) be a minimal plant representation where A, B, and C have dimensions $n \times n$, $n \times r$, and $m \times n$ respectively. An equivalent representation is any triplet $(\tilde{A}, \tilde{B}, \tilde{C})$ (with dimensions $\tilde{n} \times \tilde{n}$, $\tilde{n} \times r$, and $m \times \tilde{n}$) satisfying

$$CA^jB = \tilde{C}\tilde{A}^j\tilde{B} \quad \text{for all } j \geq 0 \tag{C-33}$$

Since (A, B, C) is minimal, (A, C) is an observable pair. Let q_i be the decoupling order of c_i, the i^{th} row of C. Suppose

$$q_1 + \ldots + q_m > n \tag{C-34}$$

so by Theorem 4.7 (A, C) is not a decouplable pair. The triplet (A, B, C) will be augmented to obtain an equivalent observable representation $(\tilde{A}, \tilde{B}, \tilde{C})$ with $\tilde{c}_i$ having the same decoupling order as c_i, and with

$$q_1 + \dots + q_m = \tilde{n} \tag{C-35}$$

First assume

$$\text{rk } C = m \tag{C-36}$$

The case rk $C < m$ will be considered later. Let $\{c_1, \dots, c_1 A^{n_1 - 1}, c_2, \dots, c_m A^{n_m - 1}\}$ be the set of n independent basis vectors obtained as described at the beginning of this appendix. It was noted earlier that

$$n_i \leq q_i \qquad i = 1, \dots, m \tag{C-37}$$

Let

$$\tilde{A} = \begin{bmatrix} A & \tilde{A}_{12} \\ \underline{0} & \tilde{A}_{22} \end{bmatrix} \quad (\tilde{n} \times \tilde{n}) \tag{C-38}$$

$$\tilde{B} = \begin{bmatrix} B \\ \underline{0} \end{bmatrix} \quad (\tilde{n} \times r) \tag{C-39}$$

$$\tilde{C} = [C, \underline{0}] \quad (m \times \tilde{n}) \tag{C-40}$$

where $\tilde{n}$ is given by (C-35). The matrices $\tilde{A}_{22}$ and $\tilde{A}_{12}$ have dimensions $(\tilde{n} - n) \times (\tilde{n} - n)$ and $n \times (\tilde{n} - n)$ respectively, where

$$\tilde{n} - n = \sum_{i=1}^{m} (q_i - n_i) \tag{C-41}$$

It is easily verified from the form of $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ that they satisfy the requirement for an equivalent representation for any $\tilde{A}_{12}$ and $\tilde{A}_{22}$. It must now be shown that $\tilde{A}_{12}$ and $\tilde{A}_{22}$ can be chosen so as to make $(\tilde{A}, \tilde{C})$ a decouplable pair.

Before selecting $\tilde{A}_{12}$ and $\tilde{A}_{22}$ a simplification can be made which will considerably reduce the amount of algebra involved. First assume that A and C are in the standard forms (C-24) to (C-30) derived in this appendix. It was shown in Section 4.3.6 that decoupling order, and thus the property of decouplability, is invariant with respect to replacement of A by (A - DC). In the present context this means that if $([\tilde{A} - \tilde{D}''\tilde{C}], \tilde{C})$ can be shown to be a decouplable pair for any $\tilde{D}''$, then $(\tilde{A}, \tilde{C})$ is also decouplable. Let

$$\tilde{D}'' = \begin{bmatrix} D'' \\ \\ \underline{0} \end{bmatrix} \quad (\tilde{n} \times m) \tag{C-42}$$

where D'' is an $n \times m$ matrix. Then

$$\tilde{A} - \tilde{D}''\tilde{C} = \begin{bmatrix} (A - D''C) & \tilde{A}_{12} \\ \\ \underline{0} & \tilde{A}_{22} \end{bmatrix} \tag{C-43}$$

