4

Fundamentals of Matrix Algebra

4.1 INTRODUCTION

In Chapter 3 matrix notation was used in writing state variable descriptions of dynamic systems. In this chapter a more careful and complete development of matrix algebra and some matrix calculus is presented from first principles. If the matrix usage in the preceding chapter posed no difficulty for the reader, then the introductory parts of this chapter can be treated as a review. The more advanced notions in this chapter, many of which are introduced in the problems, will still be worthwhile. If the earlier introduction of matrices caused the reader some uncertainty, then at least that exposure provided motivation for careful study of the present chapter. Experience shows that the similarities between matrix algebra and scalar algebra have a way of lulling the unwary into a sense of complacency. The usual scalar manipulations often seem to carry over and yield correct results. But ignoring the crucial differences will ultimately cause embarrassing or silly results.

Modeling, design, and analysis of control systems are the major subjects of this book. However, matrix theory and linear algebra are useful in almost every branch of science and engineering. The investment of time and effort that is required to work carefully through this and the next two chapters will pay dividends in deeper understanding, greater insight, and better computational skills later on.

4.2 NOTATION

Matrices are rectangular arrays of elements. The elements of a matrix are referred to as *scalars* and will be denoted by lowercase letters, a, b, α, β, etc. In order to define algebraic operations with matrices, it is necessary to restrict these scalar elements to be members of a field. A field $\mathcal{F}$ is any set of two or more elements for which the

operations of addition, multiplication, and division are defined, and for which the following axioms hold:

1. If $a \in \mathcal{F}$ and $b \in \mathcal{F}$, then $(a + b) = (b + a) \in \mathcal{F}$.
2. $(ab) = (ba) \in \mathcal{F}$.
3. There exists a unique null element $0 \in \mathcal{F}$ such that $a + 0 = a$ and $0(a) = 0$.
4. If $b \neq 0$, then $(a/b) \in \mathcal{F}$.
5. There exists a unique identity element $1 \in \mathcal{F}$ such that $1(a) = (a)1 = (a/1) = a$.
6. For every $a \in \mathcal{F}$ there is a unique negative element $-a \in \mathcal{F}$ such that $a + (-a) = 0$.
7. The associative, commutative, and distributive laws of algebra are satisfied.

Note that the set of integers does not form a field because axiom 4 is not necessarily true. Some examples of fields are the set of all rational numbers, the set of all real numbers, and the set of all complex numbers. The set of all rational polynomial functions also forms a field. Such functions are ratios of two polynomials $b(s)/a(s)$, where $a(s)$ and $b(s)$ are polynomials in a complex variable s (or z) with real or complex coefficients. Most matrices used in this book are assumed to be defined over the complex number field. For simplicity, many examples will be further restricted to real numbers, with the integers being a special subset. However, many control problems are posed in terms of transfer function matrices, so the field of rational polynomial functions is important. Matrices with polynomial elements also occur. The set of polynomial elements do not form a field because axiom 4 fails. Polynomial elements must be considered as members of the broader class of rational polynomial functions, just as integers are considered as special members of the field of rational numbers.

Boldface uppercase letters will be used to represent matrices, such as

$$\mathbf{A} = \begin{bmatrix} 42 & 16 \\ 5 & 3 \\ 8 & 1 \end{bmatrix}$$

Horizontal sets of entries such as (42 16) and (5 3) are called rows, whereas vertical sets of entries such as (42 5 8) are called columns. It will often be convenient to refer to the element in the ith row and jth column of $\mathbf{A}$ as a_{ij}. Rather than explicitly displaying all elements of $\mathbf{A}$, the shorthand notation $\mathbf{A} = [a_{ij}]$ will sometimes be used. If $\mathbf{A}$ has m rows and n columns, it is said to be an $m \times n$ (or m by n) matrix. In that case, the indices i and j in the shorthand notation indicate collectively the range of values $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. In particular, when $m = n = 1$, the matrix has a single element and is just a scalar. The subscripts are then unnecessary. If $n = 1$, the matrix has a single column and is called a *column matrix*. The column index j is then superfluous and is sometimes omitted. Similarly, when $m = 1$, the matrix is called a *row matrix*. Whenever $m = n$, the matrix is called a *square matrix*. In general, m and n can take on any finite integer values.

The four state variable system matrices $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$, the input vector $\mathbf{u}$, the output vector $\mathbf{y}$, and the state vector $\mathbf{x}$ were introduced in Chapter 3. These quantities

will receive major attention throughout the book. However, the same symbols also will be used in a more generic way in the discussions of matrix algebra.

4.3 ALGEBRAIC OPERATIONS WITH MATRICES

Matrix Equality

Matrices **A** and **B** are equal, written **A** = **B,** if and only if their corresponding elements are equal. That is, $a_{ij} = b_{ij}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. Of course, this means that equality can exist only between matrices of the same size, $m \times n$ in this case.

Matrix Addition and Subtraction

Matrix addition and subtraction are performed on an element-by-element basis. That is, if $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$ are both $m \times n$ matrices, then $\mathbf{A} + \mathbf{B} = \mathbf{C}$ and $\mathbf{A} - \mathbf{B} = \mathbf{D}$ indicate that the matrices $\mathbf{C} = [c_{ij}]$ and $\mathbf{D} = [d_{ij}]$ are also $m \times n$ matrices whose elements are given by $c_{ij} = a_{ij} + b_{ij}$ and $d_{ij} = a_{ij} - b_{ij}$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

Matrix Multiplication

Two types of multiplication can be defined. Multiplication of a matrix $\mathbf{A} = [a_{ij}]$ by an arbitrary scalar $\alpha \in \mathcal{F}$ amounts to multiplying every element in **A** by α. That is, $\alpha\mathbf{A} = \mathbf{A}\alpha = [\alpha a_{ij}]$.

Multiplication of an $m \times n$ matrix $\mathbf{A} = [a_{ij}]$ by a $p \times q$ matrix $\mathbf{B} = [b_{ij}]$ is now considered. In forming the product **AB** = **C,** it is said that **A** *premultiplies* **B** or equivalently, **B** *postmultiplies* **A.** This product is only defined when **A** has the same number of columns as **B** has rows. When this is true, that is, when $n = p$, **A** and **B** are said to be *conformable.* The elements of $\mathbf{C} = [c_{ij}]$ are then computed according to

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Clearly, the product **C** is an $m \times q$ matrix.

EXAMPLE 4.1 Let $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 8 \end{bmatrix}$, and $\mathbf{C} = \begin{bmatrix} 4 \\ -5 \end{bmatrix}$. Then

$$\mathbf{AB} = \begin{bmatrix} 2(1)+3(2) & 2(3)+3(4) & 2(5)+3(8) \\ 4(1)+5(2) & 4(3)+5(4) & 4(5)+5(8) \end{bmatrix} = \begin{bmatrix} 8 & 18 & 34 \\ 14 & 32 & 60 \end{bmatrix}$$

$$\mathbf{AC} = \begin{bmatrix} 2(4)+3(-5) \\ 4(4)+5(-5) \end{bmatrix} = \begin{bmatrix} -7 \\ -9 \end{bmatrix}, \qquad 10\mathbf{A} = \begin{bmatrix} 20 & 30 \\ 40 & 50 \end{bmatrix}$$

The products **BA, CA,** and **CB** are not defined. ■

Once the mechanics of matrix products are mastered, the notational advantages when dealing with simultaneous equations become clear. Matrices with purely nu-

meric entries provide good introductory examples, but the algebra being developed is much more general than this.

EXAMPLE 4.2 Consider the three coupled differential equations of Example 3.4. Ignoring initial conditions, the Laplace transforms are

$$(s^3 + a_1 s^2 + a_2 s + a_3)y_1(s) + a_2 s y_2(s) - a_3 y_3(s) = u_1(s)$$
$$-(a_4 s + a_5)y_1(s) + (s^2 + a_4 s + a_5)y_2(s) + 2a_4 s y_3(s) = u_2(s)$$
$$-a_6 y_1(s) + (s + a_6)y_3(s) = u_3(s)$$

By defining the 3×3 matrix with complex polynomial elements as

$$\mathbf{P}(s) = \begin{bmatrix} (s^3 + a_1 s^2 + a_2 s + a_3) & a_2 s & -a_3 \\ -(a_4 s + a_5) & (s_2 + a_4 s + a_5) & 2a_4 s \\ -a_6 & 0 & (s + a_6) \end{bmatrix}$$

and the 3×1 column vectors as

$$\mathbf{Y}(s) = \begin{bmatrix} y_1(s) \\ y_2(s) \\ y_3(s) \end{bmatrix} \qquad \mathbf{U}(s) = \begin{bmatrix} u_1(s) \\ u_2(s) \\ u_3(s) \end{bmatrix}$$

these equations are compactly written as

$$\mathbf{P}(s)\mathbf{Y}(s) = \mathbf{U}(s)$$

■

Kronecker Product

Other less frequently used definitions of products of matrices can be defined. One which will be found useful in Chapter 6 is the *Kronecker product,* written $\mathbf{A} \otimes \mathbf{B}$. Each scalar component a_{ij} of the first factor is multiplied by the entire matrix **B.** There are no conformability-like restrictions on the dimensions of the factors **A** and **B** that enter into such a product. If **A** is $n \times m$ and **B** is $p \times q$, then $\mathbf{A} \otimes \mathbf{B}$ will be of dimension $np \times mq$. Note that $\mathbf{A} \otimes \mathbf{B} \neq \mathbf{B} \otimes \mathbf{A}$, although these two products both have the same size. An example illustrates the Kronecker product definition.

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{11}b_{13} & a_{12}b_{11} & a_{12}b_{12} & a_{12}b_{13} \\ a_{11}b_{21} & a_{11}b_{22} & a_{11}b_{23} & a_{12}b_{21} & a_{12}b_{22} & a_{12}b_{23} \\ a_{21}b_{11} & a_{21}b_{12} & a_{21}b_{13} & a_{22}b_{11} & a_{22}b_{12} & a_{22}b_{13} \\ a_{21}b_{21} & a_{21}b_{22} & a_{21}b_{23} & a_{22}b_{21} & a_{22}b_{22} & a_{22}b_{23} \end{bmatrix}$$

Division

Division by a matrix, per se, is not defined. Thus it is not meaningful to "solve for" $\mathbf{Y}(s)$ in Example 4.2 by dividing out the $\mathbf{P}(s)$ matrix. An operation somewhat analogous to division, called *matrix inversion,* is discussed later.

The Null Matrix and the Unit Matrix

As a necessary part of scalar algebra, axioms 3 and 5 for number fields introduce a null element and an identity element. Correspondingly, the null matrix $\mathbf{0}$ is one that has all its elements equal to zero. Then, $\mathbf{A} + \mathbf{0} = \mathbf{A}$ and $\mathbf{0A} = \mathbf{0}$. Note, however, that the null matrix is not unique because the numbers of rows and columns it possesses can be any finite positive integers. Whenever necessary, the dimensions of the null matrix will be indicated by two subscripts, $\mathbf{0}_{mn}$. Another difference of major importance exists between the scalar zero and the null matrix. In scalar algebra, $ab = 0$ implies that either a or b or both are zero. No similar inference can be drawn from the matrix product $\mathbf{AB} = \mathbf{0}$. For a simple verification, let $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 3 & 4 \\ 3 & 4 \end{bmatrix}$ and form the product $\mathbf{AB}$. Although neither $\mathbf{A}$ nor $\mathbf{B}$ are null matrices, one or the other of these factors possesses properties which are in some sense (to be made clear later) somewhat like a zero. In matrix algebra there are varying degrees of "behaving like zero" of scalar algebra. A null matrix is a very strict, hard zero that has every property to be expected from the scalar zero. It will be seen later that matrix concepts of determinant, rank, trace, eigenvalue, singular value, and matrix norm can all be related to a matrix having some property associated with scalar zero.

The *identity,* or *unit, matrix* $\mathbf{I}$ is a square matrix with all elements zero, except those on the main diagonal ($i = j$ positions) are ones. The unit matrix is not unique because of its dimensions. When necessary, an $n \times n$ unit matrix will be denoted by $\mathbf{I}_n$. The unit matrix has algebraic properties similar to the scalar identity element—namely, if $\mathbf{A}$ is $m \times n$, then $\mathbf{I}_m \mathbf{A} = \mathbf{A}$ and $\mathbf{A}\mathbf{I}_n = \mathbf{A}$.

4.4 THE ASSOCIATIVE, COMMUTATIVE, AND DISTRIBUTIVE LAWS OF MATRIX ALGEBRA

Many of the associative, commutative, and distributive laws of scalar algebra carry over to matrix algebra, as summarized next:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}, \qquad \mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B}) = -\mathbf{B} + \mathbf{A}$$

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}, \qquad \alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$$

$$\alpha\mathbf{A} = \mathbf{A}\alpha, \qquad \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$$

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}, \qquad (\mathbf{B} + \mathbf{C})\mathbf{A} = \mathbf{BA} + \mathbf{CA}$$

One major difference exists between scalar and matrix algebra. Scalar multiplication is commutative, i.e., $ab = ba$. However, matrix multiplication is not commutative, i.e., $\mathbf{AB} \neq \mathbf{BA}$. In many cases the reversed product is not even defined because the conformability conditions are not satisfied. Even when both $\mathbf{A}$ and $\mathbf{B}$ are square so that $\mathbf{AB}$ and $\mathbf{BA}$ are both defined, they need not be equal. It is for this reason that it is necessary to distinguish between premultiplication and postmultiplication.

EXAMPLE 4.3 Let $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 1 & 8 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} -1 & 1 \\ 0 & 4 \end{bmatrix}$. Then $\mathbf{AB} = \begin{bmatrix} -2 & 14 \\ -1 & 33 \end{bmatrix}$ and $\mathbf{BA} = \begin{bmatrix} -1 & 5 \\ 4 & 32 \end{bmatrix}$. ■

4.5 MATRIX TRANSPOSE, CONJUGATE, AND THE ASSOCIATE MATRIX

The operation of matrix transposition is the interchanging of each row with the column of the same index number. If $\mathbf{A} = [a_{ij}]$, then the *transpose* of $\mathbf{A}$ is $\mathbf{A}^T = [a_{ji}]$. The matrix $\mathbf{A}$ is said to be *symmetric* if $\mathbf{A} = \mathbf{A}^T$. If $\mathbf{A} = -\mathbf{A}^T$, then $\mathbf{A}$ is *skew-symmetric*. An important property of matrix transposition of products is illustrated by

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T, \qquad (\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T, \ldots$$

The *conjugate of* $\mathbf{A}$, written $\overline{\mathbf{A}}$, is the matrix formed by replacing every element in $\mathbf{A}$ by its complex conjugate. Thus $\overline{\mathbf{A}} = [\overline{a}_{ij}]$. If all elements of $\mathbf{A}$ are real, then $\overline{\mathbf{A}} = \mathbf{A}$. If all elements are purely imaginary, then $\overline{\mathbf{A}} = -\mathbf{A}$.

The *associate matrix* of $\mathbf{A}$ is the conjugate transpose of $\mathbf{A}$. The order of these two operations is immaterial. Matrices satisfying $\mathbf{A} = \overline{\mathbf{A}}^T$ are called *Hermitian matrices*. *Skew-Hermitian* matrices satisfy $\mathbf{A} = -\overline{\mathbf{A}}^T$. For real matrices, symmetric and Hermitian mean the same thing.

4.6 DETERMINANTS, MINORS, AND COFACTORS

Determinants are defined for square matrices only. The determinant of the $n \times n$ matrix $\mathbf{A}$, written $|\mathbf{A}|$, is a scalar-valued function of $\mathbf{A}$. The familiar form of the determinants for $n = 1, 2$, and 3 are

$$n = 1 \qquad |\mathbf{A}| = a_{11}$$

$$n = 2 \qquad |\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$$

$$n = 3 \qquad |\mathbf{A}| = a_{11}a_{22}\,a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{32}a_{23}$$

There is a common pattern which can be generalized for any n. Each determinant has $n!$ terms, with each term consisting of n elements of $\mathbf{A}$, one from each row and from each column. However, the general pattern is inefficient for evaluating large determinants. Usually, a larger-order determinant is first reduced to an expression involving one or more smaller determinants. The methods of *Laplace expansion* and *pivotal condensation* can be used for this purpose. Also, the basic properties of determinants can be used to simplify the evaluation task. Some of these methods are discussed later.

Notice that a square null matrix and the matrix $\mathbf{B} = \begin{bmatrix} 3 & 4 \\ 3 & 4 \end{bmatrix}$, which was used in the discussion of the null matrix, both have zero determinants. Square matrices with zero determinants do possess some of the behaving like zero properties mentioned earlier. For example, the transfer function matrix equation from Example 4.2,

$$\mathbf{P}(s)\mathbf{Y}(s) = \mathbf{U}(s)$$

can have a nonzero output $\mathbf{Y}(s)$ even though $\mathbf{U}(s)$ is zero if the 3×3 matrix $\mathbf{P}$ has a zero determinant (for certain values of the complex variable s). This is in agreement with our concept of transfer function zeros for scalar systems. More often the input-output system transfer function would be expressed as

$$\mathbf{Y}(s) = \mathbf{H}(s)\mathbf{U}(s)$$

and the zeros of the square transfer function matrix $\mathbf{H}(s)$ could be defined as the values of s which make $|\mathbf{H}(s)| = 0$. The implication now is that nonzero inputs $\mathbf{U}(s)$ can cause zero outputs $\mathbf{Y}(s)$. After matrix inversion is introduced, it will be seen that the zeros of $\mathbf{P}(s)$ are the poles of $\mathbf{H}(s)$. However, since determinants are defined only for square matrices, they are not the most general tool for measuring when a matrix behaves in some sense like zero.

Minors

An $n \times n$ matrix $\mathbf{A}$ contains n^2 elements a_{ij}. Each of these has associated with it a unique scalar, called a *minor* M_{ij}. The minor M_{pq} is the determinant of the $n-1 \times n-1$ matrix formed from $\mathbf{A}$ by crossing out the pth row and qth column.

Cofactors

Each element a_{pq} of $\mathbf{A}$ has a *cofactor* C_{pq}, which differs from M_{pq} at most by a sign change. Cofactors are sometimes called signed minors for this reason and are given by $C_{pq} = (-1)^{p+q} M_{pq}$.

Determinants by Laplace Expansion

If $\mathbf{A}$ is an $n \times n$ matrix, any arbitrary row k can be selected and $|\mathbf{A}|$ is then given by $|\mathbf{A}| = \sum_{j=1}^{n} a_{kj} C_{kj}$. Similarly, Laplace expansion can be carried out with respect to any arbitrary column l, to obtain $|\mathbf{A}| = \sum_{i=1}^{n} a_{il} C_{il}$. Laplace expansion reduces the evaluation of an $n \times n$ determinant down to the evaluation of a string of $(n-1) \times (n-1)$ determinants, namely, the cofactors.

EXAMPLE 4.4 Given $\mathbf{A} = \begin{bmatrix} 2 & 4 & 1 \\ 3 & 0 & 2 \\ 2 & 0 & 3 \end{bmatrix}$. Three of its minors are

$$M_{12} = \begin{vmatrix} 3 & 2 \\ 2 & 3 \end{vmatrix} = 5, \qquad M_{22} = \begin{vmatrix} 2 & 1 \\ 2 & 3 \end{vmatrix} = 4, \quad \text{and} \quad M_{32} = \begin{vmatrix} 2 & 1 \\ 3 & 2 \end{vmatrix} = 1$$

The associated cofactors are

$$C_{12} = (-1)^3 5 = -5, \qquad C_{22} = (-1)^4 4 = 4, \qquad C_{32} = (-1)^5 1 = -1$$

Using Laplace expansion with respect to column 2 gives $|\mathbf{A}| = 4C_{12} = -20$. ■

Pivotal Condensation

Pivotal condensation [1], also called the method of Chio [2, 3], reduces an $n \times n$ determinant to a single $(n-1) \times (n-1)$ determinant and thus avoids the long string of

determinants encountered with Laplace expansion. Let a_{pq} be any nonzero element of **A.** This is called the *pivot element.* An $(n-1)\times(n-1)$ determinant is formed, with each of its elements obtained from a 2×2 determinant. Each 2×2 determinant contains a_{pq}, one other element from row p, one other element from column q, and the fourth element is from the fourth corner of the rectangle defined by the previous three elements. Let the $(n-1)\times(n-1)$ determinant be called $|\Delta|$. Then $|\mathbf{A}| = [1/(a_{pq})^{n-2}]|\Delta|$. Although the procedure looks complicated, in actual applications the large number of 2×2 determinants easily reduce to their numeric values. The method is best illustrated by an example.

EXAMPLE 4.5 Let **A** be a 4×4 matrix and assume that $a_{23}\neq 0$. Then

$$|\mathbf{A}| = \frac{1}{(a_{23})^2}\begin{vmatrix} \begin{vmatrix} a_{11} & a_{13}\\ a_{21} & a_{23}\end{vmatrix} & \begin{vmatrix} a_{12} & a_{13}\\ a_{22} & a_{23}\end{vmatrix} & \begin{vmatrix} a_{13} & a_{14}\\ a_{23} & a_{24}\end{vmatrix}\\ \begin{vmatrix} a_{21} & a_{23}\\ a_{31} & a_{33}\end{vmatrix} & \begin{vmatrix} a_{22} & a_{23}\\ a_{32} & a_{33}\end{vmatrix} & \begin{vmatrix} a_{23} & a_{24}\\ a_{33} & a_{34}\end{vmatrix}\\ \begin{vmatrix} a_{21} & a_{23}\\ a_{41} & a_{43}\end{vmatrix} & \begin{vmatrix} a_{22} & a_{23}\\ a_{42} & a_{43}\end{vmatrix} & \begin{vmatrix} a_{23} & a_{24}\\ a_{43} & a_{44}\end{vmatrix}\end{vmatrix}$$

Note that the pivot element a_{23} is in the same location relative to the other elements within each 2×2 determinant as it is in the original **A** matrix. ■

Useful Properties of Determinants

1. If **A** and **B** are both $n\times n$, then $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$.
2. $|\mathbf{A}| = |\mathbf{A}^T|$.
3. If all the elements in any row or in any column are zero, then $|\mathbf{A}| = 0$.
4. If any two rows of **A** are proportional, $|\mathbf{A}| = 0$. If a row is a linear combination of any number of other rows, then $|\mathbf{A}| = 0$. Similar statements hold for columns.
5. Interchanging any two rows (or any two columns) of a matrix changes the sign of its determinant.
6. Multiplying all elements of any one row (or column) of a matrix **A** by a scalar α yields a matrix whose determinant is $\alpha|\mathbf{A}|$.
7. Any multiple of a row (column) can be added to any other row (column) without changing the value of the determinant.

4.7 RANK AND TRACE OF A MATRIX

The *rank* of **A,** designated as r_A or rank(**A**), is defined as the size of the largest nonzero determinant that can be formed from **A.** A zero determinant is interpreted in terms of the zero of the number field being used. Therefore, a matrix with rational polynomial entries is considered singular only if its determinant is identically zero and not just if its determinant happens to have a zero value for certain isolated values of s or z. The same notion applies to the determination of rank for these matrices. The maximum possible

rank of an $m \times n$ matrix is obviously the smaller of m and n. If **A** takes on its maximum possible rank, it is said to be of full rank. If **A** is $n \times n$ (square) and has its maximal rank n, then the matrix is said to be *nonsingular.* We see below that nonsingular matrices can be inverted, but singular matrices have no inverse.

Nonsingular matrices do not possess any of the properties of behaving like zero, whereas singular matrices do. The generalization to nonsquare matrices is accomplished by the concept of rank. Note that null matrices always have a zero rank. Heuristically, full-rank matrices will not have zero-like behavior, whereas rank-deficient matrices, those having less than full rank, will. The amount by which they are rank-deficient, i.e.,

$$q = \min(n, m) - r_A$$

is called the *degeneracy,* or *nullity,* of the matrix **A.** This concept appears repeatedly in later chapters.

The rank of the product of two or more matrices is never more than the smallest rank of the matrices forming the product. For example, if r_A and r_B are the ranks of **A** and **B,** then $\mathbf{C} = \mathbf{AB}$ has rank r_C satisfying $0 \leq r_C \leq \min\{r_A, r_B\}$.

Let **A** be an $n \times n$ matrix. Then the *trace* of **A,** denoted by Tr(**A**), is the sum of the diagonal elements of **A,** $\text{Tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$. If **A** and **B** are conformable square matrices, then $\text{Tr}(\mathbf{A} + \mathbf{B}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B})$ and $\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA})$. From the definition of the trace it is obvious that $\text{Tr}(\mathbf{A}^T) = \text{Tr}(\mathbf{A})$. From this it follows that $\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{B}^T\mathbf{A}^T)$.

EXAMPLE 4.6 Let $\mathbf{A} = \begin{bmatrix} 1 & 5 & 8 \\ 3 & -1 & 2 \\ 4 & -4 & 6 \end{bmatrix}$. $\mathbf{B} = \begin{bmatrix} 1 & -1 & 8 \\ 3 & -3 & 2 \\ 4 & -4 & 6 \end{bmatrix}$. Then $|\mathbf{A}| = -112$, so that $r_A = 3$ and **A** is nonsingular. Also, $\text{Tr}(\mathbf{A}) = 6$. The matrix **B** has $|\mathbf{B}| = 0$, so $r_B < 3$. Crossing out column 2 and row 3 of **B** gives a 2×2 determinant with a value -22, so $r_B = 2$. The trace of **B** is $\text{Tr}(\mathbf{B}) = 4$. Forming **AB** and **BA** shows that $r_{AB} = 2$ and $r_{BA} = 2$. Also, $\text{Tr}(\mathbf{A} + \mathbf{B}) = 10 = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B})$ and $\text{Tr}(\mathbf{AB}) = 100 = \text{Tr}(\mathbf{BA})$. Note that $\text{Tr}(\mathbf{AB}) \neq \text{Tr}(\mathbf{A})\text{Tr}(\mathbf{B})$. ■

4.8 MATRIX INVERSION

The inverse of the scalar element a is $1/a$, or a^{-1}. It satisfies $a(a^{-1}) = (a^{-1})a = 1$. If an arbitrary matrix **A** is to have an analogous inverse $\mathbf{B} = \mathbf{A}^{-1}$, then the following must hold:

$$\mathbf{BA} = \mathbf{AB} = \mathbf{I}$$

Because of conformability requirements, this can never be true if **A** is not square. In addition, **A** must have a nonzero determinant, i.e., **A** must be nonsingular. When this is true, **A** has a unique inverse given by

$$\mathbf{A}^{-1} = \frac{\mathbf{C}^T}{|\mathbf{A}|}$$

where $\mathbf{C}$ is the matrix formed by the cofactors C_{ij}. The matrix $\mathbf{C}^T$ is called the *adjoint matrix,* Adj($\mathbf{A}$). Thus the inverse of a nonsingular matrix is

$$\mathbf{A}^{-1} = \text{Adj}(\mathbf{A})/|\mathbf{A}|$$

EXAMPLE 4.7 Let $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 6 & 3 \\ 1 & 3 & 5 \end{bmatrix}$. Then $\mathbf{A}^{-1}$ does not exist since $|\mathbf{A}| = 0$. Since $|\mathbf{B}| = -2$, $\mathbf{B}^{-1}$ exists and is given by $\mathbf{B}^{-1} = \begin{bmatrix} -2 & 1 \\ 3/2 & -1/2 \end{bmatrix}$. Similarly, $\mathbf{D}^{-1} = \frac{1}{70}\begin{bmatrix} 21 & -7 & 0 \\ -7 & 19 & -10 \\ 0 & -10 & 20 \end{bmatrix}$. ■

The definition of the matrix inverse just given is perfectly general. It applies to matrices whose elements are functions of time or of complex variables, such as s in Example 4.2. Thus

$$\mathbf{H}(s) = \mathbf{P}(s)^{-1} \tag{4.1}$$

As with scalar transfer functions, the poles of $\mathbf{H}(s)$ are those values of s for which elements of $\mathbf{H}$ are unbounded. The definition of the matrix inverse shows that this happens when $|\mathbf{P}(s)| = 0$. The poles of $\mathbf{H}(s)$ are the zeros of $\mathbf{P}(s)$, as mentioned earlier.

Inversion of large matrices by direct application of the above definition is tedious. Numerical techniques such as *Gaussian elimination* are often used. Matrix partitioning can also be employed to obtain a matrix inverse in terms of several smaller inverses. Another method, based on the Cayley-Hamilton theorem, is given in Chapter 8.

In many applications the entries in a matrix to be inverted are complex numbers. Although the general definition of the matrix inverse is valid for complex entries, the actual calculations become much more cumbersome. Some computer algorithms for matrix inversion are restricted to matrices with real numbers for elements. Problem 4.22 gives some partial results on inverting complex matrices using only real numbers. In other cases, the complex inverse is not the desired end result but is only an intermediate quantity that occurs while solving for $\mathbf{X}$ in simultaneous equations of the form

$$\mathbf{AX} = \mathbf{B} \quad \text{or} \quad \mathbf{XA} = \mathbf{B}$$

It is shown in Problem 4.23 that if $\mathbf{A}$ and $\mathbf{B}$ have complex entries which occur in complex conjugate pairs in a certain way, then the solution for $\mathbf{X}$ is purely real and can easily be computed using only real matrix inversion calculations.

The Inverse of a Product

Let $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots, \mathbf{W}$ be any number of conformable nonsingular matrices. Then

$$(\mathbf{ABC} \cdots \mathbf{W})^{-1} = \mathbf{W}^{-1} \cdots \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$$

Some Matrices with Special Relationships to Their Inverses

If $\mathbf{A}^{-1} = \mathbf{A}$, $\mathbf{A}$ is said to be *involutory.*
If $\mathbf{A}^{-1} = \mathbf{A}^T$, $\mathbf{A}$ is said to be *orthogonal.*
If $\mathbf{A}^{-1} = \overline{\mathbf{A}}^T$, $\mathbf{A}$ is said to be *unitary.*

4.9 PARTITIONED MATRICES

Any matrix $\mathbf{A}$ can be subdivided or partitioned into a number of smaller submatrices. If conformable matrices are partitioned in a compatible fashion, the submatrices can be treated just as if they were scalar elements when performing the operations of addition and multiplication. Of course, the order of the products is not arbitrary, as it would be with scalars.

EXAMPLE 4.8 $\mathbf{AB} = \mathbf{C}$ can be partitioned in various ways. A few of them are given next:

(a) $$\left[\begin{array}{c}\mathbf{A}_1 \\ \hline \mathbf{A}_2\end{array}\right][\mathbf{B}_1 \mid \mathbf{B}_2] = \left[\begin{array}{c|c}\mathbf{A}_1\mathbf{B}_1 & \mathbf{A}_1\mathbf{B}_2 \\ \hline \mathbf{A}_2\mathbf{B}_1 & \mathbf{A}_2\mathbf{B}_2\end{array}\right] = \left[\begin{array}{c|c}\mathbf{C}_1 & \mathbf{C}_2 \\ \hline \mathbf{C}_3 & \mathbf{C}_4\end{array}\right]$$

(b) $$\left[\begin{array}{c|c}\mathbf{A}_1 & \mathbf{A}_2 \\ \hline \mathbf{A}_3 & \mathbf{A}_4\end{array}\right]\left[\begin{array}{c}\mathbf{B}_1 \\ \hline \mathbf{B}_2\end{array}\right] = \left[\begin{array}{c}\mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_2\mathbf{B}_2 \\ \hline \mathbf{A}_3\mathbf{B}_1 + \mathbf{A}_4\mathbf{B}_2\end{array}\right] = \left[\begin{array}{c}\mathbf{C}_1 \\ \hline \mathbf{C}_2\end{array}\right]$$

(c) $$\left[\begin{array}{c|c}\mathbf{A}_1 & \mathbf{A}_2 \\ \hline \mathbf{A}_3 & \mathbf{A}_4\end{array}\right]\left[\begin{array}{c|c}\mathbf{B}_1 & \mathbf{B}_2 \\ \hline \mathbf{B}_3 & \mathbf{B}_4\end{array}\right] = \left[\begin{array}{c|c}\mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_2\mathbf{B}_3 & \mathbf{A}_1\mathbf{B}_2 + \mathbf{A}_2\mathbf{B}_4 \\ \hline \mathbf{A}_3\mathbf{B}_1 + \mathbf{A}_4\mathbf{B}_3 & \mathbf{A}_3\mathbf{B}_2 + \mathbf{A}_4\mathbf{B}_4\end{array}\right] = \left[\begin{array}{c|c}\mathbf{C}_1 & \mathbf{C}_2 \\ \hline \mathbf{C}_3 & \mathbf{C}_4\end{array}\right]$$ ■

Partitioned matrices were used without comment in Sec. 3.5 where subsystems of state variable systems were combined to obtain an overall composite state variable model. That application illustrates one possible motivation for using partitioned matrices. They allow the clustering together of groups of variables and treating the group by an identifying symbol. It is an intermediate step between displaying *all* the scalar entries and displaying the entire matrix by just a single symbol.

Partitioned matrices can be used to find an expression for the inverse of a nonsingular matrix $\mathbf{A}$. If $\mathbf{A}$ is partitioned into four submatrices, then $\mathbf{A}^{-1} = \mathbf{B}$ will also have four submatrices:

$$\mathbf{AB} = \mathbf{I} \quad \text{or} \quad \left[\begin{array}{c|c}\mathbf{A}_1 & \mathbf{A}_2 \\ \hline \mathbf{A}_3 & \mathbf{A}_4\end{array}\right]\left[\begin{array}{c|c}\mathbf{B}_1 & \mathbf{B}_2 \\ \hline \mathbf{B}_3 & \mathbf{B}_4\end{array}\right] = \left[\begin{array}{c|c}\mathbf{I} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}\end{array}\right]$$

The partitioned form implies four separate matrix equations, two of which are $\mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_2\mathbf{B}_3 = \mathbf{I}$ and $\mathbf{A}_3\mathbf{B}_1 + \mathbf{A}_4\mathbf{B}_3 = \mathbf{0}$. These can be solved simultaneously for $\mathbf{B}_1$ and $\mathbf{B}_3$. The remaining two equations give $\mathbf{B}_2$ and $\mathbf{B}_4$ and lead to the result

$$\mathbf{A}^{-1} = \left[\begin{array}{c|c}(\mathbf{A}_1 - \mathbf{A}_2\mathbf{A}_4^{-1}\mathbf{A}_3)^{-1} & -\mathbf{A}_1^{-1}\mathbf{A}_2(\mathbf{A}_4 - \mathbf{A}_3\mathbf{A}_1^{-1}\mathbf{A}_2)^{-1} \\ \hline -\mathbf{A}_4^{-1}\mathbf{A}_3(\mathbf{A}_1 - \mathbf{A}_2\mathbf{A}_4^{-1}\mathbf{A}_3)^{-1} & (\mathbf{A}_4 - \mathbf{A}_3\mathbf{A}_1^{-1}\mathbf{A}_2)^{-1}\end{array}\right]$$

Several matrix identities can be derived by starting with the reversed order, $\mathbf{BA} = \mathbf{I}$, repeating the above process, and then using the uniqueness of $\mathbf{B} = \mathbf{A}^{-1}$ to equate the various terms. One such identity, called the *matrix inversion lemma,* is particularly useful. A general form is

$$(\mathbf{A}_1 - \mathbf{A}_2\mathbf{A}_4^{-1}\mathbf{A}_3)^{-1} = \mathbf{A}_1^{-1} + \mathbf{A}_1^{-1}\mathbf{A}_2(\mathbf{A}_4 - \mathbf{A}_3\mathbf{A}_1^{-1}\mathbf{A}_2)^{-1}\mathbf{A}_3\mathbf{A}_1^{-1}$$

By letting $\mathbf{A}_1 = \mathbf{P}^{-1}$, $\mathbf{A}_2 = \mathbf{H}^T$, $\mathbf{A}_3 = \mathbf{H}$, and $\mathbf{A}_4 = -\mathbf{Q}^{-1}$, an extremely useful special form of the inversion lemma that will be encountered in recursive weighted least squares is obtained:

$$[\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{Q}\mathbf{H}]^{-1} = \mathbf{P} - \mathbf{P}\mathbf{H}^T[\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{Q}^{-1}]^{-1}\mathbf{H}\mathbf{P}$$

Diagonal, Block Diagonal, and Triangular Matrices

If the only nonzero elements of a square matrix $\mathbf{A}$ are on the main diagonal, then $\mathbf{A}$ is called a *diagonal matrix*. This is often written as $\mathbf{A} = \text{diag}[a_{11} \quad a_{22} \quad \dots \quad a_{nn}]$. For this case, $|\mathbf{A}| = a_{11}a_{22}\cdots a_{nn}$ and $\mathbf{A}^{-1} = \text{diag}[1/a_{11} \quad 1/a_{22} \quad \dots \quad 1/a_{nn}]$. The unit matrix is a special case with all $a_{ii} = 1$.

A *block diagonal,* or *quasidiagonal,* matrix is a square matrix that can be partitioned so that the only nonzero elements are contained in square submatrices along the main diagonal,

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & & & \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ & & & \mathbf{A}_k \end{bmatrix} = \text{diag}[\mathbf{A}_1 \quad \mathbf{A}_2 \quad \cdots \quad \mathbf{A}_k]$$

For this case $|\mathbf{A}| = |\mathbf{A}_1||\mathbf{A}_2|\cdots|\mathbf{A}_k|$ and $\mathbf{A}^{-1} = \text{diag}[\mathbf{A}_1^{-1} \quad \mathbf{A}_2^{-1} \quad \cdots \quad \mathbf{A}_k^{-1}]$, provided that $\mathbf{A}^{-1}$ exists.

A square matrix which has all its elements below (above) the main diagonal equal to zero is called an *upper triangular* (*lower triangular*) matrix. The determinant of any triangular matrix is the product of its diagonal elements.

4.10 ELEMENTARY OPERATIONS AND ELEMENTARY MATRICES

Three basic operations on a matrix, called *elementary operations,* are as follows:

1. The interchange of two rows (or of two columns).
2. The multiplication of every element in a given row (or column) by a scalar α.
3. The multiplication of the elements of a given row (or column) by a scalar α, and adding the result to another row (column). The original row (column) is unaltered.

It is stressed that the nature of the scalar α depends upon which number field is in use. For example, if α is a rational polynomial function, the entire discussion of elementary operations still applies without change. When these row operations are applied to the unit matrix, the resultant matrices are called elementary matrices, and are denoted as follows:

$\mathbf{E}_{p,q}$: pth and qth rows of $\mathbf{I}$ interchanged

$\mathbf{E}_p(\alpha)$: pth row of $\mathbf{I}$ multiplied by α

$\mathbf{E}_{p,q}(\alpha)$: pth row of $\mathbf{I}$ multiplied by α and added to qth row

The elementary matrices are all nonsingular. In fact,

$$|\mathbf{E}_{p,q}| = -1, \qquad |\mathbf{E}_p(\alpha)| = \alpha, \qquad |\mathbf{E}_{p,q}(\alpha)| = 1$$

The inverse of each elementary matrix is also an elementary matrix.

Premultiplication (postmultiplication) of a matrix by one of the elementary matrices performs the corresponding elementary row (column) operation on that matrix.

By performing a sequence of elementary row and column operations, any matrix of rank r can be reduced to one of the following normal forms:

$$\mathbf{I}_r, \qquad [\mathbf{I}_r \mid \mathbf{0}], \qquad \begin{bmatrix} \mathbf{I}_r \\ \hline \mathbf{0} \end{bmatrix}, \qquad \left[\begin{array}{c|c} \mathbf{I}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]$$

These are special cases of, or analogous to, matrices in the *row-reduced echelon* form. They are also sometimes called *Hermite normal forms.* These will be defined more formally in Chapter 6, where their value will be more fully appreciated. Thus elementary operations provide a practical means of computing the rank of a matrix, but they have many other uses as well.

4.11 DIFFERENTIATION AND INTEGRATION OF MATRICES

When a matrix $\mathbf{A}$ has elements which are functions of a scalar variable (such as time), differentiation and integration of the matrix are defined on an element-by-element basis. If $\mathbf{A}(t) = [a_{ij}(t)]$, then $d\mathbf{A}/dt = \dot{\mathbf{A}} = [\dot{a}_{ij}(t)]$ and $\int \mathbf{A}(\tau)\, d\tau = [\int a_{ij}(\tau)\, d\tau]$. Because of the integration rule, Laplace transforms and inverse Laplace transforms of matrices are also found on element-by-element basis. This is also true for Z-transforms of matrices.

Equation (*4.1*) could be applied to $\mathbf{P}(s)$ given in Example 4.2 to determine $\mathbf{H}(s)$ directly. However, there is an alternative approach which can—and frequently will—be used. In Example 3.4 a state variable model for the system in question was determined. The form of that model is

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{4.2a}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \tag{4.2b}$$

and the specific form of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ have been given. Assume that the initial conditions for $\mathbf{x}(t)$ at time $t = 0$ are given by $\mathbf{x}_0$. Taking the Laplace transform of Eq. (*4.2a*) and combining the two $\mathbf{X}(s)$ terms gives

$$(s\mathbf{I}_n - \mathbf{A})\mathbf{X}(s) = \mathbf{x}_0 + \mathbf{B}\mathbf{U}(s)$$

Premultiplying both sides by the inverse of $(s\mathbf{I}_n - \mathbf{A})$ gives

$$\mathbf{X}(s) = (s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{x}_0 + (s\mathbf{I}_n - A)^{-1}\mathbf{B}\mathbf{U}(s) \tag{4.3}$$

The transform of the state vector consists of the initial condition response plus the forced response. These are often called the zero input response and the zero state response, respectively. When Eq. (*4.3*) is substituted into the Laplace transform of Eq. (*4.2b*), the result is

$$\mathbf{Y}(s) = \{\mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}\}\mathbf{U}(s) + \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{x}_0 \tag{4.4}$$

Ignoring the initial condition term, the input-output transfer function matrix is given by the first term in Eq. (*4.4*),

$$\mathbf{H}(s) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \tag{4.5}$$

This is an alternative to the calculation in Eq. (*4.1*) for the transfer function matrix. In more general problems with input derivatives, the input-output expression in Example 4.2 will take the form

$$\mathbf{P}(s)\mathbf{Y}(s) = \mathbf{N}(s)\mathbf{U}(s)$$

or

$$\mathbf{Y}(s) = \mathbf{P}(s)^{-1}\mathbf{N}(s)\mathbf{U}(s) \tag{4.6}$$

where $\mathbf{N}(s)$ will be an $m \times r$ matrix of polynomials in s. As before, $\mathbf{P}$ will be an $m \times m$ matrix, where the number of inputs components in $\mathbf{U}$ is r and the number of output components in $\mathbf{Y}$ is m. The generalization of Eq. (*4.1*) for the $m \times r$ input-output transfer function matrix is

$$\mathbf{H}(s) = \mathbf{P}(s)^{-1}\mathbf{N}(s) \tag{4.7}$$

This is *one* particular form of the *matrix fraction description* (MFD) of a multiple-input, multiple-output system transfer function. It is an alternative to the state variable form of Eq. (*4.5*) [4].

Differentiation of a Determinant

Two useful rules for differentiating a determinant are

$$\frac{\partial|\mathbf{A}|}{\partial a_{ij}} = C_{ij} \qquad \text{(follows immediately from the Laplace expansion)}$$

If the $n \times n$ matrix $\mathbf{A}$ is a function of t, then $|\mathbf{A}|$ is also a function of t. Then $d|\mathbf{A}|/dt$ is just the sum of n separate determinants. The first determinant has row (or column) one differentiated, the second has row (or column) two differentiated, and so on through all n rows (columns).

4.12 ADDITIONAL MATRIX CALCULUS

4.12.1 The Gradient Operator and Differentiation with Respect to a Vector

Let $f(x_1, x_2, \ldots, x_n)$ be a scalar-valued function of n variables x_i. The variables may be, but need not be, state variables in the present discussion. For notational convenience

the dependence on n variables x_i is written as $f(\mathbf{x})$, with $\mathbf{x}$ being a vector with components x_i. The n partial derivatives of $f(\mathbf{x})$, $\partial f/\partial x_i$, will be used frequently. It is convenient to group these partials into an array and give the array a special symbol. A single-rowed array, a row vector, could be—and frequently is—used for this purpose. Here the array is arranged as a single column, that is, a column vector. This convention is an arbitrary choice. Both forms are used in the literature, and both are referred to as the *gradient vector* [1]. Three different symbols are frequently used to identify the gradient of $f(\mathbf{x})$, $\nabla_{\mathbf{x}} f = \text{grad}_{\mathbf{x}} f = df/d\mathbf{x}$. The meaning of these symbols is given by

$$\nabla_{\mathbf{x}} f = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n} \right]^T \tag{4.8}$$

The only differences which arise between the row and column vector definitions are the presence or absence of the transpose in various algebraic manipulations. Conformability requirements for matrix multiplication must always be satisfied and can be used to determine whether a row or column definition is implied.

EXAMPLE 4.9 Let $f_1(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{y}$, a bilinear function. Expanding this in terms of individual components gives $f_1(\mathbf{x}) = \sum_i \sum_j a_{ij} x_i y_j$. A typical component of the gradient is $\partial f_1/\partial x_k = \sum_i \sum_j a_{ij} (\partial x_i/\partial x_k) y_j$.

Using the independence of the x_i components gives $\partial x_i/\partial x_k = \delta_{ik} = \begin{cases} 1 \text{ if } i = k \\ 0 \text{ if } i \neq k \end{cases}$. This means that only one term in the summation over i is nonzero, so $\partial f_1/\partial x_k = \sum_j a_{kj} y_j$. This is just the kth component of the matrix product $\mathbf{A}\mathbf{y}$, so the gradient vector for this example is $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{y}) = \mathbf{A}\mathbf{y}$. ■

EXAMPLE 4.10 If $f_2(\mathbf{x}) = \mathbf{y}^T\mathbf{A}\mathbf{x}$, then $\nabla_{\mathbf{x}} f_2 \neq \mathbf{y}^T\mathbf{A}$. The gradient operator *is not* simply a canceling of the $\mathbf{x}$ vector as might be inferred from Example 4.9. By convention, the gradient is a column vector, so it cannot be equal to the row vector $\mathbf{y}^T\mathbf{A}$. The correct expression for the gradient is $\nabla_{\mathbf{x}} f_2 = \mathbf{A}^T\mathbf{y}$. ■

EXAMPLE 4.11 Let $f_3(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$, a quadratic form. In summation notation,

$$f_3(\mathbf{x}) = \sum_i \sum_j a_{ij} x_i x_j \quad \text{and} \quad \frac{\partial f_3}{\partial x_k} = \sum_i \sum_j a_{ij} \left\{ \frac{\partial x_i}{\partial x_k} x_j + x_i \frac{\partial x_j}{\partial x_k} \right\} = \sum_j a_{kj} x_j + \sum_i a_{ik} x_i$$

Returning to matrix notation $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x}$. If $\mathbf{A} = \mathbf{A}^T$, as is usual when dealing with quadratic forms, then $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$. ■

The geometrical interpretation of the gradient is often useful. To aid in visualization, the vector $\mathbf{x}$ is restricted to two components. Then for each point $\mathbf{x}$ in the plane, the function $f(\mathbf{x})$ has some prescribed value. Figure 4.1 shows such a function.

The equation $f(\mathbf{x}) = c$, with c constant, specifies a locus of points in the plane. Figure 4.2 shows the locus of points in the plane for several different values of c.

At a given point such as $\mathbf{x}_0$ in Figure 4.2, $\nabla_{\mathbf{x}} f$ is a vector normal to the curve $f(\mathbf{x}) = c$, and it points in the direction of increasing values of $f(\mathbf{x})$. The gradient defines the direction of maximum increase of the function $f(\mathbf{x})$.

The derivative of a scalar function with respect to a vector yields a vector, the gradient vector. If a vector-valued function of a vector, $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}) f_2(\mathbf{x}) \cdots f_m(\mathbf{x})]^T$, is

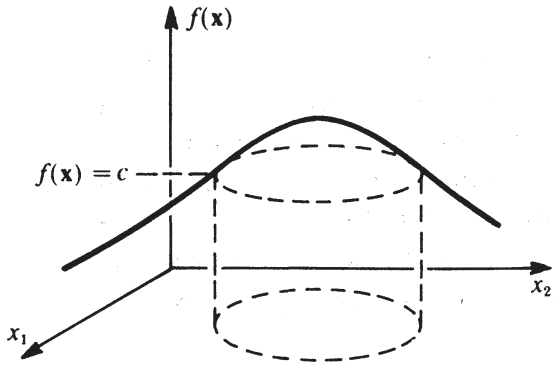


Figure 4.1

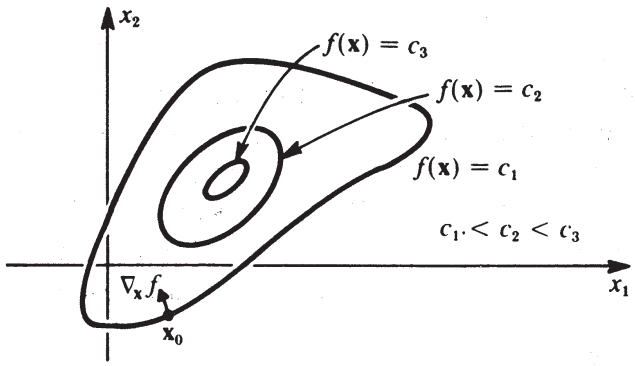


Figure 4.2

considered, the gradient of each component $f_i(\mathbf{x})$ is a column vector with the same dimension as $\mathbf{x}$. If the function $\mathbf{f}(\mathbf{x})$ is transposed to a row vector, then

$$\nabla_{\mathbf{x}}\mathbf{f}^T(\mathbf{x}) = [\nabla_{\mathbf{x}} f_1 \quad \nabla_{\mathbf{x}} f_2(\mathbf{x}) \quad \cdots \quad \nabla_{\mathbf{x}} f_m(\mathbf{x})] \tag{4.9}$$

is an $n \times m$ matrix whose columns are gradients. The transpose of this matrix will be denoted by the symbols $\nabla_{\mathbf{x}}\mathbf{f}(x)$ or simply $d\mathbf{f}/d\mathbf{x}$. That is, $d\mathbf{f}/d\mathbf{x} \triangleq [\partial f_i/\partial x_j]$ and this $m \times n$ matrix is the *Jacobian matrix*. Note that the symbol $d\mathbf{f}/d\mathbf{x}$ is just a suggestive name attached to the prescribed array of first partial derivatives. The symbol $d\mathbf{f}/d\mathbf{x}$ could just as well have been called $\partial\mathbf{f}/\partial\mathbf{x}$, and it *could* have been defined alternatively as $[\partial f_j/\partial x_i]$.

EXAMPLE 4.12 Let $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, and let the jth column of $\mathbf{A}^T$ be $\mathbf{a}_j$. Then since $f_j(\mathbf{x})$ can be written as $\mathbf{a}_j^T\mathbf{x} = \mathbf{x}^T\mathbf{a}_j$, it is immediate that $\nabla_x f_j(\mathbf{x}) = \mathbf{a}_j$, so that $d[\mathbf{A}\mathbf{x}]/d\mathbf{x} = \mathbf{A}$. ■

Let $\mathbf{g}(\mathbf{x})$ be a vector-valued function of $\mathbf{x}$ and let $f(\mathbf{g})$ be a scalar-valued function. Then the chain rule gives

$$\begin{aligned} df/dx_i &= \partial f/\partial g_1\, \partial g_1/\partial x_i + \partial f/\partial g_2\, \partial g_2/\partial x_i + \cdots + \partial f/\partial g_n\, \partial g_n/\partial x_i \\ &= [\partial g_j/\partial x_i]\, df/d\mathbf{g} \end{aligned}$$

By virtue of the convention adopted previously, the total gradient can be written as $\nabla_{\mathbf{x}} f = [d\mathbf{g}/d\mathbf{x}]^T\, df/d\mathbf{g} = [d\mathbf{g}/d\mathbf{x}]^T\, \nabla_{\mathbf{g}} f$.

EXAMPLE 4.13 Let $f(\mathbf{g}) = \mathbf{g}^T\mathbf{W}\mathbf{g}$ and let $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{y}$. Then $d\mathbf{g}/d\mathbf{x} = \mathbf{A}$ and $df/d\mathbf{g} = 2\mathbf{W}\mathbf{g}$, so that $d\{[\mathbf{A}\mathbf{x} - \mathbf{y}]^T\mathbf{W}[\mathbf{A}\mathbf{x} - \mathbf{y}]\}/d\mathbf{x} = 2\mathbf{A}^T\mathbf{W}\mathbf{g}$. ■

The preceding extends to scalar functions of several vector functions of $\mathbf{x}$. If $f = f(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}))$, then

$$df/d\mathbf{x} = [d\mathbf{g}/d\mathbf{x}]^T\, df/d\mathbf{g} + [d\mathbf{h}/d\mathbf{x}]^T df/d\mathbf{h}$$

EXAMPLE 4.14 Let $f = \mathbf{g}^T(\mathbf{x})\mathbf{h}(\mathbf{x})$ and let $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ and $\mathbf{h}(\mathbf{x}) = \mathbf{B}\mathbf{x} + \mathbf{c}$. Then

$$df/d\mathbf{x} = \mathbf{A}^T[\mathbf{B}\mathbf{x} + \mathbf{c}] + \mathbf{B}^T[\mathbf{A}\mathbf{x} + \mathbf{b}]$$ ■

The second partial derivatives of a function of a vector also arise on occasion. When $f(\mathbf{x})$ is a scalar-valued function, the matrix of all second partial derivatives, called the *Hessian matrix*, will be denoted by

$$\frac{d^2f}{dx^2}=\begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & & & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{4.10}$$

It becomes notationally awkward to continue these definitions to the second derivative of a vector function **f** with respect to a vector. This would require a three- dimensional array, with a typical element being $\partial^2 f_i/\partial x_j\, \partial x_k$.

4.12.2 Generalized Taylor Series

The Taylor series expansion is one of the most useful formulas in the analysis of nonlinear equations. The expansion of a scalar-valued function of a scalar is recalled [5]:

$$f(x_0+\delta x)=f(x_0)+\left.\frac{df}{dx}\right|_0 \delta x+\left.\frac{1}{2!}\frac{d^2f}{dx^2}\right|_0 \delta x^2+\cdots \tag{4.11}$$

The notation $\left.\dfrac{df}{dx}\right|_0$ indicates that all derivatives are evaluated at the point x_0.