Now with A and C in the form of (C-24) to (C-30) it is easy to see that D'' can be chosen to cancel all the $a_{ij\ell}$ elements in A, yielding

$$A - D''C = \begin{bmatrix} A''_{11} & \cdots & A''_{1m} \\ \vdots & & \vdots \\ A''_{m1} & \cdots & A''_{mm} \end{bmatrix} \tag{C-44}$$

with

$$A''_{ii} = \begin{bmatrix} 0 & 0 & & 0 & 0 \\ 1 & 0 & & \vdots & \vdots \\ 0 & 1 & & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & & 1 & 0 \end{bmatrix} \quad (n_i \times n_i) \tag{C-45}$$

$$A''_{ij} = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ \vdots & & \vdots & 0 & \vdots \\ 0 & \cdots & 0 & a^*_{ij} & 0 \end{bmatrix} \quad (n_i \times n_j) \tag{C-46}$$

$$A''_{ji} = \underline{0} \qquad (n_j \times n_i) \tag{C-47}$$

where $n_j \geq n_i$. Define

$$\widetilde{A}'' = \widetilde{A} - \widetilde{D}''\widetilde{C} \tag{C-48}$$

$$A'' = A - D''C \tag{C-49}$$

Now for each i for which

$$q_i - n_i > 0 \tag{C-50}$$

let there be an associated $1 \times (\tilde{n} - n)$ row vector ζ_i. These ζ_i and $\widetilde{A}_{22}$ can be chosen arbitrarily except for the following two requirements:

(i) The $(\tilde{n} - n)$ row vectors

$$\{\zeta_i, \ldots, \zeta_i \tilde{A}_{22}^{q_i - n_i - 1} \quad ; \quad \text{all i such that } q_i - n_i > 0\}$$

are linearly independent.

$$\text{(ii)} \quad \zeta_i \tilde{A}_{22}^{q_i - n_i} = -\sum_{\rho=1}^{q_i - n_i} \tilde{\alpha}_{i\rho} \zeta_i \tilde{A}_{22}^{\rho - 1} - \sum_{\ell=2}^{i=1}{}' \tilde{\omega}_{i\ell q_i} \zeta_\ell \tilde{A}_{22}^{q_i - n} \tag{C-51}$$

where the $\tilde{\alpha}_{i\rho}$ are arbitrary scalars. The $\tilde{\omega}_{i\ell q_i}$ are scalar functions of the a^*_{ij} in the A''_{ij} and will be defined later. The prime on the second summation sign in (C-51) is to indicate that the sum is to include only those ℓ for which $q_\ell - n_\ell > 0$. The summation starts at $\ell = 2$ because, as noted near the beginning of this appendix,

$$n_1 = q_1 \tag{C-52}$$

Note that (C-51) implies the eigenvalues of $\tilde{A}_{22}$ are given by the roots of the equations

$$s^{q_i - n_i} + \tilde{\alpha}_{i, q_i - n_i} s^{q_i - n_i - 1} + \ldots + \tilde{\alpha}_{i1} = 0 \tag{C-53}$$

for those i such that $q_i - n_i > 0$. Since the $\tilde{\alpha}_{i\rho}$ are arbitrary, the eigenvalues of $\tilde{A}_{22}$ are almost arbitrary. The $\tilde{A}_{12}$ matrix is constrained to satisfy the equations

$$c_i A''^{j-1} \tilde{A}_{12} = \underline{0} \qquad \text{for } j = 1, \ldots, n_i - 1 \tag{C-54}$$

$$c_i A''^{n_i - 1} \tilde{A}_{12} = \begin{Bmatrix} \zeta_i & \text{if } n_i < q_i \\ 0 & \text{if } n_i = q_i \end{Bmatrix} \tag{C-55}$$

These make up a total of n independent equations which uniquely determine $\tilde{A}_{12}$.