The Taylor expansion of a function of two variables x and y is

$$\begin{aligned} f(x_0+\delta x, y_0+\delta y)=f(x_0,y_0)&+\left.\frac{\partial f}{\partial x}\right|_0 \delta x+\left.\frac{\partial f}{\partial y}\right|_0 \delta y \\ &+\frac{1}{2!}\left[\left.\frac{\partial^2 f}{\partial x^2}\right|_0 \delta x^2+2\left.\frac{\partial^2 f}{\partial x\,\partial y}\right|_0 \delta x\,\delta y+\left.\frac{\partial^2 f}{\partial y^2}\right|_0 \delta y^2\right]+\cdots \end{aligned} \tag{4.12}$$

If the two variables x and y are used to define the vector $\mathbf{x}=[x \quad y]^T$, the preceding expansion is more compactly written as

$$f(\mathbf{x}_0+\delta\mathbf{x})=f(\mathbf{x}_0)+(\nabla_x f|_0)^T\delta\mathbf{x}+\frac{1}{2!}\delta\mathbf{x}^T\left.\frac{d^2f}{d\mathbf{x}^2}\right|_0 \delta\mathbf{x}+\cdots \tag{4.13}$$

This generalized form of the Taylor expansion is valid for any number of components of the vector **x.**

An m component vector-valued function **f**(**x**) can be viewed as m separate scalar functions. The Taylor expansion through the first two terms can be written as

$$\mathbf{f}(\mathbf{x}_0+\delta\mathbf{x})=\mathbf{f}(\mathbf{x}_0)+\nabla_x\mathbf{f}|_{\mathbf{x}_0}\delta\mathbf{x}+\cdots \tag{4.14}$$

The slight discrepancy in the gradient terms of Eqs. (*4.13*) and (*4.14*) is due to the definition of the gradient as a column vector and is the reason why the row definition is preferred by some authors. Higher terms in Eq. (*4.14*) cannot be written conveniently in matrix notation. However, the first-order terms in $\delta\mathbf{x}$ are frequently all that are used, and a good approximation results if all components of $\delta\mathbf{x}$ are sufficiently small.

The generalized Taylor series is the standard tool used in linearizing nonlinear system equations. This is utilized in Chapter 15.

4.12.3 Vectorizing a Matrix

When matrices were introduced in the previous chapter they were presented as a convenient way of arranging a number of scalar variables. No certain order was required initially, although later manipulations (e.g., matrix products) did expect certain conventions be followed. In the next chapter it will be seen that no special arrangement of elements in an abstract vector is required. We could arrange the elements around a circle if we wished. The sum of two such "circles" would be a circle, as would the product of a circle with a scalar. That is, the set of such circles is closed under the operations of addition and multiplication by a scalar. There is no compelling reason to use such a strange definition, but it could be done. However, there are times when it is more convenient to arrange elements that are traditionally in a rectangular array, and hence thought of as a matrix, into a linear column array, hence having the characteristics of a vector. The rearrangement from a rectangular array to a column is called *vectorizing* the matrix. It could be done in row order, column order, or perhaps some other scanning order. Here, the column order will be arbitrarily selected but consistently used. The capital letters used to indicate matrices will be retained, but the vectorized column form will be indicated by enclosing the letter in parenthesis. That is, if

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots \\ a_{21} & a_{22} & a_{23} & \cdots \\ a_{31} & a_{32} & a_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} \quad \text{then} \quad (\mathbf{A}) = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{12} \\ a_{22} \\ a_{32} \\ a_{13} \\ \vdots \end{bmatrix}$$

To save space, the transpose can be written $(\mathbf{A})^T = [a_{11}\ a_{21}\ a_{31}\ \cdots\ a_{12}\ a_{22}\ a_{32}\ \cdots\ a_{13}\ a_{23}\ \cdots]$. The operations of vectorizing and transposition do not commute, that is, $(\mathbf{A}^T) \neq (\mathbf{A})^T$. Why introduce such nonstandard notation? It has not been done just to make the valid point that arrangement order is arbitrary as long as it is used in a logically consistent fashion. The major reason for introducing the concept of vectorizing a matrix is that it is *convenient* in many situations. One such situation is the derivation of matrix gradient expressions, because it allows use of the already familiar vector gradient results. Another convenient application of vectorized matrices appears in Chapter 6 in the solution of a special class of linear matrix equations called Lyapunov equations.

Matrix Gradients

Let $\mathbf{A}$ be an $m \times n$ matrix and let $f(\mathbf{A})$ be a scalar-valued function of $\mathbf{A}$. Then the matrix gradient of f with respect to $\mathbf{A}$ is written as $\partial f(\mathbf{A})/\partial \mathbf{A}$. This is just a symbol for the $m \times n$ rectangular array of scalar derivatives $[\partial f(\mathbf{A})/\partial a_{ij}]$. By vectorizing $\mathbf{A}$ and applying familiar formulas for vector gradients with respect to $(\mathbf{A})$, a column arrangement of the same scalar derivatives is easily derived for many functions f. Then the definition of $(\mathbf{A})$ can be "undone" to write the matrix gradient in the more traditional rectangular matrix form. This process is now used to derive a catalog of useful matrix gradient results. If $\mathbf{A}$ and the unit matrix are both vectorized, then it is easy to see that $\text{Tr}[\mathbf{A}] = (\mathbf{A})^T(\mathbf{I})$. In this example the trace operation is an example of the function f discussed previously, and its matrix argument must now be square in this case. Then $\partial \text{Tr}[\mathbf{A}]/\partial \mathbf{A} = \partial(\mathbf{A})^T(\mathbf{I})/\partial(\mathbf{A}) = (\mathbf{I}) = \mathbf{I}$. A number of gradients of the trace of matrix products are easily derived by noting that $\text{Tr}[\mathbf{AB}] = (\mathbf{A})^T(\mathbf{B}^T) = (\mathbf{B})^T(\mathbf{A}^T) = (\mathbf{B}^T)^T(\mathbf{A}) = (\mathbf{A}^T)^T(\mathbf{B})$. Therefore, $\partial \text{Tr}[\mathbf{AB}]/\partial \mathbf{A} = \mathbf{B}^T$, $\partial \text{Tr}[\mathbf{AB}]/\partial \mathbf{B} = \mathbf{A}^T$, $\partial \text{Tr}[\mathbf{AB}]/\partial \mathbf{A}^T = \mathbf{B}$, and $\partial \text{Tr}[\mathbf{AB}]/\partial \mathbf{B}^T = \mathbf{A}$. Consider the trace of a three-term product $\text{Tr}[\mathbf{ABC}] = \text{Tr}[\mathbf{BCA}] = \text{Tr}[\mathbf{CAB}]$ and define $\mathbf{BC} \equiv \mathbf{D}$. Then, in vectorized form, $\text{Tr}[\mathbf{ABC}] = (\mathbf{A})^T(\mathbf{D}^T)$, so that $\partial \text{Tr}[\mathbf{ABC}]/\partial \mathbf{A} = (\mathbf{D}^T) = \mathbf{D}^T = \mathbf{C}^T\mathbf{B}^T$. Similarly, $\partial \text{Tr}[\mathbf{ABC}]/\partial \mathbf{B} = \mathbf{A}^T\mathbf{C}^T$ and $\partial \text{Tr}[\mathbf{ABC}]/\partial \mathbf{C} = \mathbf{B}^T\mathbf{A}^T$. The general rule that $\partial \text{Tr}[\ \ ]/\partial \mathbf{A}^T = \{\partial \text{Tr}[\ \ ]/\partial \mathbf{A}\}^T$ for any matrix $\mathbf{A}$ allows the transpose of the preceding results to be used to get the gradient with respect to $\mathbf{A}^T$, $\mathbf{B}^T$, or $\mathbf{C}^T$. Now consider two-term or three-term products where a factor is repeated, such as in $\mathbf{A}^T\mathbf{A}$, $\mathbf{ABA}$, or $\mathbf{ABA}^T$. Since $\text{Tr}[\mathbf{AA}^T] = (\mathbf{A})^T(\mathbf{A}^T)$, it is found that $\partial \text{Tr}[\mathbf{AA}^T]/\partial \mathbf{A} = 2(\mathbf{A}) = 2\mathbf{A}$. This suggests a chain-rule-like behavior in which the total gradient is the sum of the factors obtained by treating one factor at a time as variable and treating the other factors as fixed. That is, $\partial \text{Tr}[\mathbf{ABA}^T]/\partial \mathbf{A} = \partial \text{Tr}[\mathbf{AC}]/\partial \mathbf{A} + \partial \text{Tr}[\mathbf{DA}^T]/\partial \mathbf{A}$, where $\mathbf{C} = \mathbf{BA}^T$ and $\mathbf{D} = \mathbf{AB}$ are treated as constants until after the differentiation. Previous formulas can be used on each of the terms in the sum to give

$$\partial \text{Tr}[\mathbf{ABA}^T]/\partial \mathbf{A} = (\mathbf{BA}^T)^T + \mathbf{AB} = \mathbf{AB}^T + \mathbf{AB}$$

Likewise, $\partial \text{Tr}[\mathbf{ABA}]/\partial \mathbf{A} = \mathbf{A}^T\mathbf{B}^T + \mathbf{B}^T\mathbf{A}^T$. Extensions to other variations are almost limitless. For example, $\partial \text{Tr}[\mathbf{ABAC}]/\partial \mathbf{A} = \mathbf{C}^T\mathbf{A}^T\mathbf{B}^T + \mathbf{B}^T\mathbf{A}^T\mathbf{C}^T$. The previous example is a special case of this result with $\mathbf{C} = \mathbf{I}$. Also, $\partial \text{Tr}[\mathbf{ABA}^T\mathbf{C}]/\partial \mathbf{A} = \mathbf{C}^T\mathbf{AB}^T + \mathbf{CAB}$. Some formulas involving $\mathbf{A}^{-1}$ can be derived by using $\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{AA}^{-1}$ and using the previous chain rule to find $\partial \text{Tr}[\mathbf{A}^{-1}]/\partial \mathbf{A} = \partial \text{Tr}[\mathbf{A}^{-1}\mathbf{C}]/\partial \mathbf{A} + \partial \text{Tr}[\mathbf{DA}^{-1}]/\partial \mathbf{A} + \partial \text{Tr}[\mathbf{AE}]/\partial \mathbf{A}$, where $\mathbf{C} = \mathbf{AA}^{-1} = \mathbf{I}$, $\mathbf{D} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, and $\mathbf{E} = [\mathbf{A}^{-1}]^2 = \mathbf{A}^{-2}$. Therefore, $\partial \text{Tr}[\mathbf{A}^{-1}]/\partial \mathbf{A} = 2\partial \text{Tr}[\mathbf{A}^{-1}]/\partial \mathbf{A} + \mathbf{E}^T$, or $\partial \text{Tr}[\mathbf{A}^{-1}]/\partial \mathbf{A} = -\mathbf{E}^T = -[\mathbf{A}^{-2}]^T$. Similar manipulation can be used to show that

$$\partial \text{Tr}[\mathbf{BA}^{-1}\mathbf{C}]/\partial \mathbf{A} = -[\mathbf{A}^{-1}\mathbf{CBA}^{-1}]^T$$

One final matrix gradient expression where the scalar-valued function f is the determinant of its argument can be derived without use of the intermediate vectorization process. Since $|\mathbf{A}| = \Sigma a_{ij} C_{ij}$, where the Laplace expansion can be along any row or column and where C_{ij} is the ijth cofactor, it is easy to see that

$$\partial |\mathbf{A}|/\partial \mathbf{A} = [C_{ij}] = \text{Adj}[\mathbf{A}]^T$$

Applications of matrix gradients arise in several optimal control and estimation problems. A scalar cost function, such as the trace, is minimized with respect to a selectable matrix by setting the matrix gradient to zero.

REFERENCES

1. DeRusso, P. M., R. J. Roy, and C. M. Close: *State Variables for Engineers,* John Wiley, New York, 1965.
2. Hovanessian, S. A. and L. A. Pipes: *Digital Computer Methods in Engineering,* McGraw-Hill, New York, 1969.
3. Pipes, L. A.: *Applied Mathematics for Engineers and Physicists,* 3rd ed., McGraw-Hill, New York, 1971.
4. Kailath, T.: *Linear Systems,* Prentice Hall, Englewood Cliffs, N.J., 1980.
5. Taylor, A. E. and R. W. Mann: *Advanced Calculus,* Xerox, Boston, 1972.
6. Cruz, J. B., Jr., J. S. Freundenberg, and D. P. Looze: "A Relationship Between Sensitivity and Stability of Multivariable Feedback Systems," *IEEE Transactions on Automatic Control,* Vol. AC-26, No. 1, Feb. 1981, pp. 66–74.
7. Lawson, C. L. and R. J. Hanson: *Solving Least Squares Problems,* Prentice-Hall, Englewood Cliffs, N.J., 1974.
8. Kunz, K. S.: *Numerical Analysis,* McGraw-Hill, New York, 1957.

ILLUSTRATIVE PROBLEMS

Introductory Manipulations

4.1 Let

$$\mathbf{A}=\begin{bmatrix}1 & 4\\ 2 & 5\end{bmatrix},\quad \mathbf{B}=\begin{bmatrix}3 & 1\\ 1 & 3\end{bmatrix},\quad \text{and}\quad \mathbf{C}=\begin{bmatrix}42 & 16\\ 5 & 3\\ 8 & 1\end{bmatrix}$$

Compute $\mathbf{A}+\mathbf{B}$, $\mathbf{A}-\mathbf{B}$, $\mathbf{AB}$, $\mathbf{BA}$, $\mathbf{CA}$, $\mathbf{CB}$, $\mathbf{AC}$, and $\mathbf{AC}^T$.

$$\mathbf{A}+\mathbf{B}=\begin{bmatrix}1+3 & 4+1\\ 2+1 & 5+3\end{bmatrix}=\begin{bmatrix}4 & 5\\ 3 & 8\end{bmatrix},\qquad \mathbf{A}-\mathbf{B}=\begin{bmatrix}-2 & 3\\ 1 & 2\end{bmatrix}$$

$$\mathbf{AB}=\begin{bmatrix}1(3)+4(1) & 1(1)+4(3)\\ 2(3)+5(1) & 2(1)+5(3)\end{bmatrix}=\begin{bmatrix}7 & 13\\ 11 & 17\end{bmatrix},\qquad \mathbf{BA}=\begin{bmatrix}5 & 17\\ 7 & 19\end{bmatrix}$$

$$\mathbf{CA}=\begin{bmatrix}42(1)+16(2) & 42(4)+16(5)\\ 5(1)+3(2) & 5(4)+3(5)\\ 8(1)+1(2) & 8(4)+1(5)\end{bmatrix}=\begin{bmatrix}74 & 248\\ 11 & 35\\ 10 & 37\end{bmatrix},\qquad \mathbf{CB}=\begin{bmatrix}142 & 90\\ 18 & 14\\ 25 & 11\end{bmatrix}$$

$\mathbf{AC}$ is not defined because of the conformability rule: $(2\times 2)(3\times 2)$.

$$\mathbf{AC}^T=\begin{bmatrix}1 & 4\\ 2 & 5\end{bmatrix}\begin{bmatrix}42 & 5 & 8\\ 16 & 3 & 1\end{bmatrix}=\begin{bmatrix}1(42)+4(16) & 1(5)+4(3) & 1(8)+4(1)\\ 2(42)+5(16) & 2(5)+5(3) & 2(8)+5(1)\end{bmatrix}$$

$$=\begin{bmatrix}106 & 17 & 12\\ 164 & 25 & 21\end{bmatrix}$$

Multiple Variable Systems and Transfer Matrices

4.2 Consider the multiple-input, multiple-output feedback system shown in Figure 4.3, where $\mathbf{G}_1$, $\mathbf{G}_2$, $\mathbf{H}_1$, and $\mathbf{H}_2$ are transfer function matrices and $\mathbf{R}$, $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{F}_1$, and $\mathbf{F}_2$ are column matrices. If $\mathbf{R}$ has r components, $\mathbf{V}$ has m components, $\mathbf{C}$ has n components, and $\mathbf{D}$ has p components, determine the dimensions of all other matrices in the diagram.

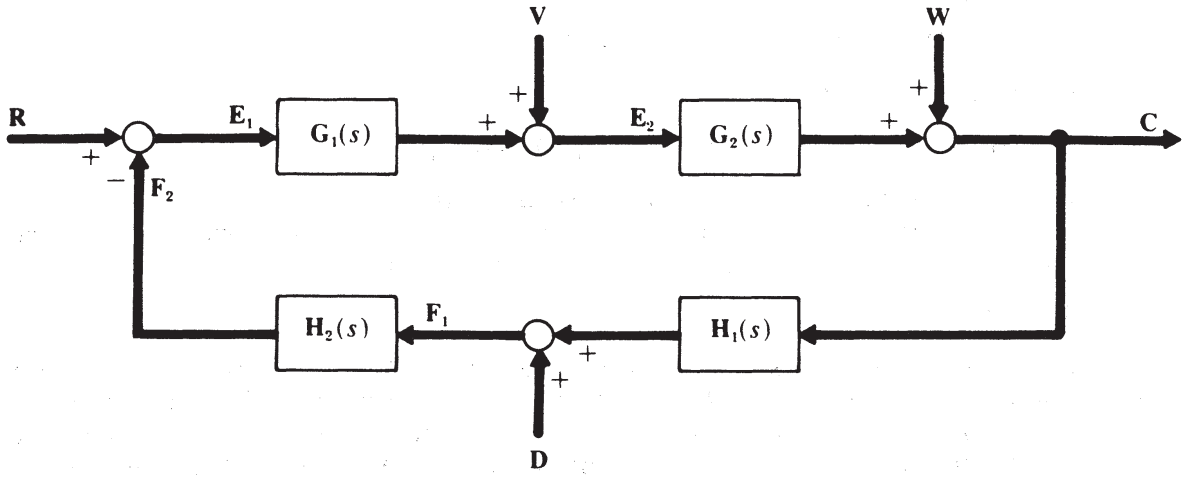


Figure 4.3

In order for $\mathbf{R} - \mathbf{F}_2 = \mathbf{E}_1$ to make sense, $\mathbf{F}_2$ and $\mathbf{E}_1$ must be $r \times 1$ matrices like $\mathbf{R}$. If $\mathbf{G}_1\mathbf{E}_1$ is to be conformable, $\mathbf{G}_1$ must have r columns. Since this product adds to $\mathbf{V}$, it must be an $m \times 1$ matrix. So $\mathbf{G}_1$ must be $m \times r$. Since $\mathbf{G}_1\mathbf{E}_1 + \mathbf{V} = \mathbf{E}_2$, $\mathbf{E}_2$ must be an $m \times 1$ matrix. Conformability requires that $\mathbf{G}_2$ have m columns, and since $\mathbf{G}_2\mathbf{E}_2 + \mathbf{W} = \mathbf{C}$ is an $n \times 1$ matrix, the transfer matrix $\mathbf{G}_2$ must be $n \times m$. Similar reasoning requires that $\mathbf{H}_1$ be a $p \times n$ matrix, $\mathbf{F}_1$ be a $p \times 1$ matrix, and $\mathbf{H}_2$ be an $r \times p$ matrix.

4.3 Referring to the system of Problem 4.2, derive the overall transfer matrix relating the input $\mathbf{R}$ to the output $\mathbf{C}$.

Ignoring all inputs except $\mathbf{R}$, this system is described by five matrix equations:

$$\mathbf{E}_1 = \mathbf{R} - \mathbf{F}_2, \qquad \mathbf{E}_2 = \mathbf{G}_1\mathbf{E}_1, \qquad \mathbf{C} = \mathbf{G}_2\mathbf{E}_2, \qquad \mathbf{F}_1 = \mathbf{H}_1\mathbf{C}, \quad \text{and} \quad \mathbf{F}_2 = \mathbf{H}_2\mathbf{F}_1$$

Depending upon the sequence of algebraic manipulations used to eliminate all terms except $\mathbf{R}$ and $\mathbf{C}$, different forms of the final result are obtained. Four different sequences are presented.

1. Eliminating $\mathbf{E}_2$ gives $\mathbf{C} = \mathbf{G}_2\mathbf{G}_1\mathbf{E}_1$ and eliminating $\mathbf{F}_1$ gives $\mathbf{F}_2 = \mathbf{H}_2\mathbf{H}_1\mathbf{C}$. Combining gives $\mathbf{C} = \mathbf{G}_2\mathbf{G}_1\mathbf{R} - \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1\mathbf{C}$ or $[\mathbf{I}_n + \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1]\mathbf{C} = \mathbf{G}_2\mathbf{G}_1\mathbf{R}$ so that

$$\mathbf{C} = [\mathbf{I}_n + \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1]^{-1}\mathbf{G}_2\mathbf{G}_1\mathbf{R}$$

The similarity with the scalar closed-loop transfer function is apparent.

2. In this sequence $\mathbf{E}_2$ is first isolated and then, in turn, related to $\mathbf{C}$ as follows: $\mathbf{E}_2 = \mathbf{G}_1(\mathbf{R} - \mathbf{F}_2)$, but

$$\mathbf{F}_2 = \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{E}_2 \quad \text{so that } [\mathbf{I}_m + \mathbf{G}_1\mathbf{H}_2\mathbf{H}_1\mathbf{G}_2]\mathbf{E}_2 = \mathbf{G}_1\mathbf{R}$$

or

$$\mathbf{E}_2 = [\mathbf{I}_m + \mathbf{G}_1\mathbf{H}_2\mathbf{H}_1\mathbf{G}_2]^{-1}\mathbf{G}_1\mathbf{R} \quad \text{so that } \mathbf{C} = \mathbf{G}_2\mathbf{E}_2 = \mathbf{G}_2[\mathbf{I}_m + \mathbf{G}_1\mathbf{H}_2\mathbf{H}_1\mathbf{G}_2]^{-1}\mathbf{G}_1\mathbf{R}$$

Note the different arrangement of terms and the size of the matrix to be inverted.

3. If $\mathbf{E}_1$ is first isolated in a similar way, we obtain

$$\mathbf{E}_1 = [\mathbf{I}_r + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1]^{-1}\mathbf{R}$$

from which

$$\mathbf{C} = \mathbf{G}_2\mathbf{G}_1[\mathbf{I}_r + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1]^{-1}\mathbf{R}$$

4. If $\mathbf{F}_1$ is first isolated, we obtain

$$\mathbf{F}_1 = [\mathbf{I}_p + \mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\mathbf{H}_2]^{-1}\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\mathbf{R}$$

Premultiplying this by $\mathbf{H}_2$ gives $\mathbf{F}_2$. Subtracting $\mathbf{F}_2$ from $\mathbf{R}$ gives

$$\mathbf{E}_1 = \{\mathbf{I}_r - \mathbf{H}_2[\mathbf{I}_p + \mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\mathbf{H}_2]^{-1}\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\}\mathbf{R}$$

Premultiplying this by $\mathbf{G}_2\mathbf{G}_1$ gives

$$\mathbf{C} = \{\mathbf{G}_2\mathbf{G}_1 - \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2[\mathbf{I}_p + \mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\mathbf{H}_2]^{-1}\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\}\mathbf{R}$$

4.4 Use the result of Problem 4.3 to establish some useful matrix identities.

In the previous problem four results of the form $\mathbf{C} = \mathbf{T}_i\mathbf{R}$ were found, with the differences contained in the four $\mathbf{T}_i$ matrices. Equating two forms for the output $\mathbf{C}$ gives

$$\mathbf{T}_i\mathbf{R} = \mathbf{T}_j\mathbf{R}$$

In general, this equation is not sufficient for concluding that $\mathbf{T}_i = \mathbf{T}_j$. However, in this case it must be true for *all* $\mathbf{R}$ that $(\mathbf{T}_i - \mathbf{T}_j)\mathbf{R} = \mathbf{0}$ so that $\mathbf{T}_i - \mathbf{T}_j$ must be the null matrix, or $\mathbf{T}_i = \mathbf{T}_j$. Therefore, the results of Problem 4.3 give the following matrix identities. They are true for any matrices which satisfy the conformability conditions, whenever the indicated inverses exist:

$$[\mathbf{I}_n + \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1]^{-1}\mathbf{G}_2\mathbf{G}_1 = \mathbf{G}_2[\mathbf{I}_m + \mathbf{G}_1\mathbf{H}_2\mathbf{H}_1\mathbf{G}_2]^{-1}\mathbf{G}_1 = \mathbf{G}_2\mathbf{G}_1[\mathbf{I}_r + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1]^{-1}$$
$$= \mathbf{G}_2\mathbf{G}_1 - \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2[\mathbf{I}_p + \mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\mathbf{H}_2]^{-1}\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1$$

4.5 Consider the system of Problem 4.2. a. Find the closed-loop transfer function matrices which relate the following input-output matrix pairs: $\mathbf{R}$ to $\mathbf{E}_1$, $\mathbf{V}$ to $\mathbf{E}_2$, $\mathbf{W}$ to $\mathbf{C}$, and $\mathbf{D}$ to $\mathbf{F}_1$. Also determine the characteristic equations. b. Discuss the matrix generalization of the return difference concept.

(a) Starting with the basic relations in Problem 4.3 and using similar manipulations leads to

$$\mathbf{E}_1 = [\mathbf{I}_r + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1]^{-1}\mathbf{R} \qquad \mathbf{C} = [\mathbf{I}_n + \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1]^{-1}\mathbf{W}$$
$$\mathbf{E}_2 = [\mathbf{I}_m + \mathbf{G}_1\mathbf{H}_2\mathbf{H}_1\mathbf{G}_2]^{-1}\mathbf{V} \qquad \mathbf{F}_1 = [\mathbf{I}_p + \mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\mathbf{H}_2]^{-1}\mathbf{D}$$

Since, for example,

$$[\mathbf{I}_r + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1]^{-1} = \frac{\text{Adj } [\mathbf{I}_r + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1]}{|\mathbf{I}_r + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1|}$$

the characteristic equation is obtained by setting the determinant in the denominator to zero. The characteristic equation is a property of the system and not the particular inputs or outputs considered. It is reasonable to expect that the following determinant identities are true (this is proven in Chapter 7, Problem 7.22, by other means):

$$|\mathbf{I}_r + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1| = |\mathbf{I}_m + \mathbf{G}_1\mathbf{H}_2\mathbf{H}_1\mathbf{G}_2| = |\mathbf{I}_n + \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1|$$
$$= |\mathbf{I}_p + \mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\mathbf{H}_2|$$

The roots of any one of these determinants constitute the poles of the multivariable system, provided no cancellation with numerator terms has occurred.

(b) The matrices contained in the preceding determinants and whose inverses appear in the preceding transfer functions are termed *return difference matrices*. These are the multivariable generalization of the return difference function in Chapter 2. In single-loop scalar problems, the return difference $R_d(s)$ is the same regardless of where the loop is broken. In

the multivariable case here, the return difference matrix is different at each point in the loop because of the order in which the factors $\mathbf{G}_i$ and $\mathbf{H}_i$ appear. In this matrix case,

$$\mathbf{F}_d(s) = \{\mathbf{I} + [\text{the product of } \mathbf{G}_i \text{ and } \mathbf{H}_i \text{ matrices in the order encountered while traversing the loop backward from the point in question}]\}$$

The *determinants* of the return difference matrices are all the same, but the matrices themselves differ from point to point. As in the scalar case, a small return difference indicates low stability margins and poor sensitivity to disturbances and model variations. Although the scalar case is unambiguous, the "size" of the return difference matrix can be measured in various ways. What is really needed is a measure of how near these matrices are to being singular. In general, the determinant is a poor measure of near singularity. The singular values of a matrix, developed in Chapter 7, are a much more meaningful measure, and the preceding four matrices will all have different singular values.

Disturbance rejection: The contribution of each of the four inputs to the output of Figure 4.3 are

$$\mathbf{C}_R = \mathbf{G}_2\mathbf{G}_1[\mathbf{I}_r + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1]^{-1}\mathbf{R}$$

$$\mathbf{C}_V = \mathbf{G}_2[\mathbf{I}_m + \mathbf{G}_1\mathbf{H}_2\mathbf{H}_1\mathbf{G}_2]^{-1}\mathbf{V}$$

$$\mathbf{C}_W = [\mathbf{I}_n + \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1]^{-1}\mathbf{W}$$

$$\mathbf{C}_D = \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2[\mathbf{I}_p + \mathbf{H}_1\mathbf{G}_2\mathbf{G}_1\mathbf{H}_2]^{-1}\mathbf{D}$$

Presumably $\mathbf{R}$ is a desired input. If the return differences are made "large," say by increasing $\mathbf{G}_1$, the outputs due to $\mathbf{V}$ and $\mathbf{W}$ disturbance inputs can be made "small." Without getting into exactly what large and small mean here, this is the essential idea behind disturbance rejection in feedback systems. Sensitivity to model errors is also reduced as the return difference matrix is made larger [6]. The return difference matrix, like its scalar counterpart, is frequency-dependent. For any real system the transfer function magnitudes will eventually go to zero as $\omega \to \infty$. Therefore, the return differences will eventually go to $\mathbf{I}$ (or 1). Robust control system design is concerned with maintaining a sufficiently high return difference over the frequency range of interest and then having it approach its asymptotic value in a graceful fashion.

4.6 A single-input, two-output feedback system has the form shown in Figure 4.3, with

$$\mathbf{G}_2\mathbf{G}_1 = \begin{bmatrix} \dfrac{1}{s+1} \\ \dfrac{1}{s+2} \end{bmatrix} \qquad \mathbf{H}_2\mathbf{H}_1 = [s \quad 1]$$

Find the characteristic equation and the transfer function matrix relating $\mathbf{R}$ to $\mathbf{C}$, using two different formulations.

Using $|\mathbf{I}_2 + \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1| = 0$ gives

$$\left| \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \dfrac{s}{s+1} & \dfrac{1}{s+1} \\ \dfrac{s}{s+2} & \dfrac{1}{s+2} \end{bmatrix} \right| = \left(1 + \frac{s}{s+1}\right)\left(1 + \frac{1}{s+2}\right) - \frac{s}{(s+1)(s+2)} = 0$$

Using $|\mathbf{I}_1 + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1| = 1 + s/(s+1) + 1/(s+2) = 0$ leads to the same characteristic equation with less effort.

Likewise, $\mathbf{C} = [\mathbf{I}_2 + \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1]^{-1}\mathbf{G}_2\mathbf{G}_1\mathbf{R}$ gives

$$\mathbf{C} = \frac{\begin{bmatrix} 1+\dfrac{1}{s+2} & -\dfrac{1}{s+1} \\ -\dfrac{s}{s+2} & 1+\dfrac{s}{s+1} \end{bmatrix}\begin{bmatrix} \dfrac{1}{s+1} \\ \dfrac{1}{s+2} \end{bmatrix}\mathbf{R}}{1+\dfrac{s}{s+1}+\dfrac{1}{s+2}}$$

Using $\mathbf{C} = \mathbf{G}_2\mathbf{G}_1[1 + \mathbf{H}_2\mathbf{H}_1\mathbf{G}_2\mathbf{G}_1]^{-1}\mathbf{R}$ gives

$$\mathbf{C} = \frac{\begin{bmatrix} \dfrac{1}{s+1} \\ \dfrac{1}{s+2} \end{bmatrix}\mathbf{R}}{1+\dfrac{s}{s+1}+\dfrac{1}{s+2}}$$

These are identical, but the second form is obtained without the requirement of matrix inversion.

4.7 Consider the four forms of the closed-loop transfer matrix derived in Problem 4.3. What are the dimensions of the matrices that need to be inverted in each form if $\mathbf{G}_1$ is 10×1000, $\mathbf{G}_2$ is 50×10, $\mathbf{H}_1$ is 1×50, and $\mathbf{H}_2$ is 1000×1?

The first form requires inverting $\mathbf{I}_n + \mathbf{G}_2\mathbf{G}_1\mathbf{H}_2\mathbf{H}_1$, which is a 50×50 matrix. Form 2 requires inverting a 10×10 matrix since $m = 10$. The third form requires inversion of a 1000×1000 matrix, since $r = 1000$. The fourth form requires only a scalar division since $p = 1$. The same possibilities exist for the size of the determinant to be used in finding the characteristic equation.

Determinants, Cramer's Rule, Rank

4.8 A 5×5 matrix decomposes into the unit matrix plus a product, as shown. Evaluate its determinant.

$$\mathbf{A} = \begin{bmatrix} 0 & -2 & -3 & -4 & -5 \\ -1 & -1 & -3 & -4 & -5 \\ 4 & 8 & 13 & 16 & 20 \\ 2 & 4 & 6 & 9 & 10 \\ 8 & 16 & 24 & 32 & 41 \end{bmatrix} = \mathbf{I}_5 + \begin{bmatrix} -1 \\ -1 \\ 4 \\ 2 \\ 8 \end{bmatrix}[1 \quad 2 \quad 3 \quad 4 \quad 5]$$

$$|\mathbf{A}| = |\mathbf{I}_5 + \mathbf{GH}| = 1 + \mathbf{HG} = 1 + [1 \quad 2 \quad 3 \quad 4 \quad 5]\begin{bmatrix} -1 \\ -1 \\ 4 \\ 2 \\ 8 \end{bmatrix} = 58$$

4.9 Does $\mathbf{A} = \begin{bmatrix} 16 & 0 & 4 & 7 \\ -3 & 8 & 8 & 2 \\ 1 & 0 & 5 & 2 \\ -7 & 6 & 5 & 4 \end{bmatrix}$ have an inverse?

$\mathbf{A}^{-1}$ exists if and only if $|\mathbf{A}| \neq 0$. To check the determinant, the method of Laplace expansion is used with respect to the second column, $|\mathbf{A}| = 8C_{22} + 6C_{42}$. The two cofactors are

$$C_{22} = (-1)^4 \begin{vmatrix} 16 & 4 & 7 \\ 1 & 5 & 2 \\ -7 & 5 & 4 \end{vmatrix} = 368 \qquad C_{42} = (-1)^6 \begin{vmatrix} 16 & 4 & 7 \\ -3 & 8 & 2 \\ 1 & 5 & 2 \end{vmatrix} = -33$$

Therefore, $|\mathbf{A}| = 8(368) - 6(33) = 2746 \neq 0$ and $\mathbf{A}^{-1}$ does exist.

4.10 Check the determinant in the previous problem using a_{31} as the pivot element.

$$|\mathbf{A}| = \frac{1}{1^2} \begin{vmatrix} \begin{vmatrix} 16 & 0 \\ 1 & 0 \end{vmatrix} & \begin{vmatrix} 16 & 4 \\ 1 & 5 \end{vmatrix} & \begin{vmatrix} 16 & 7 \\ 1 & 2 \end{vmatrix} \\ \begin{vmatrix} -3 & 8 \\ 1 & 0 \end{vmatrix} & \begin{vmatrix} -3 & 8 \\ 1 & 5 \end{vmatrix} & \begin{vmatrix} -3 & 2 \\ 1 & 2 \end{vmatrix} \\ \begin{vmatrix} 1 & 0 \\ -7 & 6 \end{vmatrix} & \begin{vmatrix} 1 & 5 \\ -7 & 5 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ -7 & 4 \end{vmatrix} \end{vmatrix} = \begin{vmatrix} 0 & 76 & 25 \\ -8 & -23 & -8 \\ 6 & 40 & 18 \end{vmatrix}$$

Using the new a_{31} as the pivot gives

$$|\mathbf{A}| = \frac{1}{6} \begin{vmatrix} \begin{vmatrix} 0 & 76 \\ 6 & 40 \end{vmatrix} & \begin{vmatrix} 0 & 25 \\ 6 & 18 \end{vmatrix} \\ \begin{vmatrix} -8 & -23 \\ 6 & 40 \end{vmatrix} & \begin{vmatrix} -8 & -8 \\ 6 & 18 \end{vmatrix} \end{vmatrix} = \frac{1}{6} \begin{vmatrix} -456 & -150 \\ -182 & -96 \end{vmatrix} = \begin{vmatrix} -76 & -25 \\ -182 & -96 \end{vmatrix} = 2746$$

4.11 **A**, **B**, $\mathbf{I}_p$, and **0** are submatrices. Show that

$$\left| \left[\begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{B} & \mathbf{I}_p \end{array} \right] \right| = |\mathbf{A}|$$

Using Laplace expansion p times with respect to the last p columns gives

$$\left| \begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{B} & \mathbf{I}_p \end{array} \right| = 1 \left| \begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{B}' & \mathbf{I}_{p-1} \end{array} \right| = 1 \cdot 1 \left| \begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{B}'' & \mathbf{I}_{p-2} \end{array} \right| = \cdots = |\mathbf{A}|$$

4.12 Show that

$$\left| \begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{B} & \mathbf{C} \end{array} \right| = |\mathbf{A}| \cdot |\mathbf{C}|$$

and that

$$\left| \begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array} \right| = |\mathbf{A}| \cdot |\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}| \quad \text{if } \mathbf{A}^{-1} \text{ exists}$$

$$= |\mathbf{D}| \cdot |\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}| \quad \text{if } \mathbf{D}^{-1} \text{ exists}$$

We have

$$\left| \left[\begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{B} & \mathbf{C} \end{array} \right] \right| = \left| \left[\begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{B} & \mathbf{I} \end{array} \right] \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline 0 & \mathbf{C} \end{array} \right] \right| = \left| \begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{B} & \mathbf{I} \end{array} \right| \cdot \left| \begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline 0 & \mathbf{C} \end{array} \right| = |\mathbf{A}| \cdot |\mathbf{C}|$$

using results of the previous problem.

If $\mathbf{A}^{-1}$ exists, the desired determinant can be multiplied by $\begin{vmatrix} \mathbf{I} & 0 \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I} \end{vmatrix} = 1$:

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = \begin{vmatrix} \mathbf{I} & 0 \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I} \end{vmatrix} \begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = \begin{vmatrix} \mathbf{A} & \mathbf{B} \\ 0 & \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \end{vmatrix} = |\mathbf{A}| \cdot |\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}|$$

If $\mathbf{D}^{-1}$ exists, use $\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = \begin{vmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ 0 & \mathbf{I} \end{vmatrix} \begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix}$ and repeat the above procedure.

4.13 Use Cramer's rule to find the solutions for x_1 and x_2, if $3x_1 + 2x_2 = 6$, and $x_1 - 5x_2 = 1$.
In matrix form,

$$\begin{bmatrix} 3 & 2 \\ 1 & -5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \end{bmatrix} \quad \text{or} \quad \mathbf{AX} = \mathbf{Y}$$

If the coefficient matrix $\mathbf{A}$ is nonsingular, Cramer's rule gives the solution for the component x_i as

$$x_i = \frac{|\mathbf{B}_i|}{|\mathbf{A}|}$$

where $\mathbf{B}_i$ is formed from $\mathbf{A}$ by replacing column i by $\mathbf{Y}$. In this example

$$x_1 = \frac{\begin{vmatrix} 6 & 2 \\ 1 & -5 \end{vmatrix}}{|\mathbf{A}|} = \frac{32}{17} \quad \text{and} \quad x_2 = \frac{\begin{vmatrix} 3 & 6 \\ 1 & 1 \end{vmatrix}}{|\mathbf{A}|} = \frac{3}{17}$$

4.14 Use Cramer's rule and partitioned matrices to prove the following theorem.

Theorem. Let $\mathbf{A}$ be an $n \times n$ matrix and let a_{ij} be any nonzero element. Define $\mathbf{B}_{ij}$ as the $n-1$ by $n-1$ matrix formed by deleting row i and column j of $\mathbf{A}$. Let $\mathbf{R}$ and $\mathbf{C}$ be $1 \times n-1$ and $n-1 \times 1$ row and column matrices formed by deleting a_{ij} from the ith row and jth column of $\mathbf{A}$. Then

$$|\mathbf{A}| = (-1)^{i+j} a_{ij} \left| \mathbf{B}_{ij} - \frac{1}{a_{ij}} \mathbf{CR} \right|$$

Proof. Consider the set of linear equations $\mathbf{AX} = \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are column matrices and $\mathbf{Y}$ is all zero except $y_i = 1$. Cramer's rule is used to solve for x_j:

$$\mathbf{x}_j = \frac{1}{|\mathbf{A}|} \left| \begin{array}{c|c|c} \mathbf{B}_1 & \begin{matrix} 0 \\ \vdots \end{matrix} & \mathbf{B}_2 \\ \hline a_{i1} \ldots & 1 & \ldots a_{in} \\ \hline \mathbf{B}_3 & \begin{matrix} \vdots \\ 0 \end{matrix} & \mathbf{B}_4 \end{array} \right| \quad \text{where} \quad \mathbf{B}_{ij} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 \\ \mathbf{B}_3 & \mathbf{B}_4 \end{bmatrix}$$

Using the Laplace expansion method with respect to column j and then rearranging gives

$$|\mathbf{A}| = \frac{(-1)^{i+j}}{x_j} |\mathbf{B}_{ij}| \tag{1}$$

In order to determine $1/x_j$, the original equation $\mathbf{AX} = \mathbf{Y}$ can be written as

$$\mathbf{B}_{ij}\,\mathbf{X}_a + \mathbf{C}x_j = \mathbf{0}$$

$$\mathbf{R}\mathbf{X}_a + a_{ij}x_j = 1$$

$\mathbf{X}_a$ is formed from $\mathbf{X}$ by deleting x_j. Thus

$$\mathbf{X}_a = -\mathbf{B}_{ij}^{-1}\,\mathbf{C}x_j$$

so that $(-\mathbf{R}\mathbf{B}_{ij}^{-1}\,\mathbf{C} + a_{ij})x_j = 1$. Using this in Eq. (1) gives

$$|\mathbf{A}| = (-1)^{i+j}|\mathbf{B}_{ij}| \cdot (a_{ij} - \mathbf{R}\mathbf{B}_{ij}^{-1}\,\mathbf{C}) = (-1)^{i+j} a_{ij} |\mathbf{B}_{ij}| \cdot \left| 1 - \frac{1}{a_{ij}} \mathbf{R}\mathbf{B}_{ij}^{-1}\,\mathbf{C} \right|$$

The determinant identities of Problem 4.5 give the final result:

$$|\mathbf{A}| = (-1)^{i+j} a_{ij} \left| \mathbf{B}_{ij} - \frac{1}{a_{ij}} \mathbf{CR} \right|$$

(A limiting process can be used to show the validity of this proof and the final result even if $|\mathbf{A}| = 0$ or $|\mathbf{B}_{ij}| = 0$.)

4.15 Use the previous theorem to evaluate $|\mathbf{A}|$, using a_{11} as the divisor.

$$|\mathbf{A}| = \begin{vmatrix} 4 & 8 & 1 & 3 \\ 2 & 5 & -1 & 3 \\ -1 & 6 & 7 & 9 \\ 1 & 1 & -3 & 3 \end{vmatrix} = 4 \left| \begin{bmatrix} 5 & -1 & 3 \\ 6 & 7 & 9 \\ 1 & -3 & 3 \end{bmatrix} - \tfrac{1}{4} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} [8 \quad 1 \quad 3] \right|$$

$$= 4 \begin{vmatrix} 1 & -3/2 & 3/2 \\ 8 & 29/4 & 39/4 \\ -1 & -13/4 & 9/4 \end{vmatrix} = 246$$

4.16 Find the rank and trace of

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 3 \\ 2 & -2 & 1 & 1 \\ 0 & 1 & 5 & 8 \\ 1 & -6 & 4 & -4 \end{bmatrix}$$

The sum of the diagonal terms gives $\text{Tr}(\mathbf{A}) = 1 - 2 + 5 - 4 = 0$. Using any one of several methods gives $|\mathbf{A}| = 338$. Since the determinant is nonzero, the rank of $\mathbf{A}$ is 4.

Matrix Inversion and Related Topics

4.17 Given

$$\begin{bmatrix} 2 & 0.5 & 2 \\ 3 & 3 & 0 \\ 1 & 0.5 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix} \quad \text{or} \quad \mathbf{AX} = \mathbf{Y}$$

find x_1, x_2, and x_3 using the definition of matrix inversion.

The solution is $\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y}$, where $|\mathbf{A}| = 6$ and

$$\text{Adj}\,\mathbf{A} = \begin{bmatrix} 6 & -6 & -1.5 \\ 0 & 2 & -0.5 \\ -6 & 6 & 4.5 \end{bmatrix}^T$$

so that

$$\mathbf{X} = \frac{1}{6} \begin{bmatrix} 6 & 0 & -6 \\ -6 & 2 & 6 \\ -1.5 & -0.5 & 4.5 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix} = \begin{bmatrix} -2 \\ \frac{7}{3} \\ \frac{35}{12} \end{bmatrix}$$

4.18 Explain the method of Gaussian elimination in terms of elementary matrices and partitioned matrices.

Gaussian elimination is used to solve $\mathbf{AX} = \mathbf{Y}$, where $\mathbf{A}$ is known, $n \times n$, and $\mathbf{Y}$ is known, $n \times p$. The $n \times p$ matrix $\mathbf{X}$ is unknown. Often $\mathbf{X}$ and $\mathbf{Y}$ are column vectors, with $p = 1$. Assuming that $\mathbf{A}^{-1}$ exists, $\mathbf{A}^{-1}\mathbf{AX} = \mathbf{A}^{-1}\mathbf{Y}$ or $\mathbf{IX} = \mathbf{A}^{-1}\mathbf{Y}$. The elementary row operations are equivalent to premultiplication by the elementary matrices. A sequence of these operations which reduces the coefficient of $\mathbf{X}$ from $\mathbf{A}$ to $\mathbf{I}$ will simultaneously change $\mathbf{Y}$ to $\mathbf{A}^{-1}\mathbf{Y}$, that is, $\mathbf{X}$. The operations are:

1. Form the $n \times n + p$ matrix $\mathbf{W}_0 = [\mathbf{A} \mid \mathbf{Y}]$.
2. Find the element in column one with maximum absolute value. Interchange that row with row one. This gives $\mathbf{W}_1 = \mathbf{E}_{1,q}\,\mathbf{W}_0 = [\mathbf{E}_{1,q}\,\mathbf{A} \mid \mathbf{E}_{1,q}\,\mathbf{Y}]$.

3. Divide the entire first row of $\mathbf{W}_1$ by w_{11}, assuming $w_{11} \neq 0$. This gives $W_2 = \mathbf{E}_1(1/w_{11})\mathbf{W}_1$. If $w_{11} = 0$, then the entire first column is zero, indicating that $\mathbf{A}$ is singular. If this happens, skip to step 6.
4. Multiply the first row of $\mathbf{W}_2$ by $-w_{21}$ and add to the second row. $\mathbf{W}_3 = \mathbf{E}_{1,2}(-w_{21})\mathbf{W}_2$.
5. Repeat step 4 using $\mathbf{E}_{1,3}(-w_{31}), \ldots, \mathbf{E}_{1,n}(-w_{n1})$ in sequence. This reduces column one to a one followed by $n-1$ zeros.
6. Find the maximum absolute value element $w_{\alpha 2}$ from column 2, rows 2 through n. Interchange that row with row 2, $\mathbf{W}_{k+1} = \mathbf{E}_{2,\alpha}\mathbf{W}_k$.
7. Divide row 2 by the current value of w_{22}. This is analogous to step 3. Repeat steps 4 and 5 until w_{22} is 1 (possibly zero if $\mathbf{A}$ is singular) and all $w_{i2} = 0$ below w_{22}.
8. Repeat steps 6 and 7 until the first n columns form an $n \times n$ upper triangular matrix with $\delta_i = 1$ or 0 for the ith diagonal. From this, $|\mathbf{A}| = (-1)^\nu \mu_1 \mu_2 \mu_3 \cdots \mu_n$, where ν is the number of row interchanges used and μ_i is the divisor used for the ith row, steps 3 and 7. If a zero is encountered on the diagonal, then $|\mathbf{A}| = 0$. When this happens, the rank of $\mathbf{A}$ is still of interest. The triangular form is a convenient starting point for reducing to one of the normal forms to determine rank. If $|\mathbf{A}| \neq 0$, continue to step 9.
9. Multiply the current $\mathbf{W}$ by $\mathbf{E}_{n,1}(-w_{1n})$, then by $\mathbf{E}_{n,2}(-w_{2n}), \ldots$. This reduces the nth column of $\mathbf{W}$ to $n-1$ zeros but leaves the n, n element unity.
10. Repeat step 9 for other columns, until $\mathbf{W}$ has the unit matrix for its first n columns. The last n columns of this final $\mathbf{W}$ matrix contain $\mathbf{A}^{-1}\mathbf{Y}$.
Note that $\mathbf{A}^{-1}$ can be found by using $\mathbf{Y} = \mathbf{I}$ when setting up $\mathbf{W}_0$.

4.19 Find $\begin{bmatrix} 0 & 3 \\ 4 & 2 \end{bmatrix}^{-1}$.

The sequence of matrices, interchanges and divisors is

$$\mathbf{W}_0 = \left[\begin{array}{cc|cc} 0 & 3 & 1 & 0 \\ 4 & 2 & 0 & 1 \end{array}\right] \xrightarrow[\nu = 1]{\text{interchange}} \left[\begin{array}{cc|cc} 4 & 2 & 0 & 1 \\ 0 & 3 & 1 & 0 \end{array}\right] \xrightarrow{\mu_1 = 4} \left[\begin{array}{cc|cc} 1 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 3 & 1 & 0 \end{array}\right]$$

$$\xrightarrow{\mu_2 = 3} \left[\begin{array}{cc|cc} 1 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 1 & \frac{1}{3} & 0 \end{array}\right] \longrightarrow \left[\begin{array}{cc|cc} 1 & 0 & -\frac{1}{6} & \frac{1}{4} \\ 0 & 1 & \frac{1}{3} & 0 \end{array}\right]$$

This problem demonstrates why row interchanges are required to avoid dividing by zero. The results are

$$|\mathbf{A}| = (-1)^\nu \mu_1 \mu_2 = -12, \qquad r_A = 2, \qquad \mathbf{A}^{-1} = \frac{1}{12}\begin{bmatrix} -2 & 3 \\ 4 & 0 \end{bmatrix}$$

4.20 Let $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 0 & 1 & 5 \end{bmatrix}$. Determine $|\mathbf{A}|$, r_A, and $\mathbf{A}^{-1}$ if it exists.

$$\mathbf{W}_0 = \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & 4 & 6 & 0 & 1 & 0 \\ 0 & 1 & 5 & 0 & 0 & 1 \end{array}\right] \xrightarrow{\nu = 1} \left[\begin{array}{ccc|ccc} 2 & 4 & 6 & 0 & 1 & 0 \\ 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 5 & 0 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{\mu_1 = 2} \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & \frac{1}{2} & 0 \\ 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 5 & 0 & 0 & 1 \end{array}\right] \longrightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 1 & 5 & 0 & 0 & 1 \end{array}\right]$$

$$\xrightarrow[\nu = 2]{\text{interchange}} \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 5 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & -\frac{1}{2} & 0 \end{array}\right]$$

At this point we see that $\mathbf{A}$ is singular, that is, $|\mathbf{A}| = 0$ and $\mathbf{A}^{-1}$ does not exist. We can therefore drop the right half of $\mathbf{W}$ and use row or column operations to give

$$\begin{bmatrix}1&2&3\\0&1&5\\0&0&0\end{bmatrix}\longrightarrow\begin{bmatrix}1&2&-7\\0&1&0\\0&0&0\end{bmatrix}\longrightarrow\begin{bmatrix}1&2&0\\0&1&0\\0&0&0\end{bmatrix}\longrightarrow\left[\begin{array}{cc|c}1&0&0\\0&1&0\\\hline 0&0&0\end{array}\right]=\left[\begin{array}{c|c}\mathbf{I}_2&0\\\hline 0&0\end{array}\right]$$

Therefore, $r_A = 2$.

4.21 Use Gaussian elimination to solve for x_1 and x_2, if $\begin{bmatrix}2&-2\\3&8\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}10\\-5\end{bmatrix}$.

$$\mathbf{W}_0=\left[\begin{array}{cc|c}2&-2&10\\3&8&-5\end{array}\right]\longrightarrow\left[\begin{array}{cc|c}3&8&-5\\2&-2&10\end{array}\right]\longrightarrow\left[\begin{array}{cc|c}1&\frac{8}{3}&-\frac{5}{3}\\2&-2&10\end{array}\right]$$

$$\longrightarrow\left[\begin{array}{cc|c}1&\frac{8}{3}&-\frac{5}{3}\\0&-\frac{22}{3}&\frac{40}{3}\end{array}\right]\longrightarrow\left[\begin{array}{cc|c}1&\frac{8}{3}&-\frac{5}{3}\\0&1&-\frac{20}{11}\end{array}\right]\longrightarrow\left[\begin{array}{cc|c}1&0&\frac{35}{11}\\0&1&-\frac{20}{11}\end{array}\right]$$

Therefore, $x_1 = \frac{35}{11}$ and $x_2 = -\frac{20}{11}$.

4.22 $\mathbf{G}$ is an $n \times n$ complex matrix. Find $\mathbf{G}^{-1}$ using only real matrix inversion routines.

Let $\mathbf{G} = \mathbf{A} + j\mathbf{B}$, with $\mathbf{A}$ and $\mathbf{B}$ real. Assuming that $\mathbf{G}^{-1}$ exists, it can also be expressed in terms of two real matrices $\mathbf{C}$ and $\mathbf{D}$, $\mathbf{G}^{-1} = \mathbf{C} + j\mathbf{D}$. The basic requirement of a matrix inverse is that

$$\mathbf{G}\mathbf{G}^{-1} = \mathbf{I} = (\mathbf{A} + j\mathbf{B})(\mathbf{C} + j\mathbf{D}) = (\mathbf{AC} - \mathbf{BD}) + j(\mathbf{AD} + \mathbf{BC})$$

Therefore, equating real parts to real parts, $\mathbf{AC} - \mathbf{BD} = \mathbf{I}$. Equating imaginary parts $\mathbf{AD} + \mathbf{BC} = \mathbf{0}$. If $\mathbf{A}^{-1}$ exists, the solution can be written as $\mathbf{C} = (\mathbf{A} + \mathbf{B}\mathbf{A}^{-1}\mathbf{B})^{-1}$ and $\mathbf{D} = -\mathbf{C}\mathbf{B}\mathbf{A}^{-1}$. The rearrangement identities of Problem 4.4 are useful in proving this. If $\mathbf{B}^{-1}$ exists but $\mathbf{A}^{-1}$ doesn't, then $[j\mathbf{G}]^{-1}$ can be sought instead. This effectively reverses the roles of $\mathbf{A}$ and $\mathbf{B}$ so the above procedure can again be used. If both $\mathbf{A}$ and $\mathbf{B}$ are singular, but $\mathbf{G}$ is not, further modifications will be necessary.

4.23 Given a set of simultaneous linear equations

$$\mathbf{X}\mathbf{A} = \mathbf{B} \tag{1}$$

where $\mathbf{A}$ and $\mathbf{B}$ are known complex-valued matrices of size $n \times n$ and $m \times n$, respectively, and where $\mathbf{X}$ is the unknown $m \times n$ matrix. Assume that columns i and $i + 1$ of $\mathbf{A}$ are complex conjugates. Assume the same for columns of $\mathbf{B}$. Show that $\mathbf{X}$ is purely real and can be computed using only real numbers from

$$\mathbf{X} = \mathbf{B}_* \mathbf{A}_*^{-1} \tag{2}$$

where $\mathbf{A}_*$ and $\mathbf{B}_*$ are formed from $\mathbf{A}$ and $\mathbf{B}$ by replacing their two complex columns by the real part and imaginary part of their respective column i.

Postmultiplying Eq. (1) by any nonsingular $n \times n$ matrix $\mathbf{T}$ and solving gives

$$\mathbf{X} = (\mathbf{BT})(\mathbf{AT})^{-1} = \mathbf{B}\mathbf{A}^{-1}$$

A particular $\mathbf{T}$ is selected that differs from the unit matrix only in the four elements defined by the intersections of rows i and j with columns i and j. The four exceptional elements form the 2×2 block

$$\mathbf{T}_i = \begin{bmatrix}\frac{1}{2}&\frac{-j}{2}\\\frac{1}{2}&\frac{j}{2}\end{bmatrix}$$

Clearly $\mathbf{T}$ satisfies the nonsingular condition. Its determinant is just $j/2$. Furthermore, $\mathbf{BT} = \mathbf{B}_*$ and $\mathbf{AT} = \mathbf{A}_*$, as demonstrated by a 2×2 case:

$$\begin{bmatrix} \alpha + j\beta & \alpha - j\beta \\ \gamma + j\delta & \gamma - j\delta \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{-j}{2} \\ \frac{1}{2} & \frac{j}{2} \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$$

Result (2) applies to matrices with any number of conjugate-pair columns, and they need not be adjacent. However, the conjugate pairs must appear in the same column numbers in both **A** and **B.** Form $\mathbf{A}_*$ and $\mathbf{B}_*$ by replacing one member of each pair by the real part and the second member by the imaginary part. Two equal real columns in **B** qualify as conjugate pairs, meaning $\mathbf{B}_*$ will have an all-zero column. This causes no problem, but the same cannot be done in **A** because two equal columns (or an all-zero column) means that **A** and $\mathbf{A}_*$ are singular. No matrix-inversion method will solve that problem.

If the original problem is to solve $\mathbf{AX} = \mathbf{B}$, then *everything* said above about *columns* must be changed to *rows*. This is equivalent to solving the transposed problem $\mathbf{X}^T\mathbf{A}^T = \mathbf{B}^T$ for $\mathbf{X}^T$.

Cholesky Decomposition

4.24 If **A** is symmetric and positive definite (see Chapter 7), it can be uniquely (except for signs) factored into $\mathbf{A} = \mathbf{S}^T\mathbf{S}$, where **S** is an upper triangular matrix. **S** is called the square root matrix of **A.** The procedure for factoring **A** is most commonly called Cholesky decomposition [7] (although it is sometimes called the method of Banachiewicz and Dwyer [8]. Deduce the algorithm for finding **S.**

The algorithm for finding the s_{ij} entries in **S** is as follows:

$$s_{11} = [a_{11}]^{1/2}; s_{1j} = a_{1j}/s_{11} \quad \text{for } j = 2, \ldots, n$$

$$s_{22} = [a_{22} - (s_{12})^2]^{1/2}$$

$$s_{2j} = [a_{2j} - s_{12}s_{1j}]/s_{22} \quad \text{for } j = 3, \ldots, n$$

$$\vdots$$

$$s_{ii} = \left[a_{ii} - \sum_{k=1}^{i-1} (s_{ki})^2\right]^{1/2} \quad \text{for } i = 2, \ldots, n$$

$$s_{ij} = \left[a_{ij} - \sum_{k=1}^{i-1} s_{ki}s_{kj}\right]/s_{ii} \quad \text{for } j = i+1, \ldots, n$$

4.25 Show how Cholesky decomposition can be used in solving simultaneous equations of the form $\mathbf{Ax} = \mathbf{y}$. Assume **y** is known and that **A** is known, symmetric, and positive definite.

Assume **S** has been found such that $\mathbf{A} = \mathbf{S}^T\mathbf{S}$. Then $\mathbf{S}^T\mathbf{Sx} = \mathbf{y}$. Define $\mathbf{Sx} = \mathbf{v}$. Then $\mathbf{S}^T\mathbf{v} = \mathbf{y}$. Because $\mathbf{S}^T$ is lower triangular, the elements of **v** can easily be found one-by-one by back substitution,

$$v_1 = y_1/s_{11}, \qquad v_2 = (y_2 - s_{12}v_1)/s_{22}, \qquad v_3 = (y_3 - s_{13}v_1 - s_{23}v_2)/s_{33}, \ldots$$

Once **v** is determined, a similar procedure can be used to find the components of **x**:

$$x_n = v_n/s_{nn}, x_{n-1} = (v_{n-1} - s_{n-1,n}v_n)/s_{n-1,n-1},$$

$$x_{n-2} = (y_{n-2} - s_{n-2,n}v_n - s_{n-2,n-1}v_{n-1})/s_{n-2,n-2}, \ldots$$

Linearizing Nonlinear Equations

4.26 The two-port nonlinear electrical device shown in Figure 4.4 is characterized by the four quantities i_1, i_2, v_1, and v_2. Use Taylor series to develop a linear model for small signal variations about a nominal operating point.

Figure 4.4

There are a variety of linear models that can be developed, depending on the form assumed for the functional relations among the four variables. There is one relationship at the input port and another at the output port. One possibility is the pair $v_1 = f_0(v_2, i_1, i_2)$ and $i_2 = g_0(i_1, v_2, v_1)$. These can be combined to yield $v_1 = f_0(v_2, i_1, g_0(i_1, v_2, v_1))$ and $i_2 = g_0(i_1, v_2, f_0(v_2, i_1, i_2))$. This shows that only two independent variables i_1 and v_2 suffice to determine v_1 and i_2. These relationships are rewritten more simply as $v_1 = f(i_1, v_2)$ and $i_2 = g(i_1, v_2)$. Let $v_2 = v_{2n} + \delta v_2$ and $i_1 = i_{1n} + \delta i_1$, where v_{2n} and i_{1n} define the nominal operating point and δv_2 and δi_1 are small variations from the nominal. Then Taylor series expansion gives

$$v_1 = f(i_{1n}, v_{2n}) + \left.\frac{\partial f}{\partial i_1}\right|_n \delta i_1 + \left.\frac{\partial f}{\partial v_2}\right|_n \delta v_2$$

$$i_2 = g(i_{1n}, v_{2n}) + \left.\frac{\partial g}{\partial i_1}\right|_n \delta i_1 + \left.\frac{\partial g}{\partial v_2}\right|_n \delta v_2$$

Obviously, $v_{1n} = f(i_{1n}, v_{2n})$ and $i_{2n} = g(i_{1n}, v_{2n})$ so that

$$\delta v_1 \triangleq v_1 - v_{1n} = h_{11}\,\delta i_1 + h_{12}\,\delta v_2$$

$$\delta i_2 \triangleq i_2 - i_{2n} = h_{21}\,\delta i_1 + h_{22}\,\delta v_2$$

where the h_{ij} terms are the partial derivatives evaluated at the nominal point. These are the hybrid or h parameters commonly used in small signal, linearized analysis of transistors and other nonlinear devices.

4.27 A tracking station measures the azimuth angle α, the elevation angle β, and the range r to an earth satellite as shown in Figure 4.5.

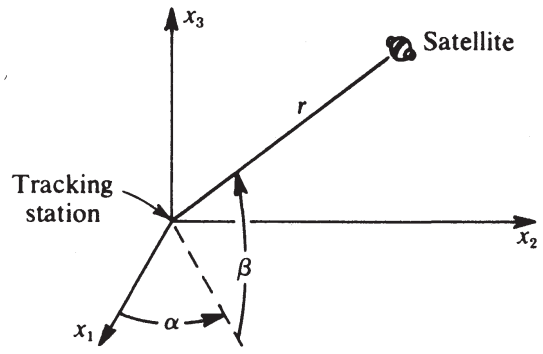


Figure 4.5

(a) Derive the nonlinear equations which relate the satellite's relative position $[x_1 \quad x_2 \quad x_3]^t$ to the measured quantities.

(b) Obtain linear equations which relate small perturbations in satellite location to small perturbations in the measurements.

(a) The station-to-satellite range magnitude is $r = \sqrt{x_1^2 + x_2^2 + x_3^2}$ and the tracking antenna angles are $\alpha = \tan^{-1}(x_2/x_1)$ and $\beta = \tan^{-1}(x_3/\sqrt{x_1^2 + x_2^2})$.

(b) Letting $\mathbf{x}(t) = \mathbf{x}_n(t) + \delta\mathbf{x}(t)$, $r(t) = r_n(t) + \delta r(t)$, $\alpha(t) = \alpha_n(t) + \delta\alpha(t)$, and $\beta(t) = \beta_n(t) + \delta\beta(t)$, the Taylor series expansion gives

$$\delta r = \sum_{i=1}^{3} \frac{\partial r}{\partial x_i}\delta x_i = \frac{1}{r_n}\mathbf{x}_n^T \delta\mathbf{x}$$

$$\delta\alpha = \sum_{i=1}^{3} \frac{\partial \alpha}{\partial x_i}\delta x_i = 1/(x_{1n}^2 + x_{2n}^2)[-x_{2n} \quad x_{1n} \quad 0]\,\delta\mathbf{x}$$

$$\delta\beta = \sum_{i=1}^{3} \frac{\partial \beta}{\partial x_i}\delta x_i = \frac{1}{r_n^2\sqrt{x_{1n}^2 + x_{2n}^2}}[-x_{1n}x_{3n} - x_{2n}x_{3n} \quad x_{1n}^2 + x_{2n}^2]\,\delta\mathbf{x}$$

Letting $\delta\mathbf{y} = [\delta r \quad \delta\alpha \quad \delta\beta]^T$ allows the preceding results to be expressed as $\delta\mathbf{y}(t) = \mathbf{C}(t)\delta\mathbf{x}(t)$, where $\mathbf{C}$ is a 3×3 matrix.

4.28 A certain process is characterized by a set of parameters $\mathbf{x}$. Measurements $\mathbf{y}$ can be made on this process, and they are related to $\mathbf{x}$ by a nonlinear algebraic equation, $\mathbf{y} = \mathbf{f}(\mathbf{x})$. The nominal values of $\mathbf{x}$ are $\mathbf{x}_n$. Describe a method of estimating the actual values of $\mathbf{x}$ based on the measurements $\mathbf{y}$.

Let $\mathbf{x} = \mathbf{x}_n + \delta\mathbf{x}$. Then $\mathbf{y} = \mathbf{f}(\mathbf{x}_n + \delta\mathbf{x}) \cong \mathbf{f}(x_n) + \left.\frac{d\mathbf{f}}{d\mathbf{x}}\right|_n \delta\mathbf{x}$.

Call $\mathbf{f}(\mathbf{x}_n) \triangleq \mathbf{y}_n$ and $\mathbf{y} - \mathbf{y}_n \triangleq \delta\mathbf{y}$. Since $\mathbf{y}$ is measured and since $\mathbf{y}_n$ can be computed from a knowledge of $\mathbf{x}_n$, $\delta\mathbf{y}$ is a known vector. The Jacobian matrix $\left.\frac{d\mathbf{f}}{d\mathbf{x}}\right|_n \triangleq \mathbf{A}$ can also be computed. The relation $\delta\mathbf{y} = \mathbf{A}\delta\mathbf{x}$ is of the form treated in Chapter 6. Because of measurement inaccuracies, redundant measurements and the least-squares technique are most commonly used to solve for $\delta\mathbf{x}$. Then the estimated parameter values are given by $\mathbf{x} = \mathbf{x}_n + \delta\mathbf{x}$.

4.29 Equation (*4.7*) expressed a matrix transfer function $\mathbf{H}(s)$ in one form of the MFD, with the inverse of a polynomial matrix as the *left* factor. This form naturally arises from systems such as the one of Example 4.2 but with input derivative terms. Show how a second MFD form can be obtained with an inverse of a polynomial matrix as the factor on the *right*.

Assume that $\mathbf{H}(s)$ is $m \times r$. If each element in $\mathbf{H}(s)$ is placed over a common denominator, the scalar polynomial $a(s)$, then

$$\mathbf{H}(s) = \mathbf{N}(s)/a(s) = \mathbf{N}(s)[\mathbf{I}_r a(s)]^{-1} = [\mathbf{I}_m a(s)]^{-1}\mathbf{N}(s)$$

This shows that both the left and right forms of the MFD are possible, but the special forms here are misleading. The numerator $\mathbf{N}(s)$ is not generally the same in both forms, and the inverted matrix is not generally diagonal, as demonstrated by the $\mathbf{P}(s)^{-1}$ in Eq. (*4.7*). To indicate the more general forms which are possible, write

$$\mathbf{H}(s) = \mathbf{P}_1(s)^{-1}\mathbf{N}_1(s) = \mathbf{N}_2(s)\mathbf{P}_2(s)^{-1}$$

These factors are clearly nonunique. Rewrite the two expressions as

$$\mathbf{P}_1(s)\mathbf{H}(s) = \mathbf{N}_1(s) \quad \text{and} \quad \mathbf{H}(s)\mathbf{P}_2(s) = \mathbf{N}_2(s)$$

Let $\mathbf{T}_1(s)$ and $\mathbf{T}_2(s)$ be any arbitrary nonsingular $m \times m$ and $r \times r$ polynomial matrices. Then

$$\mathbf{T}_1(s)\mathbf{P}_1(s)\mathbf{H}(s) = \mathbf{T}_1(s)\mathbf{N}_1(s) \quad \text{or} \quad \mathbf{H}(s) = [\mathbf{T}_1\mathbf{P}_1]^{-1}[\mathbf{T}_1\mathbf{N}_1]$$

and

$$\mathbf{H}(s)\mathbf{P}_2(s)\mathbf{T}_2(s) = \mathbf{N}_2(s)\mathbf{T}_2(s) \quad \text{or} \quad \mathbf{H}(s) = [\mathbf{N}_2\mathbf{T}_2][\mathbf{P}_2\mathbf{T}_2]^{-1}$$

Thus new factors $\mathbf{P}_1' = \mathbf{T}_1\mathbf{P}_1$ and $\mathbf{N}_1' = \mathbf{T}_1\mathbf{N}_1$ or $\mathbf{P}_2' = \mathbf{P}_2\mathbf{T}_2$ and $\mathbf{N}_2' = \mathbf{N}_2\mathbf{T}_2$ can always be created, and they will also be polynomial matrices. The matrices $\mathbf{T}_1$ and $\mathbf{T}_2$ can sometimes be generalized to include certain rational polynomial factors as long as the primed $\mathbf{P}, \mathbf{N}$ factors are still polynomial matrices (i.e., no fractions). Note that form 2 is the matrix analog of the controllable canonical form procedure, illustrated in Figure 3.9, which is generalized in Figure 4.6. The matrix equations $\mathbf{P}_2\mathbf{g} = \mathbf{u}$ and $\mathbf{N}_2\mathbf{g} = \mathbf{y}$ define an intermediate vector $\mathbf{g}$, which is often called the partial state vector. MFD form 1 is the matrix analog of the Chapter 3 observable canonical form procedure. In the scalar case, these canonical form names were attached because of the

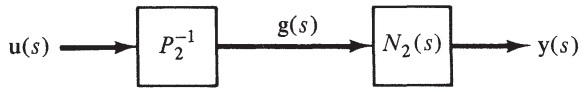


Figure 4.6

controllable or observable properties which were guaranteed to the resulting state models. Those properties are *not* guaranteed to the matrix generalizations mentioned here. Extra work is required to select "good" choices for the generalized denominator and numerator matrix factors **P** and **N**.

4.30 The transfer function

$$\mathbf{H}(s) = \begin{bmatrix} \dfrac{1}{s+1} & \dfrac{2}{(s+1)(s+2)} \\ \dfrac{1}{(s+1)(s+3)} & \dfrac{1}{s+3} \end{bmatrix}$$

$$= \frac{\begin{bmatrix} (s+2)(s+3) & 2(s+3) \\ (s+2) & (s+1)(s+2) \end{bmatrix}}{(s+1)(s+2)(s+3)}$$

can be written in either MFD form, with $\mathbf{P} = (s+1)(s+2)(s+3)\mathbf{I}_2$. The degree of the determinant of this denominator matrix is 6 in either case. Start with form 1 and use elementary row operations to find an alternative MFD which has $|\mathbf{P}_1(s)|$ with degree 4.

Elementary row operations are equivalent to premultiplying by one of the elementary matrices, which are all nonsingular. The $\mathbf{T}_1$ matrix of the previous problem is a product of elementary matrices. Premultiplying $\mathbf{P}_1$ and $\mathbf{N}_1$ is accomplished by doing row operations on $[\mathbf{P}_1 \quad \mathbf{N}_1]$. One obvious elementary operation is to divide each row by any common factors that might be present, e.g., $(s+3)$ in row 1 and $(s+2)$ in row 2 of

$$\left[\begin{array}{cc|cc} (s+1)(s+2)(s+3) & 0 & (s+2)(s+3) & 2(s+3) \\ 0 & (s+1)(s+2)(s+3) & (s+2) & (s+1)(s+2) \end{array}\right]$$

The resulting new $P(s)$ has the desired degree of 4 and

$$\mathbf{H}(s) = \begin{bmatrix} (s+1)(s+2) & 0 \\ 0 & (s+1)(s+3) \end{bmatrix}^{-1} \begin{bmatrix} (s+2) & 2 \\ 1 & (s+1) \end{bmatrix}$$

The fact that **P** is still diagonal is a peculiarity of this problem and is not a general result. Can a lower-degree $|\mathbf{P}|$ be found? This relates to the problem of finding minimal state variable realizations and is considered in Chapter 12.

4.31 Repeat Problem 4.30, starting with the second MFD form $\mathbf{H} = \mathbf{N}_2\mathbf{P}_2^{-1}$ and use elementary column operations. Postmultiplication of $\mathbf{N}_2$ and $\mathbf{P}_2$ by $\mathbf{T}_2$ is equivalent to carrying out a sequence of elementary column operations on

$$\begin{bmatrix} \mathbf{N}_2 \\ \hline \mathbf{P}_2 \end{bmatrix} = \left[\begin{array}{cc} (s+2)(s+3) & 2(s+3) \\ s+2 & (s+1)(s+2) \\ \hline (s+1)(s+2)(s+3) & 0 \\ 0 & (s+1)(s+2)(s+3) \end{array}\right]$$

The first obvious operation is to cancel a factor of $s+2$ from column 1. Then column 1 times $(s+1)(s+2)$ is subtracted from column 2 (which is the same as postmultiplying by the elementary matrix $\mathbf{E}_{1,2}(\alpha)$ of Sec. 4.10, with $\alpha = -(s+1)(s+2)$). The result of these steps is

$$\left[\begin{array}{cc} (s+3) & -s(s+3)^2 \\ 1 & 0 \\ \hline (s+1)(s+3) & -(s+1)^2(s+2)(s+3) \\ 0 & (s+1)(s+2)(s+3) \end{array}\right]$$

Now a factor of $s + 3$ can be canceled from column 2, giving

$$\mathbf{H}(s) = \begin{bmatrix} (s+3) & -s(s+3) \\ 1 & 0 \end{bmatrix} \begin{bmatrix} (s+1)(s+3) & -(s+1)^2(s+2) \\ 0 & (s+1)(s+2) \end{bmatrix}^{-1}$$

It is not generally true that the degree of all possible left- and right-form MFD determinants will be the same. The degree can be *increased* simply by selecting arbitrarily high-order polynomial factors in $\mathbf{T}_i$. There is a limit to how far the degree can be *decreased*. The desired degree 4 is achieved with this $\mathbf{P}_2(s)$. In fact, the minimal-order state variable system which has this transfer function is 4, as is discussed in Chapter 12. The determinants of both the left and right MFD forms of $\mathbf{P}$ hint that the poles or characteristic modes of the fourth-order realization will be at $s = -1, -1, -2$, and -3. The main objective of this and the preceding problem is to demonstrate certain elementary operations on polynomial matrices to achieve alternative MFD forms. The underlying questions of systematic procedures to follow, when to stop, what constitutes good forms, minimal degree, and so on are left to Chapter 12. This same transfer function is considered again in Examples 12.2 and 12.7 using other methods. Section 6.3 gives more on polynomial matrix methods.

4.32 Two subsystems are described in state variable form as

$$\dot{\mathbf{x}}_1 = \mathbf{A}_1\mathbf{x}_1 + \mathbf{B}_1\mathbf{u}_1; \quad \mathbf{y}_1 = \mathbf{C}_1\mathbf{x}_1 + \mathbf{D}_1\mathbf{u}_1; \quad r \text{ inputs, } m \text{ outputs}$$

$$\dot{\mathbf{x}}_2 = \mathbf{A}_2\mathbf{x}_2 + \mathbf{B}_2\mathbf{u}_2; \quad \mathbf{y}_2 = \mathbf{C}_2\mathbf{x}_2 + \mathbf{D}_2\mathbf{u}_2; \quad m \text{ inputs, } r \text{ outputs}$$

They are interconnected in the feedback loop shown in Figure 4.7. Use substitution and matrix algebra to derive the state variable model for the composite system with inputs $\mathbf{u}_a$ and outputs $\mathbf{y}_1$ and $\mathbf{y}_2$.

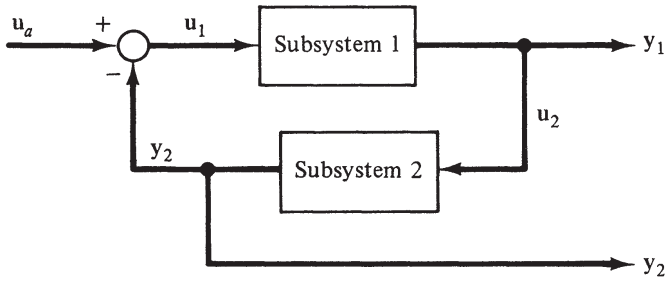


Figure 4.7

Write the output of the summing junction as

$$\mathbf{u}_1 = \mathbf{u}_a - \mathbf{y}_2 = \mathbf{u}_a - \mathbf{C}_2\mathbf{x}_2 - \mathbf{D}_2\mathbf{y}_1$$

$$= \mathbf{u}_a - \mathbf{C}_2\mathbf{x}_2 - \mathbf{D}_2[\mathbf{C}_1\mathbf{x}_1 + \mathbf{D}_1\mathbf{u}_1]$$

Combining the two $\mathbf{u}_1$ terms and premultiplying by the matrix inverse gives $\mathbf{u}_1 = [\mathbf{I}_r + \mathbf{D}_2\mathbf{D}_1]^{-1}\{\mathbf{u}_a - \mathbf{C}_2\mathbf{x}_2 - \mathbf{D}_2\mathbf{C}_1\mathbf{x}_1\}$. For convenience, let $\mathbf{L} = [\mathbf{I}_r + \mathbf{D}_2\mathbf{D}_1]^{-1}$. Then

$$\dot{\mathbf{x}}_1 = \mathbf{A}_1\mathbf{x}_1 + \mathbf{B}_1\mathbf{L}\mathbf{u}_a - \mathbf{B}_1\mathbf{L}\mathbf{C}_2\mathbf{x}_2 - \mathbf{B}_1\mathbf{L}\mathbf{D}_2\mathbf{C}_1\mathbf{x}_1$$

$$\mathbf{y}_1 = \mathbf{C}_1\mathbf{x}_1 + \mathbf{D}_1\mathbf{L}\mathbf{u}_a - \mathbf{D}_1\mathbf{L}\mathbf{C}_2\mathbf{x}_2 - \mathbf{D}_1\mathbf{L}\mathbf{D}_2\mathbf{C}_1\mathbf{x}_1$$

and

$$\dot{\mathbf{x}}_2 = \mathbf{A}_2\mathbf{x}_2 + \mathbf{B}_2\{\mathbf{C}_1\mathbf{x}_1 + \mathbf{D}_1\mathbf{L}\mathbf{u}_a - \mathbf{D}_1\mathbf{L}\mathbf{C}_2\mathbf{x}_2 - \mathbf{D}_1\mathbf{L}\mathbf{D}_2\mathbf{C}_1\mathbf{x}_1\}$$

$$\mathbf{y}_2 = \mathbf{C}_2\mathbf{x}_2 + \mathbf{D}_2\{\mathbf{C}_1\mathbf{x}_1 + \mathbf{D}_1\mathbf{L}\mathbf{u}_a - \mathbf{D}_1\mathbf{L}\mathbf{C}_2\mathbf{x}_2 - \mathbf{D}_1\mathbf{L}\mathbf{D}_2\mathbf{C}_1\mathbf{x}_1\}$$

or

$$\begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 - \mathbf{B}_1\mathbf{L}\mathbf{D}_2\mathbf{C}_1 & -\mathbf{B}_1\mathbf{L}\mathbf{C}_2 \\ \mathbf{B}_2\mathbf{C}_1 - \mathbf{B}_2\mathbf{D}_1\mathbf{L}\mathbf{D}_2\mathbf{C}_1 & \mathbf{A}_2 - \mathbf{B}_2\mathbf{D}_1\mathbf{L}\mathbf{C}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1\mathbf{L} \\ \mathbf{B}_2\mathbf{D}_1\mathbf{L} \end{bmatrix} \mathbf{u}_a$$

and

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix}=\begin{bmatrix}\mathbf{C}_1-\mathbf{D}_1\mathbf{L}\mathbf{D}_2\mathbf{C}_1 & -\mathbf{D}_1\mathbf{L}\mathbf{C}_2\\ \mathbf{D}_2\mathbf{C}_1-\mathbf{D}_1\mathbf{L}\mathbf{D}_2\mathbf{C}_1 & \mathbf{C}_2-\mathbf{D}_2\mathbf{D}_1\mathbf{L}\mathbf{C}_2\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}+\begin{bmatrix}\mathbf{D}_1\mathbf{L}\\ \mathbf{D}_2\mathbf{D}_1\mathbf{L}\end{bmatrix}\mathbf{u}_a$$

PROBLEMS

4.33 Show that every real, square matrix **A** can be written as the sum of a symmetric matrix and a skew-symmetric matrix.

4.34 Let $\mathbf{E}=[e_1 \quad e_2 \quad \cdots \quad e_n]^T$ be a column of errors in a multivariable control system. Show that the sum of the squares of the errors can be written in several forms, $e_1^2+e_2^2+\cdots+e_n^2=\mathbf{E}^T\mathbf{E}=\mathrm{Tr}(\mathbf{E}\mathbf{E}^T)$.

4.35 Consider the h-parameter model of a transistor, which is typical of many two-port devices (Figure 4.8).

$$\begin{bmatrix}v_1\\ i_2\end{bmatrix}=\begin{bmatrix}h_{11} & h_{12}\\ h_{21} & h_{22}\end{bmatrix}\begin{bmatrix}i_1\\ v_2\end{bmatrix}$$

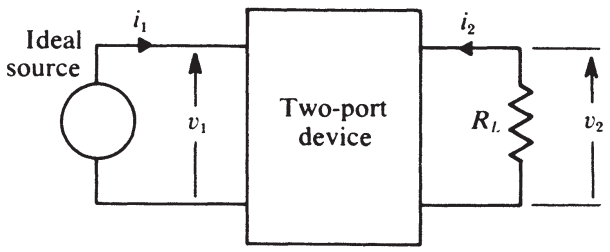


Figure 4.8

Add the third equation $v_2=-R_L i_2$ and find i_1, i_2, and v_2 if the source is an ideal voltage source v_1.

4.36 If the ideal source in the previous problem is a current source i_1, find v_1, i_2, and v_2.

4.37 Compute $|\mathbf{A}|$ using Laplace expansion, pivotal condensation, elementary operations, and the method of Problem 4.14. Draw conclusions about the effort required by each method.

$$\mathbf{A}=\begin{bmatrix}1 & 3 & -1 & 4\\ 2 & 0 & 1 & 5\\ -1 & 6 & 10 & -8\\ 0 & -2 & 7 & 1\end{bmatrix}$$

4.38 Find the inverses of

$$\mathbf{A}=\begin{bmatrix}4 & 1 & 1\\ 2 & 0 & 3\\ 1 & 1 & 5\end{bmatrix},\qquad \mathbf{B}=\begin{bmatrix}0 & -2 & -3 & -4 & -5\\ -1 & -1 & -3 & -4 & -5\\ 4 & 8 & 13 & 16 & 20\\ 2 & 4 & 6 & 9 & 10\\ 8 & 16 & 24 & 32 & 41\end{bmatrix},$$

$$\mathbf{C}=\left[\begin{array}{cc:ccc:cc}1 & 2 & 0 & 0 & 0 & 0\\ -2 & 1 & 0 & 0 & 0 & 0\\ \hdashline 0 & 0 & 5 & 7 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ \hdashline 0 & 0 & 0 & 0 & 3 & 8\\ 0 & 0 & 0 & 0 & 8 & 3\end{array}\right]$$

(*Hint*: **B** is the matrix of Problem 4.8. Use an identity from Problem 4.4).

4.39 Find the Laplace transforms of

$$\mathbf{A}(t) = \begin{bmatrix} 1 & t \\ e^{-at} & b \sin \beta t \\ t^2 & e^{-t} \cos \beta t \end{bmatrix}, \qquad \mathbf{B}(t) = \begin{bmatrix} \cosh \beta t & \sinh \beta t \\ te^{-at} & \cos \beta t \end{bmatrix}$$

4.40 Find the inverse Laplace transform of

$$\mathbf{A}(s) = \begin{bmatrix} \frac{24}{s^5} & \frac{s^2 - \beta^2}{(s^2 + \beta^2)^2} \end{bmatrix}, \qquad \mathbf{B}(s) = \begin{bmatrix} \frac{1}{(s+1)(s+2)} & \frac{s}{(s+1)(s+2)} \\ 0 & s \end{bmatrix}$$

4.41 Find the upper triangular square root matrix of

(a) $\mathbf{A} = \begin{bmatrix} 4 & -1 & 2 \\ -1 & 8 & 4 \\ 2 & 4 & 9 \end{bmatrix}$ **(b)** $\mathbf{A} = \begin{bmatrix} 4 & 6 & 1 \\ 6 & 1 & 2 \\ 1 & 2 & 2 \end{bmatrix}$

(c) $\mathbf{A} = \begin{bmatrix} 16 & 4 & 1 & -1 & 3 \\ 4 & 10 & 4 & 2 & -2 \\ 1 & 4 & 25 & 4 & 1 \\ -1 & 2 & 4 & 11 & 7 \\ 3 & -2 & 1 & 7 & 17 \end{bmatrix}$

4.42 Find both a left and right MFD form for $H(s) = \begin{bmatrix} 1/(s+1)^2 & 1/(s+1) \\ 0 & 1/(s+1) \end{bmatrix}$ Both should have $\mathbf{P}(s)$ with degree 3.

4.43 Find an MFD form with the inverse on the right and with degree of $|\mathbf{P}(s)| = 3$ for $\mathbf{H}(s) = \begin{bmatrix} 1/[(s+1)(s+2)] & 1/(s+1) \\ 1/(s+2) & 1/(s+3) \end{bmatrix}$.

4.44 Find the **A, B, C,** and **D** state matrices for a composite system of Figure 3.27 of Sec. 3.5. Subsystem 1 has the transfer function $H_1(s) = (s+5)/(s^2+3s+6)$ and subsystem 2 has $H_2(s) = (s+2)/(s^2+4s+3)$.

4.45 Use the same two subsystems as in Problem 4.44 and add two more, $H_3(s) = 10/(s+6)$ and $H_4(s) = (s-2)/(s^2+s+1)$, interconnected as in Figure 3.28. Find the state matrices for the composite system, using controllable canonical forms to describe each subsystem.

4.46 Find the composite state variable matrices for a system with the feedback topology of Problem 4.32. Use systems 1 and 2 of Problem 4.44.

4.47 Pressure drop-flow rate relations through many devices are nonlinear. For an orifice the flow rate Q and the pressure drop $P_1 - P_2$ are related by $Q = c\sqrt{P_1 - P_2}$. Derive a linear relation for the flow out of an orifice at the bottom of a tank. The tank is nominally kept filled to a height h_n. The fluid density is ρ lb-sec^2/ft^4. Thus $P_1 = \rho g h$ and $P_2 = 0$.

4.48 A navigation scheme uses a sextant to measure the angle included between the directions to two known landmarks from the position of the sextant. Let the sextant position vector be **x.** The landmark position vectors are $\mathbf{r}_1$ and $\mathbf{r}_2$.

(a) Find the nonlinear expression for the measured angle ϑ.

(b) Find the linear relation between small perturbations in θ and in **x.**

5

Vectors and Linear Vector Spaces

5.1 INTRODUCTION

Every student of introductory physics is familiar with the concept of a vector as a quantity which possesses both a magnitude and a direction. In Chapter 3 the terms state vector and state space were used in parameterizing models of dynamical systems. The state components could very well be a mixed set of physical quantities such as voltages, temperatures, and displacements. The formal procedures for picking states could even yield state components which are linear combinations of these disparate items. Arranging such a mixture of elements in a column matrix and referring to it as a vector seems inconsistent with the physical magnitude and direction concept of a vector. A primary objective of this chapter is to rectify these different notions of vectors and the vector spaces to which they belong. The discussion begins with a review of vector in the more familiar physical sense. The generalizations to the more abstract notions of vectors are then presented.

A second objective is to discuss various kinds of transformations on vectors. These topics have wide applicability in almost every branch of science and engineering. The central focus of this book is modeling and controlling physical systems. Therefore, there is interest in knowing how an initial state vector transforms into a state vector at a later time or how inputs transform to states or to outputs. Certain transformations of coordinates allow greater insight into system behavior and simplify the analysis and calculations. Some intrinsic system properties remain invariant to transformations, just as the length of a physical vector must not change when different coordinate systems are selected. Although all these topics cannot be dealt with completely in this chapter, the conceptual and computational foundations are presented.

5.2 PLANAR AND THREE-DIMENSIONAL REAL VECTOR SPACES

Many physical quantities, such as force and velocity, possess both a magnitude and a direction. Such entities are referred to as *vector* quantities. They are often represented by directed line segments or arrows. The length represents the magnitude, and the orientation indicates the direction.

If one point in the plane is defined as the origin, a unique vector can be associated with every point in the (two-dimensional) plane. The origin is defined as the zero vector, $\mathbf{0}$, and every other point can be associated with the directed line segment from the origin to the point. The same correspondence between points and directed line segments can obviously be made with points along the real line $\mathcal{R}$ (one dimension) and in three dimensions. Thus the terms point and vector can be used interchangeably.

If a coordinate system is defined in the plane, then each point can be identified by a unique pair of ordered numbers. These coordinate numbers can be written as a column matrix. However, since many different coordinate systems could be selected, a given vector could be represented by many different column matrices. A vector is *not* a column matrix but is a more basic entity which may be *represented* by a column matrix once a coordinate system is defined.

Vector Addition, Subtraction, and Multiplication by a Scalar

The coordinate-free description of vector addition is given by the parallelogram law. The sum of two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is the diagonal of the parallelogram formed with $\mathbf{v}_1$ and $\mathbf{v}_2$ as sides. Since $-\mathbf{v}_2$ is a vector with the same magnitude and orientation, but the opposite direction of $\mathbf{v}_2$, the vector difference $\mathbf{v}_1 - \mathbf{v}_2$ is just the sum of $\mathbf{v}_1$ and $-\mathbf{v}_2$. Multiplication of a vector by a scalar alters the magnitude but not the orientation. In particular, any nonzero vector $\mathbf{v}$ can be used to form a *unit* vector $\hat{\mathbf{v}}$ with the same direction as $\mathbf{v}$ by multiplying $\mathbf{v}$ by the reciprocal of its magnitude.

Whenever vectors are represented as column matrices with respect to a common coordinate system, the usual rules apply for addition of matrices and multiplication of a matrix by a scalar.

Vector Products

Products such as $\mathbf{vw}$ are not defined because of matrix conformability requirements. Three types of vector products are defined.

The *inner product* (or scalar product or dot product) of $\mathbf{v}$ and $\mathbf{w}$ is defined as $\langle \mathbf{v}, \mathbf{w} \rangle = vw \cos\theta$, where v and w are the vector magnitudes and θ is the angle included between the two vectors. When real vectors are represented in orthogonal cartesian coordinates, the inner product may be computed in terms of the components as $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T\mathbf{w} = \mathbf{w}^T\mathbf{v}$. If $\mathbf{v}$ and $\mathbf{w}$ are perpendicular, then $\theta = \pi/2$ so that $\langle \mathbf{v}, \mathbf{w} \rangle = 0$. Any two vectors which have a zero inner product are said to be *perpendicular,* or *orthogonal.* The zero vector is considered to be orthogonal to every other vector. The magnitude of a vector $\mathbf{v}$ can be expressed as $v = \langle \mathbf{v}, \mathbf{v} \rangle^{1/2}$.

The *outer product* of two real vectors $\mathbf{v} = [v_1 \quad v_2 \quad v_3]^T$ and $\mathbf{w} = [w_1 \quad w_2 \quad w_3]^T$ is defined as

$$\mathbf{v}\rangle\langle\mathbf{w} = \mathbf{v}\mathbf{w}^T = \begin{bmatrix} v_1 w_1 & v_1 w_2 & v_1 w_3 \\ v_2 w_1 & v_2 w_2 & v_2 w_3 \\ v_3 w_1 & v_3 w_2 & v_3 w_3 \end{bmatrix}$$

Since matrix multiplication is not commutative, neither is the outer product; $\mathbf{v}\mathbf{w}^T \neq \mathbf{w}\mathbf{v}^T$.

The *cross product* $\mathbf{v} \times \mathbf{w}$ is defined only in three dimensions. This product yields another vector, perpendicular to the plane of **v** and **w**. It points in the direction a right-hand screw would advance if **v** were rotated toward **w** through the smaller of the two angles θ between them. The magnitude is equal to the area of the parallelogram formed by **v** and **w**, i.e., $vw \sin\theta$.

5.3 AXIOMATIC DEFINITION OF A LINEAR VECTOR SPACE

Concepts such as directed line segments, lengths, angles, and dimensions of the space are considered to be intuitively obvious in one, two, or three dimensions. In dimensions higher than three, visualization is no longer possible. In cases such as a state vector with components made up of voltages, temperatures, and displacements, it is not yet clear what the vector characteristics are, even if there are only two or three components. For these reasons a more axiomatic definition of vectors and vector spaces is required. It will still be helpful to consider some of the general results for the particular cases of two or three dimensions. Geometrical descriptions of this nature will frequently be of use in gaining understanding of the concepts.

Linear Vector Spaces

A linear vector space $\mathscr{X}$ is a set of elements, called vectors, defined over a scalar number field $\mathscr{F}$, which satisfies the following conditions for addition and multiplication by scalars.

1. For any two vectors $\mathbf{x} \in \mathscr{X}$ and $\mathbf{y} \in \mathscr{X}$, the sum $\mathbf{x} + \mathbf{y} = \mathbf{v}$ is also a vector belonging to $\mathscr{X}$.
2. Addition is commutative: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
3. Vector addition is also associative: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$.
4. There is a zero vector, **0**, contained in $\mathscr{X}$ which satisfies $\mathbf{x} + \mathbf{0} = \mathbf{0} + \mathbf{x} = \mathbf{x}$.
5. For every $\mathbf{x} \in \mathscr{X}$ there is a unique vector $\mathbf{y} \in \mathscr{X}$ such that $\mathbf{x} + \mathbf{y} = \mathbf{0}$. This vector **y** is $-\mathbf{x}$.
6. For every $\mathbf{x} \in \mathscr{X}$ and for any scalar $a \in \mathscr{F}$, the product $a\mathbf{x}$ gives another vector $\mathbf{y} \in \mathscr{X}$. In particular, if a is the unit scalar,
$$1 \cdot \mathbf{x} = \mathbf{x} \cdot 1 = \mathbf{x}$$
7. For any scalars $a \in \mathscr{F}$ and $b \in \mathscr{F}$, and for any $\mathbf{x} \in \mathscr{X}$, $a(b\mathbf{x}) = (ab)\mathbf{x}$.

8. Multiplication by scalars is distributive,

$$(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$$

$$a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$$

The sets of all real one-, two-, or three-dimensional vectors discussed in the previous section satisfy all of these conditions and, therefore, are linear vector spaces. Elements in these familiar spaces can be represented as ordered sets of real numbers $[\alpha_1]$, $[\alpha_1 \quad \alpha_2]^T$, and $[\alpha_1 \quad \alpha_2 \quad \alpha_3]^T$, respectively. A fairly obvious generalization is to consider spaces whose elements are ordered n-tuples of real numbers $[\alpha_1 \quad \alpha_2 \cdots \alpha_n]^T$, where n is a finite integer. This space is referred to as $\mathcal{R}^n$. If the scalars α_i are allowed to be complex, then the space is referred to as $\mathcal{C}^n$. Both of these possibilities will be simultaneously covered by referring to an ordered set of n-tuples $\alpha_i \in \mathcal{F}$ as belonging to the space $\mathcal{X}^n$. Many other vector spaces can be defined. Some examples which can be verified to satisfy the required axioms are:

1. The set of all $m \times n$ matrices with elements in $\mathcal{F}$. Since valid number fields $\mathcal{F}$ include such possibilities as the real numbers, the complex numbers, or the set of rational polynomial functions with real or complex coefficients, the possibilities here are many. In particular, it is possible to define a vector space of all $m \times n$ transfer functions.
2. The set of all continuous or piecewise continuous time functions $f(t)$ on some interval $a \leq t \leq b$ or an ordered set $f_1(t), f_2(t), \ldots, f_n(t)$ of such functions.
3. The set of all polynomials of degree less than or equal to n, with coefficients belonging to the real or complex number fields $\mathcal{F}$. The zero element required by axiom 4 would be the $m \times n$ null matrix, the function which is identically zero, or the polynomial which has all its coefficients zero, respectively. A rational polynomial function could also be defined as a vector. The zero element would have all the numerator coefficients identically zero.

This short list provides just some of the possibilities and makes it clear that the notions of magnitude and direction are not required—or at least are not obvious—in the definitions of abstract vectors. The notion of a vector as an ordered n-tuple also seems not to apply in some of these examples, although that will be seen to depend upon how the notion of coordinate systems is generalized. That is, the countable Fourier expansion coefficients can be thought of as ordered components of a periodic function, defined as a vector. When the full generality required by some of these examples is implied, the vector space will be referred to as $\mathcal{X}$. For the most part, the discussions here will deal with vectors in the sense of the previous section and their generalizations to $\mathcal{X}^n$.

One fact emerges from the preceding examples that seems puzzling at first. The set of rational polynomial functions was used as a field and a vector space of $m \times n$ matrices was defined over that field in (1). Then the same rational polynomial function was itself declared a vector in (3). A comparison of the axioms used to define a field in Chapter 4 and those used here to define an abstract vector space show a great deal of

similarity. The requirements on a vector space are weaker because no inverse vector element is required. Actually *any* set of elements which qualifies as a field in the sense defined in Chapter 4 can also be used to define a vector space over itself as the field. The real line is an example. It can be considered as a one-dimensional vector space (i.e., ordered one-tuples) defined over the real number field. More complicated vector spaces can be built up as ordered sets of the simpler vectors. The simplest example is that n-tuples of reals are ordered one-tuples. The same is true of the rational polynomial functions. One of them can be treated as a vector defined over itself as the field. Or, an ordered array of them, e.g., a transfer function, can be defined as a vector. It is important to note, however, that not all vector spaces are equivalent to fields because of the requirement of the inverse element. The set of polynomials can define a vector space, but they do not constitute a field because the ratio of two such polynomials is in general not a member of the set of polynomials.

In addition to the fact that vector inverse elements are not required, it is explicitly pointed out that no notion of the *product* of two vector elements is found in the required axioms. Very frequently an additional definition of an *inner product* is imposed upon a vector space. This is extremely useful. It allows the generalization of familiar geometrical concepts, such as length or distance, and angles between vectors. Whenever this extra definition is imposed, a restricted special class of vector spaces, called *inner product spaces,* is being dealt with.

5.4 LINEAR DEPENDENCE AND INDEPENDENCE

Consider three vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ in three-dimensional space. If there exists a relation among them, such as $\mathbf{x}_3 = \alpha\mathbf{x}_1 + \beta\mathbf{x}_2$, then it is clear that $\mathbf{x}_3$ lies in the plane through $\mathbf{x}_1$ and $\mathbf{x}_2$, no matter what scalar values are attached to α and β. $\mathbf{x}_3$ is said to be dependent on $\mathbf{x}_1$ and $\mathbf{x}_2$. The notions of dependence and independence must be generalized to arbitrary sets of vectors.

Definition 5.1. Let a finite number of vectors belonging to a linear vector space $\mathscr{X}$ be denoted by $\{\mathbf{x}_i\} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. If there exists a set of n scalars, a_i, at least one of which is not zero, which satisfies $a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n = \mathbf{0}$, then the vectors $\{\mathbf{x}_i\}$ are said to be *linearly dependent.*

Definition 5.2. Any set of vectors $\{\mathbf{x}_i\}$ which is not linearly dependent is said to be *linearly independent.* That is, if $a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n = \mathbf{0}$ implies that each $a_i = 0$, then $\{\mathbf{x}_i\}$ is a set of linearly independent vectors.

EXAMPLE 5.1 Consider the set of n vectors $\mathbf{e}_i$, each of which has n components. All components of $\mathbf{e}_i$ are zero, except the ith component, which is unity. Then

$$a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \cdots + a_n\mathbf{e}_n = a_1\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix} + a_2\begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix} + \cdots + a_n\begin{bmatrix}0\\0\\\vdots\\0\\1\end{bmatrix} = \begin{bmatrix}a_1\\a_2\\\\\vdots\\a_n\end{bmatrix}$$

The only way that this sum can give the **0** vector is if each and every $a_i = 0$. Thus the set $\{\mathbf{e}_i\}$ is linearly independent. This set of $\mathbf{e}_i$ vectors represents the natural extension of the cartesian coordinate directions often used in two- and three-dimensional spaces. They will be referred to as the natural cartesian coordinates. ■

EXAMPLE 5.2 Let the components of three vectors with respect to the natural cartesian coordinates be

$$\mathbf{x}_1^T = [5 \quad 2 \quad 3], \qquad \mathbf{x}_2^T = [-1 \quad 7 \quad 4], \qquad \mathbf{x}_3^T = [14 \quad 50 \quad 36]$$

These vectors are linearly dependent because $2\mathbf{x}_1 + 3\mathbf{x}_2 - \frac{1}{2}\mathbf{x}_3 = 0$. ■

Lemma 5.1. Let $\mathcal{V} = \{\mathbf{x}_i, i = 1, n\}$ be a set of linearly dependent vectors. Then the set formed by adding any vector $\mathbf{x}_{n+1}$ to $\mathcal{V}$ is also linearly dependent.

Lemma 5.2. If a set of vectors $\{\mathbf{x}_i\}$ is linearly dependent, then one of the vectors can be written as a linear combination of the others.

Tests for Linear Dependence

Consider a set of n vectors $\{\mathbf{x}_i\}$, each having n components with respect to a given coordinate system. Let $\mathbf{A}$ be the $n \times n$ matrix which has the $\mathbf{x}_i$ vectors as columns. The set of vectors is linearly dependent if and only if $|\mathbf{A}| = 0$. The zero in this determinant test for independence must be the zero element of the number field over which the vectors are defined. In particular, if the rational polynomial functions are the field, the determinant must be identically zero for *all* values of the variable s or z or t used in defining the polynomials. It is not sufficient for the polynomial to equal zero for specific isolated values.

EXAMPLE 5.3 Use the preceding test to show that the three vectors of Example 5.2 are linearly dependent.

We find $|\mathbf{A}| = \begin{vmatrix} 5 & -1 & 14 \\ 2 & 7 & 50 \\ 3 & 4 & 36 \end{vmatrix} = 0$. Therefore, the set is linearly dependent. ■

The previous test for linear independence is not applicable when considering a set of n vectors $\{\mathbf{x}_i\}$, each of which has m components, with $m \neq n$. The matrix $\mathbf{A}$ is $m \times n$ and $|\mathbf{A}|$ is not defined. Assume the set is linearly dependent so that $\sum_{i=1}^{n} a_i \mathbf{x}_i = \mathbf{0}$ with at least one nonzero a_i. Premultiplying by $\bar{\mathbf{x}}_1^T$ gives the scalar equation

$$a_1 \bar{\mathbf{x}}_1^T \mathbf{x}_1 + a_2 \bar{\mathbf{x}}_1^T \mathbf{x}_2 + \cdots + a_n \bar{\mathbf{x}}_1^T \mathbf{x}_n = 0$$

Repeated premultiplication by $\bar{\mathbf{x}}_2^T$, then $\bar{\mathbf{x}}_3^T$, and so on gives a set of n simultaneous equations, which can be written in matrix form as

$$[\bar{\mathbf{x}}_i^T \mathbf{x}_j][\mathbf{a}] = \mathbf{0}$$

If the $n \times n$ matrix $\mathbf{G} \triangleq [\bar{\mathbf{x}}_i^T \mathbf{x}_j]$ has a nonzero determinant, then $\mathbf{G}^{-1}$ exists, and solving gives

$$\mathbf{a} = \mathrm{G}^{-1}\mathbf{0} = \mathbf{0}$$

This contradicts the assumption of at least one nonzero a_i. The matrix $\mathbf{G}$ is called the *Grammian matrix*. A necessary and sufficient condition for the set $\{\mathbf{x}_i\}$ to be linearly dependent is that $|\mathbf{G}| = 0$. An alternate means of determining linear independence is to reduce the matrix $\mathbf{A}$ to row-reduced echelon form, as mentioned in Section 4.10. This approach is convenient for computer applications and will be used in Chapter 6.

EXAMPLE 5.4 Consider two vectors defined over the complex number field, $\mathbf{x}_1 = [1+j \quad 6]^T$ and $\mathbf{x}_2 = [5+j \quad 18-12j]^T$. Show that they are linearly dependent.

$$\begin{vmatrix} 1+j & 5+j \\ 6 & 6(3-2j) \end{vmatrix} = 6\begin{vmatrix} 1+j & 5+j \\ 1 & 3-2j \end{vmatrix} = 6[(5+j)-(5+j)] \equiv 0$$

Thus the vectors are dependent. In fact, $\mathbf{x}_2 = (3-2j)\mathbf{x}_1$. Similarly, $\mathbf{x}_1 = \begin{bmatrix} (z+1)/[z(z-0.5)] \\ 1/z \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} 1/[(z+1)(z-0.5)] \\ 1/(z^2+2z+1) \end{bmatrix}$ are dependent as can be verified by (1) showing that the determinant formed with these columns is identically zero, (2) by forming the Grammian, (3) by performing elementary row and/or column operations to show that the rank is 1, or (4) by noting that $\mathbf{x}_2 = [z/(z+1)^2]\mathbf{x}_1$. Noticing the linear dependencies by inspection is not nearly so easy when the vectors—and hence the scalar proportionality factors—are defined over the complex numbers or rational polynomial functions. ■

Geometrical Significances of Linear Dependence

Two vectors can normally be used to form sides of a parallelogram. If the vectors are linearly dependent, they have the same direction, so the parallelogram degenerates to a line. It is shown in Problem 5.14 that the 2×2 Grammian determinant is equal to the square of the area of the parallelogram formed by the vectors. Thus $|\mathbf{G}| = 0$ indicates that the parallelogram has degenerated to a single line. Three vectors can normally be used to define the sides of a parallelepiped. If there is one linear dependency relation (i.e., any two of the three vectors are linearly independent but the set of three is linearly dependent), then the parallelepiped has degenerated to a plane figure and hence has zero volume. $|\mathbf{G}| = 0$ indicates this. If there are two dependency relations, the parallelepiped degenerates to a single line. Similar significance can be attached in higher dimensional cases. The number of dependency relationships among a set of vectors (or the columns of a matrix) is called the *degeneracy*, q. For an $n \times n$ matrix $\mathbf{A}$, q, n, and the rank r_A are related by

$$n = r_A + q$$

Often it is easier to determine the rank first and then use that to determine $q = n - r_A$. The degeneracy q is the key to finding eigenvectors and generalized eigenvectors for a matrix with repeated eigenvalues. This is discussed in Chapter 7.

Sylvester's Law of Degeneracy

If $\mathbf{A}$ and $\mathbf{B}$ are square conformable matrices whose product is $\mathbf{AB} = \mathbf{C}$, Sylvester's law of degeneracy can be used to place bounds on the degeneracy of $\mathbf{C}$, q_c in terms of the degeneracy of $\mathbf{A}$, q_A, and of $\mathbf{B}$, q_B:

$$\max\{q_A, q_B\} \leq q_C \leq q_A + q_B$$

If the relations between n, rank, and degeneracy are used, the following limits on the rank of **C** can be obtained:

$$r_A + r_B - n \leq r_C \leq \min\{r_A, r_B\}$$

A similar result was presented in Chapter 4 for **A** and **B** not necessarily square, but the lower limit there was $0 \leq r_C$.

5.5 VECTORS WHICH SPAN A VECTOR SPACE; BASIS VECTORS AND DIMENSIONALITY

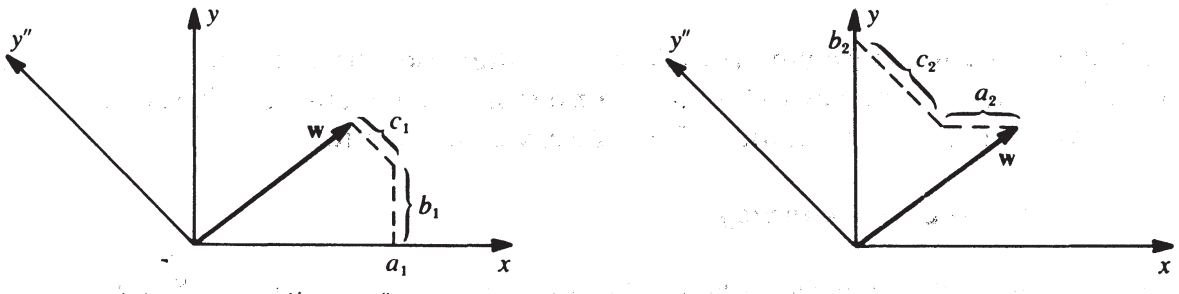
The dimension of a vector space has been referred to several times. In two or three dimensions, the concept is obvious, but in higher dimensions a precise definition must be relied upon rather than intuition. It is first necessary to define what is meant by a set of vectors which *span* a vector space.

Let $\mathscr{X}$ be a linear vector space and let $\{\mathbf{u}_i, i = 1, m\}$ be a subset of vectors in $\mathscr{X}$. The set $\{\mathbf{u}_i\}$ is said to span the space $\mathscr{X}$ if for every vector $\mathbf{x} \in \mathscr{X}$ there is at least one set of scalars $a_i \in \mathscr{F}$ which permits **x** to be expressed as a linear combination of the $\mathbf{u}_i$,

$$\mathbf{x} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_m\mathbf{u}_m = \sum_{i=1}^{m} a_i\mathbf{u}_i$$

Note that if the vectors **x** and $\mathbf{u}_i$ can be expressed as columns of n scalars with respect to a common coordinate system, then in matrix notation $\mathbf{x} = \mathbf{Ua}$, where **U** is the $n \times m$ matrix whose ith column is $\mathbf{u}_i$ and $\mathbf{a} = [a_1 \quad a_2 \ldots a_m]^T$. The scalars a_i are the components of **x** in the $\mathbf{u}_i$ coordinate direction. Vectors such as $\mathbf{u}_i$, which merely span the space, do not make a good coordinate system because there may be more vectors than necessary, and as a result the a_i coefficients are not unique.