It must now be shown that the decoupling order of $\tilde{c}_i$, the i^{th} row of $\tilde{C}$, is q_i for all $i = 1, \ldots, m$. To establish this it is necessary to develop a general expression for $\tilde{c}_i \tilde{A}''^j$. For $j \geq 1$

$$\tilde{C} \tilde{A}''^j = [CA''^j, \sum_{\rho=1}^{j} CA''^{\rho-1} \tilde{A}_{12} \tilde{A}_{22}^{j-\rho}] \tag{C-56}$$

so

$$\tilde{c}_i \tilde{A}''^j = [c_i A''^j, \sum_{\rho=1}^{j} c_i A''^{\rho-1} \tilde{A}_{12} \tilde{A}_{22}^{j-\rho}] \tag{C-57}$$

Using (C-54) this reduces to

$$\tilde{c}_i \tilde{A}''^j = [c_i A''^j, \underline{0}] \qquad \text{for } j = 1, \ldots, n_i - 1 \tag{C-58}$$

and

$$\tilde{c}_i \tilde{A}''^j = [c_i A''^j, \sum_{\rho=n_i}^{j} c_i A''^{\rho-1} \tilde{A}_{12} \tilde{A}_{22}^{j-\rho}]$$

$$\text{for } j \geq n_i \tag{C-59}$$

From the form of A'' and C it can be verified that

$$c_i A''^{\rho} = \sum_{\ell=1}^{i=1} \bar{\omega}_{i\ell\rho}\, c_\ell A''^{\rho} \qquad \text{for } \rho \geq n_i \tag{C-60}$$

with

$$\bar{\omega}_{i\ell\rho} = 0 \quad \text{if} \quad \rho \geq n_\ell \tag{C-61}$$

The scalars $\bar{\omega}_{i\ell\rho}$ are functions of the a^*_{ij} appearing in the A''_{ij}. The exact functional relationship between $\bar{\omega}_{i\ell\rho}$ and a_{ij} is not necessary to prove decouplability, but as a matter of interest the $\bar{\omega}_{i\ell\rho}$ are given by the matrix equation

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ \bar{\omega}_{21\rho} & & & \vdots \\ \vdots & & & 0 \\ \bar{\omega}_{m1\rho} & \cdots & \bar{\omega}_{m,m-1,\rho} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \bar{a}^*_{21} & & & \vdots \\ \vdots & & & 0 \\ \bar{a}^*_{m1\rho} & \cdots & \bar{a}^*_{m,m-1,\rho} & 1 \end{bmatrix}^{-1} \tag{C-62}$$

where

$$\bar{a}^*_{ij\rho} = \begin{Bmatrix} a^*_{ij} & \text{if } \rho < n_j \\ 0 & \text{if } \rho \geq n_j \end{Bmatrix} \tag{C-63}$$

Incidentally, $\bar{\omega}_{i\ell n_i}$ is equal to $\omega^*_{i\ell}$ in (C-4). When $n_i = q_i$, (C-60) reduces to

$$c_i A''^{\rho} = \underline{0} \qquad \text{for all } \rho \geq n_i \tag{C-64}$$

Then

$$\tilde{c}_i \tilde{A}''^{j} = \underline{0} \qquad \text{for all } j \geq n_i \tag{C-65}$$

which implies the decoupling order of $\tilde{c}_i$ cannot be larger than $n_i = q_i$. By Theorem 4.9 the decoupling order of $\tilde{c}_i$ cannot be smaller than q_i, so one may conclude immediately that if $n_i = q_i$ then the decoupling order for $\tilde{c}_i$ is q_i. Now consider the case where $n_i < q_i$. Post-multiplying (C-60) by $\tilde{A}_{12}$ yields

$$c_i A''^{\rho} \tilde{A}_{12} = \sum_{\ell=1}^{i-1} \bar{\omega}_{i\ell\rho} \, c_\ell \, A''^{\rho} \, \tilde{A}_{12} \quad \text{for } \rho \geq n_i \tag{C-66}$$

Equations (C-54), (C-55), and (C-61) indicate that the only nonzero terms in the above summation are those ℓ for which $\rho = n_\ell - 1$ and $n_\ell < q_\ell$. Then

$$c_i \, A''^{\rho} \, \tilde{A}_{12} = \sum_{\ell=2}^{i-1} \delta_{\rho, n_\ell - 1} \, \bar{\omega}_{i\ell\rho} \, \zeta_\ell \quad \text{for } \rho \geq n_i \tag{C-67}$$

where $\delta_{\rho, n_\ell - 1}$ is the Kronecker delta

$$\delta_{ij} = \left\{ \begin{matrix} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{matrix} \right\} \tag{C-68}$$