EXAMPLE 5.5 Consider all vectors in the plane. Then any pair of noncollinear vectors such as $\{\mathbf{x}, \mathbf{y}\}$, $\{\mathbf{x}', \mathbf{y}'\}$, and $\{\mathbf{x}'', \mathbf{y}''\}$ spans the two-dimensional space, since every vector in the plane can be represented as a combination of any one of these pairs. Another set of vectors which spans this space is $\{\mathbf{x}, \mathbf{y}, \mathbf{y}''\}$. Two ways of expressing a vector **w** in terms of these three vectors are shown in Figure 5.1. The coefficients in the linear expansion of a vector in terms of a set of spanning vectors need not be unique. There is an infinite number of possibilities in this example. ■



(*a*) $\mathbf{w} = a_1\mathbf{x} + b_1\mathbf{y} + c_1\mathbf{y}''$ (*b*) $\mathbf{w} = a_2\mathbf{x} + b_2\mathbf{y} - c_2\mathbf{y}''$

Figure 5.1

Basis Vectors

A set of basis vectors, $\mathcal{B} = \{\mathbf{v}_i\}$, for a space $\mathcal{X}$ is a subset of vectors in $\mathcal{X}$ which (1) spans the space $\mathcal{X}$ and (2) is a linearly independent set. Alternatively, a set of basis vectors is a set consisting of the minimum number of vectors required to span the space $\mathcal{X}$. There are infinitely many choices for basis vectors in a given vector space. For example, in Fig. 5.1 any *two* of the three vectors $\{\mathbf{x}, \mathbf{y}, \mathbf{y}''\}$ or any two other noncollinear vectors in the plane could be selected. Every valid basis set for a given space will contain the same number of vectors, e.g., two for the plane. Once a basis is selected for the space $\mathcal{X}$, every vector $\mathbf{x} \in \mathcal{X}$ has a *unique* representation or expansion with respect to that basis. That is, there is a unique set of scalar coefficients such that $\mathbf{x} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n$. As before, if $\mathbf{x}$ and each $\mathbf{v}_i$ can be expressed as a column of scalars, then in matrix notation $\mathbf{x} = V\mathbf{a}$. Because of the uniqueness of the relation between a given $\mathbf{x}$ and a set of coefficients $\{a_i\}$ for a given basis set, basis vectors are the natural generalization of coordinate vectors discussed in two and three dimensions. The column of scalars $\mathbf{a}$ can be viewed as the same vector $\mathbf{x}$ but expressed in a different coordinate system. It should now be clear why a column of scalars is not a vector but is only one representation of the vector in a particular set of coordinates. The vector itself is a more abstract entity, such as the directed line segment used in elementary physics, which exists independent of any coordinate system. The vector appears as a column of scalars only after a coordinate system—i.e., a basis set—is introduced.

Some basis sets are more convenient to work with than others. Most would agree that the mutually orthogonal $\mathbf{x}, \mathbf{y}$ of Fig. 5.1 would be more convenient than the other choices shown there. That set would be even more convenient if both $\mathbf{x}$ and $\mathbf{y}$ had unit length. This generalizes naturally in $\mathcal{X}^n$ to the set $\{\mathbf{e}_i\}$ discussed in Example 5.1. Whenever a specific basis is not mentioned for $\mathcal{X}^n$, this natural cartesian basis set will be implied.

Dimension of a Vector Space

Definition 5.3. The dimension of a vector space $\mathcal{X}$, written dim $(\mathcal{X})$, is equal to the number of vectors in the basis set $\mathcal{B}$. Thus an n-dimensional linear vector space has n basis vectors.

EXAMPLE 5.6

1. *Let $\mathcal{X}$ be the linear vector space consisting of all n-component vectors*

$$\mathbf{x}^T = [x_1 \quad x_2 \quad x_3 \quad \cdots \quad x_n]$$

which satisfy $x_1 = x_2 = x_3 = \cdots = x_n$. Since the basis set for this space consists of the single vector $\mathbf{v}^T = [1 \quad 1 \quad 1 \quad \cdots \quad 1]$, this space is one-dimensional.

2. *Let $\mathcal{X}$ be the linear space consisting of all polynomials of degree $n-1$ or less,* $\{f(t) | f(t) = \alpha_1 + \alpha_2 t + \alpha_3 t^2 + \cdots + \alpha_n t^{n-1}, \alpha_i \in \mathcal{F}\}$. An obvious basis set is $\{1, t, t^2, \ldots, t^{n-1}\}$. Since the basis contains n elements, dim $(\mathcal{X}) = n$, but $\mathcal{X}$ is not $\mathcal{X}^n$ as defined earlier. ■

It should be pointed out that a linear vector space can consist of a single element, the zero vector **0**. Such a space is said to be a zero-dimensional space. A field, by way of contrast, must always have at least two elements, 0 and 1.

5.6 SPECIAL OPERATIONS AND DEFINITIONS IN VECTOR SPACES

In order to generalize many of the useful concepts of familiar two- and three-dimensional spaces to n-dimensional spaces, some additional definitions are required.

Inner Product

Let $\mathscr{X}$ be an n-dimensional linear vector space defined over the scalar number field $\mathscr{F}$. If, to each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathscr{X}$, a unique scalar belonging to $\mathscr{F}$, called the inner product, is assigned, then $\mathscr{X}$ is said to be an inner product space. Various definitions for the inner product are possible. Any scalar valued function of $\mathbf{x}$ and $\mathbf{y}$ can be defined as the *inner product,* written $\langle \mathbf{x}, \mathbf{y} \rangle$, provided the following axioms are satisfied:

1. $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$ (complex conjugate property)
2. $\langle \mathbf{x}, \alpha\mathbf{y}_1 + \beta\mathbf{y}_2 \rangle = \alpha\langle \mathbf{x}, \mathbf{y}_1 \rangle + \beta\langle \mathbf{x}, \mathbf{y}_2 \rangle$ (linear, homogeneous property)
3. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ for all $\mathbf{x}$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (nonnegative length)

A commonly used definition of the complex inner product in $\mathscr{X}^n$, which is sufficiently general for our purposes, is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \bar{\mathbf{x}}^T \mathbf{y}$$

If $\mathscr{F}$ is the set of reals, then the real inner product can be defined in the same way, but the complex conjugate on $\mathbf{x}$ is then superfluous. The inner product space defined on the real scalar field is called *Euclidean space.* Unless otherwise stated, the inner product will be assumed to be the complex inner product given above.

Combining axioms 1 and 2, it is easy to show that the inner product also satisfies

$$\langle \alpha\mathbf{x}_1 + \beta\mathbf{x}_2, \mathbf{y} \rangle = \bar{\alpha}\langle \mathbf{x}_1, \mathbf{y} \rangle + \bar{\beta}\langle \mathbf{x}_2, \mathbf{y} \rangle$$

The Grammian matrix introduced in Sec. 5.4 is generally defined in terms of the inner product as $\mathbf{G} = [\langle \mathbf{x}_i, \mathbf{x}_j \rangle]$. Some further definitions of inner products are given in the problems. In particular, see Problem 5.25 for vector spaces of matrices.

Vector Norm

Axioms 1 and 3 for inner products ensure that $\langle \mathbf{x}, \mathbf{x} \rangle$ is a nonnegative real number and is zero if and only if $\mathbf{x} = \mathbf{0}$. Because of these properties, the inner product can be used to define the *length,* or *norm,* of a vector as $\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2}$. This norm will be used throughout this book unless an explicit statement to the contrary is made. It is called the quadratic norm, or in the case of real vector spaces, the Euclidean norm. In two or three dimensions, it is easy to see that this definition for the length of $\mathbf{x}$ satisfies the

conditions of Euclidean geometry. It is a generalization to n dimensions of the theorem of Pythagoras.

Many other norms can be defined; the only requirements are that $\|\mathbf{x}\|$ be a nonnegative real scalar satisfying

1. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$
2. $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ for any scalar α
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

The last condition is called the triangle inequality, for reasons which are obvious in two or three dimensions.

An important inequality, called the Cauchy-Schwarz inequality, can be expressed in terms of the norm and the absolute value of the inner product:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$$

The equality holds if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

Unit Vectors

A unit vector, $\hat{\mathbf{x}}$, is by definition a vector whose norm is unity, $\|\hat{\mathbf{x}}\| = 1$. Any nonzero vector $\mathbf{x}$ can be normalized to form a unit vector.

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$$

Metric or Distance Measure

The concept of distance between two points (vectors are used synonymously with points) in a linear vector space can be introduced by using the norm. The distance between two points $\mathbf{x}$ and $\mathbf{y}$ is defined as the scalar function

$$\rho(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

When the quadratic norm is used, this gives

$$\rho(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle^{1/2}$$

Generalized Angles in n-Dimensional Spaces

The concept of angles between vectors can be generalized to real n-dimensional spaces by extending the notion of the dot product of two- or three-dimensional spaces,

$$\mathbf{x} \cdot \mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \cdot \|\mathbf{y}\| \cos\theta$$

Thus, the cosine of the angle between $\mathbf{x} \in \mathscr{X}$ and $\mathbf{y} \in \mathscr{X}$ is

$$\cos\theta = \frac{1}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} \langle \mathbf{x}, \mathbf{y} \rangle = \langle \hat{\mathbf{x}}, \hat{\mathbf{y}} \rangle$$

It was mentioned earlier that various inner products can be defined. The particular choice of inner product dictates a specific meaning for the geometric concept of angle. Since $\langle \mathbf{x}, \mathbf{y} \rangle$ need not be real in spaces defined over the complex scalars, it is not particularly useful to try to place an interpretation upon angles in complex spaces.

Outer Product

The outer product (sometimes called the dyad product) of two vectors $\mathbf{x}$ and $\mathbf{y}$ belonging to $\mathscr{X}^n$ is

$$\mathbf{x}\rangle\langle\mathbf{y} = \mathbf{x}\bar{\mathbf{y}}^T$$

The brackets are motivated by a comparison with the usual definition for the inner product.

Multiplication of a Vector by an Arbitrary, Conformable Matrix

Since a vector in $\mathscr{X}^n$ can be represented as a column matrix with respect to a specific set of basis vectors, all the operations of matrix algebra can then be applied. In particular, premultiplication by a conformable matrix yields another column matrix, which is the representation of a vector. If the matrix multiplier is $n \times n$ and skew-symmetric, then two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathscr{R}^n$ related by $\mathbf{y} = \mathbf{A}\mathbf{x}$ can be seen to have $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. In two- or three-dimensional spaces, a zero inner product indicates that the two vectors are orthogonal (this is generalized to any vector space in the next section). Thus multiplication by a skew-symmetric matrix is somewhat akin to generalizing the cross product in that the resultant is orthogonal to $\mathbf{x}$.

Just as vectors are abstract elements that often can be represented by column matrices, matrices as used in the preceding paragraph are specific coordinate-system dependent representations of a more abstract *transformation* operator, to be discussed later.

5.7 ORTHOGONAL VECTORS AND THEIR CONSTRUCTION

Any two vectors $\mathbf{x}$ and $\mathbf{y}$ which belong to a linear vector space $\mathscr{X}$ are said to be *orthogonal* if and only if

$$\langle \mathbf{x}, \mathbf{y} \rangle = 0$$

This is the natural generalization of the geometric concept of perpendicularity. Note that this definition of orthogonality indicates that the zero vector is orthogonal to every other vector. If each pair of vectors in a given set is mutually orthogonal, then the set is said to be an *orthogonal set.* If, in addition, each vector in this orthogonal set is a unit vector, then the set is said to be *orthonormal.* Each pair of orthonormal vectors $\hat{\mathbf{v}}_i$ and $\hat{\mathbf{v}}_j$ satisfies $\langle \hat{\mathbf{v}}_i, \hat{\mathbf{v}}_j \rangle = \delta_{ij}$, where δ_{ij} is the Kronecker delta and equals 1 if $i = j$ and 0 otherwise. Orthonormal vectors are convenient choices for basis vectors. The natural set of cartesian basis vectors $\mathbf{e}_i$ of Example 5.1 is the simplest example of an orthonormal set.

The Gram-Schmidt Process

Orthonormal vectors form a convenient basis set, so it is of interest to know how to construct an orthonormal set. Given any set of n linearly independent vectors $\{\mathbf{y}_i, i = 1, n\}$, an orthonormal set $\{\hat{\mathbf{v}}_i, i = 1, n\}$ can be constructed by using the Gram-Schmidt process. The process consists of two steps. First an orthogonal set $\{\mathbf{v}_i\}$ is constructed, and second, each vector in this set is normalized. Let $\mathbf{v}_1 = \mathbf{y}_1$ and select $\mathbf{v}_2$ as the vector formed from $\mathbf{y}_2$ by subtracting out the component in the direction of $\mathbf{v}_1$. This is equivalent to requiring that $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$. Let $\mathbf{v}_2 = \mathbf{y}_2 - a\mathbf{v}_1$. Then in order to satisfy orthogonality,

$$a = \frac{\langle \mathbf{v}_1, \mathbf{y}_2 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}$$

so that

$$\mathbf{v}_2 = \mathbf{y}_2 - \frac{\langle \mathbf{v}_1, \mathbf{y}_2 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1$$

the next vector is chosen as

$$\mathbf{v}_3 = \mathbf{y}_3 - a_1\mathbf{v}_1 - a_2\mathbf{v}_2$$

and the two scalars a_i are chosen to satisfy

$$\langle \mathbf{v}_1, \mathbf{v}_3 \rangle = 0 \quad \text{and} \quad \langle \mathbf{v}_2, \mathbf{v}_3 \rangle = 0$$

This leads to

$$\mathbf{v}_3 = \mathbf{y}_3 - \frac{\langle \mathbf{v}_1, \mathbf{y}_3 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1 - \frac{\langle \mathbf{v}_2, \mathbf{y}_3 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle}\mathbf{v}_2$$

Continuing in this manner leads to the general equation

$$\mathbf{v}_i = \mathbf{y}_i - \sum_{k=1}^{i-1} \frac{\langle \mathbf{v}_k, \mathbf{y}_i \rangle}{\langle \mathbf{v}_k, \mathbf{v}_k \rangle}\mathbf{v}_k$$

After all n of the vectors $\mathbf{v}_i$ are computed, the normalization

$$\hat{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|}, \qquad i = 1, \ldots, n$$

gives the desired orthonormal set.

EXAMPLE 5.7 Construct a set of orthonormal vectors from

$$\mathbf{y}_1^T = [1 \quad 0 \quad 1], \qquad \mathbf{y}_2^T = [-1 \quad 2 \quad 1], \qquad \mathbf{y}_3^T = [0 \quad 1 \quad 2]$$

Since the Grammian gives $|\mathbf{G}| = 4$, these vectors are linearly independent.

Step 1. Let

$$\mathbf{v}_1 = \mathbf{y}_1 = [1 \quad 0 \quad 1]^T$$

$$\mathbf{v}_2 = \mathbf{y}_2 - \frac{\overset{0}{\cancel{\langle \mathbf{v}_1, \mathbf{y}_2 \rangle}}}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1 = \mathbf{y}_2$$

In this case $\mathbf{y}_1$ and $\mathbf{y}_2$ are already orthogonal.

$$\mathbf{v}_3 = \mathbf{y}_3 - \frac{\langle \mathbf{v}_1, \mathbf{y}_3 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1 - \frac{\langle \mathbf{v}_2, \mathbf{y}_3 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle}\mathbf{v}_2 = \begin{bmatrix} -\frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}^T$$

Step 2. Normalize $\mathbf{v}_i$ to get $\hat{\mathbf{v}}_i$:

$$\hat{\mathbf{v}}_i = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad \hat{\mathbf{v}}_2 = \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} = \frac{1}{\sqrt{6}}\begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, \quad \hat{\mathbf{v}}_3 = \frac{\mathbf{v}_3}{\|\mathbf{v}_3\|} = \frac{1}{\sqrt{3}}\begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$$

■

The Modified Gram-Schmidt Process

By modifying the sequence of operations slightly, a modified Gram-Schmidt process is obtained. It has superior numerical properties when the operations are carried out on a computer with finite word size. The benefits are most apparent when some vectors in the set are nearly collinear. As before, select $\mathbf{y}_1 = \mathbf{v}_1$ and $\hat{\mathbf{v}}_1 = \mathbf{v}_1/\|\mathbf{v}_1\|$. Next subtract from every $\mathbf{y}_j$, $j \geq 2$, the components in the direction of $\hat{\mathbf{v}}_1$. That is, $\mathbf{y}_j' = \mathbf{y}_j - \langle \hat{\mathbf{v}}_1, \mathbf{y}_j \rangle \hat{\mathbf{v}}_1$ for $j = 2, 3, \ldots, n$. The unit normalized version of $\mathbf{y}_2$ is selected as $\hat{\mathbf{v}}_2$. Then the components along the direction of $\hat{\mathbf{v}}_2$ are subtracted from all $\mathbf{y}_j'$,

$$\mathbf{y}_j'' = \mathbf{y}_j' - \langle \hat{\mathbf{v}}_2, \mathbf{y}_j \rangle \hat{\mathbf{v}}_2 \quad \text{for } j = 3, 4, \ldots, n$$

This continues until all n orthonormal vectors are found. Theoretically, identical results will be obtained from both versions, but practically, because of finite machine precision, some $\langle \hat{\mathbf{v}}_i, \hat{\mathbf{v}}_j \rangle$ factors will not be precisely zero and the results will differ. Matrix versions of the two construction processes are given in Problems 5.17 and 5.18.

EXAMPLE 5.8 Repeat the previous example using the modified Gram-Schmidt process.

As before $\mathbf{v}_1 = \mathbf{y}_1$ and $\hat{\mathbf{v}}_1$ is unchanged. Then $\mathbf{y}_2' = \mathbf{y}_2 - \langle \hat{\mathbf{v}}_1, \mathbf{y}_2 \rangle \hat{\mathbf{v}}_1$ as before, and $\mathbf{y}_3' = \mathbf{y}_3 - \langle \hat{\mathbf{v}}_1, \mathbf{y}_3 \rangle \hat{\mathbf{v}}_1 = [-1 \quad 1 \quad 1]^T$. Finally, $\hat{\mathbf{v}}_2 = \mathbf{y}_2'/\|\mathbf{y}_2'\|$ and $\mathbf{y}_3'' = \mathbf{y}_3' - \langle \hat{\mathbf{v}}_2, \mathbf{y}_3' \rangle \hat{\mathbf{v}}_2 = [-\frac{1}{3} \quad -\frac{1}{3} \quad \frac{1}{3}]^T$. The final orthonormal set is the same as before, but the intermediate operations are different. ■

Use of the Gram-Schmidt Process to Obtain QR *Matrix Decomposition*

It is frequently useful to express an $n \times m$ matrix $\mathbf{A}$ as a product of an orthogonal matrix $\mathbf{Q}$ (i.e., $\mathbf{Q}^{-1} = \mathbf{Q}^T$) and an upper-triangular matrix $\mathbf{R}$. The Gram-Schmidt process is one way of determining $\mathbf{Q}$ and $\mathbf{R}$ such that $\mathbf{A} = \mathbf{QR}$. Assume first that the m columns $\mathbf{a}_j$ of $\mathbf{A}$ are linearly independent. This requires that $m \leq n$. If the Gram-Schmidt process is applied to the set $\{\mathbf{a}_j\}$ to obtain the orthornormal set $\{\mathbf{t}_j\}$, the construction equations for the $\mathbf{t}_j$ vectors are

$$\mathbf{t}_1 = \alpha_{11}\mathbf{a}_1, \qquad \mathbf{t}_2 = \alpha_{12}\mathbf{a}_1 + \alpha_{22}\mathbf{a}_2, \ldots, \qquad \mathbf{t}_j = \alpha_{1j}\mathbf{a}_1 + \alpha_{2j}\mathbf{a}_2 + \cdots + \alpha_{jj}\mathbf{a}_j$$

Calculation of the α_{ij} scalar coefficients involve inner product and norm operations, as demonstrated earlier. Collectively, these construction equations can be written as the matrix equation

$$[\mathbf{t}_1 \quad \mathbf{t}_2 \quad \ldots \quad \mathbf{t}_m] = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \ldots \quad \mathbf{a}_m] \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1m} \\ 0 & \alpha_{22} & \alpha_{23} & \cdots & \alpha_{2m} \\ 0 & 0 & \alpha_{33} & \cdots & \alpha_{3m} \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & \alpha_{mm} \end{bmatrix} \tag{5.1}$$

or simply as $\mathbf{T} = \mathbf{AS}$. Although $\mathbf{T}$ need not be square, it has certain orthogonality properties. By construction, $\langle \mathbf{t}_i, \mathbf{t}_j \rangle = \delta_{ij}$. This means that $\mathbf{T}^T\mathbf{T} = \mathbf{I}_m$. However, $\mathbf{T}$ is not orthogonal in the sense defined as the end of Sec. 4.8 because $\mathbf{TT}^T \neq \mathbf{I}_n$ unless $n = m$. The $\mathbf{S}$ matrix is upper-triangular and nonsingular (because by Sylvester's law of degeneracy $\mathbf{S}$ must have rank m). Its inverse is also upper triangular. Therefore, several alternate forms are immediate.

$$\mathbf{I}_m = \mathbf{T}^T\mathbf{AS}, \qquad \mathbf{T}^T\mathbf{A} = \mathbf{S}^{-1}, \quad \text{and} \quad \mathbf{A} = \mathbf{TS}^{-1} \tag{5.2}$$

The last equation is almost in the widely used **QR** decomposition form—but not quite because $\mathbf{T}$ is not generally square and not truly orthogonal. The original columns in $\mathbf{A}$ can always be augmented with additional vectors $\mathbf{v}_k$ in such a way that the matrix $[\mathbf{A} \mid \mathbf{V}]$ has n linearly independent columns. The Gram-Schmidt process can then be applied to construct a full set of n orthonormal vectors $\{\mathbf{t}_j\}$, which can be used to define the columns of the $n \times n$ matrix $\mathbf{Q}$. This is true regardless of the size or rank of $\mathbf{A}$—that is, the earlier assumptions that $m \leq n$ and $\text{Rank}(\mathbf{A}) = m$ are no longer required.

Although the expressions in Eq. (5.2) were derived from a Gram-Schmidt construction point of view, the last version has a Gram-Schmidt *expansion* interpretation as well. This point of view is used here. If a given column $\mathbf{a}_j$ is expanded in terms of the orthonormal set $\{\mathbf{t}_j\}$, the kth expansion coefficient is $\langle \mathbf{t}_k, \mathbf{a}_j \rangle$, or $\mathbf{t}_k^T\mathbf{a}_j$. All the expansion coefficients for all $\mathbf{a}_j$ columns are contained in the matrix given by $\mathbf{Q}^T\mathbf{A} = \mathbf{R}$. The matrix $\mathbf{R}$ thus obtained will be upper-triangular (or the nonsquare generalization of upper-triangular), with exactly $r_A = \text{rank}(\mathbf{A})$ nonzero rows. Since $\mathbf{Q}$ is orthogonal, it follows that $\mathbf{A} = \mathbf{QR}$. $\mathbf{R}$ is a generalization of $\mathbf{S}^{-1}$ and $\mathbf{Q}$ is a generalization of $\mathbf{T}$ in Eq. (5.2). The following five matrices illustrate the range of possibilities. Both the original ($\mathbf{A} = \mathbf{TS}^{-1}$) and the augmented ($\mathbf{A} = \mathbf{QR}$) forms of the decomposition are shown for each matrix. Results are rounded approximations.

$m = n$, not full rank:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} [1.414 \quad 1.414]$$

$$= \begin{bmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{bmatrix} \begin{bmatrix} 1.414 & 1.414 \\ 0 & 0 \end{bmatrix}$$

$m < n$, full rank:

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 0.2673 & 0.3132 \\ 0.5345 & -0.8351 \\ 0.8018 & 0.4523 \end{bmatrix} \begin{bmatrix} 3.7417 & 1.3363 \\ 0 & 2.053 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2673 & 0.3132 & 0.9113 \\ 0.5345 & -0.8351 & 0.1302 \\ 0.8018 & 0.4523 & -0.3906 \end{bmatrix} \begin{bmatrix} 3.7417 & 1.3363 \\ 0 & 2.053 \\ 0 & 0 \end{bmatrix}$$

$m < n$, not full rank:

$$\mathbf{C} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 0.2673 \\ 0.5345 \\ 0.8018 \end{bmatrix} [3.7417 \quad 7.4833]$$

$$= \begin{bmatrix} 0.2673 & -0.9569 & -0.1139 \\ 0.5345 & 0.2455 & -0.8087 \\ 0.8018 & 0.1553 & 0.5771 \end{bmatrix} \begin{bmatrix} 3.7416 & 7.4833 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$m > n$, full rank:

$$\mathbf{D} = \begin{bmatrix} 2 & 1 & 6 \\ 1 & 4 & 8 \end{bmatrix}$$

$$= \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} \begin{bmatrix} 2.2361 & 2.6833 & 8.9443 \\ 0 & 3.1305 & 4.4721 \end{bmatrix}$$

$m > n$, not full rank:

$$\mathbf{E} = \begin{bmatrix} 1 & 4 & 7 & 3 \\ 2 & 0 & 2 & 1 \\ 3 & 4 & 9 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2673 & 0.7715 \\ 0.5345 & -0.6172 \\ 0.8018 & 0.1543 \end{bmatrix} \begin{bmatrix} 3.7417 & 4.2762 & 1.0156 & 4.543 \\ 0 & 3.7033 & 5.5549 & 2.3156 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2673 & 0.7715 & 0.5774 \\ 0.5345 & -0.6172 & 0.5774 \\ 0.8018 & 0.1543 & -0.5774 \end{bmatrix} \begin{bmatrix} 3.7417 & 4.2762 & 1.0156 & 4.543 \\ 0 & 3.7033 & 5.5549 & 2.3156 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The determination of a **QR** decomposition is not generally a hand calculation. It is very worthwhile to have a computer algorithm for this purpose. The **QR** decomposition procedure provides a good way of determining the rank of a matrix. It can be adapted to solving the eigenvalue problem of Chapter 7. Finally, in Chapter 12 it provides an easy way of determining a minimal-dimension state model from an arbitrary state model.

5.8 VECTOR EXPANSIONS AND THE RECIPROCAL BASIS VECTORS

Every vector $\mathbf{x} \in \mathcal{X}$ has a unique expansion

$$\mathbf{x} = \sum_{i=1}^{n} a_i \mathbf{v}_i$$

with respect to the basis set $\mathcal{B} = \{\mathbf{v}_i, i = 1, n\}$. Taking the inner product of $\mathbf{v}_j$ and $\mathbf{x}$ gives

$$\langle \mathbf{v}_j, \mathbf{x} \rangle = \left\langle \mathbf{v}_j, \sum_{i=1}^{n} a_i \mathbf{v}_i \right\rangle = \sum_{i=1}^{n} a_i \langle \mathbf{v}_j, \mathbf{v}_i \rangle$$

If the basis set is orthonormal so that $\langle \mathbf{v}_j, \mathbf{v}_i \rangle = \delta_{ij}$, then the jth expansion coefficient is $a_j = \langle \mathbf{v}_j, \mathbf{x} \rangle$.

EXAMPLE 5.9 Use the set of orthonormal vectors generated in Example 5.7 as basis vectors and find the three coefficients a_i which allow $\mathbf{z} = [4 \quad -8 \quad 1]^T$ to be written as

$$\mathbf{z} = a_1 \hat{\mathbf{v}}_1 + a_2 \hat{\mathbf{v}}_2 + a_3 \hat{\mathbf{v}}_3$$

Because $\{\hat{\mathbf{v}}_i\}$ is an orthonormal set,

$$a_1 = \langle \hat{\mathbf{v}}_1, \mathbf{z} \rangle, \qquad a_2 = \langle \hat{\mathbf{v}}_2, \mathbf{z} \rangle \quad \text{and} \quad a_3 = \langle \hat{\mathbf{v}}_3, \mathbf{z} \rangle$$

so that

$$\mathbf{z} = \frac{5}{\sqrt{2}}\hat{\mathbf{v}}_1 - \frac{19}{\sqrt{6}}\hat{\mathbf{v}}_2 + \frac{5}{\sqrt{3}}\hat{\mathbf{v}}_3 \qquad \blacksquare$$

Basis vector expansions can be used to gain insight into the state equations of a time-invariant dynamic system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}$$

At any given time instant, $\mathbf{x} \in \Sigma$, the n-dimensional state space. Let $\{\mathbf{v}_i, i = 1, \ldots, n\}$ be a set of constant basis vectors for Σ. Then at any time instant there exists a set of unique scalars α_i such that $\mathbf{x} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n$, or $\mathbf{x} = \mathbf{V}\boldsymbol{\alpha}$. Since all $\mathbf{v}_i$ are constant, the time variations of $\mathbf{x}$ must be contained in the expansion coefficients α_i, and thus $\dot{\mathbf{x}} = \mathbf{V}\dot{\boldsymbol{\alpha}}$. Substitution into the state equations gives

$$\mathbf{V}\dot{\boldsymbol{\alpha}} = \mathbf{A}\mathbf{V}\boldsymbol{\alpha} + \mathbf{B}\mathbf{u} \quad \text{or} \quad \dot{\boldsymbol{\alpha}} = \mathbf{V}^{-1}\mathbf{A}\mathbf{V}\boldsymbol{\alpha} + \mathbf{V}^{-1}\mathbf{B}\mathbf{u} \quad \text{and} \quad \mathbf{y} = \mathbf{C}\mathbf{V}\boldsymbol{\alpha} + \mathbf{D}\mathbf{u}$$

By defining $\mathbf{A}' = \mathbf{V}^{-1}\mathbf{A}\mathbf{V}$, $\mathbf{B}' = \mathbf{V}^{-1}\mathbf{B}$, $\mathbf{C}' = \mathbf{C}\mathbf{V}$, and $\mathbf{x}' \equiv \boldsymbol{\alpha}$, it is seen that the change of basis vectors from the original set, which might have been the natural cartesian set, to $\{\mathbf{v}_i\}$ has created a different state variable model for the same system. In Chapter 3 various forms of the state variable models were derived. Is it possible that all the varieties which were presented can be related by a simple change of basis? That this is not always so is now demonstrated. Consider a system whose input output transfer function is $H(s) = (s + 1)/(s^2 + 3s + 2)$. The controllable canonical form of the state equations are obtained from Figure 5.2 as

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u \qquad y = [1 \quad 1]\mathbf{x}$$

The observable canonical form is obtained from Figure 5.3 as

$$\dot{\mathbf{x}}' = \begin{bmatrix} -3 & 1 \\ -2 & 0 \end{bmatrix}\mathbf{x}' + \begin{bmatrix} 1 \\ 1 \end{bmatrix}u \qquad y = [1 \quad 0]\mathbf{x}'$$

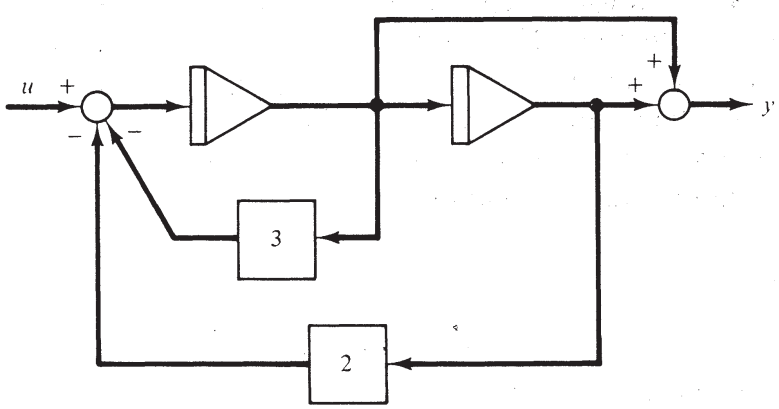


Figure 5.2

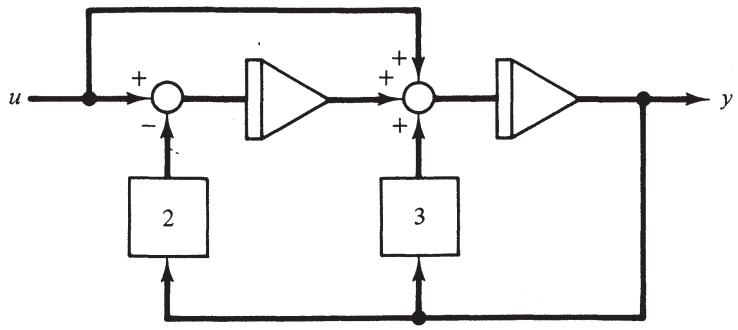


Figure 5.3

It is to be shown that no basis set (or equivalently, no nonsingular matrix $\mathbf{V}$) exists for which all three transformations from $\mathbf{A}, \mathbf{B}, \mathbf{C}$ to $\mathbf{A}', \mathbf{B}', \mathbf{C}'$ will be true. Consider first $\mathbf{V}^{-1}\mathbf{AV} = \mathbf{A}'$, or $\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}\mathbf{V} = \mathbf{V}\begin{bmatrix} -3 & 1 \\ -2 & 0 \end{bmatrix}$. Expanding $\mathbf{V}$ into components v_{ij} and writing out the matrix products shows that v_{11} must equal v_{22} but that $\mathbf{V}$ is otherwise unrestricted so far. From the expanded form of $\mathbf{CV} = \mathbf{C}'$ it is found that equality requires in addition that $v_{12} = v_{21}$. Using both of these restrictions in $\mathbf{V}^{-1}\mathbf{B} = \mathbf{B}'$ converted to $\mathbf{B} = \mathbf{VB}'$ gives

$$\begin{bmatrix} v_{11} & v_{12} \\ v_{12} & v_{11} \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

which is an impossible contradiction. These two particular state models are *not* related by a simple change of basis vectors.

Can some other type of transformation be found which relates the primed and unprimed state models? To answer this, differentiate both forms of the y equation, giving

$$\dot{y} = \mathbf{C}\dot{\mathbf{x}} = \mathbf{CAx} + \mathbf{CB}u \quad \text{and} \quad \dot{y} = \mathbf{C}'\dot{\mathbf{x}}' = \mathbf{C}'\mathbf{A}'\mathbf{x}' + \mathbf{C}'\mathbf{B}'u$$

Grouping the differentiated and undifferentiated y equations together and combining all u terms gives

$$\begin{bmatrix} \mathbf{C}' \\ \mathbf{C}'\mathbf{A}' \end{bmatrix}\mathbf{x}' = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \end{bmatrix}\mathbf{x} + \begin{bmatrix} \mathbf{CB} - \mathbf{C}'\mathbf{B}' \\ 0 \end{bmatrix}u$$

Call the 2×2 matrix coefficients of $\mathbf{x}$ and $\mathbf{x}'$, $\mathbf{Q}$ and $\mathbf{Q}'$, respectively, and let the column coefficient of u be $\mathbf{W}$. $\mathbf{Q}'$ is invertible in this case, so $\mathbf{x}' = [\mathbf{Q}']^{-1}\{\mathbf{Qx} + \mathbf{W}u\}$. Comparing this with the transformation $\mathbf{x}' = \mathbf{Vx}$, which represents a change of basis, the presence of the $\mathbf{W}u$ term is an obvious difference. However, the important difference is that $[\mathbf{Q}']^{-1}\mathbf{Q}$ can never be represented by a nonsingular $\mathbf{V}$, since $\mathbf{Q}$ is singular for this system. Many variants of the state equations can be related by a change of basis vectors. The particular system models examined here cannot because of a failure in a basic property, to be examined in detail in Chapter 11.

Reciprocal Basis Vectors

When the basis set $\mathcal{B} = \{\mathbf{v}_i\}$ is not orthonormal, the preceding simple results no longer hold, but every vector $\mathbf{z} \in \mathcal{X}$ still has a unique expansion

$$\mathbf{z} = \sum_{i=1}^{n} a_i \mathbf{v}_i$$

Another set of n vectors, called the *reciprocal basis vectors* $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$, is introduced to facilitate finding the expansion coefficients. These reciprocal or dual basis vectors are defined by n^2 equations, each of the form

$$\langle \mathbf{r}_i, \mathbf{v}_j \rangle = \delta_{ij}$$

In matrix form this set of equations becomes

$$\mathbf{RB} = \mathbf{I}$$

where $\mathbf{B}$ is the $n \times n$ matrix whose columns are $\mathbf{v}_i$ and $\mathbf{R}$ is the $n \times n$ matrix whose *rows* are $\bar{\mathbf{r}}_i^T$. Thus $\mathbf{R} = \mathbf{B}^{-1}$, so the reciprocal basis vector $\mathbf{r}_i$ is the conjugate transpose of the ith row of $\mathbf{B}^{-1}$. With the reciprocal basis vectors available, it is apparent that the expansion coefficients are given by $a_i = \langle \mathbf{r}_i, \mathbf{z} \rangle$ so that

$$\mathbf{z} = \sum_{i=1}^{n} \langle \mathbf{r}_i, \mathbf{z} \rangle \mathbf{v}_i$$

EXAMPLE 5.10 Let $\mathbf{v}_1 = [1 \quad 0]^T$ and $\mathbf{v}_2 = [-1 \quad 1]^T$. Express the vector $\mathbf{z} = [3 \quad 3]^T$ in terms of this basis set.

First, the reciprocal basis set is found from

$$\mathbf{R} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

so that $\mathbf{r}_1 = [1 \quad 1]^T$ and $\mathbf{r}_2 = [0 \quad 1]^T$. The coefficients are

$$a_1 = \langle \mathbf{r}_1, \mathbf{z} \rangle = 6 \quad \text{and} \quad a_2 = \langle \mathbf{r}_2, \mathbf{z} \rangle = 3$$

so that $\mathbf{z} = 6\mathbf{v}_1 + 3\mathbf{v}_2$. A sketch of the $\mathbf{r}_i$ and $\mathbf{v}_i$ vectors may be informative. ■

The matrix $\mathbf{B}$ of basis vectors need not always be square. For example, the basis set for a two-dimensional subspace of a four-dimensional space would consist of two $\mathbf{b}_i$ vectors, each with four components. $\mathbf{B}$ is of dimension 4×2 and has no inverse in the usual sense. There will be two reciprocal basis vectors $\mathbf{r}_i$ also, and using their conjugate transposes as rows, $\mathbf{R}$ is of dimension 2×4. There are two ways for determining $\mathbf{R}$. One could augment the columns in $\mathbf{B}$ with two more columns $\mathbf{B}_a$ so that $[\mathbf{B} \mid \mathbf{B}_a]$ is square and invertible. If both columns in $\mathbf{B}_a$ are selected to be orthogonal to all columns in $\mathbf{B}$ (by forcing $\mathbf{B}^T\mathbf{B}_a = [0]$), then $[\mathbf{B} \quad \mathbf{B}_a]^{-1} = \begin{bmatrix} \mathbf{R} \\ \mathbf{R}_a \end{bmatrix}$. That is, the conjugate transposes of the desired reciprocal basis vectors are found in the first two rows of the augmented inverse. The second method of finding $\mathbf{R}$ is to notice that $\mathbf{R} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ will have the desired orthogonality property $\mathbf{RB} = \mathbf{I}$. In fact, these two methods give the same result. The direct expression for $\mathbf{R}$ is called the *left pseudo-inverse* of $\mathbf{B}$ and appears in many applications involving projections, approximations, and least-squares solutions, as is seen in the next chapter.

EXAMPLE 5.11 Find the reciprocal basis vectors for the basis set $\mathbf{b}_1 = [1 \quad 1 \quad 0 \quad 1]^T$ and $b_2 = [2 \quad 1 \quad 1 \quad 0]^T$. Then use them to find the components of the vector $\mathbf{y} = [3 \quad 0 \quad 1 \quad 2]^T$ along

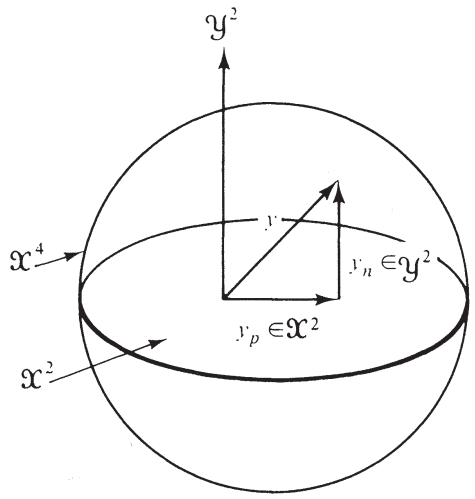


Figure 5.4

these two basis directions. That is, find the projection of $\mathbf{y}$ onto the two-dimensional space spanned by $\mathbf{b}_1$ and $\mathbf{b}_2$.

Using the augmentation approach first, we must find two independent solutions of

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Two solutions are $\mathbf{b}_3 = [0.5 \quad 0 \quad -1 \quad 0.5]^T$ and $\mathbf{b}_4 = [0.5 \quad -1 \quad 0 \quad 0.5]^T$. Using these to form $\mathbf{B}_a$ and inverting gives

$$\left[\begin{array}{c} \mathbf{R} \\ \hline \mathbf{R}_a \end{array}\right] = \left[\begin{array}{cccc} 0 & \frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & 0 & \frac{1}{3} & -\frac{1}{3} \\ \hline \frac{1}{3} & 0 & -\frac{2}{3} & -\frac{1}{3} \\ \frac{1}{3} & -\frac{2}{3} & 0 & \frac{1}{3} \end{array}\right]$$

Direct calculation of the second method shows that $(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ gives the first two rows, namely, $\mathbf{R}$, so $\mathbf{r}_1 = [0 \quad \frac{1}{3} \quad -\frac{1}{3} \quad \frac{2}{3}]^T$ and $\mathbf{r}_2 = [\frac{1}{3} \quad 0 \quad \frac{1}{3} \quad -\frac{1}{3}]^T$. The expansion coefficients of the vector $\mathbf{y}$ along the basis vectors $\mathbf{b}_1$ and $\mathbf{b}_2$ are $\langle \mathbf{r}_1, \mathbf{y} \rangle = 1$ and $\langle \mathbf{r}_2, \mathbf{y} \rangle = \frac{2}{3}$. Note that both these calculations are given by the matrix product $\mathbf{Ry}$. The resulting vector, the projection of $\mathbf{y}$ onto the space of $\{\mathbf{b}_1, \mathbf{b}_2\}$, is $\mathbf{y}_p = [\frac{7}{3} \quad \frac{5}{3} \quad \frac{2}{3} \quad 1]^T$ when it is expressed in terms of the same basis vectors as the original $\mathbf{y}$. This vector could just as well be referred to in component form as $[1 \quad \frac{2}{3}]^T$, provided it is understood that the basis vectors being used are $\mathbf{b}_1$ and $\mathbf{b}_2$. The component of $\mathbf{y}$ which is normal to the space of $\{\mathbf{b}_1, \mathbf{b}_2\}$ is given by $\mathbf{y}_n = \langle \mathbf{r}_3, \mathbf{y} \rangle \mathbf{b}_3 + \langle r_4, \mathbf{y} \rangle \mathbf{b}_4 = -\frac{1}{3}\mathbf{b}_3 + \frac{5}{3}\mathbf{b}_4 = [\frac{2}{3} \quad -\frac{5}{3} \quad \frac{1}{3} \quad 1]^T$. The expansion coefficients can also be computed from $\mathbf{R}_a\,\mathbf{y} = [-\frac{1}{3} \quad \frac{5}{3}]^T$. The original vector has been decomposed into components $\mathbf{y} = \mathbf{y}_p + \mathbf{y}_n$, and it is easily verified that $\mathbf{y}_p$ and $\mathbf{y}_n$ are orthogonal. With a little imagination, Figure 5.4 represents the decomposition of a four-dimensional space $\mathcal{X}^4$ into two separate orthogonal two-dimensional spaces, $\mathcal{X}^2$ and $\mathcal{Y}^2$. The projection of $\mathbf{y}$ onto each subspace is also shown. These notions are formalized in the next section. ■

5.9 LINEAR MANIFOLDS, SUBSPACES, AND PROJECTIONS

Let $\mathscr{X}$ be a linear vector space defined over the number field $\mathscr{F}$. A nonempty subset, $\mathscr{M}$, of $\mathscr{X}$ is called a *linear manifold* if for each vector $\mathbf{x}$ and $\mathbf{y}$ in $\mathscr{M}$, the combination $\alpha\mathbf{x} + \beta\mathbf{y}$ is also in $\mathscr{M}$ for arbitrary $\alpha, \beta \in \mathscr{F}$. The zero vector, of necessity, is included in every linear manifold.

A closed linear manifold is called a *subspace.* In finite dimensional spaces, there is no distinction between linear manifolds and subspaces, because every finite dimensional manifold is closed.

A subspace of an n-dimensional linear vector space $\mathscr{X}^n$ is itself a linear vector space contained within $\mathscr{X}^n$, but with dimension $m \leq n$. A *proper* subspace has $m < n$.

EXAMPLE 5.12 The spaces $\mathscr{X}^2$ and $\mathscr{Y}^2$ of Example 5.11 are both two-dimensional subspaces of the four-dimensional space $\mathscr{X}^4$.

The space defined in Problem 5.11 is a three-dimensional subspace of $\mathscr{X}^5$, and the first space defined in Example 5.6 is a one-dimensional subspace of $\mathscr{X}^n$. In general, since $\mathscr{X}^n$ has n basis vectors, deleting any one of the basis vectors leaves a basis set for an $n-1$ dimensional subspace, deleting two allows the definition of an $n-2$ dimensional subspace, etc. Note that $\mathbf{0}$ must be an element of every subspace. If it is the only element, then that subspace is zero dimensional. ■

Starting with one vector space it is possible to define other spaces, called *subspaces,* by selecting subsets of the basis vectors. The process can also go the other way. Starting with two linear vector spaces $\mathscr{U}$ and $\mathscr{V}$ defined over the same number field $\mathscr{F}$, a new vector space $\mathscr{X}$ can be constructed from their sum:

$$\mathscr{X} = \mathscr{U} + \mathscr{V}$$

This means that every vector $\mathbf{x}$ in $\mathscr{X}$ can be written as

$$\mathbf{x} = \mathbf{u} + \mathbf{v}, \qquad \mathbf{u} \in \mathscr{U}, \qquad \mathbf{v} \in \mathscr{V}$$

If there is one and only one pair $\mathbf{u}, \mathbf{v}$ for each $\mathbf{x}$, then $\mathscr{X}$ is called the *direct sum* of $\mathscr{U}$ and $\mathscr{V}$, written

$$\mathscr{X} = \mathscr{U} \oplus \mathscr{V}$$

This implies that the only vector common to both $\mathscr{U}$ and $\mathscr{V}$ is $\mathbf{0}$. In this case,

$$\dim(\mathscr{X}) = \dim(\mathscr{U}) + \dim(\mathscr{V})$$

In Example 5.11 $\mathscr{X}^4 = \mathscr{X}^2 \oplus \mathscr{Y}^2$. This is a direct sum because the basis vectors $\mathbf{b}_3$ and $\mathbf{b}_4$ were constructed to be orthogonal to both $\mathbf{b}_1$ and $\mathbf{b}_2$, and therefore every $\mathbf{u} \in \mathscr{Y}^2$ is orthogonal to every $\mathbf{v} \in \mathscr{X}^2$. As a simple example of a sum (as opposed to a direct sum) of two spaces, define $\mathscr{U}$ as the linear space with basis $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ of Example 5.11. Then $\mathscr{X}^4 = \mathscr{U} + \mathscr{Y}^2$. The fact that both spaces have one basis vector in common prevents a unique decomposition of vectors and causes $\dim(\mathscr{X}^4) \neq \dim(\mathscr{U}) + \dim(\mathscr{Y}^2)$.

Regardless of whether a space $\mathscr{X}$ was constructed as a sum of two spaces $\mathscr{U}$ and $\mathscr{V}$ or if $\mathscr{U}$ and $\mathscr{V}$ were selected as subspaces of $\mathscr{X}$, each vector $\mathbf{x}$ can be written as $\mathbf{x} = \mathbf{u} + \mathbf{v}$. Then $\mathbf{u}$ is called the projection of $\mathbf{x}$ on $\mathscr{U}$ and $\mathbf{v}$ is the projection of $\mathbf{x}$ on $\mathscr{V}$.

The Projection Theorem

Let $\mathcal{X}^n$ be an n-dimensional vector space, and let $\mathcal{U}$ be a subspace of dimension $m < n$. Then for every $\mathbf{x} \in \mathcal{X}^n$ there exists a vector $\mathbf{u} \in \mathcal{U}$, called the projection of $\mathbf{x}$ on $\mathcal{U}$, which satisfies

$$\langle \mathbf{x} - \mathbf{u}, \mathbf{y} \rangle = 0$$

for every vector $\mathbf{y} \in \mathcal{U}$. This says that $\mathbf{w} = \mathbf{x} - \mathbf{u}$ is orthogonal to $\mathbf{y}$. In other words, $\mathbf{u}$ is the orthogonal projection of $\mathbf{x}$ on $\mathcal{U}$, and $\mathbf{w}$ is orthogonal to the subspace $\mathcal{U}$.

For proofs of the projection theorem, see References 1 and 2.

For a given $\mathbf{x}$ there is a unique projection $\mathbf{u}$, but there are infinitely many $\mathbf{x}$ vectors which have the same projection. The set of all vectors in $\mathcal{X}^n$ which are orthogonal to $\mathcal{U}$ forms an $n - m$ dimensional subspace of $\mathcal{X}^n$, called the *orthogonal complement* of $\mathcal{U}$, written $\mathcal{U}^\perp$. Every $\mathbf{w} \in \mathcal{U}^\perp$ is orthogonal to every $\mathbf{y} \in \mathcal{U}$. The set of all vectors which are orthogonal to $\mathcal{U}^\perp$ is the subspace $\mathcal{U}$, that is, $(\mathcal{U}^\perp)^\perp = \mathcal{U}$. The spaces $\mathcal{X}^2$ and $\mathcal{Y}^2$ of Example 5.11 are orthogonal complements of one another. By using any subspace $\mathcal{U}$ and its orthogonal complement, an n-dimensional space can be expressed as the direct sum $\mathcal{X}^n = \mathcal{U} \oplus \mathcal{U}^\perp$. Each vector $\mathbf{x} \in \mathcal{X}^n$ can be written uniquely as $\mathbf{x} = \mathbf{u} + \mathbf{v}$. It is easy to show that $\|\mathbf{x}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$ because of the orthogonal nature of this decomposition.

The projection theorem and related concepts can be used to develop the theory of least squares estimation and the theory of *generalized* or *pseudo-inverses* of non-square or singular matrices. Some of these applications appear in the next chapter.

5.10 PRODUCT SPACES

Let $\mathcal{X}$ and $\mathcal{Y}$ be arbitrary linear vector spaces defined over the field $\mathcal{F}$. Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. Then the *product space* $\mathcal{X} \times \mathcal{Y}$ is defined as all ordered pairs of vectors $(\mathbf{x}, \mathbf{y})$. It can be verified that the product space satisfies the conditions of Sec. 5.3 and is therefore a linear vector space. Let $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{y}_1)$ and $\mathbf{z}_2 = (\mathbf{x}_2, \mathbf{y}_2)$ belong to $\mathcal{X} \times \mathcal{Y}$. Then addition and scalar multiplication are defined by $\mathbf{z}_1 + \mathbf{z}_2 = (\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}_1 + \mathbf{y}_2) = \mathbf{z}_2 + \mathbf{z}_1$ and $\alpha \mathbf{z}_1 = (\alpha \mathbf{x}_1, \alpha \mathbf{y}_1)$. The zero vector in $\mathcal{X} \times \mathcal{Y}$ is the ordered pair of zero elements $\mathbf{0} \in \mathcal{X}$ and $\mathbf{0} \in \mathcal{Y}$.

Product spaces can be formed as the product of any number of spaces. The familiar Euclidean three-dimensional space is a product space formed from products of the real line $\mathcal{R}^1$, $\mathcal{R}^3 = \mathcal{R}^1 \times \mathcal{R}^1 \times \mathcal{R}^1$. Another common product space is formed from n products of the space of square integrable functions, $\mathcal{L}_2[a, b] \times \mathcal{L}_2[a, b] \times \cdots \times \mathcal{L}_2[a, b]$. Each element in this space is of the form $(f_1(t), f_2(t), \ldots, f_n(t))$, where $f_i(t) \in \mathcal{L}_2[a, b]$. Elements in this product space are usually written more simply as n component vectors $\mathbf{f}(t)$.

The spaces used in forming a product space need not be the same type of spaces. If $\mathcal{R}^1$ is considered as a vector space with elements t and if $\mathbf{x} \in \mathcal{X}^m$, $\mathbf{y} \in \mathcal{Y}$, then elements of $\mathcal{R}^1 \times \mathcal{X}^m \times \mathcal{Y}$ are $\mathbf{z} = (t, \mathbf{x}, \mathbf{y})$. Product spaces were used in Chapter 3 in the definitions of a dynamic system. It should now be clear that a number of diverse objects can be grouped together and treated as components of a single vector in a product space.

For example, two time points t_0 and t_1, each in some segment τ of the real line, an initial state vector $\mathbf{x}(t_0) \in \Sigma$ and a segment of input vector functions $\mathbf{u}_{[t0,t1]} \in \mathcal{U}$ can be used to define a point or vector $\mathbf{p} = (t_0, t_1, \mathbf{x}(t_0), \mathbf{u}_{[t0,t1]})$ which belongs to the product space $\tau \times \tau \times \Sigma \times \mathcal{U}$. One requirement of a dynamical system is that there exist a *unique* mapping $\mathbf{x}(t_1) = \mathbf{g}(\mathbf{p})$.

5.11 TRANSFORMATIONS OR MAPPINGS

The concepts of functions, which were introduced in Sec. 3.1, are now generalized to abstract vector spaces. Let $\mathcal{X}$ and $\mathcal{Y}$ be linear vector spaces (not necessarily distinct), which are defined over the same scalar number field $\mathcal{F}$. If for each vector $\mathbf{x} \in \mathcal{X}$ there is associated, according to some rule, a vector $\mathbf{y} \in \mathcal{Y}$, then that "rule" defines a mapping of $\mathbf{x}$ into $\mathbf{y}$. This mapping rule is referred to as a transformation (or an operator or a function). This relationship is expressed by

$$\mathcal{A} : \mathcal{X} \rightarrow \mathcal{Y}$$

The transformation is $\mathcal{A}$ and the mapping rule is $\mathcal{A}(\mathbf{x}) = \mathbf{y}$. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are called the domain and codomain of $\mathcal{A}$, respectively. The domain is often written as $\mathcal{D}(\mathcal{A})$, and the range of $\mathcal{A}$ is $\mathcal{A}(\mathcal{X})$ or $\mathcal{R}(\mathcal{A})$. Obviously $\mathcal{R}(\mathcal{A})$ is contained within or equal to $\mathcal{Y}$. This is written as $\mathcal{R}(\mathcal{A}) \subseteq \mathcal{Y}$. In general, $\mathcal{A}$ maps $\mathcal{X}$ *into* $\mathcal{Y}$, but if the equality holds, it maps $\mathcal{X}$ *onto* $\mathcal{Y}$. Again, if $\mathcal{A}(\mathbf{x}) = \mathbf{y}$, $\mathbf{y}$ is called the image of $\mathbf{x}$ or $\mathbf{x}$ is the pre-image of $\mathbf{y}$. The transformation $\mathcal{A}$ is said to be one-to-one if

$$\mathbf{x}_1 \neq \mathbf{x}_2 \Rightarrow \mathcal{A}(\mathbf{x}_1) \neq \mathcal{A}(\mathbf{x}_2)$$

or equivalently, if

$$\mathcal{A}(\mathbf{x}_1) = \mathcal{A}(\mathbf{x}_2) \Rightarrow \mathbf{x}_1 = \mathbf{x}_2$$

If $\mathcal{A}$ is both one-to-one and onto, then for each $\mathbf{y} \in \mathcal{Y}$ there is a unique pre-image $\mathbf{x} \in \mathcal{X}$, and an inverse transformation $\mathcal{A}^{-1}$ maps $\mathbf{y}$ into $\mathbf{x}$. In this case, $\mathcal{A}(\mathbf{x}) = \mathbf{y}$ and $\mathcal{A}^{-1}(\mathbf{y}) = \mathbf{x}$, so $\mathcal{A}^{-1}(\mathcal{A}(\mathbf{x})) = \mathbf{x}$. Thus

$$\mathcal{A}^{-1}\mathcal{A} = \mathcal{I}$$

is the identity transformation which maps each vector in its domain into itself.