Then (C-59) becomes

$$\tilde{c}_i \, \tilde{A}''^{j} = \Big[c_i A''^{j}, \; c_i \, A''^{n_i - 1} \, \tilde{A}_{12} \, \tilde{A}_{22}^{j-n_i} + \sum_{\rho = n_i + 1}^{j} c_i A''^{\rho - 1} \, \tilde{A}_{12} \, \tilde{A}_{22}^{j-\rho} \Big]$$

$$= \left[c_i A''^{j},\ \zeta_i \tilde{A}_{22}^{n-n_i} \right.$$

$$\left. + \sum_{\rho=n_i+1}^{j} \sum_{\ell=2}^{i-1}{}' \delta_{\rho n_\ell} \bar{\omega}_{i\ell,\rho-1} \zeta_\ell \tilde{A}_{22}^{j-\rho} \right]$$

$$= \left[c_i A''^{j},\ \zeta_i \tilde{A}_{22}^{j-n_i} \right.$$

$$\left. + \sum_{\ell=2}^{i-1} \tilde{\omega}_{i\ell j} \zeta_\ell \tilde{A}_{22}^{j-n_\ell} \right] \qquad \text{for } j \geq n_i \tag{C-69}$$

where the Kronecker delta was used to eliminate the summation over ρ and

$$\tilde{\omega}_{i\ell j} = \begin{Bmatrix} \bar{\omega}_{i\ell,n_\ell-1} & \text{if } n_i + 1 \leq n_\ell \leq j \\ 0 & \text{otherwise} \end{Bmatrix} \tag{C-70}$$

Letting $j = q_i$ and using (C-51), (C-69) becomes

$$\bar{c}_i \tilde{A}''^{q_i} = \left[c_i A''^{q_i},\ \zeta_i \tilde{A}_{22}^{q_i-n_i} + \sum_{\ell=2}^{i-1} \tilde{\omega}_{i\ell q_i} \zeta_\ell \tilde{A}_{22}^{q_i-n_\ell} \right]$$

$$= \left[c_i A''^{q_i},\ - \sum_{\rho=1}^{q_i-n_i} \tilde{\alpha}_{i\rho} \zeta_i \tilde{A}_{22}^{\rho-1} \right] \tag{C-71}$$

Now define the following set of $1 \times n$ row vectors:

If $n_i = q_i$ let

$$\tilde{v}_{ij} = [c_i A''^{j-1}, \underline{0}] \quad \text{for all } j \geq 1 \tag{C-72}$$

If $n_i < q_i$ let

$$\tilde{v}_{ij} = \begin{cases} [c_i A''^{j-1}, \underline{0}] & \text{for } 1 \leq j \leq n_i \\ [\underline{0}, \zeta_i A_{22}^{j-n_i-1}] & \text{for } j > n_i \end{cases} \tag{C-73}$$

Then if $n_i = q_i$

$$\tilde{c}_i \tilde{A}''^{j-1} = \tilde{v}_{ij} \quad \text{for all } j \geq 1 \tag{C-74}$$

If $n_i < q_i$

$$\tilde{c}_i \tilde{A}''^{j-1} = \tilde{v}_{ij} \quad \text{for } 1 \leq j \leq n_i \tag{C-75}$$

and for $j > n_i$

$$\tilde{c}_i \tilde{A}''^{j-1} = [c_i A''^{j-1}, 0] + \tilde{v}_{ij} + \sum_{\ell=2}^{i-1} \tilde{\omega}_{i\ell, j-1} \tilde{v}_{\ell j} \tag{C-76}$$

Using (C-60) this becomes

$$\tilde{c}_i \tilde{A}''^{j-1} = \tilde{v}_{ij} + \sum_{\rho=1}^{i-1} \bar{\omega}_{i\rho, j-1} \tilde{v}_{\rho j} + \sum_{\ell=2}^{i-1}{}' \tilde{\omega}_{i\ell, j-1} \tilde{v}_{\ell j} \tag{C-77}$$