The *null space* $\mathcal{N}(\mathcal{A})$ of the transformation $\mathcal{A}$ is the set of all vectors $\mathbf{x} \in \mathcal{X}$, which are mapped into the zero vector in $\mathcal{Y}$:

$$\mathcal{N}(\mathcal{A}) \triangleq \{\mathbf{x} \in \mathcal{X} | \mathcal{A}(\mathbf{x}) = \mathbf{0}\}$$

Linear Transformations

A transformation $\mathcal{A} : \mathcal{X} \rightarrow \mathcal{Y}$ is said to be *linear* if the following two conditions are satisfied:

1. For any $\mathbf{x}_1$ and $\mathbf{x}_2 \in \mathcal{X}$, $\mathcal{A}(\mathbf{x}_1 + \mathbf{x}_2) = \mathcal{A}(\mathbf{x}_1) + \mathcal{A}(\mathbf{x}_2)$.
2. For any $\mathbf{x} \in \mathcal{X}$ and any scalar $\alpha \in \mathcal{F}$, $\mathcal{A}(\alpha\mathbf{x}) = \alpha\mathcal{A}(\mathbf{x})$.

Although nonlinear functions or transformations arise in connection with nonlinear control systems, major emphasis in this book is on linear systems or linear approximations to nonlinear ones. For this reason the rest of this chapter is devoted to linear transformations.

By far the most useful linear transformation for the purposes of this book is one whose domain and codomain are finite dimensional vector spaces. Every linear transformation of this type can be represented as a matrix, once suitable bases are selected. This is seen as follows.

Consider the linear transformation $\mathcal{A}: \mathcal{X}^n \to \mathcal{X}^m$ such that for $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{X}^m$, $\mathbf{y} = \mathcal{A}(\mathbf{x})$. Let $\{\mathbf{v}_i, i = 1, \ldots, n\}$ and $\{\mathbf{u}_i, i = 1, \ldots, m\}$ be basis sets for $\mathcal{X}^n$ and $\mathcal{X}^m$, respectively. Then $\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$ and the linearity properties of $\mathcal{A}$ give

$$\mathbf{y} = \sum_{i=1}^{n} \alpha_i \mathcal{A}(\mathbf{v}_i) = [\mathcal{A}(\mathbf{v}_1) \mid \mathcal{A}(\mathbf{v}_2) \mid \cdots \mid \mathcal{A}(\mathbf{v}_n)] \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \tag{5.3}$$

The vectors $\mathcal{A}(\mathbf{v}_i)$ are images of the basis vectors $\mathbf{v}_i$ under the transformation $\mathcal{A}$. Since $\mathbf{y}$ and each $\mathcal{A}(\mathbf{v}_i)$ belong to $\mathcal{X}^m$, they have unique expansions with respect to the basis set $\{\mathbf{u}_i\}$,

$$\mathbf{y} = \sum_{j=1}^{m} \beta_j \mathbf{u}_j \quad \text{and} \quad \mathcal{A}(\mathbf{v}_i) = \sum_{j=1}^{m} a_{ji} \mathbf{u}_j \tag{5.4}$$

Combining equations (5.3) and (5.4) gives

$$\mathbf{y} = \sum_{j=1}^{m} \beta_j \mathbf{u}_j = \sum_{i=1}^{n} \alpha_i \left[\sum_{j=1}^{m} a_{ji} \mathbf{u}_j \right]$$

Interchanging the order of summation and using the fact that the expansion coefficients β_j are unique lead to

$$\beta_j = \sum_{i=1}^{n} a_{ji} \alpha_i, \quad j = 1, 2, \ldots, m \tag{5.5}$$

Let $[\mathbf{x}]_{\mathbf{v}} \triangleq [\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_n]^T$ and $[\mathbf{y}]_{\mathbf{u}} \triangleq [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_m]^T$ be the coordinate representations of the vectors $\mathbf{x}$ and $\mathbf{y}$ with respect to the basis sets $\{\mathbf{v}_i\}$ and $\{\mathbf{u}_i\}$, respectively. (When the natural cartesian basis of Example 5.1 is used, this cumbersome notation is not necessary since then $\mathbf{x} = [x_1 \quad x_2 \quad \cdots \quad x_n]^T$ and $\mathbf{y} = [y_1 \quad y_2 \quad \cdots \quad y_m]^T$.) Regardless of which basis sets are selected, the transformation $\mathbf{y} = \mathcal{A}(\mathbf{x})$, or equivalently the set of Eqs. (5.5), can be represented by the matrix equation $[\mathbf{y}]_{\mathbf{u}} = \mathbf{A}[\mathbf{x}]_{\mathbf{v}}$. The matrix $\mathbf{A}$ is $m \times n$, and a typical element a_{ji} is seen to be the jth component (with respect to the basis $\{\mathbf{u}_i\}$) of the image of $\mathbf{v}_i$. The particular matrix representation $\mathbf{A}$ for $\mathcal{A}$ obviously depends on the choice of basis in both $\mathcal{X}^n$ and $\mathcal{X}^m$. Changing either basis set changes the resultant representation $\mathbf{A}$. However, many properties of $\mathcal{A}$ are independent of the particular representation $\mathbf{A}$. For example, the rank of $\mathcal{A}$ equals the rank of $\mathbf{A}$ regardless of which representation $\mathbf{A}$ is used. This is also the dimension of the range space of $\mathcal{A}$:

$$\text{rank}\,(\mathcal{A}) = r_A = \dim\,(\mathcal{R}(\mathcal{A}))$$

The range of $\mathcal{A}$ is frequently referred to as the column space of **A.**

Change of Basis

Consider the n-dimensional linear vector space $\mathcal{X}^n$. Let $\{\mathbf{v}_i, i = 1, \ldots, n\}$ and $\{\mathbf{v}_i', i = 1, \ldots, n\}$ be two basis sets. Each vector $\mathbf{x} \in \mathcal{X}^n$ can be expressed with respect to either basis; for example,

$$\mathbf{x} = \sum_{j=1}^{n} x_j \mathbf{v}_j = \sum_{i=1}^{n} x_i' \mathbf{v}_i'$$

where x_j and x_i' are scalar components. Since the basis vectors themsevles belong to $\mathcal{H}^n$, one set ca be expressed in terms of the other. For example,

$$\mathbf{v}_j = \sum_{i=1}^{n} b_{ij} \mathbf{v}_i'$$

Using this result to eliminate $\mathbf{v}_j$ in the expression for $\mathbf{x}$ gives

$$\sum_{j=1}^{n} x_j \sum_{i=1}^{n} b_{ij} \mathbf{v}_i' = \sum_{i=1}^{n} x_i' \mathbf{v}_i' \quad \text{or} \quad \sum_{i=1}^{n} \left(\sum_{j=1}^{n} b_{ij} x_j - x_i' \right) \mathbf{v}_i' = \mathbf{0}$$

Linear independence of the set $\{\mathbf{v}_i'\}$ requires that

$$\sum_{j=1}^{n} b_{ij} x_j = x_i'$$

The component vectors $[\mathbf{x}]_v$ and $[\mathbf{x}]_{v'}$ are thus related by a matrix multiplication:

$$[\mathbf{x}]_{v'} = [\mathbf{B}][\mathbf{x}]_v$$

A change of basis is seen to be equivalent to a matrix multiplication. The effect of a change of basis on the representation of a linear transformation is now considered. Let $\mathcal{A}$ map vectors in $\mathcal{X}^n$ into other vectors also in $\mathcal{X}^n$: $\mathcal{A} : \mathcal{X}^n \rightarrow \mathcal{X}^n$, where $\mathcal{A}(\mathbf{x}) = \mathbf{y}$. Let **A** be the representation of $\mathcal{A}$ when the basis $\{\mathbf{v}_i\}$ is used, and let $\mathbf{A}'$ be the representation when $\{\mathbf{v}_i'\}$ is used. The relation between **A** and $\mathbf{A}'$ is to be found. When using the unprimed basis set, the transformation is represented as

$$\mathbf{A}[\mathbf{x}]_v = [\mathbf{y}]_v$$

When using the primed basis set,

$$\mathbf{A}'[\mathbf{x}]_{v'} = [\mathbf{y}]_{v'}$$

But it was shown earlier that coordinate representations of any vector with respect to two sets of basis vectors are related by

$$[\mathbf{x}]_{v'} = [\mathbf{B}][\mathbf{x}]_v \qquad [\mathbf{y}]_{v'} = [\mathbf{B}][\mathbf{y}]_v$$

Thus

$$\mathbf{A}'[\mathbf{B}][\mathbf{x}]_v = [\mathbf{B}][\mathbf{y}]_v$$

The matrix **B** which represents the change of basis always has an inverse (see Problem 5.27), so

$$[\mathbf{B}^{-1}]\mathbf{A}'[\mathbf{B}][\mathbf{x}]_v = [\mathbf{y}]_v$$

This is the representation of the transformation in the unprimed system. The two representations for $\mathscr{A}$ are related by

$$\mathbf{B}^{-1}\mathbf{A}'\mathbf{B} = \mathbf{A}$$

This relationship between $\mathbf{A}'$ and $\mathbf{A}$ is called a *similarity transformation.* Any two matrices which are related by a similarity transformation are said to be *similar matrices.* In the present context similar matrices are representations of a linear transformation with respect to different basis vectors.

If both basis sets $\{\mathbf{v}_i\}$ and $\{\mathbf{v}_i'\}$ are orthonormal, then it can be shown (see Problem 5.28) that the matrix **B** is an orthogonal matrix. That is,

$$\mathbf{B}^{-1} = \mathbf{B}^T$$

In this case the two representations of $\mathscr{A}$ are related by an *orthogonal transformation,*

$$\mathbf{B}^T\mathbf{A}'\mathbf{B} = \mathbf{A}$$

Operations with Linear Transformations

Every linear transformation on finite dimensional spaces can be represented as a matrix. It is natural to expect that algebraic operations with linear transformations are governed by rules much like those of matrix algebra. Let $\mathscr{X}^n$, $\mathscr{X}^m$, and $\mathscr{X}^p$ be linear vector spaces defined over the same scalar number field $\mathscr{F}$. Then if

$$\mathscr{A}_1 : \mathscr{X}^n \to \mathscr{X}^m, \qquad \mathscr{A}_2 : \mathscr{X}^n \to \mathscr{X}^m \quad \text{and} \quad \mathscr{A}_1(\mathbf{x}) = \mathbf{y}_1, \qquad \mathscr{A}_2(\mathbf{x}) = \mathbf{y}_2$$

then

$$(\mathscr{A}_1 + \mathscr{A}_2)(\mathbf{x}) = \mathscr{A}_1(\mathbf{x}) + \mathscr{A}_2(\mathbf{x}) = \mathbf{y}_1 + \mathbf{y}_2$$

If $\mathscr{A}_1 : \mathscr{X}^n \to \mathscr{X}^m$ and $\mathscr{A}_2 : \mathscr{X}^m \to \mathscr{X}^p$, then $\mathscr{A}_2\mathscr{A}_1 : \mathscr{X}^n \to \mathscr{X}^p$ and, in general, $\mathscr{A}_1\mathscr{A}_2$ is not defined. Thus linear transformations are distributive but not commutative.

A norm can be defined for linear transformations, as follows. If

$$\mathscr{A}(\mathbf{x}) = \mathbf{y}$$

then $\|\mathbf{y}\| = \|\mathscr{A}(\mathbf{x})\|$. If there exists a finite number K such that $\|\mathscr{A}(\mathbf{x})\| \leq K\|\mathbf{x}\|$ for all $\mathbf{x}$, the linear transformation is said to be *bounded.* (Every linear transformation on finite dimensional spaces is bounded). Assuming that $\mathscr{A}$ is bounded, the norm of $\mathscr{A}$, written $\|\mathscr{A}\|$, is the smallest value of K which provides such a bound. Alternatively, $\|\mathscr{A}\|$ is the least upper bound (supremum or sup) of $\|\mathscr{A}(\mathbf{x})\|/\|\mathbf{x}\|$ for nonzero $\mathbf{x}$. Two equivalent formulas are

$$\|\mathscr{A}\| = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathscr{A}(\mathbf{x})\|}{\|\mathbf{x}\|} \quad \text{or} \quad \|\mathscr{A}\| = \sup_{\|\mathbf{x}\| = 1} \|\mathscr{A}(\mathbf{x})\| \tag{5.6}$$

There are various possible choices for the norm of the vector $\mathcal{A}(\mathbf{x})$. The most familiar is the Euclidean norm of Sec. 5.6, but see also Problems 5.33 and 5.34. Each particular choice induces a different form for $\|\mathcal{A}\|$. If $\mathbf{A}$ is the matrix representation for a finite dimensional transformation, $\mathcal{A}$, and if the quadratic vector norm is used, then

$$\|\mathcal{A}\|^2 = \max_{\|\mathbf{x}\| = 1} \{\bar{\mathbf{x}}^T \bar{\mathbf{A}}^T \mathbf{A}\mathbf{x}\}$$

Some properties satisfied by the norm of a linear transformation are

$$\|\mathcal{A}(\mathbf{x})\| \leq \|\mathcal{A}\| \cdot \|\mathbf{x}\| \quad \text{for all } \mathbf{x}$$

$$\|\mathcal{A}_1 + \mathcal{A}_2\| \leq \|\mathcal{A}_1\| + \|\mathcal{A}_2\|$$

$$\|\mathcal{A}_1 \mathcal{A}_2\| \leq \|\mathcal{A}_1\| \cdot \|\mathcal{A}_2\|$$

$$\|\alpha \mathcal{A}\| = |\alpha| \cdot \|\mathcal{A}\|$$

As defined earlier, the norm of every linear transformation is a nonnegative number, and is zero only for a null transformation, i.e., a transformation which maps every vector into the zero vector.

A particular class of linear transformations is that which maps vectors into the one-dimensional vector space formed by the scalar number field. These transformations are called *linear functionals* [3].

5.12 ADJOINT TRANSFORMATIONS

Let $\mathcal{A} : \mathcal{X}_1 \rightarrow \mathcal{X}_2$ be a linear transformation, where $\mathcal{X}_1$ and $\mathcal{X}_2$ are inner product spaces, with inner products $\langle , \rangle_1$ and $\langle , \rangle_2$, respectively. For each $\mathbf{x} \in \mathcal{X}_1$, $\mathcal{A}(\mathbf{x}) = \mathbf{y} \in \mathcal{X}_2$. If $\mathbf{z}$ is an arbitrary vector in $\mathcal{X}_2$, then the inner product $\langle \mathbf{z}, \mathbf{y} \rangle_2 = \langle \mathbf{z}, \mathcal{A}(\mathbf{x}) \rangle_2$ is well defined and can be used to define the *adjoint transformation* $\mathcal{A}^* : \mathcal{X}_2 \rightarrow \mathcal{X}_1$, according to $\langle \mathbf{z}, \mathcal{A}(\mathbf{x}) \rangle_2 = \langle \mathcal{A}^*(\mathbf{z}), \mathbf{x} \rangle_1$. It can be shown for finite dimensional spaces that $\mathcal{A}^*$ is also a linear transformation, i.e., if $\mathcal{A}^*(\mathbf{z}_1) = \mathbf{w}_1$ and $\mathcal{A}^*(\mathbf{z}_2) = \mathbf{w}_2$, then $\mathcal{A}^*(\alpha_1 \mathbf{z}_1 + \alpha_2 \mathbf{z}_2) = \alpha_1 \mathbf{w}_1 + \alpha_2 \mathbf{w}_2$ for arbitrary scalars α_1 and $\alpha_2 \in \mathcal{F}$. This follows from the linearity properties of the inner product.

EXAMPLE 5.13 Let $\mathcal{A}$ be a transformation from an n-dimensional vector space to an m-dimensional space, with the usual definition of the complex inner products,

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle \triangleq \bar{\mathbf{x}}_1^T \mathbf{x}_2, \qquad \langle \mathbf{y}_1, \mathbf{y}_2 \rangle \triangleq \bar{\mathbf{y}}_1^T \mathbf{y}_2$$

The operator $\mathcal{A}$ can be represented by an $m \times n$ matrix $\mathbf{A}$, so that

$$\langle \mathbf{z}, \mathcal{A}(\mathbf{x}) \rangle = \bar{\mathbf{z}}^T (\mathbf{A}\mathbf{x}) = \overline{(\bar{\mathbf{A}}^T \mathbf{z})}^T \mathbf{x} = \langle \bar{\mathbf{A}}^T \mathbf{z}, \mathbf{x} \rangle$$

For this example $\mathcal{A}^*$ is represented by the matrix $\bar{\mathbf{A}}^T$. ■

The adjoint transformation defined by the inner product should not be confused with the adjoint matrix defined and used in Chapter 4. Adjoint transformations appear in several roles in modern control theory, and some of these will be developed later. Only a few properties of adjoint transformations are presented here.

If $\mathcal{A}:\mathcal{X}_1\rightarrow\mathcal{X}_2$, then $\mathcal{A}^*:\mathcal{X}_2\rightarrow\mathcal{X}_1$. Also, $\mathcal{A}^*\mathcal{A}:\mathcal{X}_1\rightarrow\mathcal{X}_1$ and $\mathcal{A}\mathcal{A}^*:\mathcal{X}_2\rightarrow\mathcal{X}_2$. If $\mathcal{X}_1=\mathcal{X}_2$ and $\mathcal{A}=\mathcal{A}^*$, then $\mathcal{A}$ is said to be *self-adjoint.* In all cases, it can be shown that $\|\mathcal{A}\|=\|\mathcal{A}^*\|$, and that $(\mathcal{A}^*)^*=\mathcal{A}$. It is clear that $\mathcal{A}^*\mathcal{A}$ is generally not equal to $\mathcal{A}\mathcal{A}^*$. Those particular transformations for which $\mathcal{A}^*\mathcal{A}=\mathcal{A}\mathcal{A}^*$ are said to be *normal transformations.*

Let $\mathcal{A}:\mathcal{X}_1\rightarrow\mathcal{X}_2$ be an arbitrary linear transformation. Then the linear vector spaces $\mathcal{X}_1$ and $\mathcal{X}_2$ can be written as direct sums

$$\begin{aligned}\mathcal{X}_1&=\mathcal{N}(\mathcal{A})\oplus\overline{\mathcal{R}}(\mathcal{A}^*)\\ \mathcal{X}_2&=\mathcal{N}(\mathcal{A}^*)\oplus\overline{\mathcal{R}}(\mathcal{A})\end{aligned}\tag{5.7}$$

where $\mathcal{N}(\cdot)$ and $\mathcal{R}(\cdot)$ are the null space and range of the indicated tranformations. $\overline{\mathcal{R}}(\cdot)$ denotes the closure of the range $\mathcal{R}(\cdot)$, that is, $\mathcal{R}(\cdot)$ plus the limit of all convergent sequences of elements in $\mathcal{R}(\cdot)$. In finite dimensional spaces every subspace is closed, so that $\overline{\mathcal{R}}(\cdot)=\mathcal{R}(\cdot)$. Equation (5.7) constitutes an *orthogonal* decomposition of $\mathcal{X}_1$ into two linear subspaces. That is, for any vector $\mathbf{x}\in\mathcal{N}(\mathcal{A})$ and any vector $\mathbf{y}\in\mathcal{R}(\mathcal{A}^*)$, $\langle\mathbf{x},\mathbf{y}\rangle=0$. Equation (5.7) also provides an orthogonal decomposition for $\mathcal{X}_2$. Additional results for abstract transformations and their adjoints are found in the problems for this chapter. More concrete applications, where the operators are just matrices, are found in Sec. 5.13 and throughout the next chapter.

5.13 SOME FINITE-DIMENSIONAL TRANSFORMATIONS

Every linear transformation from one finite-dimensional space to another finite-dimensional space can be represented as a matrix. Within this general category, a few special transformations are now discussed.

Rotations

A particular transformation that frequently arises in control applications is a pure rotation. This can often be viewed in two ways. The result can be considered as a new vector obtained by rotating the original vector, or it can be considered as the same vector expressed in terms of a new coordinate system which is rotated with respect to the original coordinate system. The latter point of view is adopted for the time being, and the treatment is restricted to real, three-dimensional space, $\mathcal{R}^3$. Let $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}$ be an orthonormal basis set, and more specifically, let it define a right-handed cartesian coordinate system. Let $\{\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3\}$ be another right-handed cartesian coordinate system.

The set $\{\mathbf{x}_i\}$ might represent an orthogonal triad fixed to an aerospace vehicle or a tracking antenna. The set $\{\mathbf{y}_i\}$ might represent an inertially fixed coordinate system. These two sets can be brought into coincidence by a sequence of rotations. The most familiar set of angles of rotation are the Euler angles [4], although the present discussion applies to any sequence of finite rotations such as those of Figure 5.5. A rotation θ about the $\mathbf{x}_2$ axis rotates $\mathbf{x}_1$ and $\mathbf{x}_3$ into $\mathbf{x}_1'$ and $\mathbf{x}_3'$, and leaves $\mathbf{x}_2'=\mathbf{x}_2$. A rotation ψ about $\mathbf{x}_1'$ gives $\mathbf{x}_1''=\mathbf{x}_1'$ and $\mathbf{x}_2'',\mathbf{x}_3''$. The final rotation ϕ about $\mathbf{x}_3''$ gives $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3=\mathbf{x}_3''$.

Let $\mathbf{z}$ be an arbitrary vector and let $[\mathbf{z}]$, $[\mathbf{z}]'$, $[\mathbf{z}]''$, and $[\mathbf{z}]'''$ be its coordinate

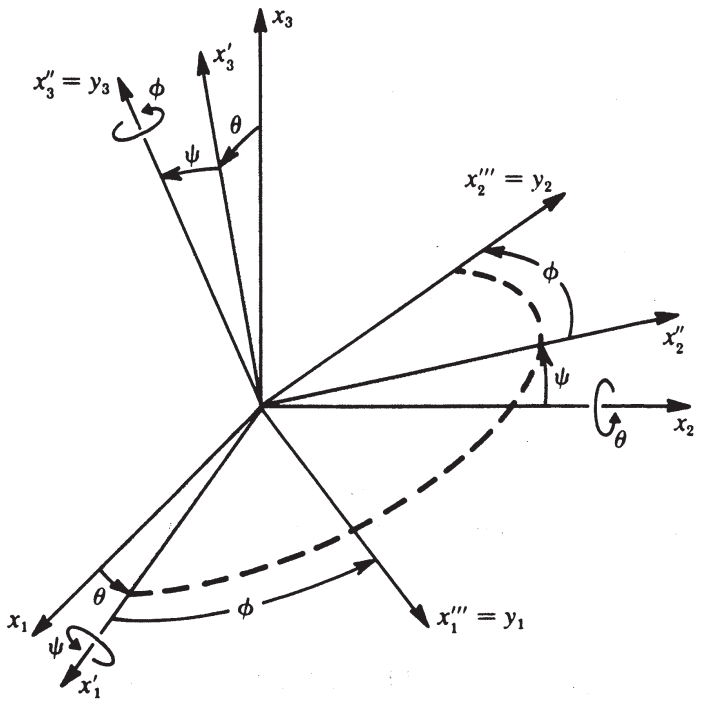


Figure 5.5

representations in the $\{\mathbf{x}_i\}$, $\{\mathbf{x}_i'\}$, $\{\mathbf{x}_i''\}$, and $\{\mathbf{y}_i\}$ coordinate systems, respectively. It is easily verified that

$$[\mathbf{z}]' = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}[\mathbf{z}], \qquad [\mathbf{z}]'' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & \sin\psi \\ 0 & -\sin\psi & \cos\psi \end{bmatrix}[\mathbf{z}]'$$

$$[\mathbf{z}]''' = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}[\mathbf{z}]'' \tag{5.8}$$

The overall transformation from $[\mathbf{z}]$ to $[\mathbf{z}]'''$ is given by the product of the three transformation matrices.

A symbolic method of representing coordinate rotations has been developed [5, 6]. These resolver-like diagrams, called Piograms, make it possible to write vector components in the new coordinate system without resorting to successive matrix multiplications (see Problem 5.37).

Reflections

The relation between a vector $\mathbf{x}$ and its reflection $\mathbf{x}_r$ at a plane surface defined by a unit normal $\mathbf{n}$ is $\mathbf{x}_r = \mathbf{x} - 2\langle \mathbf{n}, \mathbf{x}\rangle \mathbf{n}$. That is, $\mathbf{x}$ and $\mathbf{x}_r$ are equal except for a sign change in the component along $\mathbf{n}$. This can be rewritten as

$$\mathbf{x}_r = \mathbf{x} - 2\mathbf{n}\rangle\langle\mathbf{n}\mathbf{x} = [\mathbf{I} - 2\mathbf{n}\rangle\langle\mathbf{n}]\mathbf{x}$$

The matrix $\mathbf{A}_r = [\mathbf{I} - 2\mathbf{n}\rangle\langle\mathbf{n}]$ is the general representation of a reflection transformation. It is characterized by the fact that $|\mathbf{A}_r| = -1$, as verified by using the results of Problem 4.5.

Projections

A simple example of a transformation $\mathcal{A}(\mathbf{x})$ which maps $\mathbf{x}$ into its orthogonal projection on a hyperplane with a unit normal vector $\mathbf{n}$ is

$$\mathbf{A}_p = [\mathbf{I} - \mathbf{n}\rangle\langle\mathbf{n}]$$

This is fairly obvious since $\mathbf{A}_p\,\mathbf{x} = \mathbf{x} - \langle\mathbf{n}, \mathbf{x}\rangle\mathbf{n}$ has the effect of subtracting out the component of $\mathbf{x}$ along $\mathbf{n}$. It is easily verified that $\mathbf{A}_p\,\mathbf{A}_p = \mathbf{A}_p^2 = \mathbf{A}_p$. In general, any linear transformation which satisfies

$$\mathcal{A}^2 = \mathcal{A}$$

is a projection, although it need not be an orthogonal projection as in the above case. It is always possible to express a linear vector space as a direct sum $\mathcal{X} = \mathcal{U} \oplus \mathcal{V}$, where $\mathcal{U}$ and $\mathcal{V}$ are nonvoid subspaces of $\mathcal{X}$. This means that for each $\mathbf{x} \in \mathcal{X}$ there is one and only one way of writing

$$\mathbf{x} = \mathbf{u} + \mathbf{v}, \quad \text{where } \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}$$

A transformation $\mathcal{P}$ satisfying $\mathcal{P}(\mathbf{x}) = \mathbf{u}$ is said to be the projection on $\mathcal{U}$ along $\mathcal{V}$.

A Practical Application. Many control problems involve coordinate rotations. Some involve projections of vector quantities onto a sensor and others involve reflections. A typical kind of pointing and tracking example from geometrical optics is now given to demonstrate all three.

EXAMPLE 5.14 Suppose that an earth resource satellite consists of a steerable plane mirror and an imaging focal plane. The image of a right angle formed by the square corner of a Nebraska cornfield is to be captured on the focal plane. This image will be skewed or distorted—that is, the edges of the field will no longer appear orthogonal in general. Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be unit vectors at the corner of the field, and let $\mathbf{v}_1'$ and $\mathbf{v}_2'$ be their images on the focal plane. Find expressions for these images and then evaluate their inner product to show nonorthogonality.

There are four coordinate systems involved in this problem, the ground-fixed system $\{x, y, z\}$, the satellite coordinate system, the mirror coordinates $\{x_m, y_m, z_m\}$, and the focal plane coordinates $\{x_f, y_f, z_f\}$. Figure 5.6 shows these and defines the satellite position with respect to the corner in terms of the azimuth angle ψ and zenith angle β and the slant range R. If $\mathbf{z}_m$ is the normal coordinate to the mirror, then the vector $\mathbf{z}_m$ can be written in terms of components in the $\{x, y, z\}$ system as

$$\mathbf{z}_m = \mathbf{T}_{GM}\begin{bmatrix}0\\0\\1\end{bmatrix} = \mathbf{T}_{GS}\,\mathbf{T}_{SM}\begin{bmatrix}0\\0\\1\end{bmatrix}$$

where $\mathbf{T}_{GM}$, $\mathbf{T}_{GS}$, and $\mathbf{T}_{SM}$ are 3×3 rotation matrices that transform vectors from mirror-to-gound, satellite-to-ground, and mirror-to-satellite, respectively. Note that $\mathbf{v}_1$ and $\mathbf{v}_2$ are assumed aligned with the ground x and y axes, respectively. The apparent reflections of $\mathbf{v}_1$ and $\mathbf{v}_2$ are given by

$$\mathbf{v}_1'' = [\mathbf{I} - 2\mathbf{z}_m\,\mathbf{z}_m^T]\mathbf{v}_1 = \mathbf{A}_r\,\mathbf{v}_1$$

$$\mathbf{v}_2'' = \mathbf{A}_r\,\mathbf{v}_2$$

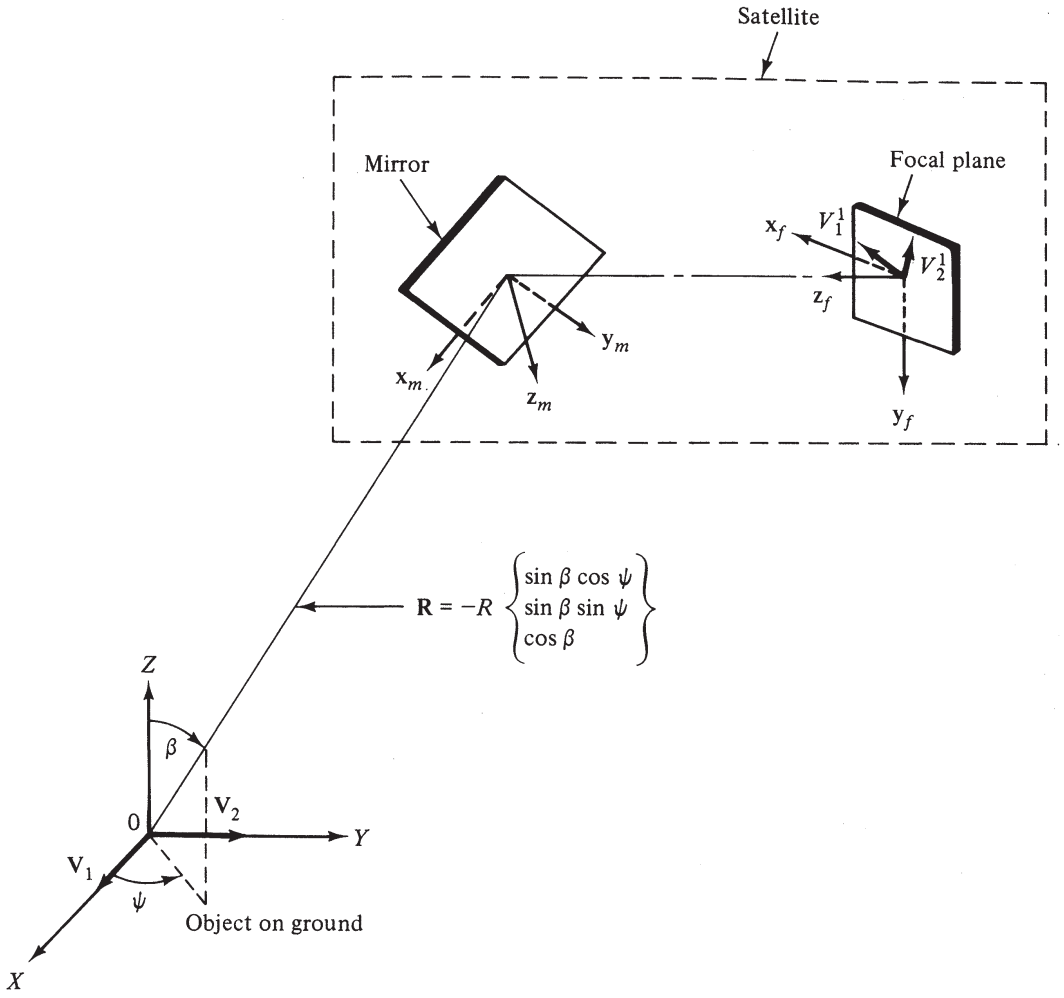


Figure 5.6

where $\mathbf{A}_r$ is the reflection matrix for the mirror. The reflected images are still expressed in ground coordinates. Let the normal to the focal plane be the vector $\mathbf{z}_f$ and assume that the x_f and y_f directions in the focal plane are suitably defined. When expressed in the focal plane coordinates, the reflected images of the two vectors are

$$\mathbf{v}_1''' = \mathbf{T}_{FG}\,\mathbf{A}_r\,\mathbf{v}_1 \quad \text{and} \quad \mathbf{v}_2 = \mathbf{T}_{FG}\,\mathbf{A}_r\,\mathbf{v}_2$$

where $\mathbf{T}_{FS}$ is the transformation from satellite to focal plane coordinates and where $\mathbf{T}_{FG} = \mathbf{T}_{FS}\,\mathbf{T}_{SG} = \mathbf{T}_{FS}\,\mathbf{T}_{GS}^T$. Note that because all the coordinate frames are orthogonal, the transformation matrices are orthogonal, so $\mathbf{T}_{SG} = \mathbf{T}_{GS}^{-1} = \mathbf{T}_{GS}^T$. The triply primed vectors are still three-dimensional. The focal plane images are the projection of these onto the focal plane:

$$\mathbf{v}_1' = [\mathbf{I} - \mathbf{n}\mathbf{n}^T]\mathbf{v}_1''' = A_p\,\mathbf{v}_1''' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v}_1'''$$

and $\mathbf{v}_2' = \mathbf{A}_p\,\mathbf{v}_2'''$, where $\mathbf{n} = [0 \quad 0 \quad 1]^T$. The projection matrix has been named $\mathbf{A}_p$. In general there would be a lens system between the mirror and the focal plane. It merely scales the vectors without changing their directions, so this complexity is neglected here. The true physical angle θ

on the ground is assumed to be 90° here, but in general it is given by $\cos(\theta) = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle$. The apparent angle on the focal plane is found from $\cos(\theta_f) = \langle \mathbf{v}_1', \mathbf{v}_2' \rangle / \|\mathbf{v}_1'\| \|\mathbf{v}_2'\|$. ■

EXAMPLE 5.15 In order to use the relations of the previous example the various transformation matrices must be known. For simplicity the satellite coordinates are assumed aligned with the ground-fix coordinates so that $\mathbf{T}_{GS} = \mathbf{I}$. The optical focal plane is assumed fixed to the vehicle with an orientation which gives

$$\mathbf{T}_{FS} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

That leaves only the mirror orientation to be specified. However, it cannot be arbitrarily specified. The mirror must be steered so that the correct scene is reflected and projected upon the focal plane. Treat this as an open-loop control problem and determine $\mathbf{T}_{SM}$ so that the satellite-to-ground vector $\mathbf{R}$ is projected onto the focal plane origin.

Note that $\mathbf{T}_{SM}$ is needed to find the vector $\mathbf{z}_m$, which in turn is used to calculate $\mathbf{A}_r$. The desired $\mathbf{A}_r$ matrix will now be found directly. The $\mathbf{R}$ vector, after reflection, must be entirely along the normal to the focal plane, so

$$[0 \quad 0 \quad R]^T = \mathbf{T}_{FG}\, \mathbf{A}_r\, \mathbf{R} \tag{5.9}$$

is required. In order to determine the unknown mirror orientation matrix $\mathbf{A}_r$, two more independent equations are needed. One can be obtained by specifying how the x_f, y_f axes are rotated about the focal plane normal. One way of doing this is to force the unit vector $\mathbf{u}$, which is normal to both $\mathbf{R}$ and the ground x axis, to project along the $-y_f$ axis. $\mathbf{u} = \mathbf{R} \times \mathbf{x} / \|\mathbf{R} \times \mathbf{x}\|$ and then

$$[0 \quad -1 \quad 0]^T = \mathbf{T}_{FG}\, \mathbf{A}_r\, \mathbf{u} \tag{5.10}$$

A third independent equation is available from the cross product of Eq. (5.9) and (5.10):

$$[0 \quad 0 \quad R]^T \times [0 \quad -1 \quad 0]^T = \mathbf{T}_{FG}\, \mathbf{A}_r (\mathbf{R} \times \mathbf{u}) \tag{5.11}$$

Since $\mathbf{T}_{GS} = \mathbf{I}$, $\mathbf{T}_{FS} = \mathbf{T}_{FG}$. Thus Eqs. (5.9), (5.10) and (5.11) can be combined into one matrix equation and solved to give

$$\mathbf{A}_r = \begin{bmatrix} 0 & 0 & R \\ -R & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} [\mathbf{R} \mid \mathbf{u} \mid \mathbf{R} \times \mathbf{u}]^{-1}$$

Using the definition of $\mathbf{A}_r$, the orientation of the mirror normal vector can be found. The required gimbal angles for the mirror can then be calculated and used as the open-loop commands to the two axes of the mirror-drive servos. Closed-loop error-nulling controllers would normally be used in an actual system. The purpose here was to demonstrate that rotations, reflections, and projections are useful in real control problems. ■

EXAMPLE 5.16 For the system of the previous two examples, suppose the satellite is located with respect to the desired corner at $\beta = 30°$, $\psi = 45°$, and a slant range of 100 nautical miles. Compute the skew in the 90° corner.

With these values,

$$\mathbf{A}_r = \begin{bmatrix} -0.9524 & 0.13606 & 0.33328 \\ 0.30855 & 0.35998 & 0.88177 \\ 0.11783 & -0.94265 & 0.33673 \end{bmatrix} \quad \text{(rounded)}$$

The focal plane images are found to be $\mathbf{v}_1' = [0.95242 \quad -0.11783]^T$ and $\mathbf{v}_2' = [-0.13606 \quad 0.94265]^T$, so that the inner product gives $\theta_f = 105.266°$. The skew (distortion from the true

angle) is 15.266°. As the angle β approaches zero (direct overhead viewing) the skew approaches zero for all ψ. The skew also decreases to zero if the corner is viewed from above either the x axis or the y axis, i.e., for ψ either 0° or 90°. Table 5.1 gives results for a few representative combinations.

TABLE 5.1

ψ	β	θ_f
45	30	105.266
45	20	96.903
45	10	91.741
45	0	90
0	30	90

■

5.14 SOME TRANSFORMATIONS ON INFINITE DIMENSIONAL SPACES

Most of the analysis of lumped-parameter systems in modern control theory can be considered in terms of a finite dimensional linear space, the state space. Consequently, the major emphasis is on finite dimensional transformations. However, transformations on infinite dimensional spaces do arise, and two of the more important ones are mentioned here.

It is recalled that the dimension of a space is equal to the number of elements in its basis set. The set of all periodic functions with period π is an example of an infinite dimensional space and its basis could be selected as the functions $\{\sin nt, n = 0, 1, \ldots\}$. The expansion with respect to this basis is the Fourier series. The set of all continuous functions, or of integrable functions, or of all square integrable functions are other examples of infinite dimensional spaces. A space is not necessarily infinite dimensional just because its elements are functions of time. For example, the space of all polynomials of degree 3 or less—i.e., $\{f(t) | f(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3, \alpha_i \in \mathcal{F}\}$—is a four-dimensional space.

Integral Relations

The integral form of a system's input-output equation was mentioned in Sec. 1.5 and a particular example is used in Problem 5.36. For a linear system the input and output can be related by

$$\mathbf{y}(T) = \int_{-\infty}^{T} \mathbf{W}(T, \tau)\mathbf{u}(\tau)\, d\tau$$

where $\mathbf{y}(T)$ is the $m \times 1$ output vector at time T, $\mathbf{u}(t)$ is an $r \times 1$ input vector for each value of t, and $\mathbf{W}(t, \tau)$ is the $m \times r$ weighting matrix. At each time T, $\mathbf{y}(T)$ is a vector in an m-dimensional space. A particular input function, $u(t), t \in (-\infty, T]$ can be considered as an element of the (infinite dimensional) input function space $\mathcal{U}$. The input-output integral represents a transformation $\mathcal{A} : \mathcal{U} \rightarrow \mathcal{Y}^m$, where $\mathcal{A}(u) = \mathbf{y}(T)$.

The equation for the Laplace transform

$$\mathbf{y}(s) = \int_0^{\infty} e^{-st} y(t)\, dt$$

provides another example of a linear transformation on infinite dimensional spaces. The domain of these transformations must be suitably restricted so that the indicated operations "make sense." In other words, nonintegrable functions cannot be integrated, and functions which cannot be bounded by some exponential function do not have Laplace transforms.

Differential Relations

A linear differential equation can be considered to be a transformation, but again infinite dimensional spaces (function spaces) are involved. The simplest case

$$\frac{dx}{dt} = u$$

maps a function $x(t)$ into another function $u(t)$. Of course, the domain of this transformation must be restricted to the class of functions which are differentiable. Other restrictions may be necessary as well. Perhaps only those functions for which $x(0) = 0$ are considered. This constitutes an initial condition. The relation

$$\left[\mathbf{I}\frac{d}{dt} - \mathbf{A}\right]\mathbf{x}(t) = \mathbf{B}\mathbf{u}(t)$$

is another example of a differential transformation which maps $\mathbf{x}(t)$ into $\mathbf{Bu}(t)$. Although this equation appears repeatedly in the modern formulation of control problems, it will not be necessary to consider it as an abstract mapping on function spaces. Rather, this brief section dealing with transformations on function spaces is intended only to hint at a direction for an abstract treatment of all linear transformations. If the function spaces are Hilbert spaces (i.e., complete inner product spaces), then the results parallel the finite dimensional results to a large degree [2]. In general, however, there will be some major differences. Every finite dimensional linear vector space is complete, and every transformation on finite dimensional spaces is bounded. These are not generally true in the infinite dimensional case. For example, the space of square integrable functions $\mathcal{L}_2[a, b]$ of Problem 5.22 is complete. But the space of all continuous functions $C[a, b)$ is not complete because a sequence of continuous functions may converge to a discontinuous function. Differential operators are examples of *unbounded* linear transformations. Some of the other differences arise because of the greater variety of definitions that can be given for norms and distance measures in function spaces. A complete treatment of these topics can be found in texts on functional analysis [1, 3, 7].

REFERENCES

1. Friedman, B.: *Principles and Techniques of Applied Mathematics,* John Wiley, New York, 1956.
2. Halmos, P. R.: *Finite Dimensional Vector Spaces,* 2d ed., D. Van Nostrand, Princeton, N.J., 1958.
3. Taylor, A. E.: *Functional Analysis,* John Wiley, New York, 1958.

4. Goldstein, H.: *Classical Mechanics,* Addison-Wesley, Reading, Mass., 1959.
5. Pio, R. L.: "Symbolic Representation of Coordinate Transformations," *IEEE Transactions on Aerospace and Navigational Electronics,* Vol. ANE-11, No. 2, June 1964, pp. 128–134.
6. Pio, R. L.: "Euler Angle Transformations," *IEEE Transactions on Automatic Control,* Vol. AC-11, No. 4, Oct. 1966, pp. 707–715.
7. Kolomogorov, A. N. and S. V. Fomin: *Elements of the Theory of Functional Analysis,* Vol. 1 (1957) and Vol. 2 (1961), Graylock Press, Albany, N.Y. (Translated from the Russian editions).
8. Strang, G.: *Linear Algebra and Its Applications,* Academic Press, New York, 1980.

ILLUSTRATIVE PROBLEMS

Vectors in Two and Three Dimensions

5.1 If two vectors $\mathbf{v}^T = [3 \quad -5 \quad 6]$ and $\mathbf{w}^T = [\alpha \quad 2 \quad 2]$ are known to be orthogonal, what is α?

Assuming that the given components are expressed with respect to a common coordinate system, orthogonality requires

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \mathbf{w} = 0 \quad \text{or} \quad 3\alpha - 10 + 12 = 0$$

so that $\alpha = -\frac{2}{3}$.

5.2 If $\mathbf{v}^T = [3 \quad -5 \quad 6]$ and $\mathbf{w}^T = [5 \quad 8]$, find $\langle \mathbf{v}, \mathbf{w} \rangle$ and the two outer products.

Since $\mathbf{v}$ and $\mathbf{w}$ have different numbers of components and hence belong to different dimensional spaces, their inner product is not defined. The outer products are, however,

$$\mathbf{v}\mathbf{w}^T = \begin{bmatrix} 3 \\ -5 \\ 6 \end{bmatrix} [5 \quad 8] = \begin{bmatrix} 15 & 24 \\ -25 & -40 \\ 30 & 48 \end{bmatrix} = (\mathbf{w}\mathbf{v}^T)^T$$

5.3 Consider the nonzero vector $\mathbf{v}$ with complex components $\mathbf{v} = \begin{bmatrix} j \\ 1 \end{bmatrix}$. Compute $\mathbf{v}^T\mathbf{v}$ and $\bar{\mathbf{v}}^T\mathbf{v}$.

Vector multiplication gives

$$\mathbf{v}^T\mathbf{v} = (j)(j) + 1 = 0$$

$$\bar{\mathbf{v}}^T\mathbf{v} = (-j)(j) + 1 = 2$$

Therefore, if the real form of the inner product $\langle \mathbf{v}, \mathbf{v} \rangle = \mathbf{v}^T\mathbf{v}$ is used to define length, nonzero vectors can have zero "length." When the complex form of the inner product $\langle \mathbf{v}, \mathbf{v} \rangle = \bar{\mathbf{v}}^T\mathbf{v}$ is used, this cannot happen.

5.4 Find the component of $\mathbf{v}^T = [2 \quad -3 \quad -4]$ in the direction of the vector $\mathbf{w}^T = [1 \quad 2 \quad 1]$.

First find the unit vector $\hat{\mathbf{w}}$ in the desired direction:

$$\|\mathbf{w}\| = \langle \mathbf{w}, \mathbf{w} \rangle^{1/2} = \sqrt{6}$$

$$\hat{\mathbf{w}} = \frac{1}{\sqrt{6}}\mathbf{w}$$

Then form the inner product:

$$\langle \mathbf{v}, \hat{\mathbf{w}} \rangle = \frac{1}{\sqrt{6}}(2 - 6 - 4) = \frac{-1}{\sqrt{6}}8$$

This result indicates that $\mathbf{v}$ has a component along the negative $\mathbf{w}$ direction and its magnitude is $8/\sqrt{6}$.

5.5 Find the cross products $\mathbf{v} \times \mathbf{w}$ and $\mathbf{w} \times \mathbf{v}$ if $\mathbf{v}^T = [v_1 \quad v_2 \quad v_3]$ and $\mathbf{w}^T = [w_1 \quad w_2 \quad w_3]$ are real vectors.

If we let $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ be unit vectors along the three mutually orthogonal coordinate axes, the cross product can be computed using the following determinant:

$$\mathbf{v} \times \mathbf{w} = \left\| \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{bmatrix} \right\|$$

Using Laplace expansion with respect to row one gives

$$\mathbf{v} \times \mathbf{w} = \mathbf{e}_1(v_2 w_3 - v_3 w_2) - \mathbf{e}_2(v_1 w_3 - v_3 w_1) + \mathbf{e}_3(v_1 w_2 - v_2 w_1)$$

Since $\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, $\mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, the column matrix representations are

$$\mathbf{v} \times \mathbf{w} = \begin{bmatrix} v_2 w_3 - v_3 w_2 \\ v_3 w_1 - v_1 w_3 \\ v_1 w_2 - v_2 w_1 \end{bmatrix}$$

$$\mathbf{w} \times \mathbf{v} = \begin{bmatrix} v_3 w_2 - v_2 w_3 \\ v_1 w_3 - v_3 w_1 \\ v_2 w_1 - v_1 w_2 \end{bmatrix}$$

5.6 Show that $\mathbf{v} \times \mathbf{w}$ of the previous problem can be written as the product of a skew-symmetric matrix and $\mathbf{w}$, or another skew-symmetric matrix and $\mathbf{v}$.

Direct matrix multiplication verifies that

$$\mathbf{v} \times \mathbf{w} = \begin{bmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 0 & w_3 & -w_2 \\ -w_3 & 0 & w_1 \\ w_2 & -w_1 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

5.7 From the results of the preceding problem, what can be said about the product $\mathbf{x}^T\mathbf{A}\mathbf{x}$ if $\mathbf{x}$ is a real three-component vector and $\mathbf{A}$ is skew-symmetric?

Any 3×3 skew-symmetric matrix could be used to define a 3×1 vector, and then $\mathbf{A}\mathbf{x}$ would represent a cross product. Therefore, $\mathbf{A}\mathbf{x}$ is a vector perpendicular to $\mathbf{x}$. Therefore, $\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle = \mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{0}$ for every real $\mathbf{x}$. In fact, if $\mathbf{A}$ is skew-symmetric of arbitrary dimension, the result $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{0}$ is true for any real, conformable vector $\mathbf{x}$.

Defining a Vector Space

5.8 Does the set of all vectors in the first and fourth quadrants of the plane form a vector space?

No. The first four conditions of Sec. 5.3, page 159, hold. However, if $\mathbf{x}$ is in the first or fourth quadrant, $-\mathbf{x}$ is in the second or third quadrant, so condition 5 is not satisfied and neither is 6 when negative scalars are considered.

5.9 Does the set of all three-dimensional vectors inside a sphere of finite radius constitute a linear vector space?

No. If $\mathbf{x}$ and $\mathbf{y}$ are inside the sphere, $\mathbf{x} + \mathbf{y}$ need not be. If a is sufficiently large, $a\mathbf{x}$ will also extend outside the sphere. Conditions 1 and 6 are not satisfied.

5.10 Consider all vectors defined by points in a plane passing through a three-dimensional space. Is this set a linear vector space?

This is a linear vector space if and only if the plane passes through the origin. If it does not, condition 4 is not satisfied.

5.11 What is the dimension of the space $\mathscr{X}$ defined as the set of all linear combinations of

$$\mathbf{x}_1^T = [1 \quad 2 \quad 3 \quad 4 \quad 5]$$
$$\mathbf{x}_2^T = [1 \quad 0 \quad 0 \quad 0 \quad 1]$$
$$\mathbf{x}_3^T = [0 \quad 1 \quad 1 \quad 0 \quad 0]$$

The manner in which $\mathscr{X}$ is defined guarantees that $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ span this space. Since the Grammian gives $|\mathbf{G}| = 98$, the set is linearly independent and thus constitutes a basis set. Since there are three vectors in the basis set, the dimension of $\mathscr{X}$, written dim ($\mathscr{X}$), is three even though every vector in $\mathscr{X}$ has five components.

Linear Dependence, Independence, and Degeneracy

5.12 Prove that the addition of the zero vector $\mathbf{0}$ to any set of linearly independent vectors yields a set of linearly dependent vectors.

Let $\mathscr{V} = \{\mathbf{x}_i, i = 1, n\}$ be a set of linearly independent vectors. This means that $a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n = \mathbf{0}$ requires $a_i = 0, i = 1, n$. Select $a_{n+1} \neq 0$. Then

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n + a_{n+1}\mathbf{0} = \mathbf{0}$$

so the set of vectors $\{\mathbf{0}, \mathbf{x}_i, i = 1, n\}$ is a linearly dependent set.

5.13 Use the Grammian to test the following vectors for linear dependence:

$$\mathbf{x}_1^T = [1 \quad 1 \quad 0 \quad 0], \qquad \mathbf{x}_2^T = [1 \quad 1 \quad 1 \quad 1], \qquad \mathbf{x}_3^T = [0 \quad 0 \quad 1 \quad 1]$$

The Grammian determinant is

$$|\mathbf{G}| = \left|\begin{bmatrix} 2 & 2 & 0 \\ 2 & 4 & 2 \\ 0 & 2 & 2 \end{bmatrix}\right| = 0$$

Hence the three vectors are linearly dependent. Note that the Grammian is symmetric for vectors with real components. In general, it is a Hermitian matrix.

5.14 Consider two real vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ expressed as 2×1 column vectors in terms of the natural coordinate vectors. Show that $|\mathbf{G}|$ is the square of the area of the parallelogram which has $\mathbf{x}_1$ and $\mathbf{x}_2$ as sides.

We have

$$|\mathbf{G}| = \begin{vmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_2 \\ \mathbf{x}_2^T\mathbf{x}_1 & \mathbf{x}_2^T\mathbf{x}_2 \end{vmatrix} = \begin{vmatrix} x_1^2 & x_1x_2\cos\theta \\ x_1x_2\cos\theta & x_2^2 \end{vmatrix}$$

where x_1 and x_2 are the magnitudes of $\mathbf{x}_1$ and $\mathbf{x}_2$ and θ is the included angle. The definition of the inner product of Sec. 5.2 has been used in arriving at this result. Expanding gives

$$|\mathbf{G}| = x_1^2x_2^2(1 - \cos^2\theta) = x_1^2x_2^2\sin^2\theta$$

Considering x_1 as the base of the parallelogram, the height is $x_2\sin\theta$, so the result is proven.

5.15 What are the rank and degeneracy of the following matrices?

(a) $\mathbf{A} = \begin{bmatrix} 6 & 2 & 4 \\ 2 & 0 & 2 \\ 1 & -1 & 2 \end{bmatrix}$ **(b)** $\mathbf{B} = \begin{bmatrix} 4 & 3 & 7 & 1 \\ 2 & 6 & 2 & 10 \\ 8 & 6 & 14 & 2 \\ 1 & 3 & 1 & 5 \end{bmatrix}$

(c) $\mathbf{C} = \begin{bmatrix} 6 & -4 & -4 & -9 \\ 24 & 3 & 0 & -9 \\ -14 & 3 & 4 & 12 \\ 48 & 25 & 16 & 9 \end{bmatrix}$

(a) $|\mathbf{A}| = 0$ so that $r_A < 3$. Picking the submatrix $\mathbf{A}_1 = \begin{bmatrix} 6 & 2 \\ 2 & 0 \end{bmatrix}$ gives $|\mathbf{A}_1| = -4 \neq 0$, so $r_A = 2$. The degeneracy is $q_A = n - r_A$ or $q_A = 1$. The one linear dependency relation between columns can be written as $\mathbf{x}_2 = \mathbf{x}_1 - \mathbf{x}_3$.

(b) $|\mathbf{B}| = 0$ (row 2 is twice row 4). Therefore $r_B < 4$. Any 3×3 matrix $\mathbf{B}_1$ formed by crossing out a row and column also has $|\mathbf{B}_1| = 0$ since row 3 is twice row 1. Therefore $r_B < 3$. It is easy to find a nonzero 2×2 determinant, so $r_B = 2$ and $q_B = 4 - 2 = 2$.

(c) $|\mathbf{C}| = 0$ as does the determinant of every 3×3 submatrix. In fact, there are just two linearly independent column vectors

$$\mathbf{x}_1 = [1 \quad 3 \quad -2 \quad 5]^T \quad \text{and} \quad \mathbf{x}_2 = [-1 \quad 0 \quad 1 \quad 4]^T$$

which can be used to generate $\mathbf{C}$. Then column 1 of $\mathbf{C}$ is $\mathbf{c}_1 = 8\mathbf{x}_1 + 2\mathbf{x}_2$. Likewise, $\mathbf{c}_2 = 1\mathbf{x}_1 + 5\mathbf{x}_2$, $\mathbf{c}_3 = 4\mathbf{x}_2$, and $\mathbf{c}_4 = -3\mathbf{x}_1 + 6\mathbf{x}_2$. In this case $r_C = 2$ and $q_C = 2$.