Now define the $m \times n$ matrices

$$\tilde{V}_j = \begin{bmatrix} \tilde{v}_{ij} \\ \cdot \\ \cdot \\ \cdot \\ \tilde{v}_{mj} \end{bmatrix} \tag{C-78}$$

From the form of (C-74), (C-75), and (C-77) it can be verified that for any j

$$\begin{bmatrix} \tilde{C} \\ \tilde{C}A'' \\ \cdot \\ \cdot \\ \cdot \\ \tilde{C}A''^{j-1} \end{bmatrix} = \hat{T}_{Vj} \begin{bmatrix} \tilde{V}_1 \\ \tilde{V}_2 \\ \cdot \\ \cdot \\ \cdot \\ \tilde{V}_j \end{bmatrix} \tag{C-79}$$

where $\hat{T}_{Vj}$ is an $(m \cdot j) \times (m \cdot j)$ triangular matrix of the form

$$\hat{T}_{Vj} = \begin{bmatrix} 1 & & 0 & \cdots & 0 \\ & \ddots & \vdots & & \vdots \\ & & \ddots & & \vdots \\ \cdots & & t_{j\ell\rho} & \ddots & 0 \\ & & \vdots & & 1 \end{bmatrix} \tag{C-80}$$

The lower left half of $\hat{T}_{Vj}$ is made up of the $\omega_{i\rho j}$ and $\omega_{i\ell j}$ in (C-77). For present purposes the significant feature of $\hat{T}_{Vj}$ is its triangular form. From (C-71)

$$\tilde{c}_i \tilde{A}''^{q_i} = [\, c_i A''^{q_i}\, ,\, 0] - \sum_{\rho=1}^{q_i - n_i} \tilde{\alpha}_{i\rho}\, \tilde{v}_{i,n_i+\rho} \tag{C-81}$$

Because of the special form of A''

$$c_i A''^{q_i} = \underline{0}$$

so

$$\tilde{c}_i \tilde{A}''^{q_i} = - \sum_{\rho=1}^{q_i-n_i} \tilde{\alpha}_{i\rho} \tilde{v}_{i,n_i+\rho} \tag{C-82}$$

This implies that $\tilde{c}_i \tilde{A}''^{q_i}$ is linearly dependent on the rows of the matrix

$$\begin{bmatrix} \tilde{V}_1 \\ \cdot \\ \cdot \\ \cdot \\ \tilde{V}_{q_i} \end{bmatrix}$$

Since

$$\begin{bmatrix} \tilde{V}_1 \\ \cdot \\ \cdot \\ \cdot \\ \tilde{V}_{q_i} \end{bmatrix} = \hat{T}_{Vq_i} \begin{bmatrix} \tilde{C} \\ \cdot \\ \cdot \\ \cdot \\ \tilde{C}\tilde{A}''^{q_i-1} \end{bmatrix} \tag{C-83}$$

this also implies $\tilde{c}_i \tilde{A}''^{q_i}$ is linearly dependent on the rows of the matrix

$$\begin{bmatrix} \tilde{C} \\ \cdot \\ \cdot \\ \cdot \\ \tilde{C}\tilde{A}''^{q_i-1} \end{bmatrix}$$

Therefore

$$\mathrm{rk} \begin{bmatrix} \tilde{C} \\ \cdot \\ \cdot \\ \cdot \\ \tilde{C}\tilde{A}''^{q_i-1} \\ \tilde{c}_i \tilde{A}''^{q_i} \end{bmatrix} = \mathrm{rk} \begin{bmatrix} \tilde{C} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \tilde{C}\tilde{A}''^{q_i-1} \end{bmatrix} \tag{C-84}$$

which shows that the decoupling order of $\tilde{c}_i$ is no larger than q_i. Since by Theorem 4.9 the decoupling order of $\tilde{c}_i$ cannot be less than