Gram-Schmidt Process

5.16 Use the Gram-Schmidt process to construct a set of orthonormal vectors from

$$\mathbf{x}_1 = \begin{bmatrix} 1+j \\ 1-j \\ j \end{bmatrix}, \qquad \mathbf{x}_2 = \begin{bmatrix} 2j \\ 1-2j \\ 1+2j \end{bmatrix}, \qquad \mathbf{x}_3 = \begin{bmatrix} 1 \\ j \\ 5j \end{bmatrix}$$

The Grammian is first used to verify linear independence of the set $\{\mathbf{x}_i\}$. The complex form of the inner product must be used:

$$\mathbf{G} = [\langle \mathbf{x}_i, \mathbf{x}_j \rangle] = [\bar{\mathbf{x}}_i^T \mathbf{x}_j] = \begin{bmatrix} 5 & 7 & 5 \\ 7 & 14 & 8+4j \\ 5 & 8-4j & 27 \end{bmatrix}$$

The determinant is $|\mathbf{G}| = 377 \neq 0$; therefore, the $\mathbf{x}_i$ are linearly independent.

Step 1. Construct an orthogonal set of $\mathbf{v}_i$:

$$\mathbf{v}_1 = \mathbf{x}_1$$

$$\mathbf{v}_2 = \begin{bmatrix} 2j \\ 1-2j \\ 1+2j \end{bmatrix} - \frac{7}{5}\begin{bmatrix} 1+j \\ 1-j \\ j \end{bmatrix} = \frac{1}{5}\begin{bmatrix} -7+3j \\ -2-3j \\ 5+3j \end{bmatrix}$$

Anticipating their need in advance, the products $\langle \mathbf{v}_2, \mathbf{v}_2 \rangle = 21/5$ and $\langle \mathbf{v}_2, \mathbf{x}_3 \rangle = 1 + 4j$ are computed.

$$\mathbf{v}_3 = \begin{bmatrix} 1 \\ j \\ 5j \end{bmatrix} - \frac{5}{5}\begin{bmatrix} 1+j \\ 1-j \\ j \end{bmatrix} - \frac{5(1+4j)}{21(5)}\begin{bmatrix} -7+3j \\ -2-3j \\ 5+3j \end{bmatrix} = \frac{1}{21}\begin{bmatrix} 19+4j \\ -31+53j \\ 7+61j \end{bmatrix}$$

Step 2. Normalize to obtain

$$\hat{\mathbf{v}}_1 = \frac{1}{\sqrt{5}}\begin{bmatrix} 1+j \\ 1-j \\ j \end{bmatrix}, \quad \hat{\mathbf{v}}_2 = \frac{1}{\sqrt{105}}\begin{bmatrix} -7+3j \\ -2-3j \\ 5+3j \end{bmatrix}$$

Using $\langle \mathbf{v}_3, \mathbf{v}_3 \rangle = \dfrac{7917}{441}$ gives

$$\hat{\mathbf{v}}_3 = \sqrt{\frac{441}{7917}}\,\mathbf{v}_3 = \frac{1}{\sqrt{7917}}\begin{bmatrix} 19+4j \\ -31+53j \\ 7+61j \end{bmatrix}$$

It is a good exercise in the use of the complex inner product to verify that these results satisfy $\langle \hat{\mathbf{v}}_i, \hat{\mathbf{v}}_j \rangle = \delta_{ij}$.

5.17 Given a set of independent, real vectors $\{\mathbf{y}_i\}$, show that the Gram-Schmidt process for generating the orthonormal set $\{\hat{\mathbf{v}}_i\}$ can be expressed as a recursive matrix calculation:

$$\mathbf{T}_1 = \mathbf{I} \quad \text{(the initial condition)}$$

$$\mathbf{v}_i = \mathbf{T}_i \mathbf{y}_i \quad \text{(removal of components along previously computed } \mathbf{v}_k\text{)}$$

$$\hat{\mathbf{v}}_i = \mathbf{v}_i / \|\mathbf{v}_i\| \quad \text{(normalization step)}$$

$$\mathbf{T}_{i+1} = \mathbf{T}_i - \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T \quad \text{(the recursion step)}$$

The Gram-Schmidt formulas from Sec. 5.7 can be written as

$$\begin{aligned}
\mathbf{v}_{i+1} &= \mathbf{y}_{i+1} - \sum_{k=1}^{i} \langle \hat{\mathbf{y}}_k, \mathbf{y}_{i+1} \rangle \hat{\mathbf{u}}_k \\
&= \mathbf{y}_{i+1} - \sum_{k=1}^{i} \hat{\mathbf{u}}_k \langle \hat{\mathbf{y}}_k, \mathbf{y}_{i+1} \rangle \\
&= \left[\mathbf{I} - \sum_{k=1}^{i} \hat{\mathbf{u}}_k \langle \hat{\mathbf{y}}_k \right] \mathbf{y}_{i+1} \\
&= \left[\underbrace{\mathbf{I} - \sum_{k=1}^{i} \hat{\mathbf{u}}_k \langle \hat{\mathbf{y}}_k}_{\mathbf{T}_i} - \hat{\mathbf{v}}_i \langle \hat{\mathbf{v}}_i \right] \mathbf{y}_{i+1} \\
&= [\mathbf{T}_i - \hat{\mathbf{v}}_i \langle \hat{\mathbf{v}}_i] \mathbf{y}_{i+1} \\
&= \mathbf{T}_{i+1} \mathbf{y}_{i+1}
\end{aligned}$$

This proves the validity of the recursion from step i to step $i + 1$. Since it is true for $i = 1$, this constitutes a proof by induction. Notice that each of the sequence of $\mathbf{T}_i$ operators is a projection operator, since $\mathbf{T}_i \mathbf{T}_i = \mathbf{T}_i$. This is also easily proven by induction. It is obviously true for $\mathbf{T}_1 = \mathbf{I}$. At a general step,

$$\begin{aligned}
\mathbf{T}_{i+1} \mathbf{T}_{i+1} &= [\mathbf{T}_i - \hat{\mathbf{v}}_i \langle \hat{\mathbf{v}}_i][\mathbf{T}_i - \hat{\mathbf{v}}_i \langle \hat{\mathbf{v}}_i] = \mathbf{T}_i \mathbf{T}_i - \hat{\mathbf{v}}_i \langle \hat{\mathbf{v}}_i \mathbf{T}_i - \hat{\mathbf{v}}_i \langle \hat{\mathbf{v}}_i \mathbf{T}_i - \hat{\mathbf{v}}_i \langle \mathbf{v}_i, \hat{\mathbf{v}}_i \rangle \langle \hat{\mathbf{v}}_i \\
&= \mathbf{T}_i \mathbf{T}_i - \hat{\mathbf{v}}_i \langle \hat{\mathbf{v}}_i = \mathbf{T}_i - \hat{\mathbf{v}}_i \langle \hat{\mathbf{v}}_i = \mathbf{T}_{i+1}
\end{aligned}$$

In this calculation the facts that $\langle \hat{\mathbf{v}}_i, \hat{\mathbf{v}}_i \rangle = 1$ and $\mathbf{T}_{i+1} \hat{\mathbf{v}}_i = 0$ were used.

5.18 Give a matrix version of the modified Gram-Schmidt process and contrast it with the results of the previous problem.

Assume that $\mathbf{y}_1$ is selected as $\mathbf{v}_1$, as before. Then the modified Gram-Schmidt process immediately subtracts the components along $\mathbf{v}_1$ from all other $\mathbf{y}_i$ vectors, for $i = 2, \ldots, n$. This can be accomplished using the projection operator $\mathbf{P}_1 = \mathbf{I} - \hat{\mathbf{v}}_1 \langle \hat{\mathbf{v}}_1$. That is, all the $\mathbf{y}_i$ vectors are replaced by $\mathbf{y}_i' = \mathbf{P}_1 \mathbf{y}_i$, for $i = 2, \ldots, n$. These are the projections on a subspace normal to $\hat{\mathbf{v}}_1$. Then $\mathbf{y}_2'$ is selected as $\mathbf{v}_2$ and normalized to $\hat{\mathbf{v}}_2$, and the whole process is repeated. If at any step a vector is found with $\|\mathbf{y}_i'\| = 0$ (in practice, less than epsilon), then the original $\mathbf{y}_i$ is a linear combination of the previously calculated $\{\hat{\mathbf{v}}_j, j = 1, \ldots, i - 1\}$. In that case, the ith vector is skipped and the process continues with the next $\mathbf{y}_{i+1}$ vector. The pseudocode for this calculation might look as follows:

```
Rank = 0
For i = 1 to m
  If ||y_i|| < ε increment i and test next vector
  If ||y_i|| ≥ ε then
      Rank = rank + 1
      If rank = n, quit. The entire set of n has been found.
```

$\hat{\mathbf{v}}_i = \mathbf{y}_i/\|\mathbf{y}_i\|$
$\mathbf{P} = \mathbf{I} - \hat{\mathbf{v}}_i\langle\hat{\mathbf{v}}_i$
For $j = i + 1$ to m
$\mathbf{y}_j \leftarrow \mathbf{P}\mathbf{y}_j$ (Replace $\mathbf{y}_j$ by its projection)
Increment i and repeat

Upon completion, the number of orthonormal vectors will equal the rank of the matrix **A**, whose columns are the vectors $\mathbf{y}_i$. Often it is desired to find a full set of n orthonormal vectors, even though $n > m$ or rank $(\mathbf{A}) < n$ for some other reason. This can be done by selecting a sufficient number of extra column vectors with components chosen randomly and appending these to the matrix **A**.

The difference between the modified and unmodified Gram-Schmidt processes is that here each vector $\mathbf{y}_i$ is modified to $\mathbf{y}_i'$ repeatedly, but each vector $\hat{\mathbf{v}}_j$ is used only once in a projection. In the unmodified version, each vector $\mathbf{y}_i$ is adjusted only once, but the vectors $\hat{\mathbf{v}}_j$ are used repeatedly in the ever-more complicated $\mathbf{T}_i$ projection operators.

Geometry in n-Dimensional Spaces

5.19 The equation of a plane in n dimensions is $\langle \mathbf{c}, \mathbf{x}\rangle = a$, where **c** is the normal to the plane and a is a scalar constant. Find the point on the plane nearest the origin and find the distance to this point.

Any **x** can be decomposed into a component $\mathbf{x}_n$ normal to the plane plus $\mathbf{x}_p$ parallel to the plane. Then $\langle \mathbf{c}, \mathbf{x}\rangle = \langle \mathbf{c}, \mathbf{x}_n\rangle + \langle \mathbf{c}, \mathbf{x}_p\rangle = a$ for every **x** terminating on the plane. Since **c** and $\mathbf{x}_p$ are orthogonal, $\langle \mathbf{c}, \mathbf{x}_p\rangle = 0$. Since **c** and $\mathbf{x}_n$ are parallel, $\mathbf{x}_n = \pm\|\mathbf{x}_n\|\mathbf{c}/\|\mathbf{c}\|$. Then $\langle \mathbf{c}, \mathbf{x}_n\rangle = \pm\langle \mathbf{c}, \mathbf{c}\rangle\|\mathbf{x}_n\|/\|\mathbf{c}\| = a$. The + or − sign must be selected to agree with the sign of a. Solving gives the minimum distance as $\|\mathbf{x}_n\| = |a|/\|\mathbf{c}\|$ and the closest point is $\mathbf{x}_n = a\mathbf{c}/\langle \mathbf{c}, \mathbf{c}\rangle$.

5.20 Find the minimum distance from the origin to a point (x_1, x_2) on the line $6x_1 + 2x_2 = 4$, and find the coordinates of that point.

The normal to the line is $\mathbf{c} = [6 \quad 2]^T$, and so $\|\mathbf{c}\| = \sqrt{40}$. The results of Problem 5.19 apply, and the minimum distance is $\|\mathbf{x}_n\| = 4/\sqrt{40} = 2/\sqrt{10}$. The point nearest the origin is

$$\mathbf{x}_n = \frac{1}{5}\begin{bmatrix}3\\1\end{bmatrix}$$

5.21 Generalize the concepts of lines, planes, spheres, cones, and convex sets to n-dimensional Euclidean spaces.

The generalization of a line is the set of all vectors satisfying

$$\mathbf{x} = a\mathbf{v} + \mathbf{k} \tag{1}$$

where **v** and **k** are constant vectors and a is a scalar.

The generalization of a plane is called a *hyperplane*. An $n - 1$ dimensional hyperplane consists of the set of n-dimensional vectors **x** satisfying

$$\langle \mathbf{c}, \mathbf{x}\rangle = a \tag{2}$$

where **c** and a are a constant vector and scalar, respectively. Since a subspace always contains the **0** vector, equation (1) represents a subspace only if **k** is zero. Equation (2) represents a subspace only if a is zero.

Points on or inside a hypersphere of radius R are defined by the set of all **x** satisfying

$$\langle \mathbf{x}, \mathbf{x}\rangle \leq R^2$$

A right circular cone of semi-vertex angle θ, with its axis in the direction of a unit vector **n**, and with vertex at the origin, consists of the set of all **x** satisfying

$$\langle \mathbf{n}, \mathbf{x}\rangle/\|\mathbf{x}\| = \cos\theta$$

If $\theta = \pi/2$, the cone degenerates to a hyperplane containing the origin.

If the line segments connecting every two points in a set contain only points in the set, the set is *convex*. A convex set of vectors in n-dimensional space is a set for which the vector

$$\mathbf{z} = a\mathbf{x}_1 + (1 - a)\mathbf{x}_2$$

belongs to the set for every $\mathbf{x}_1, \mathbf{x}_2$ in the set and for every real scalar satisfying $0 \leq a \leq 1$.

Some Generalizations

5.22 Let $\mathscr{L}_2[a, b]$ be the linear space consisting of all real square integrable functions of t, that is, all functions $f(t)$ satisfying $\int_a^b f(\tau)^2\, d\tau < \infty$. Define a suitable inner product, norm, and metric.

The three requirements for an inner product can be shown to be satisfied by

$$\langle f, g \rangle = \int_a^b f(\tau) g(\tau)\, d\tau \quad \text{where} \quad f, g \in \mathscr{L}_2[a, b]$$

As in other cases, a norm can always be defined as

$$\|f\| = \langle f, f \rangle^{1/2}$$

and the metric or distance measure between two functions can be defined in terms of the norm,

$$\rho(f, g) = \|f - g\|$$

The inner product space defined by this set of functions and the inner product definition is a complete infinite dimensional linear vector space. A space $\mathscr{X}$ is *complete* if every convergent (Cauchy) sequence of elements in $\mathscr{X}$ converges to a limit which is also in $\mathscr{X}$. Any infinite dimensional inner product space which is complete is called a *Hilbert space.*

5.23 Use the results of the previous problem to prove that the mean value of a real function is always less than or equal to its root-mean-square (rms) value.

We are to prove that

$$\frac{1}{T}\int_0^T f(\tau)\, d\tau \leq \left[\frac{1}{T}\int_0^T f^2(\tau)\, d\tau\right]^{1/2}$$

For any functions f and $g \in \mathscr{L}_2[0, T]$,

$$\|f - g\| \geq 0$$

or

$$\langle f - g, f - g \rangle \geq 0 \quad \text{or} \quad \langle f, f \rangle \geq 2\langle f, g \rangle - \langle g, g \rangle$$

Choosing the particular function $g = \text{constant} = \frac{1}{T}\int_0^T f(\tau)\, d\tau$ gives

$$\int_0^T f^2\, dt \geq \frac{1}{T}\left[\int_0^T f\, dt\right]^2$$

Dividing both sides by T and taking the square root gives the desired result.

5.24 Is it always necessary to define the norm in terms of the inner product?

No. Linear spaces can be defined with a norm, but without any mention of an inner product. Such spaces are called *normed linear spaces.* If they are infinite dimensional spaces, and are complete, then they are usually referred to as *Banach spaces.*

Two examples of other valid norms for finite dimensional spaces whose elements $\mathbf{x}$ are ordered n-tuples of scalars belonging to $\mathscr{F}$ are

$$\|\mathbf{x}\| = \max_i \{|x_1|, |x_2|, \ldots, |x_n|\}$$

and

$$\|\mathbf{x}\|_p = [|x_1|^p + |x_2|^p + \cdots + |x_n|^p]^{1/p}$$

where p is real, $1 \leq p \leq \infty$. The quadratic norm is a special case with $p = 2$. All the required axioms for a norm are satisfied for these examples.

5.25 Let $\mathscr{X}$ be a linear space consisting of all $n \times n$ matrices defined over $\mathscr{F}$. (Let $n = 2$ for simplicity.) Show that

(a) $\langle \mathbf{A}, \mathbf{B} \rangle = \text{Tr}\,(\overline{\mathbf{A}}^T \mathbf{B})$ is a valid inner product; and

(b) an orthonormal basis for this space is the set of matrices

$$\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$

(a) The inner product of two elements must yield a scalar. Obviously the trace gives a scalar. In addition, the three axioms must be satisfied:

(i) $\langle \mathbf{A}, \mathbf{B} \rangle = \overline{\langle \mathbf{B}, \mathbf{A} \rangle}$ but

$$\langle \mathbf{A}, \mathbf{B} \rangle = \text{Tr}\,(\overline{\mathbf{A}}^T \mathbf{B})$$

and

$$\overline{\langle \mathbf{B}, \mathbf{A} \rangle} = \overline{\text{Tr}\,(\overline{\mathbf{B}}^T \mathbf{A})} = \text{Tr}\,(\mathbf{B}^T \overline{\mathbf{A}}) = \text{Tr}\,(\mathbf{B}^T \overline{\mathbf{A}})^T = \text{Tr}\,(\overline{\mathbf{A}}^T \mathbf{B})$$

(ii) $\langle \mathbf{A}, \alpha\mathbf{B}_1 + \beta\mathbf{B}_2 \rangle = \alpha\langle \mathbf{A}, \mathbf{B}_1 \rangle + \beta\langle \mathbf{A}, \mathbf{B}_2 \rangle$ but

$$\langle \mathbf{A}, \alpha\mathbf{B}_1 + \beta\mathbf{B}_2 \rangle = \text{Tr}\,[\overline{\mathbf{A}}^T(\alpha\mathbf{B}_1 + \beta\mathbf{B}_2)] = \alpha\ \text{Tr}\,(\overline{\mathbf{A}}^T \mathbf{B}_1) + \beta\ \text{Tr}\,(\overline{\mathbf{A}}^T \mathbf{B}_2)$$
$$= \alpha\langle \mathbf{A}, \mathbf{B}_1 \rangle + \beta\langle \mathbf{A}, \mathbf{B}_2 \rangle$$

(iii) $\langle \mathbf{A}, \mathbf{A} \rangle \geq 0$ for all $\mathbf{A}$ and equals zero if and only if $\mathbf{A} = \mathbf{0}$. Let $\mathbf{A} = [a_{ij}]$. Then

$$\langle \mathbf{A}, \mathbf{A} \rangle = \text{Tr}\,(\mathbf{A}^T \mathbf{A}) = \bar{a}_{11} a_{11} + \bar{a}_{21} a_{21} + \bar{a}_{12} a_{12} + \bar{a}_{22} a_{22}$$
$$= |a_{11}|^2 + |a_{21}|^2 + |a_{12}|^2 + |a_{22}|^2$$

This is obviously nonnegative and can vanish only if $a_{ij} = 0$ for all i and j.

(b) First show that the set is orthonormal. $\mathbf{G} = [\langle \mathbf{V}_i, \mathbf{V}_j \rangle]$, where $\mathbf{V}_i$ are the four indicated matrices. Simple calculation shows that $\mathbf{G}$ is the 4×4 unit matrix, and thus the set is orthonormal. They span the space, since every 2×2 matrix $[a_{ji}]$ can be written as

$$\mathbf{A} = a_{11}\mathbf{V}_1 + a_{12}\mathbf{V}_2 + a_{21}\mathbf{V}_3 + a_{22}\mathbf{V}_4$$

An orthonormal set which spans the space is an orthonormal basis set.

5.26 Let $\mathscr{X}$ be an n-dimensional linear vector space with $\mathbf{x} \in \mathscr{X}$. Discuss the transformation $\mathscr{A}(\mathbf{x}) = \|\mathbf{x}\|$.

This function maps $\mathscr{X}$ into the real line, $\mathscr{A} : \mathscr{X} \to \mathscr{R}^1$. In this case the range of $\mathscr{A}$ is the nonnegative half line, so $\mathscr{R}(\mathscr{A}) \neq \mathscr{R}^1$ and the mapping is not onto. It is not one-to-one either, because many different vectors can have the same norm. Thus, given a value for $\|\mathbf{x}\|$, it is not possible to determine which vector $\mathbf{x}$ is the pre-image, i.e., $\mathscr{A}^{-1}$ does not exist. This transformation is not a linear transformation since, in general,

$$\mathscr{A}(\mathbf{x}_1 + \mathbf{x}_2) = \|\mathbf{x}_1 + \mathbf{x}_2\| \neq \|\mathbf{x}_1\| + \|\mathbf{x}_2\| = \mathscr{A}(\mathbf{x}_1) + \mathscr{A}(\mathbf{x}_2)$$

The null space of this transformation consists of the single vector $\mathbf{x} = \mathbf{0}$, by the properties of the norm.

Change of Basis

5.27 A vector $\mathbf{x}$ is represented by the $n \times 1$ column matrix $[\mathbf{x}]_v$ with respect to the basis $\{\mathbf{v}_i\}$ and by the $n \times 1$ column matrix $[\mathbf{x}]_{v'}$ with respect to another basis $\{\mathbf{v}_i'\}$. Show that the matrix $\mathbf{B}$ satisfying $[\mathbf{x}]_{v'} = \mathbf{B}[\mathbf{x}]_v$ always has an inverse.

Expand a typical member of the first basis in terms of the second basis, $\mathbf{v}_j = \sum_{i=1}^{n} b_{ij}\mathbf{v}_j'$. Likewise, a typical $\mathbf{v}_j'$ can be expanded as $\mathbf{v}_j' = \sum_{i=1}^{n} c_{ij}\mathbf{v}_i$. Thus $\mathbf{v}_i' = \sum_{k=1}^{n} c_{ki}\mathbf{v}_k$. Eliminating $\mathbf{v}_i'$ from the first equation gives

$$\mathbf{v}_j = \sum_{i=1}^{n} b_{ij} \sum_{k=1}^{n} c_{ki}\mathbf{v}_k = \sum_{i=1}^{n}\sum_{k=1}^{n} b_{ij}c_{ki}\mathbf{v}_k.$$

Because the vectors $\mathbf{v}_j$ are linearly independent, this requires

$$\sum_{i=1}^{n} b_{ij}c_{ki} = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases}$$

Letting $\mathbf{B} = [b_{ij}]$ and $\mathbf{C} = [c_{ij}]$, this can be written as $\mathbf{CB} = \mathbf{I}$. If the preceding procedure is modified slightly, by eliminating $\mathbf{v}_i$ from the second equation, we obtain $\mathbf{BC} = \mathbf{I}$. Taken together, the last two results imply that $\mathbf{C} = \mathbf{B}^{-1}$.

5.28 Show that if $\{\mathbf{v}_i\}$ and $\{\mathbf{v}_i'\}$ are both real, orthonormal basis sets, then $\mathbf{B}^{-1} = \mathbf{B}^T$.

Results of the previous problem give $\mathbf{B}^{-1} = \mathbf{C}$. But $c_{ij} = \langle \mathbf{r}_i, \mathbf{v}_j' \rangle$, where $\{\mathbf{r}_i\}$ are the reciprocal bases associated with $\{\mathbf{v}_i\}$. Since $\{\mathbf{v}_i\}$ is an orthonormal set, $\mathbf{r}_i = \mathbf{v}_i$ and $c_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j' \rangle$. Similarly, $b_{ij} = \langle \mathbf{r}_i', \mathbf{v}_j \rangle$ in the general case, where $\{\mathbf{r}_i'\}$ is the set of reciprocal bases for $\{\mathbf{v}_i'\}$. For the orthonormal case, $b_{ij} = \langle \mathbf{v}_i', \mathbf{v}_j \rangle$. Interchanging subscripts gives $b_{ji} = \langle \mathbf{v}_j', \mathbf{v}_i \rangle$. For the case of real vectors this shows that $b_{ji} = c_{ij}$, or $\mathbf{B}^T = \mathbf{C} = \mathbf{B}^{-1}$.

Adjoint Transformations

5.29 Consider the linear transformation $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$, with the adjoint transformation $\mathcal{A}^*$. Prove that $\|\mathcal{A}\| = \|\mathcal{A}^*\|$.

The definition of the adjoint requires that $\langle \mathbf{y}, \mathcal{A}(\mathbf{x}) \rangle = \langle \mathcal{A}^*(\mathbf{y}), \mathbf{x} \rangle$ for all $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$. The Cauchy-Schwarz inequality and the definition of $\|\mathcal{A}^*\|$ gives

$$|\langle \mathbf{y}, \mathcal{A}(\mathbf{x}) \rangle| = |\langle \mathcal{A}^*(\mathbf{y}), \mathbf{x} \rangle| \leq \|\mathcal{A}^*(\mathbf{y})\| \cdot \|\mathbf{x}\| \leq \|\mathcal{A}^*\| \cdot \|\mathbf{y}\| \cdot \|\mathbf{x}\|$$

This must be true for all $\mathbf{y}, \mathbf{x}$ including the particular pair related by $\mathcal{A}(\mathbf{x}) = \mathbf{y}$. Using this gives

$$\langle \mathcal{A}(x), \mathcal{A}(x) \rangle \leq \|\mathcal{A}^*\| \cdot \|\mathcal{A}(\mathbf{x})\| \cdot \|\mathbf{x}\| \quad \text{or} \quad \frac{\|\mathcal{A}(\mathbf{x})\|}{\|\mathbf{x}\|} \leq \|\mathcal{A}^*\|$$

for all $\mathbf{x} \neq \mathbf{0}$. Therefore,

$$\|\mathcal{A}^*\| \geq \|\mathcal{A}\| \tag{1}$$

Similarly, if particular $\mathbf{x}, \mathbf{y}$ pairs are chosen such that $\mathbf{x} = \mathcal{A}^*(\mathbf{y})$, then

$$|\langle \mathcal{A}^*(\mathbf{y}), \mathcal{A}^*(\mathbf{y}) \rangle| \leq \|\mathbf{y}\| \cdot \|\mathcal{A}(x)\| \leq \|\mathbf{y}\| \cdot \|\mathbf{x}\| \cdot \|\mathcal{A}\|$$

or

$$\|\mathcal{A}^*(y)\|^2 \leq \|\mathbf{y}\| \cdot \|\mathcal{A}^*(y)\| \cdot \|\mathcal{A}\| \quad \text{or} \quad \frac{\|\mathcal{A}^*(\mathbf{y})\|}{\|\mathbf{y}\|} \leq \|\mathcal{A}\| \quad \text{for all } \mathbf{y} \neq \mathbf{0}$$

Therefore, $\|\mathcal{A}\| \geq \|\mathcal{A}^*\|$. This, together with Eq. (*1*), gives $\|\mathcal{A}\| = \|\mathcal{A}^*\|$.

5.30 Let $\mathcal{X}$ be the set of all n-component real functions defined over $[t_0, t_f]$ which have continuous first derivatives and let $\mathcal{Y}$ be the set of all continuous functions defined over the same interval. Find the adjoint of $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ if $\mathcal{A}(\mathbf{x}) = (d/dt)\mathbf{x} - \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an $n \times n$ matrix.

The inner product for $\mathcal{X}$ and $\mathcal{Y}$ takes the form of equation (*1*). The adjoint is defined by equation (*2*):

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = \int_{t_0}^{t_f} \mathbf{x}_1^T(\tau)\mathbf{x}_2(\tau)\, d\tau \tag{1}$$

$$\langle \mathbf{y}, \mathscr{A}(\mathbf{x}) \rangle = \langle \mathscr{A}^*(\mathbf{y}), \mathbf{x} \rangle \tag{2}$$

Integrating the left-hand side of Eq. (2) gives

$$\int_{t_0}^{t_f} \mathbf{y}^T(\tau)\left[\frac{d\mathbf{x}}{dt} - \mathbf{A}\mathbf{x}(\tau)\right] d\tau = \mathbf{y}^T(\tau)\mathbf{x}(\tau)\Big|_{t_0}^{t_f} - \int_{t_0}^{t_f}\left[\frac{d\mathbf{y}^T}{dt} + \mathbf{y}(\tau)^T\mathbf{A}\right]\mathbf{x}(\tau)\, d\tau$$

The term involving the limits of integration, $\mathbf{y}^T(t_f)\mathbf{x}(t_f) - \mathbf{y}^T(t_0)\mathbf{x}(t_0)$, could be made to vanish by specifying appropriate boundary conditions. Ignoring this term, the remaining term is in the form of the right-hand side of Eq. (2). Therefore, $\mathscr{A}^*(\mathbf{y}) = -(d\mathbf{y}/dt) - \mathbf{A}^T\mathbf{y}(t)$. We conclude that the formal adjoint of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is $\dot{\mathbf{y}} = -\mathbf{A}^T\mathbf{y}$. This adjoint differential equation arises frequently in optimal control theory and in other applications.

5.31 Consider the equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with $\mathbf{x}(t_0) = \mathbf{x}_0$. Show that $\mathbf{y}^T(t)\mathbf{x}(t)$ is a constant for all time, where $\mathbf{y}$ satisfies the adjoint equation $\dot{\mathbf{y}} = -\mathbf{A}^T\mathbf{y}$.

A necessary and sufficient condition that $\mathbf{y}^T\mathbf{x}$ be constant is that $(d/dt)(\mathbf{y}^T\mathbf{x}) = 0$. Using the chain rule for differentiating gives $(d/dt)(\mathbf{y}^T\mathbf{x}) = \dot{\mathbf{y}}^T\mathbf{x} + \mathbf{y}^T\dot{\mathbf{x}}$. Substituting in for $\dot{\mathbf{y}}$ and $\dot{\mathbf{x}}$ gives

$$(d/dt)(\mathbf{y}^T\mathbf{x}) = -\mathbf{y}^T\mathbf{A}\mathbf{x} + \mathbf{y}^T\mathbf{A}\mathbf{x} = 0$$

5.32 The input-output equation $\mathbf{y}(t) = \int_{t_0}^{\infty} \mathbf{W}(t, \tau)\,\mathbf{u}(\tau)\, d\tau$ defines a linear transformation from the space of r-component square integrable functions $\mathbf{u}(t) \in \mathscr{U}$ to the set of m-component continuous functions $\mathbf{y}(t) \in \mathscr{Y}$. The functions are defined over (t_0, ∞). Consider $\mathscr{A} : \mathscr{U} \rightarrow \mathscr{Y}$ and find $\mathscr{A}^*$, if the inner product for each space has the form

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = \int_{t_0}^{\infty} \mathbf{x}_1^T(t)\mathbf{x}_2(t)\, dt$$

The transformation $\mathscr{A}$ is $\mathscr{A}(\mathbf{u}) = \int_{t_0}^{\infty} \mathbf{W}(t, \tau)\mathbf{u}(\tau)\, d\tau$ and $\langle \mathbf{z}, \mathscr{A}(\mathbf{u}) \rangle = \langle \mathscr{A}^*(\mathbf{z}), \mathbf{u} \rangle$. Using the inner product definition gives

$$\langle \mathbf{z}, \mathscr{A}(\mathbf{u}) \rangle = \int_{t_0}^{\infty} \mathbf{z}^T(t) \int_{t_0}^{\infty} \mathbf{W}(t, \tau)\mathbf{u}(\tau)\, d\tau\, dt = \int_{t_0}^{\infty}\int_{t_0}^{\infty} \mathbf{z}^T(t)\mathbf{W}(t, \tau)\, dt\, \mathbf{u}(\tau)\, d\tau$$

Therefore, $\mathscr{A}^*(\mathbf{z}) = \int_{t_0}^{\infty} \mathbf{W}^T(t, \tau)\mathbf{z}(t)\, dt$. The weighting matrix for the adjoint equation is the transpose of $\mathbf{W}$ and the integration variable is t rather than τ as in the original transformation. This operator is self-adjoint if $\mathbf{W}(t, \tau) = \mathbf{W}^T(\tau, t)$.

Matrix Norms

5.33 Consider a linear transformation which maps vectors from one finite dimensional space to another. Let $\mathbf{A}$ be its matrix representation. If $\|\mathbf{x}\| = \max_i |x_i|$, find $\|\mathbf{A}\|$.

With this definition for the vector norm,

$$\|\mathbf{A}\mathbf{x}\| = \max_i \left|\sum_j a_{ij}x_j\right| \le \left\{\max_i \left|\sum_j a_{ij}\right|\right\}\left\{\max_j |x_j|\right\} \le \left\{\max_i \sum_j |a_{ij}|\right\}\|\mathbf{x}\|$$

The norm of $\mathbf{A}$ must satisfy $\|\mathbf{A}\mathbf{x}\| \le \|\mathbf{A}\|\cdot\|\mathbf{x}\|$. We see that $\max_i \sum_j |a_{ij}|$ qualifies as a bound for $\|\mathbf{A}\mathbf{x}\|/\|\mathbf{x}\|$, with $\|\mathbf{x}\| \neq 0$. To show that it is the *least* upper bound, and hence is equal to the norm, we must show that no smaller bound is possible. This can be done by demonstrating that the bound is actually attained for some $\mathbf{x}$.

Let i^* be the row i, which maximizes the above sum. Let

$$x_j = \begin{cases} 0 & \text{if } a_{i^*j} = 0 \\ \dfrac{\bar{a}_{i^*j}}{a_{i^*j}} & \text{otherwise} \end{cases}$$

For this choice, $\|\mathbf{x}\| = 1$ and

$$\max_i \left| \sum_j a_{ij} x_j \right| = \left| \sum_j \frac{a_{i^*j} \bar{a}_{i^*j}}{|a_{i^*j}|} \right| = \sum_j |a_{i^*j}|$$

Therefore, $\|\mathbf{A}\| = \max_i \sum_j |a_{ij}|$.

5.34 Let the vector norm be defined by $\|\mathbf{x}\| = \sum_i |x_i|$ and find a bound for $\|\mathbf{A}\|$.

Beginning with the given definition for the norm, a series of manipulations gives

$$\|\mathbf{Ax}\| = \sum_i \left| \sum_j a_{ij} x_j \right| \le \sum_i \left| \sum_j a_{ij} \right| \cdot \left| \sum_j x_j \right| \le \sum_i \sum_j |a_{ij}| \sum_j |x_j| = \sum_i \sum_j |a_{ij}| \cdot \|\mathbf{x}\|$$

Since $\|\mathbf{Ax}\| \le \|\mathbf{x}\| \sum_i \sum_j |a_{ij}|$, the double summation term is an upper bound for $\|\mathbf{Ax}\|/\|\mathbf{x}\|$, $\|\mathbf{x}\| \neq 0$.

Since $\|\mathbf{A}\|$ is the least such bound, $\|\mathbf{A}\| \le \sum_i \sum_j |a_{ij}|$.

Miscellaneous Applications

5.35 Find the projection of $\mathbf{y} = [1 \quad -3 \quad 4 \quad 2 \quad 8]^T$ on the subspace spanned by

$$\mathbf{x}_1 = [1 \quad 2 \quad -3 \quad 1 \quad 0]^T \quad \text{and} \quad \mathbf{x}_2 = [0 \quad 1 \quad 3 \quad 3 \quad 1]^T$$

The dimension of the subspace is two, since $|\mathbf{G}| = 284 \neq 0$. An orthonormal basis is constructed:

$$\hat{\mathbf{v}}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} = \frac{1}{\sqrt{15}} [1 \quad 2 \quad -3 \quad 1 \quad 0]^T$$

$$\hat{\mathbf{v}}_2 = \frac{\mathbf{x}_2 - \langle \mathbf{x}_2, \hat{\mathbf{v}}_1 \rangle \hat{\mathbf{v}}_1}{\|\mathbf{x}_2 - \langle \mathbf{x}_2, \hat{\mathbf{v}}_1 \rangle \hat{\mathbf{v}}_1\|} = \frac{1}{\sqrt{4260}} \begin{bmatrix} 4 \\ 23 \\ 33 \\ 49 \\ 15 \end{bmatrix}$$

The projection of $\mathbf{y}$ on this subspace is

$$\mathbf{y}_p = \langle \hat{\mathbf{v}}_1, \mathbf{y} \rangle \hat{\mathbf{v}}_1 + \langle \hat{\mathbf{v}}_2, \mathbf{y} \rangle \hat{\mathbf{v}}_2$$

$$= -\sqrt{15}\, \hat{\mathbf{v}}_1 + \frac{285}{\sqrt{4260}} \hat{\mathbf{v}}_2 = \begin{bmatrix} -1 \\ -2 \\ 3 \\ -1 \\ 0 \end{bmatrix} + \frac{285}{4260} \begin{bmatrix} 4 \\ 23 \\ 33 \\ 49 \\ 15 \end{bmatrix} = \frac{1}{284} \begin{bmatrix} -208 \\ -131 \\ 1479 \\ 647 \\ 285 \end{bmatrix}$$

The projection $\mathbf{y}_p$ is the closest vector in the subspace to $\mathbf{y}$, in the sense that

$$\|\mathbf{y} - \mathbf{y}_p\|^2 \le \|\mathbf{y} - \mathbf{z}\|^2$$

for every $\mathbf{z}$ in the subspace. These results are directly related to the problem of least squares approximations, considered in the next chapter.

5.36 Consider the dc motor of Problem 2.5, with transfer function

$$\frac{\Omega(s)}{V(s)} = \frac{K'}{s + a}$$

If the motor is initially at rest, $\omega(0) = 0$, find the input $v(t)$ which gives an angular velocity $\omega(T) = 100$ at a fixed time T, while minimizing a measure of the input energy,

$$J = \int_0^T v^2(\tau)\, d\tau$$

The input-output relationship is written in terms of the *system weighting function*. Since $W(t, 0) = \mathcal{L}^{-1}\{K'/(s + a)\} = K'e^{-at}$, the weighting function is $W(t, \tau) = K'e^{-a(t-\tau)}$. Then

$$\omega(T) = \int_0^T W(T, \tau)v(\tau)\,d\tau$$

This is in the form of an inner product, $100 = \langle W(T, \tau), v(\tau)\rangle$, so the Cauchy-Schwarz inequality can be used to give

$$100 = |\langle W(T, \tau), v(\tau)\rangle| \leq \|W(T, \tau)\| \cdot \|v(\tau)\|$$

The minimum value of $\|v(\tau)\|$ is obtained when the equality holds. Therefore,

$$\|v(\tau)\| = \frac{100}{\|W(T, \tau)\|} = \frac{100}{\left\{\int_0^T [K'e^{-a(T-\tau)}]^2\,d\tau\right\}^{1/2}} = \frac{100\sqrt{2a}}{K'[1 - e^{-2aT}]^{1/2}}$$

The equality holds if and only if $v(t)$ and $W(T, t)$ are linearly dependent. This means $v(t) = kW(T, t)$ for some scalar k. Comparing gives $\|v\| = |k|\,\|W\| = 100/\|W\|$ and so

$$|k| = \frac{100}{\|W\|^2} \quad \text{and} \quad v_{\text{optimal}}(t) = \frac{100}{\|W\|^2}W(T, t) = \frac{200ae^{-a(T-t)}}{K'[1 - e^{-2aT}]}$$

5.37 A satellite position vector has components (X, Y, Z) with respect to an inertially fixed coordinate system with origin at the earth's center. Determine the components of this position vector as measured by a tracking station at longitude L degrees east and latitude λ degrees north, if the angle between the X_I inertial axis and the 0° longitudinal meridian is ϕ degrees and the earth's radius is R_e. See Figure 5.7.

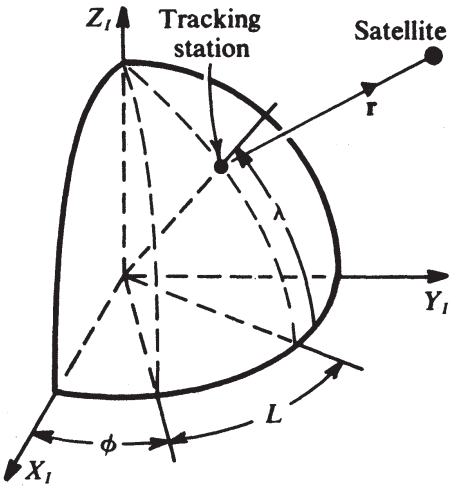


Figure 5.7

The coordinate transformations can be represented by the Piogram of Figure 5.8. The measured range vector has the following components in the Up-East-North coordinate system. (These components can be read directly from the Piogram by using a few standard conventions [5, 6].) The origin is shifted by subtracting R_e from the Up component:

$$\mathbf{r} = \begin{bmatrix} X[\cos(\phi + L)\cos\lambda] + Y\sin(\phi + L)\cos\lambda + Z\sin\lambda - R_e \\ -X\sin(\phi + L) + Y\cos(\phi + L) \\ -X\cos(\phi + L)\sin\lambda - Y\sin(\phi + L)\sin\lambda + Z\cos\lambda \end{bmatrix}$$

5.38 Demonstrate how the spatial attitude orientation of a vehicle (aircraft, satellite, etc.) can be determined by sighting two known stars.

The attitude of the vehicle will be characterized by the 3×3 transformation matrix $\mathbf{T}_{BE}$, which relates a set of orthonormal body fixed axes $\{x, y, z\}$ to a fixed orthonormal inertial

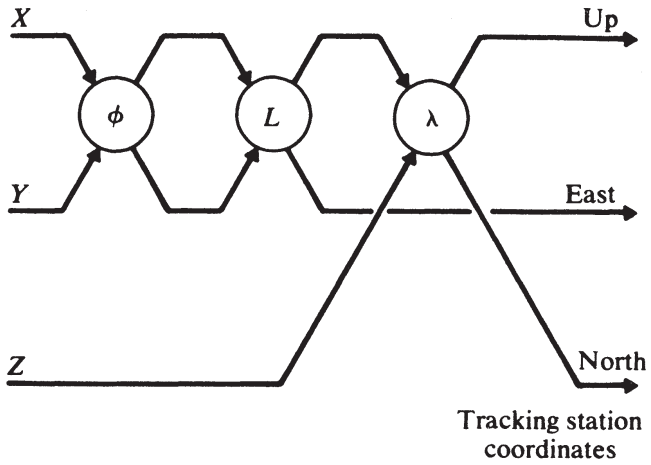


Figure 5.8

coordinate system $\{X, Y, Z\}$. This means a vector $\mathbf{v}$ expressed in the two coordinate systems satisfies

$$\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \mathbf{T}_{BE} \begin{bmatrix} v_X \\ v_Y \\ v_Z \end{bmatrix}$$

$\mathbf{T}_{BE}$ is to be found. The unit vectors which point toward the two stars are assumed available from star catalogs, in inertial components. That is,

$$\hat{\mathbf{u}} = [u_X \quad u_Y \quad u_Z]$$

$$\hat{\mathbf{v}} = [v_X \quad v_Y \quad v_Z]$$

The pointing direction of a telescope mounted in the vehicle is described by two gimbal angles α and β as shown in Figure 5.9.

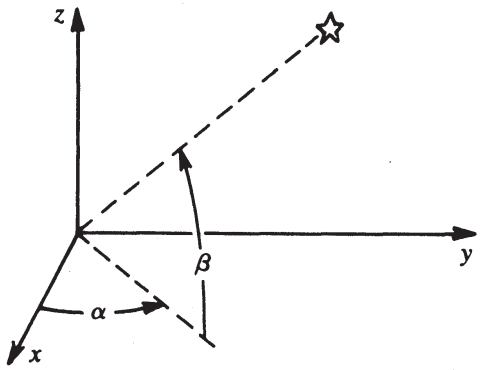


Figure 5.9

The unit vectors to the given stars are expressed in vehicle fixed coordinates as

$$\hat{\mathbf{u}}_B = \begin{bmatrix} \cos\alpha_1 \cos\beta_1 \\ \sin\alpha_1 \cos\beta_1 \\ \sin\beta_1 \end{bmatrix}, \qquad \hat{\mathbf{v}}_B = \begin{bmatrix} \cos\alpha_2 \cos\beta_2 \\ \sin\alpha_2 \cos\beta_2 \\ \sin\beta_2 \end{bmatrix}$$

where α_1, β_1 and α_2, β_2 are the pointing angles for the two stars. $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are known, and $\hat{\mathbf{u}}_B$ and $\hat{\mathbf{v}}_B$ are available from measurements.

In order to determine $\mathbf{T}_{BE}$, one more relation is necessary. A vector normal to the plane defined by $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ is constructed using the cross product. Thus

$$\mathbf{w} \triangleq \hat{\mathbf{u}} \times \hat{\mathbf{v}} \quad \text{and} \quad \mathbf{w}_B \triangleq \hat{\mathbf{u}}_B \times \hat{\mathbf{v}}_B$$

Now $\hat{\mathbf{u}}_B = \mathbf{T}_{BE}\hat{\mathbf{u}}$, $\hat{\mathbf{v}}_B = \mathbf{T}_{BE}\hat{\mathbf{v}}$, and $\mathbf{w}_B = \mathbf{T}_{BE}\mathbf{w}$. These can be combined into

$$[\hat{\mathbf{u}}_B \quad \hat{\mathbf{v}}_B \quad \mathbf{w}_B] = \mathbf{T}_{BE}[\hat{\mathbf{u}} \quad \hat{\mathbf{v}} \quad \mathbf{w}]$$

If the vectors to the two stars are linearly independent, then the 3×3 matrix $[\hat{\mathbf{u}} \quad \hat{\mathbf{v}} \quad \mathbf{w}]$ is nonsingular and

$$\mathbf{T}_{BE} = [\hat{\mathbf{u}}_B \quad \hat{\mathbf{v}}_B \quad \mathbf{w}_B][\hat{\mathbf{u}} \quad \hat{\mathbf{v}} \quad \mathbf{w}]^{-1}$$

PROBLEMS

Linear Independence, Orthonormal Basis Vectors, and Reciprocal Basis Vectors

5.39 Consider $\mathbf{x}_1 = [1 \quad 2 \quad 3]^T$, $\mathbf{x}_2 = [1 \quad -2 \quad 3]^T$, $\mathbf{x}_3 = [0 \quad 1 \quad 1]^T$.
(a) Show that this set is linearly independent.
(b) Generate an orthonormal set using the Gram-Schmidt procedure.

5.40 Considering $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ of Problem 5.39 as a basis set, find the reciprocal basis set.

5.41 Express the vector $\mathbf{z} = [6 \quad 4 \quad -3]^T$ in terms of the orthonormal basis set $\{\hat{\mathbf{v}}_i\}$ of Problem 5.39.

5.42 Express the vector $\mathbf{z} = [6 \quad 4 \quad -3]^T$ in terms of the original basis set $\{\mathbf{x}_i\}$ of Problem 5.39 by using the reciprocal basis vectors $\{\mathbf{r}_i\}$ found in Problem 5.40.

5.43 Find the reciprocal basis set if the basis vectors are $\mathbf{x}_1 = [4 \quad 2 \quad 1]^T$, $\mathbf{x}_2 = [2 \quad 6 \quad 3]^T$, $\mathbf{x}_3 = [1 \quad 3 \quad 5]^T$.

5.44 Given $\mathbf{x}_1 = [1 \quad 1 \quad 1]^T$, $\mathbf{x}_2 = [1 \quad -1 \quad 1]^T$, $\mathbf{x}_3 = [1 \quad 0 \quad 0]^T$. Use these as basis vectors and find reciprocal basis vectors. Also, express $\mathbf{z} = [6 \quad 3 \quad 1]^T$ in terms of the basis vectors.

5.45 Show that an orthonormal basis set and the corresponding set of reciprocal basis vectors are the same.

5.46 Verify that the following four $\mathbf{y}_i$ vectors are linearly independent by computing their Grammian. Then use them to construct an orthonormal set using the Gram-Schmidt process. Then use the orthonormal vectors to expand the vector $\mathbf{x} = [12.3 \quad 9.8 \quad -4.03 \quad 33.33]^T$.

The given **y** *vectors*

$$\mathbf{y}_1 = \begin{bmatrix} 4.4400002\text{E}+01 \\ 1.2800000\text{E}+01 \\ 1.5000000\text{E}+00 \\ -2.1000000\text{E}+01 \end{bmatrix}, \qquad \mathbf{y}_2 = \begin{bmatrix} 7.7700000\text{E}+00 \\ 2.1500000\text{E}+01 \\ 1.0000000\text{E}+01 \\ 0.0000000\text{E}+00 \end{bmatrix},$$

$$\mathbf{y}_3 = \begin{bmatrix} -3.3329999\text{E}+00 \\ 4.1250000\text{E}+00 \\ 6.6670001\text{E}-01 \\ 1.0000000\text{E}+00 \end{bmatrix}, \qquad \mathbf{y}_4 = \begin{bmatrix} 9.1250000\text{E}+00 \\ 2.1222000\text{E}+00 \\ -3.0500000\text{E}+00 \\ 4.4400001\text{E}+00 \end{bmatrix}$$

5.47 Compute the Grammian for the following vectors and draw conclusions about their linear independence.

The given **y** *vectors*

$$\mathbf{y}_1 = \begin{bmatrix} 1.0000000\text{E}+00 \\ 1.0000000\text{E}+00 \\ 1.0000000\text{E}+00 \\ 1.0000000\text{E}+00 \end{bmatrix}, \qquad \mathbf{y}_2 = \begin{bmatrix} 1.0001000\text{E}+00 \\ 9.9989998\text{E}-01 \\ 1.0000000\text{E}+00 \\ 1.0000000\text{E}+00 \end{bmatrix}, \qquad \mathbf{y}_3 = \begin{bmatrix} -2.0000000\text{E}+00 \\ -1.9999000\text{E}+00 \\ -2.0000000\text{E}+00 \\ -2.0000000\text{E}+00 \end{bmatrix}$$

5.48 Find the orthogonal projection of the vectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{bmatrix} -4 \\ 2 \\ -8 \\ 3 \\ 9 \end{bmatrix}$$

on the subspace spanned by the following three vectors.

The given **y** *vectors*

$$\mathbf{y}_1 = \begin{bmatrix} 1.0000000\text{E}+01 \\ 5.0000000\text{E}+00 \\ -5.0000000\text{E}+00 \\ 1.0000000\text{E}+00 \\ 6.0000000\text{E}+00 \end{bmatrix}, \qquad \mathbf{y}_2 = \begin{bmatrix} 1.1000000\text{E}+01 \\ 4.0000000\text{E}+00 \\ -1.1000000\text{E}+01 \\ -2.0000000\text{E}+00 \\ 6.0000000\text{E}+00 \end{bmatrix},$$

$$\mathbf{y}_3 = \begin{bmatrix} 1.0000000\text{E}+00 \\ 0.0000000\text{E}+00 \\ -1.0000000\text{E}+00 \\ -1.0000000\text{E}+00 \\ -1.0000000\text{E}+00 \end{bmatrix}$$

5.49 Determine the dimension of the vector space spanned by $\mathbf{x}_1 = [1 \quad 2 \quad 2 \quad 1]^T$, $\mathbf{x}_2 = [1 \quad 0 \quad 0 \quad 1]^T$, $\mathbf{x}_3 = [3 \quad 4 \quad 4 \quad 3]^T$.

5.50 Find the minimum distance from the origin to the plane $2x_1 + 3x_2 - x_3 = -5$ and find coordinates of the point on the plane nearest the origin.

5.51 Let $\mathbf{c} = [1 \quad 2 \quad -1]^T$ and $\mathbf{y} = [2 \quad 5 \quad 3]^T$. Find the projection of **y** that is parallel to the family of planes defined by $\langle \mathbf{c}, \mathbf{x} \rangle = \text{constant}$.

5.52 Show that the various Fourier series expansion formulas are special cases of the general expansion formula in an infinite dimensional linear inner product space:

$$\mathbf{x} = \sum_{i=1}^{\infty} \langle \mathbf{r}_i, \mathbf{x} \rangle \mathbf{v}_i$$

5.53 Under what conditions does $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{A} \mathbf{y}$ define a valid inner product for an n-dimension vector space defined over the real number field?

5.54 If $\mathcal{X}^m$ and $\mathcal{X}^n$ are m- and n-dimensional linear vector spaces, respectively, then the product space $\mathcal{X}^m \times \mathcal{X}^n$ is itself a linear vector space consisting of all ordered pairs of $\mathbf{x} \in \mathcal{X}^m$, $\mathbf{y} \in \mathcal{X}^n$. That is,

$$\mathcal{X}^m \times \mathcal{X}^n = \{(\mathbf{x}, \mathbf{y}); \mathbf{x} \in \mathcal{X}^m, \mathbf{y} \in \mathcal{X}^n\}$$

If $\mathcal{X}^m$ has an inner product $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_m$ and $\mathcal{X}^n$ has an inner product $\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_n$, show that the appropriate inner product for $\mathcal{X}^m \times \mathcal{X}^n$ is

$$\left\langle \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{bmatrix} \right\rangle = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_m + \langle \mathbf{y}_1, \mathbf{y}_2 \rangle_n$$

5.55 Let $\mathcal{L}_2^n$ be the linear space consisting of all complex valued square integrable n component vector functions of a scalar variable $t \in [a, b]$, $\mathbf{f}(t) = [f_i(t)]$, where $i = 1, n$. Define an appropriate inner product and norm for this space.

5.56 Prove the Cauchy-Schwarz inequality given on page 167.

5.57 Let $\mathcal{R}^n$ be an n-dimensional Euclidean space with an orthonormal basis $\mathcal{B} = \{\mathbf{v}_i, i = 1, n\}$. Prove that for any $\mathbf{x} \in \mathcal{R}^n$,

$$\|\mathbf{x}\|^2 \geq \sum_{i=1}^{m} |\langle \mathbf{v}_i, \mathbf{x} \rangle|^2$$

where the summation is over any subset of m basis vectors. This is called Bessel's inequality. If $m = n$, the equality holds.

5.58 Consider the linear transformation $\mathcal{A} : \mathcal{R}^3 \to \mathcal{R}^3$. The basis vectors are selected as

$$\mathbf{v}_1 = [1 \quad 0 \quad 1]^T, \qquad \mathbf{v}_2 = [1 \quad 1 \quad 0]^T, \qquad \mathbf{v}_3 = [1 \quad 1 \quad 1]^T$$

The images of the basis vectors under the transformation $\mathcal{A}$ are

$$\mathbf{u}_1 = [2 \quad -1 \quad 3]^T, \qquad \mathbf{u}_2 = [-1 \quad -1 \quad 2]^T, \qquad \mathbf{u}_3 = [1 \quad 1 \quad 5]^T$$

with respect to the same basis. What is the matrix representation for $\mathcal{A}$?

5.59 The coordinate representations of a real vector $\mathbf{x}$ with respect to two different sets of orthonormal basis vectors are related by

$$[\mathbf{x}]' = \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} [\mathbf{x}]$$

Verify that $\sum_{i=1}^{3} x_i^2 = \sum_{i=1}^{3} (x_i')^2$. Would this be true for nonorthonormal bases?

5.60 If $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are real unit vectors in three-dimensional space with an included angle α, and if $\mathbf{w} = \hat{\mathbf{u}} \times \hat{\mathbf{v}}$, find an expression for $\mathbf{A}^{-1} = [\hat{\mathbf{u}} \quad \hat{\mathbf{v}} \quad \mathbf{w}]^{-1}$.