q_i, it may be concluded that it is, in fact, equal to q_i. To establish that $(\tilde{A}'', \tilde{C})$ is decouplable it is only necessary to show that this pair is observable. Because of requirement (i) on the ζ_i and the fact that the n row vectors $\{c_1, \ldots, c_1 A''^{n_1-1}, c_2, \ldots, c_m A''^{n_m-1}\}$ are linearly independent, it follows that the $\tilde{n}$ row vectors $\{\tilde{v}_{11}, \ldots, \tilde{v}_{1q_1}, \tilde{v}_{21}, \ldots, \tilde{v}_{mq_m}\}$ are likewise linearly independent. This means

$$\mathrm{rk} \begin{bmatrix} \tilde{V}_1 \\ \cdot \\ \cdot \\ \cdot \\ \tilde{V}_{\tilde{n}} \end{bmatrix} = \tilde{n} \tag{C-85}$$

And by (C-79) this implies

$$\mathrm{rk} \begin{bmatrix} \tilde{C} \\ \tilde{C}\tilde{A}'' \\ \cdot \\ \cdot \\ \cdot \\ \tilde{C}\tilde{A}''^{\tilde{n}-1} \end{bmatrix} = \tilde{n} \tag{C-86}$$

This shows that $(\tilde{A}'', \tilde{C})$ is an observable pair, and is therefore decouplable. Consequently, $(\tilde{A}, \tilde{C})$ is also decouplable.

When

$$\mathrm{rk}\ C < m \tag{C-87}$$

the development proceeds in a similar way except that for the dependent rows of C the associated ζ_i appear in $\tilde{C}$. To clarify this, suppose

$$\mathrm{rk}\ C = m' \tag{C-88}$$

and assume the first m' rows of C are independent. Partition C so that

$$C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \tag{C-89}$$

where C_1 is $m' \times n$ and

$$\text{rk } C_1 = m' \tag{C-90}$$

The rows of C_2 are dependent on the rows of C_1. Now $\tilde{A}$ and $\tilde{B}$ have the same forms as previously given in (C-38) and (C-39), but $\tilde{C}$ has the form

$$\tilde{C} = \begin{bmatrix} C_1 & \underline{0} \\ C_2 & \tilde{C}_{22} \end{bmatrix} \tag{C-91}$$

The rows of $\tilde{C}_{22}$ are chosen to be linearly independent. They play the same role as the ζ_i in the previous development. Note that this makes

$$\text{rk } \tilde{C} = m \tag{C-92}$$

It is again easily verified that this is an equivalent representation. Now A and C_1 can be put into the standard forms (C-24) to (C-30). A simplification similar to the previous case is achieved by taking

$$\tilde{D}'' = \begin{bmatrix} C''_{11} & \underline{0} \\ \underline{0} & \underline{0} \end{bmatrix} \tag{C-93}$$

where D''_{11} is $n \times m'$ and is selected so that $(A - D''_{11} C_1)$ has the form of A'' given by (C-44) to (C-48) (except that (C-44) has m'^2 blocks instead of m^2). Then

$$\tilde{A} - \tilde{D}''\tilde{C} = \begin{bmatrix} A - D''_{11} C_1 & \tilde{A}_{12} \\ \underline{0} & \tilde{A}_{22} \end{bmatrix}$$

$$= \begin{bmatrix} A'' & \tilde{A}_{12} \\ \underline{0} & \tilde{A}_{22} \end{bmatrix} \qquad \text{(C-94)}$$

From this point on, the development follows the previous case with $n_i = 0$ for the rows of C_2.

REFERENCES

1. Athans, M., and Falb, P. L., Optimal Control, McGraw-Hill, 1966.

2. Balakrishnan, A. V., "New Computing Technique in System Identification, Journal of Computer and System Sciences, Vol. 2, No. 1, June 1968, pp. 102-116.

3. Booth, T. L., Sequential Machines and Automata Theory, Wiley, 1968.

4. Brockett, R. W., Finite Dimensional Linear Systems, Wiley, 1970.

5. Chien, T. T., (Ph.D. Thesis in progress), M.I.T.

6. Crawford, B. S., "Operation and Design of Multi-Jet Spacecraft Control System", Sc.D. Thesis, M.I.T., 1968.

7. Gantmacher, F. R., The Theory of Matrices (Vol. 1), Chelsea Publishing Co., 1959.

8. Gilmore, J. P., "A Non-Orthogonal Gyro Configuration", M.S. Thesis, M.I.T., 1967.

9. Kautz, W. H., "Fault Testing and Diagrams in Combinational Digital Circuits", IEEE Transactions on Computers, Vol. C-17 No. 4, April 1968, pp. 352-366.