5.61 Derive a matrix differential equation for the time rate of change $\dot{\mathbf{T}}_{BE}$ for the transformation of Problem 5.38.

5.62 A mirror lies in the plane defined by $-2x_1 + 3x_2 + x_3 = 0$. Find the reflected image of $\mathbf{y} = [4 \quad -2 \quad 3]^T$ and also find the orthogonal projection of $\mathbf{y}$ on the plane of the mirror.

5.63 Show that $\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}$ gives the orthogonal projection of $\mathcal{X}$ onto $\mathcal{R}(\mathcal{A}^*)$.

5.64 If $\mathbf{A}$ is the matrix representation of an operator which acts on $\mathbf{x} \in \mathcal{X}$, and if $\|\mathbf{x}\|^2 = \bar{\mathbf{x}}^T\mathbf{x}$, show that $\|\mathbf{A}\| \leq [\sum_i \sum_j |a_{ij}|^2]^{1/2}$.

5.65 The vector norm is defined by $\|\mathbf{x}\|^2 = \bar{\mathbf{x}}^T\mathbf{x}$ and $\mathbf{Ax} = \mathbf{y}$.
(a) Show that $\|\mathbf{A}\|^2 = \max\{\gamma_i\}$, where $\{\gamma_i\}$ is the set of eigenvalues for $\bar{\mathbf{A}}^T\mathbf{A}$.
(b) Show that if $\mathbf{A}$ is normal, $\|\mathbf{A}\| = \max_i |\lambda_i|$, where $\{\lambda_i\}$ is the set of eigenvalues for $\mathbf{A}$.

5.66 Let $\mathcal{X}$ and $\mathcal{Y}$ be linear vector spaces whose elements are summable n-component vector sequences, $\{\mathbf{x}(k), k = 0, 1, 2, \ldots\}$ and $\{\mathbf{y}(k), k = 0, 1, 2, \ldots\}$. A linear transformation $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ is defined by the first-order difference equation $\mathcal{A}(\mathbf{x}) = \mathbf{x}(k+1) - \mathbf{A}_k\mathbf{x}(k)$. Find $\mathcal{A}^*$ if the inner product is defined as $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = \sum_{k=0}^{\infty} \mathbf{x}_1(k)^T\mathbf{x}_2(k)$.

6

Simultaneous Linear Equations

6.1 INTRODUCTION

The task of solving a set of simultaneous linear algebraic equations is frequently encountered by engineers and scientists in all fields. Many problems of estimation, control, system identification, pole-placement, and optimization depend on the solution of simultaneous equations. The properties of controllability and observability of linear systems are conditions which directly relate to the ability to solve a set of simultaneous equations. The stability and natural modes of a system are determined by the solution of an eigenvalue problem, which involves solution of simultaneous equations. This chapter uses the matrix theory and linear algebra of the last two chapters to study this class of problems. Several important applications are also introduced. The material in this chapter will be used in every chapter in the rest of this book.

6.2 STATEMENT OF THE PROBLEM AND CONDITIONS FOR SOLUTIONS

Consider the set of simultaneous linear algebraic equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= y_m \end{aligned}$$

In matrix notation this is simply

$$\mathbf{Ax} = \mathbf{y} \tag{6.1}$$

where the elements of a_{ij} of the $m \times n$ matrix $\mathbf{A}$ are known, as are the scalar components y_i of the $m \times 1$ vector $\mathbf{y}$. The $n \times 1$ vector $\mathbf{x}$ contains the unknowns which are

to be determined if possible. Any vector, say $\mathbf{x}_1$, which satisfies all m of these equations is called a *solution.* Not every set of simultaneous equations has a solution. The *augmented matrix,* defined by $\mathbf{W} = [\mathbf{A} \mid \mathbf{y}]$, indicates whether or not solutions exist. In fact,

1. If $r_W \neq r_A$, no solution exists. The equations are *inconsistent.*
2. If $r_W = r_A$, at least one solution exists.
   (a) If $r_W = r_A = n$, there is a *unique* solution for $\mathbf{x}$.
   (b) If $r_W = r_A < n$, then there is an infinite set of solution vectors.

It is clearly impossible for r_A to exceed n, so the only possibilities are that there are no solutions, or exactly one solution, or an infinity of solutions. In order to explain fully the basis for these results, two linear vector spaces and the mappings between them must be studied. Let $\mathcal{X}^n$ be the space of all n-dimensional $\mathbf{x}$ vectors and let $\mathcal{X}^m$ be the space of all m-dimensional $\mathbf{y}$ vectors. The matrix $\mathbf{A}$ can be considered as a concrete example of an operator which maps members of $\mathcal{X}^n$ into members of $\mathcal{X}^m$. As discussed in Chapter 5, there is another operator, the adjoint operator $\mathbf{A}^*$, which maps elements of $\mathcal{X}^m$ back into $\mathcal{X}^n$. In the present case the adjoint operator is just the conjugate transpose of $\mathbf{A}$, i.e., $\mathbf{A}^* = \overline{\mathbf{A}}^T$. It is very useful to know that the two spaces under discussion can each be written as an orthogonal sum of two subspaces. First $\mathcal{X}^m$ is considered. The primary subspace of interest is the range space of $\mathbf{A}$, $\mathcal{R}(\mathbf{A})$. This is the space of all $\mathbf{y}$ vectors which are images of some $\mathbf{x}$ vector. Since a column-partitioned version of Eq. (*6.1*) is

$$x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + \cdots + x_n \mathbf{a}_n = \mathbf{y}$$

it is seen that $\mathcal{R}(\mathbf{A})$ is actually made up of all possible linear combinations of the columns $\mathbf{a}_j$ of $\mathbf{A}$. For this reason $\mathcal{R}(\mathbf{A})$ is also called the *column space* of $\mathbf{A}$, written as $L(\mathbf{a}_j)$. It should already be clear that for a particular $\mathbf{A}$ and $\mathbf{y}$ of Eq. (*6.1*), if $\mathbf{y} \notin L(\mathbf{a}_j)$, there is no $\mathbf{x}$ solution. Saying that $\mathbf{y} \in L(\mathbf{a}_j)$ is equivalent to saying that $\mathbf{y}$ is a linear combination of the columns of $\mathbf{A}$ and hence rank($\mathbf{A}$) = rank($\mathbf{W}$). This is the condition which is necessary for the existence of *at least one* solution $\mathbf{x}$. The columns $\mathbf{a}_j$ span the column space directly from the definition. If, moreover, these n columns form a *basis* for it, then the dimension of $L(\mathbf{a}_j)$ is n, meaning the following:

1. There is a *unique* set of n x_j coefficients for each $\mathbf{y}$ in $L(\mathbf{a}_j)$.
2. $L(\mathbf{a}_j) \equiv \mathcal{X}^n$, and hence $m = n$ is required.
3. Rank($\mathbf{A}$) = rank($\mathbf{W}$) = n.

The critical difference between *solutions* for Eq. (*6.1*) and *one unique solution* hinges upon whether the columns of $\mathbf{A}$ merely span the column space or form a basis set.

The orthogonal complement of a linear vector space such as $L(\mathbf{a}_j)$ is another linear vector space, denoted by $L(\mathbf{a}_j)^\perp$. $\mathcal{X}^m$ can be written as the direct sum $\mathcal{X}^m = L(\mathbf{a}_j) \oplus L(\mathbf{a}_j)^\perp$. *The space* $\mathcal{X}^m$ has been decomposed into two orthogonal subspaces as promised, but a more descriptive explanation of the orthogonal complement will be given shortly. First attention is directed to $\mathcal{X}^n$. From among all $\mathbf{x} \in \mathcal{X}^n$, those for which

$\mathbf{Ax} = \mathbf{0}$ form the *null space* of **A**, written $\mathcal{N}(\mathbf{A})$. Using the orthogonal complement of $\mathcal{N}(\mathbf{A})$, we can write the direct sum $\mathcal{X}^n = \mathcal{N}(\mathbf{A}) \oplus \mathcal{N}(\mathbf{A})^{\perp}$. The two orthogonal complement spaces which have been introduced can be given more meaning by considering the adjoint operator $\overline{\mathbf{A}}^T$ mapping from $\mathcal{X}^m$ to $\mathcal{X}^n$. Let $\mathbf{c}_j$ be the columns of $\overline{\mathbf{A}}^T$, that is, the conjugates of the *rows* of **A**. Then the adjoint mapping $\overline{\mathbf{A}}^T\mathbf{y} = \mathbf{x}$ can be written in column-partitioned form as

$$y_1\mathbf{c}_1 + y_2\mathbf{c}_2 + \cdots + y_m\mathbf{c}_m = \mathbf{x}$$

All **x** vectors in the range space of $\overline{\mathbf{A}}^T$ are linear combinations of the columns $\mathbf{c}_j$; hence $\mathcal{R}(\overline{\mathbf{A}}^T)$ is frequently called the *row-space* of **A**, sometimes written $L(\mathbf{c}_j)$. Problem 6.21 shows this to be precisely the same space as the orthogonal complement of the null space of **A**, so $\mathcal{X}^n$ is the direct sum of the null space of **A** and the row space of **A**.

Returning to the space $\mathcal{X}^m$, those vectors **y** for which $\overline{\mathbf{A}}^T\mathbf{y} = \mathbf{0}$ form the null space of $\overline{\mathbf{A}}^T$. This space is frequently called the *left null space* of **A** because the conjugate transpose of the previous equation gives $\overline{\mathbf{y}}^T\mathbf{A} = \mathbf{0}$. Vectors **y** satisfying this relation are called the *left null vectors* of **A**, and the left null space of **A** is the space of all left null vectors. This space is precisely the same space introduced earlier as the orthogonal complement of $L(\mathbf{a}_j)$. Therefore, $\mathcal{X}^m$ is the direct sum of the column space of **A** and the left null space of **A**.

To summarize the results of this section, it has been found that every $m \times n$ matrix **A** has four important vector spaces associated with it. These are

The column space $L(\mathbf{a}_j) \equiv \mathcal{R}(\mathbf{A})$
The null space $\mathcal{N}(\mathbf{A})$
The row space $L(\mathbf{c}_i) \equiv \mathcal{R}(\overline{\mathbf{A}}^T)$
The left null space $\mathcal{N}(\overline{\mathbf{A}}^T)$

The primary vector spaces $\mathcal{X}^n$ and $\mathcal{X}^m$ can be written as direct sums, as suggested pictorially in Figure 6.1. For a particular matrix **A**, some of these subspaces may be zero-dimensional—that is, contain only the zero vector.

6.3 THE ROW-REDUCED ECHELON FORM OF A MATRIX

The rank of certain matrices plays a vital role in the above discussion and in many other contexts in modern control theory. An efficient method of determining the rank of a matrix is to put the matrix into *row-reduced echelon* form [1]. This form is obtained

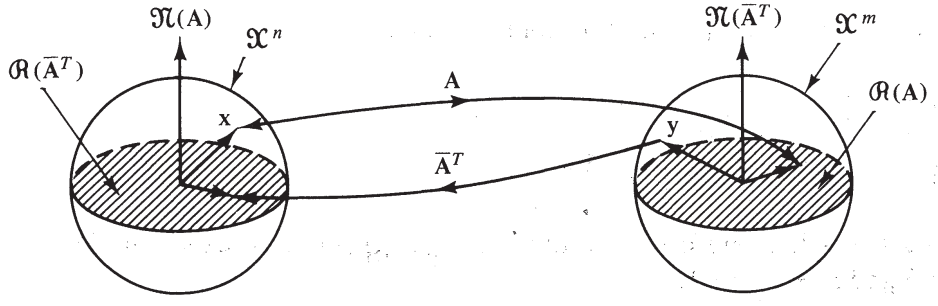


Figure 6.1

by using elementary row operations, just as described in Problem 4.18, page 147, for Gaussian elimination. Elementary row operations are performed until the first nonzero term in each row is unity and all terms below these ones are zero. This form of the matrix is the *echelon form,* but the matrix is not yet in row-reduced echelon form. Even in this intermediate form, the rank of the matrix is obvious by inspection. It is the number of nonzero rows in the matrix. Recall from Chapter 4 that the elementary matrices are nonsingular, and therefore multiplication of a matrix by them does not change the rank of that matrix.

Additional row operations are carried out until the leading ones in each row are the only nonzero terms in their respective columns. That is, any nonzero terms *above* the leading ones of the echelon form are now removed. The result is the desired row-reduced echelon form. Every matrix has a *unique* row-reduced echelon form. Some texts refer to this as the Hermite normal form of the matrix. The rank of a matrix is certainly obvious from its row-reduced echelon form, but its usefulness goes far beyond that. Any nonsingular **A** matrix just reduces to the unit matrix **I.** Therefore, aside from confirming the rank, it seems that all the information in **A** is "lost" by this reduction. The usefulness of the technique usually comes about by applying the reduction to some matrix other than just the coefficient matrix **A** in a set of simultaneous equations. If it is applied to the composite matrix $\mathbf{W} = [\mathbf{A} \mid \mathbf{y}]$ defined before, and if the resultant row-reduced echelon form is called $\mathbf{W}' = [\mathbf{A}' \mid \mathbf{y}']$, then

$$\text{rank } \mathbf{W} = \text{rank } \mathbf{W}' \quad \text{and} \quad \text{rank } \mathbf{A} = \text{rank } \mathbf{A}'$$

so that inspection of $\mathbf{W}'$ reveals instantly which of the previous categories applies. Further, when solutions do exist they are obtained directly from $\mathbf{W}'$ with little or no additional effort.

EXAMPLE 6.1 In the following nine situations a set of simultaneous equations of the form $\mathbf{Ax} = \mathbf{y}$ is being considered. In each case the **W** matrix containing **A** and **y** is displayed, followed by the row-reduced echelon form $\mathbf{W}'$. These are then used to draw conclusions about the original set of equations. Note that in cases 1, 2, and 3 the number of equations, m, is equal to the number of unknowns in **x**, n. In cases 4 and 5 m is less than n and in cases 6 through 9 m is greater than n.

1.

$$\mathbf{W} = \left[\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & 1 & 1 & -1 \\ 1 & 0 & 1 & 5 \end{array}\right] \qquad \mathbf{W}' = \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

From the reduced form, $r_A = 2$ and $r_W = 3$, so no solutions exist.

2.

$$\mathbf{W} = \left[\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & 1 & 1 & -1 \\ 1 & 0 & 1 & 3 \end{array}\right] \qquad \mathbf{W}' = \left[\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

From this, $r_A = 2 = r_W < n = 3$. Therefore, an infinite number of solutions exist, and $\mathbf{W}'$ tells us that $\mathbf{x}_1 + \mathbf{x}_3 = 3$ and $\mathbf{x}_2 + \mathbf{x}_3 = -1$.

3.

$$\mathbf{W} = \left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 7 \\ 3 & 2 & 1 & 1 \end{array}\right] \qquad \mathbf{W}' = \left[\begin{array}{ccc|c} 1 & 0 & 0 & 2.125 \\ 0 & 1 & 0 & -4.5 \\ 0 & 0 & 1 & 3.625 \end{array}\right]$$

Thus $r_A = r_W = n = 3$, so a unique solution exists and it is just $\mathbf{y}'$.

4.

$$\mathbf{W} = \left[\begin{array}{ccc|c} 1 & -1 & 2 & 8 \\ -1 & 2 & 0 & 2 \end{array}\right] \qquad \mathbf{W}' = \left[\begin{array}{ccc|c} 1 & 0 & 4 & 18 \\ 0 & 1 & 2 & 10 \end{array}\right]$$

Here $r_A = r_W = 2 < n$, so solutions exist but are not unique. They all must satisfy $\mathbf{x}_1 + 4\mathbf{x}_3 = 18$ and $\mathbf{x}_2 + 2\mathbf{x}_3 = 10$.

5.

$$\mathbf{W} = \left[\begin{array}{cccc|c} 1 & 2 & 3 & 1 & 1 \\ -4 & 5 & 1 & 9 & 2 \\ -2 & 8 & 6 & 10 & 3 \end{array}\right] \qquad \mathbf{W}' = \left[\begin{array}{cccc|c} 1 & 0 & 1 & -1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

This case has $r_A = 2, r_W = 3$, so no solutions exist.

6.

$$\mathbf{W} = \left[\begin{array}{cc|c} 1 & 2 & 2 \\ 3 & 4 & 3 \\ 5 & 6 & 4 \end{array}\right] \qquad \mathbf{W}' = \left[\begin{array}{cc|c} 1 & 0 & -1 \\ 0 & 1 & 1.5 \\ 0 & 0 & 0 \end{array}\right]$$

Since $r_A = r_W = 2 = n$, there is a unique solution given by the first two components of $\mathbf{y}'$ namely $\mathbf{x} = [-1 \quad 1.5]^T$.

7.

$$\mathbf{W} = \left[\begin{array}{cc|c} 1 & 2 & 2 \\ 3 & 4 & 3 \\ 5 & 6 & -4 \end{array}\right] \qquad \mathbf{W}' = \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]$$

Note that this is the same as case 6 except for the sign of one component of **y**. The results are quite different. Since $r_A = 2$ and $r_W = 3$, there is no solution.

8.

$$\mathbf{W} = \left[\begin{array}{ccc|c} 1 & 3 & 5 & 3 \\ 1 & 4 & 6 & 3.5 \\ -1 & 5 & 3 & 1 \\ -1 & 4 & 2 & 0.5 \\ 1 & 3 & 5 & 3 \end{array}\right] \qquad \mathbf{W}' = \left[\begin{array}{ccc|c} 1 & 0 & 2 & 1.5 \\ 0 & 1 & 1 & 0.5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Here, even though there are more equations than unknowns, there are still solutions—in fact, an infinite number—all of which satisfy $\mathbf{x}_1 + 2\mathbf{x}_3 = 1.5$ and $\mathbf{x}_2 + \mathbf{x}_3 = 0.5$.

9. Changing only the **y** vector of the previous case to $[3 \quad 3 \quad 1 \quad 1 \quad 3]^T$ leads to

$$\mathbf{W}' = \left[\begin{array}{ccc|c} 1 & 0 & 2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Since $r_A = 2$ and $r_W = 3$, there is no solution. In addition to demonstrating the various categories of simultaneous equations and illustrating row-reduced echelon forms, this

example is intended to show that characterizations strictly in terms of numbers of equations and numbers of unknowns are clearly inadequate. ■

6.3.1 Applications to Polynomial Matrices

The previous examples of obtaining row-reduced echelon matrices by use of elementary operations all involved matrices defined over the real numbers. This can be applied to complex numbers as well. The notions of elementary matrices $\mathbf{E}_p(\alpha)$ and $\mathbf{E}_{p,q}(\alpha)$ in Sec. 4.10 are valid for α belonging to *any* scalar number field, including the rational polynomial functions. However, for many purposes, when dealing with polynomial matrices it is desirable to redefine slightly the elementary matrices and hence the allowed elementary operations. The purpose is to ensure that the results of our restricted elementary operations remain polynomial matrices (and not matrices of *ratios* of polynomials). We call the modified elementary matrices the *polynomial-restricted* elementary matrices, defined as follows:

1. The elementary matrix $\mathbf{E}_{p,q}$ to interchange rows or columns is not a function of α and requires no change.
2. The row (or column) multiplier $\mathbf{E}_p(\alpha)$ will be restricted to either real or complex α. Disallowing polynomial α ensures that $\mathbf{E}_p(\alpha)^{-1}$ is still an elementary matrix in the restricted sense. If α were a polynomial, the inverse would involve *ratios* of polynomials.
3. The matrix $\mathbf{E}_{p,q}(\alpha)$, which adds α times the pth row (column) to the qth row (column), does allow α to be a polynomial. Note that $|\mathbf{E}_{p,q}(\alpha)|$ is not a function of α, so that the inverse remains a polynomial-restricted elementary matrix.

The polynomial-restricted elementary operations (row or column) can be carried out by pre- or postmultiplication by the appropriate $\mathbf{E}$ matrix, as before. The notion of the row-reduced echelon form also must be modified in the case of polynomial matrices. The more general term Hermite form will be used to denote a matrix in which

1. The first nonzero entry in a row is a *monic* polynomial, that is, a polynomial in which the coefficient of the highest power is unity.
2. All terms in the same column and *below* this leading monic polynomial are zero.
3. All terms in this same column and *above* the leading term are polynomials of degree less than the degree of the leading monic polynomial.

Note that this reduces to the previous definition of the row-reduced echelon form when the leading monic polynomials are all of degree zero, i.e., just the scalar 1. It is understood that lower-degree polynomials are scalar zeros in this case. A more standard interpretation of a polynomial of degree less than zero (such as degree of -1) would lead us outside the realm of polynomials and into rational polynomial functions, which is what we are trying to avoid. Three examples of polynomial matrices in Hermite normal form are

$$\begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 15 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} s^2+2s+1 & 3 & s \\ 0 & s+1 & s^2+2s+1 \\ 0 & 0 & s^3+4s \end{bmatrix}, \quad \begin{bmatrix} s & 5s+1 \\ 0 & s^2+3s \\ 0 & 0 \end{bmatrix}$$

EXAMPLE 6.2 Use polynomial-restricted elementary operations to reduce $\mathbf{P}(s) = \begin{bmatrix} s+1 & s & 5 \\ s-1 & s^2+3s+2 & s \end{bmatrix}$ to Hermite normal form. A sequence of elementary row operations modifies $\mathbf{P}(s)$ successively to

$$\begin{bmatrix} s+1 & s & 5 \\ -2 & s^2+2s+2 & s-5 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -0.5(s^2+2s+2) & -0.5(s-5) \\ s+1 & s & 5 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 1 & -0.5(s^2+2s+2) & -0.5(s-5) \\ 0 & s^3+3s^2+6s+2 & s^2-4s+5 \end{bmatrix}$$

Some obvious intermediate steps have not been given. The actual sequence of elementary operations used was

$$\mathbf{E}_2(2)\mathbf{E}_{1,2}(-s-1)\mathbf{E}_{1,2}\mathbf{E}_2(-0.5)\mathbf{E}_{1,2}(-1) = \begin{bmatrix} 0.5 & -0.5 \\ -s+1 & s+1 \end{bmatrix}$$

The order of application of the elementary matrices was right to left; that is, $\mathbf{E}_{1,2}(-1)$ was used first, then $\mathbf{E}_2(-0.5)$, and so on. ■

The elementary operations used in the reduction can be systematically applied [2]. The first nonzero term in each row is usually the diagonal term. On a column-by-column basis, the terms below the diagonal must be reduced to zero. For the general column j, assume that at some point in the reduction process polynomials $p_1(s)$ and $p_2(s)$ are in the jj and ij positions, with $i > j$. Row interchanges can be used to ensure that degree $(p_1) \leq$ degree(p_2). Standard long division gives $p_2(s)/p_1(s) = q(s) + r(s)/p_1(s)$, where q and r are quotient and remainder polynomials. Therefore, $p_2(s) - p_1(s)q(s) = r(s)$. Hence, premultiplication of the matrix by $\mathbf{E}_{j,i}(-q(s))$ will reduce the ij term from $p_2(s)$ to the lower-degree $r(s)$. This procedure can be applied to each nonzero term below the diagonal. A row interchange can then be used to bring the minimum degree nonzero remainder to the jj diagonal, and the whole column-reduction process can be repeated. If a constant remainder (polynomial of degree zero) is ever found, that term is placed on the diagonal, normalized to unity, and then used immediately to reduce all other terms in that column to zero. If all terms below the diagonal are reduced to zero while the diagonal term remains a finite-degree polynomial, then the terms above the diagonal need not be zero. However, they must be reduced to polynomials of a lower degree than the diagonal. This can be done using the same long-division procedure as before, leaving only the remainder terms above the diagonal. Although simple in concept, this reduction to Hermite normal form can be algebraically tedious. To demonstrate this, the reader should verify that

$$\mathbf{P}(s) = \begin{bmatrix} 3s+2 & 6 \\ s & s \\ s^2-1 & s+5 \end{bmatrix} \quad \text{can be reduced to} \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

by a series of 11 elementary row operations.

6.3.2 Application to Matrix Fraction Description of Systems

The polynomial-restricted elementary matrices can be used systematically to reduce MFDs of transfer function matrices $\mathbf{H}(s)$. (See Eq. (*4.7*) and Problems 4.29 through 4.31.) The left-divisor form $\mathbf{H}(s) = \mathbf{P}_1^{-1}(s)\mathbf{N}_1(s)$ can be written as $\mathbf{P}_1(s)\mathbf{H}(s) = \mathbf{N}_1(s)$. Premultiplying by a sequence of elementary matrices leaves $\mathbf{H}(s)$ unchanged and therefore is equivalent to performing elementary row operations on $\mathbf{W} \equiv [\mathbf{P}_1(s) \,\vdots\, \mathbf{N}_1(s)]$ to obtain $[\mathbf{T}(s)\mathbf{P}_1(s) \,\vdots\, \mathbf{T}(s)\mathbf{N}_1(s)] = [\underline{\mathbf{P}}_1(s) \,\vdots\, \underline{\mathbf{N}}_1(s)]$. The matrix $\mathbf{T}(s)$ is the product of elementary matrices and thus is invertible. Its inverse is still a polynomial matrix because of the restrictions placed on the polynomial-restricted elementary matrices. Therefore, $[\mathbf{P}_1(s) \,\vdots\, \mathbf{N}_1(s)] = [\mathbf{T}^{-1}\underline{\mathbf{P}}_1 \,\vdots\, \mathbf{T}^{-1}\underline{\mathbf{N}}_1]$. This shows that the matrix $\mathbf{T}^{-1}$ is a common factor of both $\mathbf{P}_1$ and $\mathbf{N}_1$, on the left. More commonly, $\mathbf{T}$ is called a *left common divisor* of $\mathbf{P}_1$ and $\mathbf{N}_1$. It is clear that the common divisor cancels, leaving a reduced MFD with the same original transfer function,

$$\mathbf{H}(s) = [\mathbf{T}^{-1}\underline{\mathbf{P}}_1]^{-1}[\mathbf{T}^{-1}\underline{\mathbf{N}}_1] = \underline{\mathbf{P}}_1^{-1}\underline{\mathbf{N}}_1$$

Similarly, starting with the right-hand divisor form of the MFD

$$\mathbf{H}(s) = \mathbf{N}_2(s)\mathbf{P}_2^{-1}(s)$$

it is clear that postmultiplication of $\mathbf{H}(s)\mathbf{P}_2(s) = \mathbf{N}_2(s)$ by elementary matrices leaves $\mathbf{H}(s)$ unchanged and is equivalent to elementary column operations on

$$\mathbf{Y} \equiv \begin{bmatrix} \mathbf{N}_2(s) \\ \hline \mathbf{P}_2(s) \end{bmatrix}$$

This leads to the notion of right common divisors for $\mathbf{N}_2$ and $\mathbf{P}_2$. In either case, these concepts are generalizations of the notion of pole-zero cancellations in scalar transfer functions. When the elementary row (column) operations are carried out on $\mathbf{W}$ (or $\mathbf{Y}$) to the limit—i.e., until the Hermite normal form is reached—the product $\mathbf{T}$ of the elementary matrices used will represent the *greatest* common left (or right) divisor. The resulting transfer function representations $\mathbf{H}(s) = \underline{\mathbf{P}}_1^{-1}(s)\underline{\mathbf{N}}_1(s) = \underline{\mathbf{N}}_2(s)\underline{\mathbf{P}}_2^{-1}(s)$ are maximally reduced in the sense that greatest common divisors have been removed and no further common polynomial factors can be canceled. This is a very important consideration in several instances when analyzing multiple-input-output systems within the transfer function and polynomial matrix domain. This will be useful in Chapter 13. The major emphasis in this book is on state variable representation of systems. Even here, the polynomial matrix representations often play a role in obtaining appropriate state models, as is shown in Chapter 12.

6.4 SOLUTION BY PARTITIONING

It is assumed in this section that $r_A = r_W$, so that one or more solutions exist. By definition, the $m \times n$ coefficient matrix $\mathbf{A}$ contains a nonsingular $r_A \times r_A$ matrix. The original equations $\mathbf{Ax} = \mathbf{y}$ can always be rearranged and partitioned into

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \hline \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \hline \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \hline \mathbf{y}_2 \end{bmatrix}$$

where $\mathbf{A}_1$ is $r_A \times r_A$ and nonsingular. Depending on the relation between m, n, and r_A, some of the terms in the partitioned equation will not be required. For example, if $m = n = r_A$, then $\mathbf{A}_1 = \mathbf{A}$, $\mathbf{x}_1 = \mathbf{x}$, and $\mathbf{y}_1 = \mathbf{y}$. The general case is treated here, and then

$$\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 = \mathbf{y}_1 \quad \text{or} \quad \mathbf{x}_1 = \mathbf{A}_1^{-1}[\mathbf{y}_1 - \mathbf{A}_2\mathbf{x}_2]$$

The *degeneracy* of $\mathbf{A}$ is $q_A = n - r_A$. The values of the q_A components of $\mathbf{x}_2$ are completely arbitrary and generate the q_A parameter family of solutions for $\mathbf{x}$ mentioned on page 208, case 2(b). If $r_A = n$, as in case 2(a), then $\mathbf{A}_2$, $\mathbf{A}_4$, and $\mathbf{x}_2$ will not be present in the partitioned equation. In that case the unique solution is $\mathbf{x} = \mathbf{A}_1^{-1}\mathbf{y}_1$. If in addition $m = n$, then $\mathbf{A}_3$ and $\mathbf{y}_2$ will not be present and $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$. This is the simple case mentioned in Chapter 4, and $\mathbf{x}$ could be computed by using Cramer's rule or various matrix inversion techniques. However, Gaussian elimination or similar reduction techniques are more efficient for large values of n.

EXAMPLE 6.3 Consider once more the situation of case 2 in Example 6.1, and find all solutions $\mathbf{x}$ for

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \\ 3 \end{bmatrix}$$

Here $r_A = 2$ (rows 1 and 3 are identical) and $r_W = 2$ also. An infinite set of solutions exists. Let

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{x}_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \mathbf{y}_1 = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

$$\mathbf{A}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \mathbf{x}_2 = x_3 \qquad \mathbf{y}_2 = 3$$

Then

$$\mathbf{x}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \left\{ \begin{bmatrix} 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} x_3 \right\} = \begin{bmatrix} 3 - x_3 \\ -1 - x_3 \end{bmatrix}$$

The one-parameter family of solutions is $\mathbf{x} = [3 - x_3 \quad -1 - x_3 \quad x_3]^T$ with x_3 arbitrary. ■

6.5 A GRAM-SCHMIDT EXPANSION METHOD OF SOLUTION

The set of m simultaneous equations in n unknowns $\mathbf{x}$ is again considered.

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{6.1}$$

No special assumptions are made at the outset about r_A and r_W relative to each other or to m and n. By definition, there are r_A independent a_j columns in the matrix $\mathbf{A}$. These vectors can be used as a basis set for a linear vector space $L(a_j)$ called the *column space* of $\mathbf{A}$. The number of vectors in this basis set could be n or any smaller positive integer in a given case. In addition to these r_A vectors a_j, the $\mathbf{y}$ vector is considered, giving a set of $r_A + 1$ vectors. The Gram-Schmidt procedure is used on this set to form an orthonormal basis set $\{\hat{\mathbf{v}}_j\}$. The only possible exception is the last vector $\hat{\mathbf{v}}_{r_A+1}$. Since $\mathbf{y}$ may be linearly dependent on the columns $\mathbf{a}_j$, it might not be possible to form a nonzero

vector from $\mathbf{y}$ which is orthogonal to all the $\mathbf{a}_j$ vectors. If $\mathbf{y}$ is linearly dependent on the $\mathbf{a}_j$ vectors, then the unnormalized vector $\hat{\mathbf{v}}_{r_A+1}$ will automatically come out zero during the Gram-Schmidt construction. Since the zero vector is orthogonal to every other vector, an orthogonal set of $\{\hat{\mathbf{v}}_j\}$ can thus be constructed in all cases. Each vector in the set is a unit vector with the possible exception of a zero vector as the last entry. Form the $m \times (r_A + 1)$ matrix $\mathbf{V}$ from the set. Premultiplying Eq. (*6.1*) by V^T is equivalent to premultiplying the previously defined $\mathbf{W}$ matrix. The result is

$$\mathbf{V}^T\mathbf{W} = \left[\begin{array}{ccccccc|c}
\langle\hat{\mathbf{v}}_1,\mathbf{a}_1\rangle & \langle\hat{\mathbf{v}}_1,\mathbf{a}_2\rangle & & \cdots & \langle\hat{\mathbf{v}}_1,\mathbf{a}_r\rangle & \langle\hat{\mathbf{v}}_1,\mathbf{a}_{r+1}\rangle \cdots & \langle\hat{\mathbf{v}}_1,\mathbf{a}_n\rangle & \langle\hat{\mathbf{v}}_1,\mathbf{y}\rangle \\
0 & \langle\hat{\mathbf{v}}_2,\mathbf{a}_2\rangle & & \cdots & \langle\hat{\mathbf{v}}_2,\mathbf{a}_r\rangle & \langle\hat{\mathbf{v}}_2,\mathbf{a}_{r+1}\rangle \cdots & \langle\hat{\mathbf{v}}_2,\mathbf{a}_n\rangle & \langle\hat{\mathbf{v}}_2,\mathbf{y}\rangle \\
0 & 0 & \langle\hat{\mathbf{v}}_3,\mathbf{a}_3\rangle & \cdots & \langle\hat{\mathbf{v}}_3,\mathbf{a}_r\rangle & \langle\hat{\mathbf{v}}_3,\mathbf{a}_{r+1}\rangle \cdots & \langle\hat{\mathbf{v}}_3,\mathbf{a}_n\rangle & \langle\hat{\mathbf{v}}_3,\mathbf{y}\rangle \\
\vdots & \vdots & \vdots & & & & & \\
0 & 0 & 0 & \cdots & \langle\hat{\mathbf{v}}_r,\mathbf{a}_r\rangle & \langle\hat{\mathbf{v}}_r,\mathbf{a}_{r+1}\rangle \cdots & \langle\hat{\mathbf{v}}_r,\mathbf{a}_n\rangle & \langle\hat{\mathbf{v}}_r,\mathbf{y}\rangle \\
\hline
0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \langle\hat{\mathbf{v}}_{r+1},\mathbf{y}\rangle
\end{array}\right]$$

(r rows; $r+1$ rows)

(r columns; $n-r$ columns; $n+1$ columns)

In writing this semitriangular form it is assumed that the first r columns of $\mathbf{A}$ are the r_A independent ones used in the Gram-Schmidt process. The entire last row of the above matrix will be zero with the possible exception of the very last term $\langle\hat{\mathbf{v}}_{r+1},\mathbf{y}\rangle$. If $\mathbf{y}$ is dependent on the columns of $\mathbf{A}$, then this term will be zero, since then $\hat{\mathbf{v}}_{r+1}$ is exactly zero. This is the case for which there are solutions, since then $r_A = r_W$. When this last inner product is not zero, $r_W > r_A$, so no solutions exist. The last inner product is the component of $\mathbf{y}$ normal to the column space of $\mathbf{A}$. There is no $\mathbf{x}$ vector which will cause $\mathbf{Ax}$ to equal this part of $\mathbf{y}$. The vector $\mathbf{y}$ can be decomposed into a component $\mathbf{y}_p$ parallel to $L(\mathbf{a}_j)$ and a component $\mathbf{y}_e$ normal to $L(\mathbf{a}_j)$, $\mathbf{y} = \mathbf{y}_p + \mathbf{y}_e$. See Figure 6.2.

The best that can be done by choice of $\mathbf{x}$ is to force $\mathbf{Ax} = \mathbf{y}_p$. The unavoidable error committed in doing this is the residual $\mathbf{y} - \mathbf{Ax} = \mathbf{y}_p + \mathbf{y}_e - \mathbf{Ax} = \mathbf{y}_e$. The length or norm of this residual error is the lower corner element in $\mathbf{W}'$, namely $\|\mathbf{y}_e\| = \langle\hat{\mathbf{v}}_{r+1},\mathbf{y}\rangle$. Thus, a glance at $\mathbf{W}'$ tells whether or not solutions exist, and if they do not, then the magnitude of the smallest possible error is also given.

The solution which satisfies $\mathbf{y}$, or $\mathbf{y}_p$ if necessary, is found from $\mathbf{W}'$ as in Section

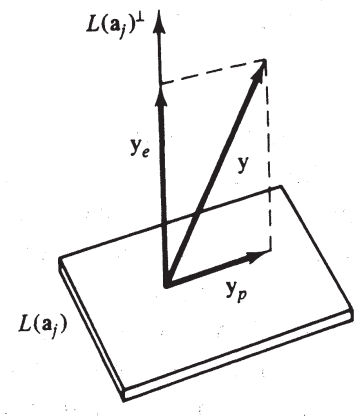


Figure 6.2

6.3, except the final row is ignored. This represents r_A equations in n unknowns. At least one solution always exists, and it will be unique if and only if $r_A = n$. When more than one solution exists, some additional criterion may be used to select one particular solution. These underdetermined problems are discussed further in Section 6.7. In cases where $\mathbf{y}$ has a nonzero component normal to $L(\mathbf{a}_j)$, the procedure given here leads to the least-squares solution (or solutions). This topic is pursued further in Section 6.8.

EXAMPLE 6.4 Analyze the following set of equations using the Gram-Schmidt expansion method (GSE).

$$\begin{bmatrix} 1 & 3 & 2 \\ 2 & 5 & 3 \\ 3 & 7 & 4 \\ 4 & 9 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}$$

In this example it is easy to see that column 3 of $\mathbf{A}$ is the difference between columns 2 and 1, so $\mathbf{A}$ has rank 2. Columns 1 and 2 of $\mathbf{A}$ are used, along with $\mathbf{y}$, to form the following orthonormal set:

THE ORTHONORMAL BASIS SET

$$\mathbf{V} = \begin{bmatrix} 1.8257418\text{E}-01 & 8.1649667\text{E}-01 & 3.6514840\text{E}-01 \\ 3.6514837\text{E}-01 & 4.0824848\text{E}-01 & -1.8257420\text{E}-01 \\ 5.4772252\text{E}-01 & 5.8400383\text{E}-07 & -7.3029685\text{E}-01 \\ 7.3029673\text{E}-01 & -4.0824789\text{E}-01 & 5.4772240\text{E}-01 \end{bmatrix}$$

Then $\mathbf{V}^T\mathbf{W} = \mathbf{W}'$.

THE EXPANSION COEFFICIENT VECTOR(S)

$$\mathbf{W}' = \begin{bmatrix} 5.4772258\text{E}+00 & 1.2780193\text{E}+01 & 7.3029675\text{E}+00 & 3.6514840\text{E}-01 \\ 3.9339066\text{E}-06 & 8.1650591\text{E}-01 & 8.1650186\text{E}-01 & 4.0824819\text{E}-01 \\ -9.5367432\text{E}-07 & -2.3841858\text{E}-06 & -1.1920929\text{E}-06 & 1.6431677\text{E}+00 \end{bmatrix}$$

Since the first three entries in the last row and the first entry in row 2 are theoretically zero, a "machine zero" can be defined. Here any number of magnitude less than 4×10^{-6} is set to zero.

From this it is seen that

1. *There is no solution to the given set of equations, since $r_A = 2$ and $r_W = 3$.* The equations are inconsistent.

2. *The best that can be done is to satisfy the column space portion of the equations.* Let this be called a projected solution. If this is done the residual error will have the norm

$$\|\mathbf{Ax} - \mathbf{y}\| = 1.6431677$$

3. *There are an infinite number of projected solutions, all of which will give exactly this same residual error norm.* They all must satisfy (rounded)

$$0.8165(x_2 + x_3) = 0.40825$$

$$5.4772x_1 + 12.7802x_2 + 7.3030x_3 = 0.36515$$

If the row-reduced-echelon (RRE) method of Section 6.3 is applied to this problem instead, the resulting matrix $\mathbf{W}'$ is

$$\mathbf{W}' = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

This also indicates that the equations are inconsistent but gives no clue about how closely a solution can be approached. It also indicates, apparently, that if one row of inconsistent equations could be ignored, an infinite set of solutions would exist, and they would all satisfy $x_1 - x_3 = 0$ and $x_2 + x_3 = 0$. Although the RRE method has yielded somewhat less information than the GSE method, it has yielded the one-dimensional null space spanned by $\mathbf{e} = [1 \quad -1 \quad 1]^T$. It can be shown that any constant α times $\mathbf{e}$ can be added to any projected solution $\mathbf{x}_p$ found from the GSE method (or any other method) and the result will still be a projected solution. That is, $\mathbf{x}_p + \alpha\mathbf{e} = \mathbf{x}$ is also a projected solution. ■

6.6 HOMOGENEOUS LINEAR EQUATIONS

The set of homogeneous equations $\mathbf{Ax} = \mathbf{0}$ always has at least one solution, $\mathbf{x} = \mathbf{0}$. This is true because $\mathbf{A}$ and $\mathbf{W}$ always have the same rank. However, $\mathbf{x} = \mathbf{0}$ is called the *trivial solution.* In order for *nontrivial solutions* to exist, it must be true that $r_A < n$. Of course, if one such nontrivial solution exists, there will be an infinite set of solutions with $n - r_A$ free parameters. The methods of the previous sections apply to the homogeneous case without modification.

It is pointed out that any set of nonhomogeneous equations $\mathbf{Ax} = \mathbf{y}$ can always be written as an equivalent set of homogeneous equations:

$$[\mathbf{A} \mid \mathbf{y}]\begin{bmatrix} \mathbf{x} \\ \hline -1 \end{bmatrix} = \mathbf{0} \quad \text{or} \quad \mathbf{W}\begin{bmatrix} \mathbf{x} \\ \hline -1 \end{bmatrix} = \mathbf{0}$$

EXAMPLE 6.5 Find all nontrivial solutions to the equations $\mathbf{Ax} = \mathbf{0}$, if $\mathbf{A} = \begin{bmatrix} 0 & 2 & 1 \\ 0 & 2 & 1 \\ 0 & -4 & -2 \end{bmatrix}$. The matrix $\mathbf{W}$ and its RRE form $\mathbf{W}'$ are

THE MATRIX W

$$\begin{bmatrix} 0.0000000\text{E}+00 & 2.0000000\text{E}+00 & 1.0000000\text{E}+00 & 0.0000000\text{E}+00 \\ 0.0000000\text{E}+00 & 2.0000000\text{E}+00 & 1.0000000\text{E}+00 & 0.0000000\text{E}+00 \\ 0.0000000\text{E}+00 & -4.0000000\text{E}+00 & -2.0000000\text{E}+00 & 0.0000000\text{E}+00 \end{bmatrix}$$

RANK OF W IS 1: THE HERMITE FORM W' FOLLOWS

$$\begin{bmatrix} 0.0000000\text{E}+00 & 1.0000000\text{E}-00 & 5.0000000\text{E}-01 & 0.0000000\text{E}+00 \\ 0.0000000\text{E}+00 & 0.0000000\text{E}-00 & 0.0000000\text{E}-00 & 0.0000000\text{E}+00 \\ 0.0000000\text{E}+00 & 0.0000000\text{E}-00 & 0.0000000\text{E}-00 & 0.0000000\text{E}+00 \end{bmatrix}$$

Therefore, all nontrivial solutions are linear combinations of the following two vectors, which constitute a basis set for the null space of $\mathbf{A}$.

DEGENERACY OF A IS 2; NULL SPACE BASIS IS

$$\begin{bmatrix} -1.0000000\text{E}+00 \\ 0.0000000\text{E}+00 \\ 0.0000000\text{E}+00 \end{bmatrix} \begin{bmatrix} 0.0000000\text{E}+00 \\ 5.0000000\text{E}-01 \\ -1.0000000\text{E}+00 \end{bmatrix}$$

■

6.7 THE UNDERDETERMINED CASE

When the matrix **A** has $m < n$, there is no possibility of a unique solution **x.** The underdetermined case, which has an infinite number of solutions ($r_A = r_W$), is discussed here. The methods of Sections 6.3, 6.4, or 6.5 can be used to find the family of solutions. This section presents a method for singling out one particular solution, the solution with the minimum norm, $\|\mathbf{x}\|$. Problem 6.47 suggests that the procedure can be generalized to various other weighted norms.

The Minimum Norm Solution

Consider the equation $\mathbf{Ax} = \mathbf{y}$, where **A** is $m \times n$ with $m < n$ and with $r_A = r_W$. If $r_A < m$, some rows of **W** are linearly dependent. This means that some of the original equations are redundant and can be deleted without losing information. Assume that these deletions have been made and as a result $r_A = r_W = m$. The conjugate transpose of the m rows of any **A** can be used to define the n-component vectors $\{\mathbf{c}_i, i = 1, \ldots, m\}$. These vectors belong to $\mathcal{X}^n$. The space spanned by the set of $\mathbf{c}_i$ vectors is called the *row space* $L(\mathbf{c}_i)$ of **A.** In general, $L(\mathbf{c}_i)$ will be a subspace of $\mathcal{X}^n$, and here $r_A = m$ means that it is an m-dimensional subspace with the $\mathbf{c}_i$ vectors forming a basis. The space $\mathcal{X}^n$, which contains all possible **x** vectors, can be written as the direct sum

$$\mathcal{X}^n = L(\mathbf{c}_i) \oplus L(\mathbf{c}_i)^{\perp}$$

Every vector in $\mathcal{X}^n$ can be written

$$\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 \quad \text{where } \mathbf{x}_1 \in L(\mathbf{c}_i), \mathbf{x}_2 \in L(\mathbf{c}_i)^{\perp}$$

Figure 6.3 illustrates this decomposition for $n = 3$ and $m = 2$.
The norm of **x** satisfies

$$\|\mathbf{x}\|^2 = \|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2$$

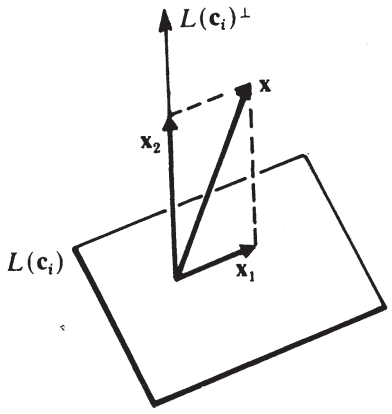


Figure 6.3

Since $\mathbf{x}_2 \in L(\mathbf{c}_i)^{\perp}$, $\langle \mathbf{c}_i, \mathbf{x}_2 \rangle = 0$ for each $\mathbf{c}_i$, so $\mathbf{A}\mathbf{x}_2 = \mathbf{0}$. Thus

$$\mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{x}_1 + \mathbf{x}_2) = \mathbf{A}\mathbf{x}_1 = \mathbf{y}$$

For every $\mathbf{x}_1 \in L(\mathbf{c}_i)$, $\mathbf{x}_1 = \sum_{i=1}^{m} \alpha_i \mathbf{c}_i = \overline{\mathbf{A}}^T \boldsymbol{\alpha}$. Using $\mathbf{A}\mathbf{x}_1 = \mathbf{y}$ gives

$$\mathbf{A}\overline{\mathbf{A}}^T \boldsymbol{\alpha} = \mathbf{y}$$

But $\mathbf{A}\overline{\mathbf{A}}^T$ is an $m \times m$ matrix with rank m and is therefore nonsingular. Solving for $\boldsymbol{\alpha}$ gives $\boldsymbol{\alpha} = (\mathbf{A}\overline{\mathbf{A}}^T)^{-1}\mathbf{y}$ and therefore $\mathbf{x}_1 = \overline{\mathbf{A}}^T \boldsymbol{\alpha} = \overline{\mathbf{A}}^T(\mathbf{A}\overline{\mathbf{A}}^T)^{-1}\mathbf{y}$. This $\mathbf{x}_1$ is the unique $\mathbf{x} \in L(\mathbf{c}_i)$, which satisfies $\mathbf{A}\mathbf{x} = \mathbf{y}$. From the norm relations, it is clear that $\mathbf{x}_1$ is the minimum norm solution, since any other solution must have a component in $L(\mathbf{c}_i)^{\perp}$ and this would increase the norm. The minimum norm solution then is

$$\mathbf{x} = \overline{\mathbf{A}}^T(\mathbf{A}\overline{\mathbf{A}}^T)^{-1}\mathbf{y}$$

This result can also be derived by using Lagrange multipliers [3] and straightforward minimization of $\|\mathbf{x}\|^2$ subject to the constraint $\mathbf{A}\mathbf{x} - \mathbf{y} = \mathbf{0}$. Although it is not pursued here, many other classes of problems amount to using the minimum (or maximum) of a cost function to single out one desirable solution from the infinite set available in the underdetermined case. If the cost function is linear, a linear programming problem results. The minimum norm problem is an example of quadratic programming. Other nonlinear programming problems can also be posed.

EXAMPLE 6.6 The minimum norm solution of $\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$ is

$$\mathbf{x}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 4 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 8 \end{bmatrix}$$

■

EXAMPLE 6.7 Find the minimum norm solution to the projected problem derived in Example 6.4. From the previous analysis the projected problem is

$$\begin{bmatrix} 5.4772 & 12.7802 & 7.3030 \\ 0 & 0.8165 & 0.8165 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.3651 \\ 0.4082 \end{bmatrix}$$

A direct calculation of $\mathbf{x} = \mathbf{A}^T[\mathbf{A}\mathbf{A}^T]^{-1}\mathbf{y}$ can be made in this simple case. Alternately one can form

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} 246.6670 & 16.3979 \\ 16.3979 & 1.3333 \end{bmatrix}$$

and then solve $\mathbf{A}\mathbf{A}^T\mathbf{x}' = \mathbf{y}$ for $\mathbf{x}'$. Then, finally, $\mathbf{x} = \mathbf{A}^T\mathbf{x}'$. The advantage of doing this is that a matrix inversion is not directly needed, and a routine such as the RRE package can instead be used to solve for $\mathbf{x}'$. This is the approach used here.

$$\mathbf{W} = \left[\begin{array}{cc|c} 246.6670 & 16.3979 & 0.3651 \\ 16.3979 & 1.3333 & 0.4082 \end{array}\right] \qquad \mathbf{W}' = \left[\begin{array}{cc|c} 1 & 0 & -0.10346 \\ 0 & 1 & 1.57862 \end{array}\right]$$

Therefore, $\mathbf{x}' = \begin{bmatrix} -0.10346 \\ 1.57862 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} -0.5667 \\ -0.0334 \\ 0.5334 \end{bmatrix}$. ■

6.8 THE OVERDETERMINED CASE

When there are more equations than unknowns, the $m \times n$ coefficient matrix $\mathbf{A}$ has $m > n$. If the equations are inconsistent, no solution exists. This situation often arises because of inaccuracies in measuring the components of the $\mathbf{y}$ vector, or because the relationship assumed to exist between $\mathbf{x}$ and $\mathbf{y}$, as expressed by $\mathbf{A}$, is oversimplified or wrong. *Approximate* solution vectors $\mathbf{x}$ are desired in this case. Three approaches are presented. The first method ignores some equations and places total reliance on those remaining. The second method (least squares) places equal reliance on all equations with the hope that the errors will average out. The third method (weighted least squares) uses all of the equations but weights some more heavily than others. An alternative computational procedure (recursive weighted least squares) is also given for obtaining the latter two approximations.

A considerable amount of information, which is useful for the overdetermined case, has already been given. The GSE method of Section 6.5 applies to this case, as already demonstrated. When the problem is overdetermined, $\langle \hat{\mathbf{v}}_{r+1}, \mathbf{y} \rangle \neq 0$. The so-called projected solution is then sought. If the projected solution is nonunique, the minimum norm solution is often singled out, as was done in Example 6.4 and continued in Example 6.7. This combination of GSE plus minimum norm solution always gives a solution to Eq. (*6.1*). It is true for the underdetermined, overdetermined, or uniquely determined cases. The only difficulty that might remain on a machine solution is the ability to recognize the difference between 0 and a very small number, or the difference between vectors that are linearly dependent or nearly linearly dependent. The notion of machine zero was introduced, and an example of how it can be determined on a given problem has been given. The GSE–minimum norm solution combination has many things in common with the method of *singular value decomposition,* but there are also some unique differences.

In this section a more traditional approach to the overdetermined problem is presented. It is assumed that $\mathbf{A}$ is of full rank n and that $m > n$.

Ignore Some Equations

If a subset of n equations is selected and the remaining $m - n$ are ignored, an approximate solution can be obtained. The basis for ignoring certain equations is a subjective matter. Perhaps certain equations are more reliable for one reason or another. Perhaps other results are obviously "wild points" and can be discarded. If $\mathbf{A}_1$ is a nonsingular $n \times n$ matrix formed by deleting rows from $\mathbf{A}$ and if $\mathbf{y}_1$ is the $n \times 1$ vector obtained from $\mathbf{y}$ by deleting the corresponding elements, then a result which satisfies n of the original equations is

$$\mathbf{x} = \mathbf{A}_1^{-1} \mathbf{y}_1$$

Least-Squares Approximate Solution

If all the equations are used correctly, errors may tend to average out, and a good approximation for $\mathbf{x}$ results. Since no one $\mathbf{x}$ can satisfy all the simultaneous equations, it is inappropriate to write the equality $\mathbf{Ax} = \mathbf{y}$. Rather, an $n \times 1$ error vector $\mathbf{e}$ is introduced:

$$\mathbf{e} = \mathbf{y} - \mathbf{Ax}$$

The least-squares approach yields the one $\mathbf{x}$ which minimizes the sum of the squares of the e_i components. That is, $\mathbf{x}$ is chosen to minimize

$$\|\mathbf{e}\|^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{Ax})^T(\mathbf{y} - \mathbf{Ax})$$

The vectors $\mathbf{e}$, $\mathbf{y}$, and $\mathbf{Ax}$ all belong to $\mathcal{X}^m$. But $\mathbf{Ax}$ belongs to the *column space* of $\mathbf{A}$, the space spanned by the columns $\mathbf{a}_j$ of $\mathbf{A}$, denoted by $L(\mathbf{a}_j)$. $\mathcal{X}^m$ is decomposed, as shown in Figure 6.1, into

$$\mathcal{X}^m = L(\mathbf{a}_j) \oplus L(\mathbf{a}_j)^\perp$$

The error has a unique decomposition:

$$\mathbf{e} = \mathbf{e}_1 + \mathbf{e}_2, \qquad \mathbf{e}_1 \in L(\mathbf{a}_j), \qquad \mathbf{e}_2 \in L(\mathbf{a}_j)^\perp$$

(The vector $\mathbf{e}_2$ is the $\mathbf{y}_e$ vector of Section 6.5). The norm of $\mathbf{e}$ satisfies $\|\mathbf{e}\|^2 = \|\mathbf{e}_1\|^2 + \|\mathbf{e}_2\|^2$.