10. Kohavi, Z., and Lavalle, P., "Design of Sequential Machines with Fault Detection Capabilities", IEEE Transactions on Electronic Computers, Vol. EC-16, No. 4, August 1967, pp. 473-484.

11. Kohavi, Z., Switching Theory and Finite Automata, McGraw-Hill, 1970.

12. Kokotovic, P., and Rutman, R., "Sensitivity of Automatic Control Systems (Survey)", Automation and Remote Control, Vol. 26, No. 4, April 1965, pp. 727-749.

13. Kokotovich, P., "Method of Sensitivity Points in the Investigation and Optimization of Linear Control Systems", Automation and Remote Control, Vol. 25, No. 12, December 1964, pp. 1512-1518.

14. Lion, P. M., "Rapid Identification of Linear and Nonlinear Systems", AIAA Journal, Vol. 5, No. 10, October 1967, pp. 1835-1842.

15. Luenberger, D. G., "Observing the State of a Linear System", IEEE Transactions on Military Electronics, Vol. MIL-8, No. 2, April 1964, pp. 74-80.

16. Norkin, K. B., "Search Methods for Adjusting Controlled Models in Problems of Determining the Parameters of Objects", Automation and Remote Control, Vol. 29, No. 11, November 1968, pp. 1779-1784.

17. Parks, P. C., "Liapunov Redesign of Model Reference Adaptive Control Systems", IEEE Transactions on Automatic Control, Vol. AC-11, No. 3, July 1966, pp. 362-367.

18. Pierce, W. H., Failure Tolerant Computer Design, Academic Press, 1965.

19. Porter, B., and Woodhead, M. A., "Performance of Optimal Control Systems When Some of the State Variables are Not Measurable", International Journal on Control, Vol. 8, No. 2, August 1968, pp. 191-195.

20. Potter, J. E., "A Guidance-Navigation Separation Theorem", Report RE-11, Experimental Astronomy Laboratory, M.I.T., August 1964.

21. Rockwell, D., (Ph.D. Thesis in progress), M.I.T.

22. Sakharov, M. P., "Simplification of Sensitivity Models in the Design of Self-Adjusting and Adaptive Systems", Automation and Remote Control, Vol. 29, No. 5, May 1968, pp. 743-747.

23. Simon, J. D., and Mitter, S. K., "A Theory of Modal Control", Information and Control, Vol. 3, No. 4, October 1968, pp. 316-353.

24. Tryon, J. G., "Quadded Logic", Redundancy Techniques for Computing Systems, (Wilcox and Mann, Ed.), Spartan Books, 1962.

25. Tsypkin, Ya. Z., "Adaption, Training, and Self-Organization in Automatic Systems", Automation and Remote Control, Vol. 27, No. 1, January 1966, pp. 16-51.

26. Wonham, W. M., "On Pole Assignment in Multi-input Controllable Linear Systems", Technical Report 67-2, Brown University, 1967.

27. Young, P. C., "An Instrumental Variable Method for Real-Time Identification of a Noisy Process", Automatica, Vol. 6, 1970, pp. 271-287.

BIOGRAPHICAL NOTE

Richard Vernon Beard was born on October 10, 1942, in Virginia, Illinois, where he graduated from Virginia High School in 1960. He attended Purdue University from 1960 to 1965, earning a B.S. degree "with highest distinction" in Engineering Sciences in 1964, and an M.S. degree in Engineering Sciences in 1965. After two years of military service, he entered M.I.T. in 1967 as a full-time graduate student. All his graduate studies have been supported by Nation Science Foundation Fellowships. He is a member of Tau Beta Pi.

Mr. Beard is married to the former Joann Loring of Virginia, Illinois.