Since $\mathbf{y}$ is given and since $\mathbf{Ax} \in L(\mathbf{a}_j)$, the choice of $\mathbf{x}$ cannot affect $\mathbf{e}_2$. The least-squares solution vector $\mathbf{x}$ is the one for which $\|\mathbf{e}_1\|^2 = 0$, so $\mathbf{e}_1 = 0$. This means that the projection of $\mathbf{y}$ on $L(\mathbf{a}_j)$, call it $\mathbf{y}_p$, must equal $\mathbf{Ax} = \sum_{j=1}^{n} x_j \mathbf{a}_j$. Since $r_A = n$, the columns $\mathbf{a}_j$ form a *basis* for $L(\mathbf{a}_j)$ so that $\mathbf{y}_p = \sum_{j=1}^{n} \alpha_j \mathbf{a}_j$. Because of the uniqueness of this expansion, $\alpha_j = x_j$, that is, $\mathbf{x} = \boldsymbol{\alpha}$. The set of n reciprocal basis vectors $\mathbf{r}_i$ is defined by $\langle \mathbf{r}_i, \mathbf{a}_j \rangle = \delta_{ij}$, or in matrix form

$$\underset{(n \times m)}{\mathbf{R}} \cdot \underset{(m \times n)}{\mathbf{A}} = \mathbf{I}$$

Since $\mathbf{A}$ is not square, it cannot be inverted to find $\mathbf{R}$. It is still true that $\alpha_j = \langle \mathbf{r}_j, \mathbf{y} \rangle = x_j$ so that

$$\boldsymbol{\alpha} = \mathbf{Ry} = \mathbf{x} \tag{6.2}$$

Therefore, $\mathbf{Ax} = \mathbf{ARy} = \mathbf{y}_p$. Using $\mathbf{e}_2 = \mathbf{y} - \mathbf{y}_p$ and the fact that $\langle \mathbf{a}_j, \mathbf{e}_2 \rangle = 0$ gives $\mathbf{A}^T[\mathbf{y} - \mathbf{ARy}] = \mathbf{0}$, or

$$\mathbf{Ry} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \tag{6.3}$$

Combining Eqs. (6.2) and (6.3) gives the least-squares solution

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$$

The amount of error in this approximate solution is indicated by

$$\|\mathbf{e}\|^2 = \|\mathbf{e}_2\|^2 = \mathbf{y}^T[\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T]\mathbf{y} = \|\mathbf{y} - \mathbf{Ax}\|^2$$

Recall that the square root of this quantity was given directly in the GSE method.

The matrix $\mathbf{R} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is a particular example of the *generalized* or *pseudo-inverse* [5] of **A**, written $\mathbf{A}^\dagger$. If $\mathbf{A}^{-1}$ exists, then $\mathbf{A}^\dagger = \mathbf{A}^{-1}$ and $\|\mathbf{e}\|^2 = 0$. The minimum norm solution of Section 6.7 provides another example of the pseudo-inverse that was appropriate to those circumstances, namely, $\mathbf{A}\dagger = \overline{\mathbf{A}}^T(\mathbf{A}\overline{\mathbf{A}}^T)^{-1}$. The general solutions to the $n \times n$ nonsingular case, the underdetermined minimum norm case, and the overdetermined least-squares case can all be expressed in terms of the pseudo-inverse as $\mathbf{x} = \mathbf{A}^\dagger\mathbf{y}$.

Weighted Least-Squares Approximation to the Solution‡

Ignoring some equations or placing equal reliance on all equations represents two extremes. If some equations are more reliable than others, but all equations are to be retained, a weighted least-squares approximation can be used. That is, **x** should minimize $\mathbf{e}^T\mathbf{R}^{-1}\mathbf{e} = (\mathbf{y} - \mathbf{A}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x})$. $\mathbf{R}^{-1}$ is symmetric, $m \times m$, nonsingular, and often diagonal. Those familiar with random processes should know that **R** is generally selected as the covariance matrix for the noise on the vector **y**. Smaller values of r_{ii} will cause e_i^2 to be smaller, and the ith equation is more nearly satisfied. If a norm $\|\mathbf{e}\|^2_{\mathbf{R}^{-1}} = \mathbf{e}^T\mathbf{R}^{-1}\mathbf{e}$ is defined, the method of orthogonal projections immediately leads to

$$\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{R}^{-1}\mathbf{y}$$

If $r_A = n$, as assumed here, the $n \times n$ matrix $\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A}$ is nonsingular and the weighted least-squares solution is

$$\mathbf{x} = (\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{R}^{-1}\mathbf{y}$$

Notice that if **A** is not full rank, the required inverse will not exist, signaling that the least-squares solution is not unique.

The least-squares and weighted least-squares formulas can also be derived simply by setting $\partial\|\mathbf{e}\|^2/\partial x_i = 0$, using results of Sec. 4.12.

Recursive Weighted Least-Squares Solutions

The preceding sections dealt with what is commonly called "batch least squares," because all data equations are treated in one batch. A recursive method of using each new set of data as it is received is now presented.

Assume that a set of m equations

$$\mathbf{y}_k = \mathbf{A}\mathbf{x} + \mathbf{e}$$

has been used to obtain a weighted least-squares estimate for **x**, denoted by $\mathbf{x}_k$:

$$\mathbf{x}_k = (\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{R}^{-1}\mathbf{y}_k$$

As is often the case, assume that an additional set of relations

$$\mathbf{y}_{k+1} = \mathbf{H}_{k+1}\mathbf{x} + \mathbf{e}_{k+1}$$

‡ The matrix **R** in this section is unrelated to the matrix of reciprocal basis vectors of previous sections.

then becomes available. It is desired to obtain a new estimate for $\mathbf{x}$, denoted as $\mathbf{x}_{k+1}$, which combines both sets of data and minimizes

$$J = [\mathbf{e}^T \quad \mathbf{e}_{k+1}^T] \left[\begin{array}{c|c} \mathbf{R}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{R}_{k+1}^{-1} \end{array}\right] \left[\begin{array}{c} \mathbf{e} \\ \hline \mathbf{e}_{k+1} \end{array}\right]$$

$\mathbf{R}_{k+1}^{-1}$ is the weighting matrix, analogous to $\mathbf{R}^{-1}$ but applied to the new data $\mathbf{y}_{k+1}$. It is not necessary to reprocess the whole set of equations involving $[\mathbf{y}_k \mid \mathbf{y}_{k+1}]$ in order to determine $\mathbf{x}_{k+1}$. It is shown in Problem 6.12, using partitioned matrices and a matrix inversion identity, that

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{K}_k[\mathbf{y}_{k+1} - \mathbf{H}_{k+1}\mathbf{x}_k]$$

where $\mathbf{K}_k = \mathbf{P}_k\mathbf{H}_{k+1}^T[\mathbf{H}_{k+1}\mathbf{P}_k\mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}]^{-1}$ and $\mathbf{P}_k \triangleq (\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A})^{-1}$, which is available from the computation of $\mathbf{x}_k$. If still other sets of equations are to be incorporated, the above relations can be used recursively. A new matrix, $\mathbf{P}_{k+1}$, is then needed and is given by

$$\mathbf{P}_{k+1} = [\mathbf{P}_k^{-1} + \mathbf{H}_{k+1}^T\mathbf{R}_{k+1}^{-1}\mathbf{H}_{k+1}]^{-1} \tag{6.4}$$

Using the matrix inversion lemma, page 132, this can also be written as

$$\mathbf{P}_{k+1} = \mathbf{P}_k - \mathbf{P}_k\mathbf{H}_{k+1}^T[\mathbf{H}_{k+1}\mathbf{P}_k\mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}]^{-1}\mathbf{H}_{k+1}\mathbf{P}_k \tag{6.5}$$

The latter form is often more convenient. For example, if $\mathbf{y}_{k+1}$ is a scalar, then matrix inversion is not required, just a scalar division.

Data Deweighting

A common occurrence is that data are received sequentially over time. As each new group of data is received, it is used to improve the estimate of $\mathbf{x}$. If this process is carried out over a sufficient number of steps, one will find that the $\mathbf{P}_{k+1}$ matrix has decreased to very small values due to the repeated addition of a nonnegative term to its inverse in Eq. (6.4). This in turn will cause the value of $\mathbf{K}_{k+1}$ to become small. This means that the corrections made to $\mathbf{x}_k$ in order to determine $\mathbf{x}_{k+1}$ get small, independent of what new or surprising information may be contained in the latest measurements. In order to prevent the recursive estimator from failing to respond adequately to new data (called going to sleep), some form of data deweighting is often used. Two types will be presented here, additive deweighting and multiplicative deweighting. In both cases the $\mathbf{P}$ matrix is prevented from getting too small. The easiest way to change the former algorithm is to introduce another matrix $\mathbf{M}_k$, given by

$$\mathbf{M}_k = \mathbf{P}_k/\beta \quad \text{with } \beta < 1\text{; this is multiplicative deweighting}$$

or

$$\mathbf{M}_k = \mathbf{P}_k + \mathbf{Q} \quad \text{with } \mathbf{Q} \text{ a positive definite matrix; this is additive deweighting}$$

The gain is now computed as

$$\mathbf{K}_k = \mathbf{M}_k\mathbf{H}_{k+1}^T[\mathbf{H}_{k+1}\mathbf{M}_k\mathbf{H}_{k+1}^T + \mathbf{R}]^{-1}$$

and the new $\mathbf{P}_{k+1}$ is given by either Eq. (*6.4*) or (*6.5*), but with $\mathbf{P}_k$ on the right-hand side replaced everywhere by $\mathbf{M}_k$. The formula for updating the estimate of $\mathbf{x}$ remains the same. If $\beta = 1$ or $\mathbf{Q} = 0$, both of these deweighting schemes revert to the original algorithm. Values used for the so-called forgetting factor β depend on how much deweighting is desired. A concept called asymptotic sample length, ASL, is a measure of how much past data are having a significant effect on the current estimate of $\mathbf{x}$. A relation between ASL and β is

$$\text{ASL} = 1/(1 - \beta)$$

Therefore, the commonly used values of β between 0.999 and 0.95 correspond to asymptotic sample lengths of 1000 past measurements down to 20. Although $\mathbf{Q}$ is often selected as a diagonal matrix with small diagonal elements, an idea of the appropriate magnitudes can be obtained by assuming that $\mathbf{Q} = \alpha\mathbf{P}$, with α a scalar. Then comparison of the two forms of deweighting shows that $\alpha = (1/\beta) - 1$. Therefore, to get an ASL of 1000, α would be 0.001 or $\mathbf{Q}$ should be about 0.1% of $\mathbf{P}$. Of course, since $\mathbf{P}$ is changing, this kind of comparison is not perfect. It does indicate roughly that a small $\mathbf{Q}$ matrix can be an effective deweighting scheme. Figure 6.4 shows the relative weight applied to past measurements when forming the current estimate of $\mathbf{x}$. The curve is normalized so that the current measurement weight is one. This is computed by using

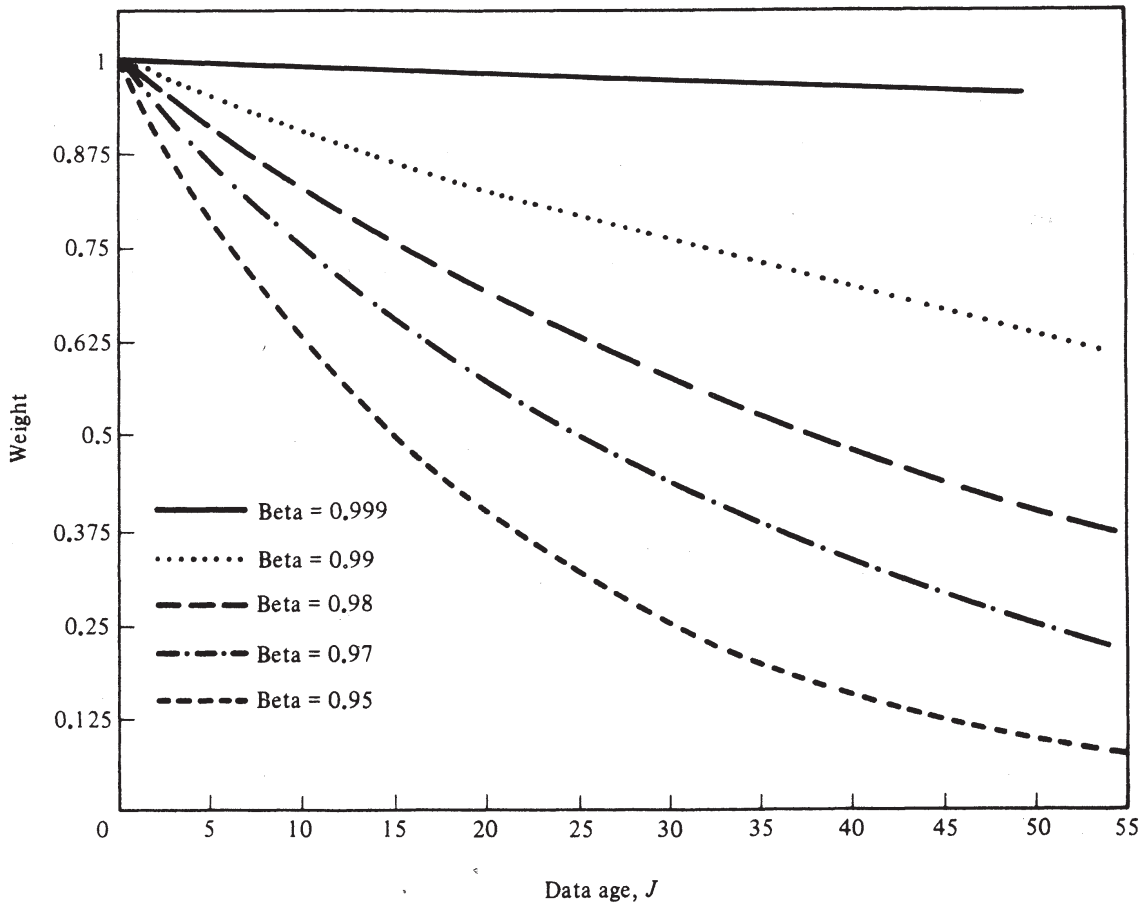


Figure 6.4 Deweighting for common forgetting factors

the recursive estimation algorithm backwards in time for a scalar **x.** Qualitatively similar deweighting occurs in the vector case, but it is harder to normalize and display.

If the matrices **P, Q,** and **R** are given the appropriate statistical interpretation as covariance matrices of certain signals, then the recursive equations constitute a simple example of the discrete *Kalman filtering* equations. The Kalman filter is used extensively in modern control theory in order to estimate the internal (state) variables of a linear system based on noisy measurements of the output variables [6, 7]. A deterministic state estimation procedure, the *observer,* is presented in Chapter 13. The least-squares approach finds many other applications in modern control theory as well.

6.9 TWO BASIC PROBLEMS IN CONTROL THEORY

Consider the discrete-time state equations of Chapter 3,

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

$$\mathbf{y}(k+1) = \mathbf{C}\mathbf{x}(k+1) + \mathbf{D}\mathbf{u}(k+1)$$

Let the initial conditions for the state vector be $\mathbf{x}(0)$. Then the states at succeeding time points are

$$\mathbf{x}(1) = \mathbf{A}\mathbf{x}(0) + \mathbf{B}\mathbf{u}(0)$$

$$\mathbf{x}(2) = \mathbf{A}\mathbf{x}(1) + \mathbf{B}\mathbf{u}(1)$$

$$= \mathbf{A}[\mathbf{A}\mathbf{x}(0) + \mathbf{B}\mathbf{u}(0)] + \mathbf{B}\mathbf{u}(1) = \mathbf{A}^2\mathbf{x}(0) + \mathbf{A}\mathbf{B}\mathbf{u}(0) + \mathbf{B}\mathbf{u}(1)$$

Continuing in this fashion the state at a general time point k is found to be

$$\mathbf{x}(k) = \mathbf{A}^k\mathbf{x}(0) + \mathbf{B}\mathbf{u}(k-1) + \mathbf{A}\mathbf{B}\mathbf{u}(k-2) + \mathbf{A}^2\mathbf{B}\mathbf{u}(k-3) + \cdots + \mathbf{A}^{k-2}\mathbf{B}\mathbf{u}(1) + \mathbf{A}^{k-1}\mathbf{B}\mathbf{u}(0)$$

6.9.1 A CONTROL PROBLEM

One of the two basic control problems is the determination of a sequence of control inputs $\mathbf{u}(i)$ which will transfer a known initial state vector $\mathbf{x}(0)$ to the origin of state space at some finite time point k. It is convenient to stack up all the unknown control vectors into one composite vector $\mathbf{U}$ defined as

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}(k-1) \\ \mathbf{u}(k-2) \\ \mathbf{u}(k-3) \\ \vdots \\ \mathbf{u}(1) \\ \mathbf{u}(0) \end{bmatrix}$$

Also define $\mathbf{P} = [\mathbf{B} \mid \mathbf{AB} \mid \mathbf{A}^2\mathbf{B} \mid \ldots \mid \mathbf{A}^{k-1}\mathbf{B}]$. Then $\mathbf{x}(k) = \mathbf{A}^k\mathbf{x}(0) + \mathbf{PU}$. Setting the final state to zero gives a set of simultaneous linear equations for the unknown controls **U**:

$$\mathbf{PU} = -\mathbf{A}^k\mathbf{x}(0)$$

Under what conditions will there be a control sequence **U** that will drive a *given* $\mathbf{x}(0)$ to zero? How about an *arbitrary* $\mathbf{x}(0)$? If there is a solution vector **U**, is it unique? The composite partitioned matrix **P** plays the role of the general matrix **A**, which was discussed throughout most of this chapter. If the state vector has n components and if each $\mathbf{u}(i)$ has r components, then **B** is $n \times r$, as are **AB** and any power of **A** times **B**. There are k partitions of this type in **P**, so **P** is an $n \times (kr)$ matrix; of course, **U** has kr unknown components. If the $n \times 1$ vector $-\mathbf{A}^k\mathbf{x}(0)$ happens to be a linear combination of the columns of **P**, then a solution for **U** will exist. In order for this to be true for any arbitrary vector $-\mathbf{A}^k\mathbf{x}(0)$, and hence for any arbitrary $\mathbf{x}(0)$ initial condition, it is necessary that the range space of **P** must span the n-dimensional state space. This means that there are n linearly independent vectors among the columns of **P**, and therefore the rank of **P** must be n. If **P** has full rank n, there will be solutions, but are they unique? And what about the final time index k? If $k = 1$, then $\mathbf{P} \equiv \mathbf{B}$ and $\text{rank}(\mathbf{P}) \leq r$. If the value of r is less than n, it is clear that no solutions can exist for arbitrary $\mathbf{x}(0)$, although they may for certain $\mathbf{x}(0)$. As more time is allowed (i.e., as k increases) the matrix **P** contains more columns, and its rank may increase. It is shown in Chapter 8 that the rank of **P** will never increase beyond the value achieved for $k = n$, the number of states, because further partitions will always be linear combinations of the first n. For this reason the test of whether controls can be found which will drive an arbitrary $\mathbf{x}(0)$ to zero in finite time is equivalent to testing whether $\text{rank}[\mathbf{B} \mid \mathbf{AB} \mid \mathbf{A}^2\mathbf{B} \mid \ldots \mid \mathbf{A}^{n-1}\mathbf{B}] = n$. If a particular system passes this test, it is said to be *controllable*. This is pursued in more detail in Chapter 11. Assume that this controllability condition is met. In general the sequence of controls making up **U** is not unique for a given final time k. Many optional control sequences might all bring the final state to the origin. One desirable solution might be the minimum norm solution of Section 6.7. Let the subscript k indicate explicitly how many control cycles are being used in **U**. Then

$$\mathbf{U}_k = -\mathbf{P}_k^T[\mathbf{P}_k\,\mathbf{P}_k^T]^{-1}\mathbf{A}^k\mathbf{x}(0)$$

Even though the rank of $\mathbf{P}_k$ will not increase for $k > n$, the amount of control effort required to drive the initial state to zero will change in general. The minimum norm squared is

$$\|\mathbf{U}_k\|^2 = \mathbf{x}(0)^T(\mathbf{A}^k)^T[\mathbf{P}_k\,\mathbf{P}_k^T]^{-1}\mathbf{A}^k\mathbf{x}(0)$$

The $\mathbf{A}^k$ terms could increase or decrease with k depending upon the system stability. The term inside the inverse is

$$\mathbf{PP}^T = \mathbf{BB}^T + \mathbf{ABB}^T\mathbf{A}^T + \cdots + \mathbf{A}^{k-1}\mathbf{BB}^T[\mathbf{A}^{k-1}]^T$$

It would be expected to grow with k as more terms are added to the sum. Thus the inverse itself will diminish. It seems intuitive that if the system is stable so that its

unforced state is decaying toward zero, not much control effort will be required to help the state reach zero if k is large. This is borne out by the preceding results.

6.9.2 A State Estimation Problem

The same discrete-time system is considered, but now the inputs $\mathbf{u}(i)$ are all assumed known. The question to be addressed is whether the *unknown* initial state vector $\mathbf{x}(0)$ can be determined from knowledge of the sequence of output vectors $\mathbf{y}(j)$ for $j = 0, 1, \ldots, k$. If the first $k+1$ output vectors are stacked up into one composite vector $\mathbf{Y}_k = [\mathbf{y}(0)^T \quad \mathbf{y}(1)^T \quad \ldots \quad \mathbf{y}(k)^T]^T$, the preceding results can be used to write

$$\mathbf{Y}_k = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^k \end{bmatrix} \mathbf{x}(0) + \begin{bmatrix} \mathbf{D} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{CB} & \mathbf{D} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{CAB} & \mathbf{CB} & \mathbf{D} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & & & & \\ \mathbf{CA}^{k-1}\mathbf{B} & \mathbf{CA}^{k-2}\mathbf{B} & & & \cdots & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{u}(0) \\ \mathbf{u}(1) \\ \mathbf{u}(2) \\ \vdots \\ \mathbf{u}(k) \end{bmatrix}$$

Since **A, B, C,** and **D** as well as all the input terms $\mathbf{u}(i)$ are assumed known, they can be brought to the left side of the equation to define a new vector $\mathbf{Y}_k'$ or all $\mathbf{u}(i)$ can be assumed zero without loss of generality. This leaves a set of simultaneous linear equations relating outputs in $\mathbf{Y}_k$ (or $\mathbf{Y}_k'$) and the unknown initial state $\mathbf{x}(0)$. Define the matrix $\mathbf{Q}_k^T = [\mathbf{C}^T \mid \mathbf{A}^T\mathbf{C}^T \mid (\mathbf{A}^2)^T\mathbf{C}^T \mid \ldots \mid (\mathbf{A}^k)^T\mathbf{C}^T]$. The simultaneous equations then become $\mathbf{Y}_k = \mathbf{Q}_k\,\mathbf{x}(0)$. If each $\mathbf{y}(i)$ vector has m components, the $\mathbf{Q}_k$ matrix will be of dimension $mk \times n$. The maximum rank is n. Just as with the previous matrix **P,** it is shown in Chapter 8 that the rank of $\mathbf{Q}_k$ will not increase for values of k larger than $n-1$. If $\mathbf{Q}_k$ achieves its full rank n, then the $n \times n$ matrix $\mathbf{Q}_k^T\,\mathbf{Q}_k$ will be invertible. If this is true, then

$$\mathbf{x}(0|k) = [\mathbf{Q}_k^T\,\mathbf{Q}_k]^{-1}\,\mathbf{Q}_k^T\,\mathbf{Y}_k$$

Is this the unique solution for $\mathbf{x}(0)$, or is it merely a least-squares approximation for $\mathbf{x}(0)$? If the vector $\mathbf{Y}_k$ belongs to the n-dimensional range space of $\mathbf{Q}_k$, then this is indeed the solution. This is just another way of saying that even though there may be more equations than unknowns, they are *not* inconsistent. However, since the vector $\mathbf{Y}_k$ has mk components, it is very possible that $\mathbf{Y}_k$ might not lie in the n-dimensional range subspace, perhaps because of measurement errors or modeling errors in **A, B, C,** or **D.** In that case the equation for $\mathbf{x}(0|k)$ is a least-squares approximate solution for $\mathbf{x}(0)$ based on the k available data points. Note that the existence of a unique least-squares answer also requires that $\mathbf{Q}_k$ be of full rank n. The condition that

$$[\mathbf{C}^T \mid \mathbf{A}^T\mathbf{C}^T \mid (\mathbf{A}^2)^T\mathbf{C}^T \mid \ldots \mid (\mathbf{A}^{n-1})^T\mathbf{C}^T] \quad \text{has rank} \quad n$$

is necessary if $\mathbf{x}(0)$ is to be found from the observed data. A system which meets this criterion is said to be *observable.* Chapter 11 looks into this further. The intent here is to stress the importance of the theory of simultaneous linear equations. Problem 6.17 develops the equations for recursively updating the estimate of $\mathbf{x}(0)$ each time a new noisy measurement becomes available. This gives an improved estimate $\mathbf{x}(0|k+1)$ by using the newest measurement $\mathbf{y}(k+1)$ to add a correction to $\mathbf{x}(0|k)$. Problem 6.18

modifies these results so that estimates of the *current* state $\mathbf{x}(k)$ are recursively computed, rather than estimates of the initial state.

6.10 LYAPUNOV EQUATIONS

Another type of equation which occurs in control theory is

$$\mathbf{XA} + \mathbf{BX} = \mathbf{C} \tag{6.6}$$

where **A, B,** and **C** are known matrices and **X** is a matrix of unknowns. Assume that the dimensions of **A, B, C,** and **X** are $m \times m$, $n \times n$, $n \times m$, and $n \times m$, respectively. This is usually called the *Lyapunov equation.* It is linear in the unknowns, but it is of an entirely different type than those that have been treated previously. For small-dimensional problems it is not difficult to expand the equation into component form and thus obtain simultaneous equations in the unknown x_{ij} components. However, a more general procedure can also be developed using the concept of vectorized matrices, which was presented in Sec. 4.12. The columns of the unknown matrix **X** are stacked into a single column, referred to as **(X)**. The two matrix products involving **X** can be epxressed in terms of the vectorized **(X)** by using the Kronecker product of Chapter 4. Specifically,

$$(\mathbf{XA}) = [\mathbf{A}^T \otimes \mathbf{I}_n](\mathbf{X})$$

$$(\mathbf{BX}) = [\mathbf{I}_m \otimes \mathbf{B}](\mathbf{X})$$

so that the total vectorized equation can be written

$$\{[\mathbf{A}^T \otimes \mathbf{I}_n] + [\mathbf{I}_m \otimes \mathbf{B}]\}(\mathbf{X}) = (\mathbf{C})$$

Of course, the vectors **(X)** and **(C)** are $nm \times 1$ columns, and the coefficient matrix **Q** created with the two Kronecker products is of size $nm \times nm$. While the size of the problem has seemingly been multiplied, what has been accomplished is the positioning of both unknown **X** terms on the same side of a known square matrix **Q.** If **Q** has an inverse, the unique solution for **X** is expressible, still in vectorized form, as

$$(\mathbf{X}) = \{[\mathbf{A}^T \otimes \mathbf{I}_n] + [\mathbf{I}_m \otimes \mathbf{B}]\}^{-1}(\mathbf{C}) = \mathbf{Q}^{-1}(\mathbf{C})$$

The matrix **X** is then recreated by undoing the vectorization process.

EXAMPLE 6.8 Let $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -18 & -27 & -10 \end{bmatrix}$, and $\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 0 \end{bmatrix}$. Find the 3×2 **X** matrix which satisfies Eq. (6.6).

The vectorized form of the equations is

$$\begin{bmatrix} 0 & 1 & 0 & -2 & 0 & 0 \\ 0 & 0 & 1 & 0 & -2 & 0 \\ -18 & -27 & -10 & 0 & 0 & -2 \\ 1 & 0 & 0 & -3 & 1 & 0 \\ 0 & 1 & 0 & 0 & -3 & 1 \\ 0 & 0 & 1 & -18 & -27 & -13 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

The 6×6 matrix $\mathbf{Q}$ has a determinant value of 6720. Solving and putting the x_{ij} components into their traditional rectangular array gives

$$\mathbf{X} \approx \begin{bmatrix} -1.407143 & -0.4785714 \\ 0.042857 & -0.0285714 \\ 1.942857 & 0.8714286 \end{bmatrix}$$

■

EXAMPLE 6.9 Repeat the previous example if the matrix $\mathbf{B}$ is changed to

$$\mathbf{B} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 18 & -9 & -8 \end{bmatrix}$$

Using the two Kronecker products forms a 6×6 coefficient matrix $\mathbf{Q}$, which is singular of rank 5. The preceding solution process is not possible, and no unique solution exists. Do nonunique solutions exist? The rank of $\mathbf{W} = [\mathbf{Q} \mid (\mathbf{C})]$ is 6, showing that the equations are now inconsistent, and *no* solution exists. ■

The conditions under which solutions to Eq. (*6.6*) exist are now stated without proof. The concept of *eigenvalues* must be anticipated from the next chapter. Let $\{\lambda_i, i = 1, \ldots, m\}$ be the eigenvalues of $\mathbf{A}$. Let $\{\mu_j, j = 1, \ldots, n\}$ be the eigenvalues of $\mathbf{B}$. Then a unique solution exists for Eq. (*6.6*) if and only if

$$\lambda_i + \mu_j \neq 0 \quad \text{for all } i, j \text{ pairs}$$

In the two preceding examples $\mathbf{A}$ has eigenvalues of -1 and -2. In Example 6.8 $\mathbf{B}$ has eigenvalues of -1, -3, and -6, so the conditions for a unique solution are satisfied. In Example 6.9 the modified matrix $\mathbf{B}$ has eigenvalues of 1, -3, and -6. Now $\lambda_1 + \mu_1 = 0$, so that the conditions for solutions are not satisfied. A special case of Eq. (*6.6*), which commonly occurs in stability, random processes, and optimal control problems, has $\mathbf{A}$ and $\mathbf{B}$ both $n \times n$ and transposes of each other. Then $\mathbf{X}$ and $\mathbf{C}$ must also be $n \times n$ matrices. Since $\mathbf{A}$ and $\mathbf{A}^T$ have the same eigenvalues, the existence conditions fail to be satisfied only when $\mathbf{A}$ has one or more pairs of eigenvalues positioned symmetrically with respect to the $j\omega$ axis, such as $-\alpha$ and $+\alpha$ for some scalar α.

If $\mathbf{C}$ is symmetric, then $\mathbf{X}$ will also be symmetric. In this case the vectorized equations will contain unnecessary redundancies in both $(\mathbf{C})$ and $(\mathbf{X})$, which could be removed. The corresponding rows of $\mathbf{Q}$ can then be removed, and the sum of the columns which multiply $\mathbf{X}_{ij}$ and $\mathbf{X}_{ji}$ is used. This gives a reduced set of equations for the $n(n+1)/2$ unknowns. This need not be done, however. The full set of n^2 equations is still solvable, and the resulting $\mathbf{X}$ will be symmetric (except possibly for rounding error).

EXAMPLE 6.10 Let $\mathbf{A} = \begin{bmatrix} 0 & -4 \\ 1 & -2 \end{bmatrix}$, $\mathbf{B} = \mathbf{A}^T$, and $\mathbf{C} = \begin{bmatrix} 0 & 0 \\ 0 & -5 \end{bmatrix}$. Then the solution for $\mathbf{X}$ is found as $\mathbf{X} = \text{diag}\{0.3125, \quad 1.25\}$, which is symmetric, as promised. In fact, $\mathbf{X}$ is positive definite, a property which is defined in the next chapter. This is related to the fact that both eigenvalues of $\mathbf{A}$ have negative real parts and to the properties of $\mathbf{C}$. These issues are clarified in later chapters. ■

REFERENCES

1. Strang, G.: *Linear Algebra and Its Applications,* Academic Press, New York, 1980.
2. Kailath, T.: *Linear Systems,* Prentice Hall, Englewood Cliffs, N.J., 1980.
3. Taylor, A. E. and R. W. Mann: *Advanced Calculus,* Xerox, Boston, Mass., 1972.
4. Forsythe, G. E., M. A. Malcolm, and C. Moler: *Computer Methods for Mathematical Computations,* Prentice Hall, Englewood Cliffs, N.J., 1977.
5. Penrose, R.: "A Generalized Inverse for Matrices," *Proceedings of the Cambridge Philosophical Society,* Vol. 51, Part 3, 1955, pp. 406–413.
6. Kalman, R. E.: "A New Approach to Linear Filtering and Prediction Problems," *Trans. of the ASME, Journal of Basic Engineering,* Vol. 82, 1960, pp. 35–45.
7. Maybeck, P. S.: *Stochastic Models, Estimation and Control,* Vol. 1, Academic Press, New York, 1979.
8. Meditch, J. S., *Stochastic Optimal Linear Estimation and Control,* McGraw-Hill, New York, 1969.

ILLUSTRATIVE PROBLEMS

6.1 Use arguments in $\mathcal{X}^n$ to draw conclusions about solutions to $\mathbf{Ax} = \mathbf{y}$.

Let the rows of $\mathbf{A}$ be considered as the conjugate transpose of n component vectors $\mathbf{c}_i$. Then the set of simultaneous equations is equivalent to m scalar equations of the form

$$\langle \mathbf{c}_i, \mathbf{x} \rangle = y_i \quad \text{where } i = 1, \ldots, m$$

Each of these equations defines an $n - 1$ dimensional hyperplane in $\mathcal{X}^n$, with a normal $\mathbf{c}_i$.

The existence of a solution means that there exists a vector $\mathbf{x}$ that simultaneously terminates in all m of the hyperplanes. If the set $\{\mathbf{c}_i\}$ is linearly independent, so that $r_A = m$, the intersection of m hyperplanes of dimension $n - 1$ defines an $n - m$ dimensional hyperplane. Every vector $\mathbf{x}$ terminating in this hyperplane is a solution. If $m = n$, the hyperplane is of zero dimension, i.e., a single point, and defines a unique solution. Obviously, whenever $r_A = m$, $r_W = m$ also.

If $r_A < m$, two or more of the $n - 1$ dimensional hyperplanes are either parallel or they coincide. If parallel but distinct, they never intersect and the equations are inconsistent. It can be shown that these geometrical conditions are equivalent to the algebraic conditions given in terms of the $\mathbf{W}$ matrix.

Homogeneous Equations

6.2 Let $\mathbf{A}$ be an $n \times n$ matrix with $r_A = n - 1$. Show that a nontrivial solution to $\mathbf{Ax} = \mathbf{0}$ can be selected as any nonzero column of the matrix Adj $\mathbf{A}$.

For any $n \times n$ matrix, $\mathbf{A}[\text{Adj}(\mathbf{A})] = \mathbf{I}_n|\mathbf{A}|$. Any column j of this equation can be singled out and written as $\mathbf{A}[\text{Adj}(\mathbf{A})]_j = [\mathbf{I}_n]_j|\mathbf{A}|$. If $r_A < n$, the determinant is zero, and $\mathbf{A}$ times *any* column of Adj($\mathbf{A}$) then gives a zero vector. Selecting a particular column j which is not identically zero thus gives a nontrivial solution.

6.3 Use the results of Problem 6.2 to find a nontrivial solution for

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}$$

Since $r_A = 2$, the degeneracy is $n - r_A = 1$, so a one-parameter family of nontrivial solutions exists. It is found by computing $\text{Adj}\,\mathbf{A} = \begin{bmatrix} 2 & 2 & -1 \\ 0 & 0 & 0 \\ -2 & -2 & 1 \end{bmatrix}$. Thus $\mathbf{x} = k\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, k \neq 0$, generates the set of all nontrivial solutions.

6.4 If an $n \times n$ matrix has rank $r_A < n$, it can be shown that $\mathbf{Ax} = \mathbf{0}$ has $q = n - r_A$ linearly independent solutions. They may be chosen as linearly independent columns of

$$\frac{d^{q-1}}{d\epsilon^{q-1}}\{\text{Adj}[\mathbf{A} - \mathbf{I}\epsilon]\}\bigg|_{\epsilon=0}$$

Find nontrivial solutions for this problem when $\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ -2 & -2 & -2 \end{bmatrix}$.

The rank of $\mathbf{A}$ is 1, so $q = 2$ and

$$\lim_{\epsilon \to 0} \frac{d}{d\epsilon}\{\text{Adj}[\mathbf{A} - \mathbf{I}\epsilon]\} = \begin{bmatrix} 0 & 1 & 1 \\ 2 & 1 & 2 \\ -2 & -2 & -3 \end{bmatrix}$$

There are just two linearly independent columns, and nontrivial solutions are $\mathbf{x}_1 = [0 \quad 2 \quad -2]^T$, $\mathbf{x}_2 = [1 \quad 1 \quad -2]^T$ or any linear combination of these two. $\mathbf{x}_1$ and $\mathbf{x}_2$ form a basis for the two-dimensional subspace defined by $\langle \mathbf{c}, \mathbf{x} \rangle = 0$, where $\mathbf{c}$ is the transpose of any row of $\mathbf{A}$.

6.5 Find all nontrivial solutions to

$$\begin{bmatrix} 1 & 3 & 5 \\ 1 & 4 & 6 \\ -1 & 5 & 3 \\ -1 & 4 & 2 \\ 1 & 3 & 5 \end{bmatrix}\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

This is the same $\mathbf{A}$ matrix as appeared in Example 6.1(8). The RRE form for $\mathbf{W}$ is the same as given in that example, except that the last column is all zeros. Therefore, all nontrivial solutions must satisfy $x_1 + 2x_3 = 0$ and $x_2 + x_3 = 0$. Thus $\mathbf{x} = a[-2 \quad -1 \quad 1]^T$, for any scalar a, constitutes the one parameter family of nontrivial solutions. From $\mathbf{W}'$ it can be seen that the rank of $\mathbf{A}$ is 2 and the number of unknowns is $n = 3$, so the degeneracy or nullity is $q = 3 - 2 = 1$. This is the dimension of the null space of $\mathbf{A}$.

6.6 Find all nontrivial solutions to

$$\begin{bmatrix} 4 & -2 & 3 \\ 1 & 3 & 1 \\ 1 & 3 & 1 \end{bmatrix}\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

For this problem the RRE form of $\mathbf{W}$ is

$$\mathbf{W}' = \left[\begin{array}{ccc|c} 1 & 0 & 0.78571427 & 0 \\ 0 & 1 & 0.07142857 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Therefore, $x_1 + 0.78571427x_3 = 0$ and $x_2 + 0.07142857x_3 = 0$. All nontrivial solutions must be proportional to

$$\mathbf{x} = [11 \quad 1 \quad -14]^T$$

6.7 Do nontrivial solutions exist for the following?

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

Using the RRE method, or just computing its determinant, shows that the rank of **A** is 3. Its degeneracy is zero, it is nonsingular, and so the only solution to this problem is the trivial solution $\mathbf{x} = \mathbf{0}$.

Minimum Norm Solutions

6.8 Find the minimum norm solution for $[1 \quad 2]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1$.

Identifying $\mathbf{A} = [1 \quad 2]$, the minimum norm solution is

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \left\{ [1 \quad 2] \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}^{-1} (1) = \frac{1}{5} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

6.9 A specified amount of constant current, i, must be delivered to the ground point of Figure 6.5. Specify v_1, v_2, and v_3 so that the total energy dissipated in the resistors is minimized.

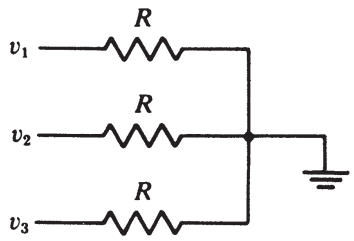


Figure 6.5

It is required that

$$v_1/R + v_2/R + v_3/R = i \quad \text{or} \quad \frac{1}{R}[1 \quad 1 \quad 1]\mathbf{v} = i$$

The total energy dissipated per unit time is

$$\dot{\varepsilon} = \frac{1}{R}[v_1^2 + v_2^2 + v_3^2] = \frac{1}{R}\|\mathbf{v}\|^2$$

The desired solution is thus the minimum norm solution.

$$\mathbf{v} = \frac{1}{R} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \left\{ \frac{1}{R^2} [1 \quad 1 \quad 1] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}^{-1} i = \frac{R}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} i$$

Frequently, the norm can be given a physical interpretation of energy or power. This is one reason why minimum norm solutions are often sought for underdetermined problems.

6.10 A small microcomputer has five terminals connected to it. The fraction of the time devoted to each terminal is x_i, so

$$x_1 + x_2 + x_3 + x_4 + x_5 = 1$$

The programmer at terminal 2 types four times as fast as the programmer at 1. They are both typing in the same code and must finish at the same time so $x_1 = 4x_2$. Both terminals 3 and 4 are sending mail files to 5, so $x_3 + x_4 = x_5$. Find the minimum norm solution for allocating CPU time.

In matrix form the constraints are

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & -4 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

The minimum norm solution is

$$\mathbf{x} = \mathbf{A}^T[\mathbf{A}\mathbf{A}^T]^{-1}\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -4 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 5 & -3 & 1 \\ -3 & 17 & 0 \\ 1 & 0 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.28436 \\ 0.07109 \\ 0.15602 \\ 0.15602 \\ 0.32739 \end{bmatrix}$$

6.11 Find the shortest four-dimensional vector from the origin to the four-dimensional hyperplane described by

$$5x_1 - 2x_2 + x_3 + 7x_4 = 12$$

This is the same as asking for the minimum norm solution to

$$[5 \quad -2 \quad 1 \quad 7]\mathbf{x} = 12$$

Therefore,

$$\mathbf{x} = \begin{bmatrix} 5 \\ -2 \\ 1 \\ 7 \end{bmatrix} [79]^{-1}(12) = \begin{bmatrix} 0.7595 \\ -0.3038 \\ 0.1519 \\ 1.0633 \end{bmatrix}$$

Least Squares, Weighted Least Squares, and Recursive Least Squares

6.12 Consider the set of simultaneous linear equations

$$\left[\begin{array}{c} \mathbf{y}_k \\ \hline \mathbf{y}_{k+1} \end{array}\right] = \left[\begin{array}{c} \mathbf{A} \\ \hline \mathbf{H}_{k+1} \end{array}\right]\mathbf{x} + \left[\begin{array}{c} \mathbf{e} \\ \hline \mathbf{e}_{k+1} \end{array}\right]$$

Find the vector $\mathbf{x}$ which minimizes

$$J = [\mathbf{e}^T \quad \mathbf{e}_{k+1}^T]\left[\begin{array}{c|c} \mathbf{R}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{R}_{k+1}^{-1} \end{array}\right]\left[\begin{array}{c} \mathbf{e} \\ \hline \mathbf{e}_{k+1} \end{array}\right]$$

The weighted least-squares estimate is

$$\mathbf{x}_{k+1} = \left\{[\mathbf{A}^T \mid \mathbf{H}_{k+1}^T]\left[\begin{array}{c|c} \mathbf{R}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{R}_{k+1}^{-1} \end{array}\right]\left[\begin{array}{c} \mathbf{A} \\ \hline \mathbf{H}_{k+1} \end{array}\right]\right\}^{-1}[\mathbf{A}^T \mid \mathbf{H}_{k+1}^T]\left[\begin{array}{c|c} \mathbf{R}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{R}_{k+1}^{-1} \end{array}\right]\left[\begin{array}{c} \mathbf{y}_k \\ \hline \mathbf{y}_{k+1} \end{array}\right]$$

$$= [\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A} + \mathbf{H}_{k+1}^T\mathbf{R}_{k+1}^{-1}\mathbf{H}_{k+1}]^{-1}[\mathbf{A}^T\mathbf{R}^{-1}\mathbf{y}_k + \mathbf{H}_{k+1}^T\mathbf{R}_{k+1}^{-1}\mathbf{y}_{k+1}]$$

Defining $\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A} = \mathbf{P}_k^{-1}$ and using the matrix inversion identity of Sec. 4.9 gives

$$\mathbf{x}_{k+1} = \{\mathbf{P}_k - \mathbf{P}_k\mathbf{H}_{k+1}^T[\mathbf{H}_{k+1}\mathbf{P}_k\mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}]^{-1}\mathbf{H}_{k+1}\mathbf{P}_k\}\{\mathbf{A}^T\mathbf{R}^{-1}\mathbf{y}_k + \mathbf{H}_{k+1}^T\mathbf{R}_{k+1}^{-1}\mathbf{y}_{k+1}\}$$

Note that $\mathbf{P}_k\mathbf{A}^T\mathbf{R}^{-1}\mathbf{y}_k = \mathbf{x}_k$ is the weighted least-squares solution when only the first group of equations is used. Therefore,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{P}_k\mathbf{H}_{k+1}^T[\mathbf{H}_{k+1}\mathbf{P}_k\mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}]^{-1}\mathbf{H}_{k+1}\mathbf{x}_k$$
$$+ \mathbf{P}_k\mathbf{H}_{k+1}^T\{\mathbf{I} - [\mathbf{H}_{k+1}\mathbf{P}_k\mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}]^{-1}\mathbf{H}_{k+1}\mathbf{P}_k\mathbf{H}_{k+1}^T\}\mathbf{R}_{k+1}^{-1}\mathbf{y}_{k+1}$$

The unit matrix in the last equation is written as

$$\mathbf{I} = [\mathbf{H}_{k+1}\mathbf{P}_k\mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}]^{-1}[\mathbf{H}_{k+1}\mathbf{P}_k\mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}]$$

This step is analogous to finding the common denominator in scalar algebra and leads to

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{P}_k \mathbf{H}_{k+1}^T [\mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}]^{-1} \{\mathbf{y}_{k+1} - \mathbf{H}_{k+1} \mathbf{x}_k\}$$

If this recursive process is to be continued, then an expression for $\mathbf{P}_{k+1}$ is needed. If $\mathbf{A}$ is replaced by $\left[\begin{array}{c} \mathbf{A} \\ \hline \mathbf{H}_{k+1} \end{array}\right]$ and if $\mathbf{R}^{-1}$ is replaced by $\left[\begin{array}{c|c} \mathbf{R}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{R}_{k+1}^{-1} \end{array}\right]$, then the definition for $\mathbf{P}_k$ is modified to read

$$\mathbf{P}_{k+1} = \left\{ [\mathbf{A}^T \mid \mathbf{H}_{k+1}^T] \left[\begin{array}{c|c} \mathbf{R}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{R}_{k+1}^{-1} \end{array}\right] \left[\begin{array}{c} \mathbf{A} \\ \hline \mathbf{H}_{k+1} \end{array}\right] \right\}^{-1}$$

$$= [\mathbf{A}^T \mathbf{R}^{-1} \mathbf{A} + \mathbf{H}_{k+1}^T \mathbf{R}_{k+1}^{-1} \mathbf{H}_{k+1}]^{-1} = [\mathbf{P}_k^{-1} + \mathbf{H}_{k+1}^T \mathbf{R}_{k+1}^{-1} \mathbf{H}_{k+1}]^{-1}$$

6.13 A tracking station measures $\dot{r}$, the time derivative of the range to a satellite, every second. The measurements are noisy. Find the least-squares fit to a straight line.

The measurements are assumed to fit the equation $\dot{r}(t) = a + bt + e(t)$, where a and b are to be determined. Measurement times are $t = 1, 2, \ldots, k$:

$$\begin{bmatrix} \dot{r}_1 \\ \dot{r}_2 \\ \vdots \\ \dot{r}_k \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & k \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_k \end{bmatrix}$$

The least-squares solution gives

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} k & \sum_{j=1}^{k} j \\ \sum_{j=1}^{k} j & \sum_{j=1}^{k} j^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{k} \dot{r}_i \\ \sum_{j=1}^{k} j \dot{r}_j \end{bmatrix}$$

Using the identities $\sum_{j=1}^{k} j = \dfrac{k(k+1)}{2}$, $\sum_{j=1}^{k} j^2 = \dfrac{k(k+1)(2k+1)}{6}$, and carrying out the matrix inversion gives

$$a = \frac{2(2k+1) \sum_{j=1}^{k} \dot{r}_j - 6 \sum_{j=1}^{k} j\dot{r}_j}{k(k-1)}, \quad b = \frac{-6(k+1) \sum_{j=1}^{k} \dot{r}_j + 12 \sum_{j=1}^{k} j\dot{r}_j}{k(k^2-1)}$$

If $k = 1$, the results are indeterminate, indicating that an infinite number of lines can be passed through a single point.

6.14 Investigate the following equations:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 2 \\ 3 \\ 14 \end{bmatrix}$$

The RRE form of $\mathbf{W}$ is $\mathbf{W}' = \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array}\right]$.

Since the rank of $\mathbf{A}$ is 2 and the rank of $\mathbf{W}$ is 3, the equations are inconsistent. Least-squares solutions are investigated by several methods.

(a) The normal equation $\mathbf{A}^T \mathbf{A}\mathbf{x} = \mathbf{A}^T \mathbf{y}$ is in this case

$$\begin{bmatrix} 35 & 44 \\ 44 & 56 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 81 \\ 100 \end{bmatrix}$$

Straightforward matrix inversion could be used, or the RRE method as follows:

$$\mathbf{W} = \left[\begin{array}{cc|c} 35 & 44 & 81 \\ 44 & 56 & 100 \end{array}\right] \longrightarrow \left[\begin{array}{cc|c} 1 & 0 & 5.6666 \\ 0 & 1 & -2.6666 \end{array}\right] \quad \text{so} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5.6666 \\ -2.6666 \end{bmatrix}$$

(b) Cholesky decomposition gives

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 5.91608 & 0 \\ 7.43736 & 0.82808 \end{bmatrix}\begin{bmatrix} 5.91608 & 7.43736 \\ 0 & 0.82808 \end{bmatrix} = \mathbf{S}^T\mathbf{S}$$

$$\text{Solving } \mathbf{S}^T\mathbf{v} = \begin{bmatrix} 81 \\ 100 \end{bmatrix} \text{ gives } \mathbf{v} = \begin{bmatrix} 13.69150 \\ -2.20818 \end{bmatrix}$$

Then solving $\mathbf{Sx} = \mathbf{v}$ gives the same answer for $\mathbf{x}$ as in part (a).

(c) Using the GSE method, the two columns of the original $\mathbf{A}$ matrix, plus the vector $\mathbf{y}$, are used to generate an orthonormal basis set which is shown as columns of the matrix $\mathbf{V}$.

$$\mathbf{V} = \begin{bmatrix} 1.6903085E-01 & 8.9708525E-01 & 4.0824756E-01 \\ 5.0709254E-01 & 2.7602646E-01 & -8.1649655E-01 \\ 8.4515423E-01 & -3.4503257E-01 & 4.0824878E-01 \end{bmatrix}$$

Then $\mathbf{V}^T\mathbf{Ax} = \mathbf{V}^T\mathbf{y}$ can be compactly written as $\mathbf{V}^T\mathbf{W} = \mathbf{W}'$. Rounding terms of order 10^{-6} to zero gives

$$\mathbf{W}' = \left[\begin{array}{cc|c} 5.91608 & 7.43736 & 13.69150 \\ 0 & 0.82808 & -2.20821 \\ 0 & 0 & 4.08249 \end{array}\right].$$

Compare this with the Cholesky results.

The last row indicates that the norm of the residual error is $\|\mathbf{y}_e\| = 4.08256$. Ignoring the last row and solving the first two obviously again give the same answer.

6.15 After seven semesters of college a student surmises that his cumulative grade point average (GPA) is a cubic function of the number of semesters completed. His record to date is

Semester, s	1	2	3	4	5	6	7	8
Cum. GPA	2.5	3.1	2.9	2.8	2.8	3.0	3.1	??

Find the coefficients for the least-squares fit to a cubic. Also determine the residual error norm. Then use the cubic to predict his GPA on graduation day after semester 8.

It is assumed that GPA $= a_0 + a_1 s + a_2 s^2 + a_3 s^3$. In matrix form this student's data and postulated model are

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \\ 1 & 5 & 25 & 125 \\ 1 & 6 & 36 & 216 \\ 1 & 7 & 49 & 343 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 2.5 \\ 3.1 \\ 2.9 \\ 2.8 \\ 2.8 \\ 3.0 \\ 3.1 \end{bmatrix}$$

Using the GSE method leads to the following $\mathbf{W}'$ matrix, rounded off.

$$\mathbf{W}' = \left[\begin{array}{cccc|c} 2.646 & 10.583 & 52.915 & 296.324 & 7.635 \\ 0 & 5.292 & 42.322 & 291.033 & 0.283 \\ 0 & 0 & 9.165 & 109.982 & -0.033 \\ 0 & 0 & 0 & 14.700 & 0.327 \\ \hline 0 & 0 & 0 & 0 & 0.284 \end{array}\right]$$

From this the coefficients are determined as

$$\mathbf{a} = [1.828 \quad 0.993 \quad -0.270 \quad 0.022]^T \quad \text{and} \quad \|\mathbf{y}_e\| = 0.284$$

Using these coefficients the estimated GPA after eight semesters is

$$1.828 + 8(0.993) - 64(0.27) + 512(0.022) = 3.756.$$

6.16 Use the recursive least-squares algorithm to estimate $\mathbf{x}$ from Problem 6.14. Start with an initial estimate of $\mathbf{x} = \mathbf{0}$ and set $\mathbf{P}_0 = \text{diag}[10000, 10000]$. Use $\mathbf{R} = 10$ and $\mathbf{Q} = 0$ (no deweighting). Also add the following additional equations to be processed:

$$0 = 3x_1 + 6x_2, \quad 10 = 2x_1 + x_2$$

The recursive calculations are rounded off and tabulated in the order performed.

k	$\mathbf{M}_k$	$\mathbf{H}_k$	$\mathbf{y}_k$	$\mathbf{K}_k$	$\mathbf{y}_k - \mathbf{H}_k\mathbf{x}_{k-1}$	$\mathbf{x}_k$	$\mathbf{P}_k$
0	—	—	—	—	—	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}10000 & 0\\0 & 10000\end{pmatrix}$
1	$\begin{pmatrix}10000 & 0\\0 & 10000\end{pmatrix}$	(1 2)	2	$\begin{pmatrix}0.2\\0.4\end{pmatrix}$	2	$\begin{pmatrix}0.4\\0.8\end{pmatrix}$	$\begin{pmatrix}8000 & -4000\\-4000 & 2000\end{pmatrix}$
2	$\begin{pmatrix}8000 & -4000\\-4000 & 2000\end{pmatrix}$	(3 4)	3	$\begin{pmatrix}0.993\\-0.495\end{pmatrix}$	−1.4	$\begin{pmatrix}-0.990\\1.493\end{pmatrix}$	$\begin{pmatrix}49.628 & -34.474\\-34.474 & 24.815\end{pmatrix}$
3	$\begin{pmatrix}49.628 & -34.474\\-34.474 & 24.815\end{pmatrix}$	(5 6)	14	$\begin{pmatrix}0.664\\-0.415\end{pmatrix}$	9.992	$\begin{pmatrix}5.648\\-2.652\end{pmatrix}$	$\begin{pmatrix}23.245 & -18.264\\-18.264 & 14.529\end{pmatrix}$
4	$\begin{pmatrix}23.245 & -18.264\\-18.264 & 14.529\end{pmatrix}$	(3 6)	0	$\begin{pmatrix}-0.470\\0.382\end{pmatrix}$	−1.033	$\begin{pmatrix}6.134\\-3.046\end{pmatrix}$	$\begin{pmatrix}4.507 & -3.037\\-3.037 & 2.155\end{pmatrix}$
5	$\begin{pmatrix}4.507 & -3.037\\-3.037 & 2.155\end{pmatrix}$	(2 1)	10	$\begin{pmatrix}0.331\\-0.217\end{pmatrix}$	0.779	$\begin{pmatrix}6.392\\-3.216\end{pmatrix}$	$\begin{pmatrix}2.526 & -1.739\\-1.739 & 1.304\end{pmatrix}$

Notice that the estimate of $\mathbf{x}$ after three measurements is not exactly the same as was found in Problem 6.14. The difference is due to the initial estimates used for $\mathbf{x}$ and $\mathbf{P}$.

Recursive Weighted Least Squares with Discrete-Time Systems

6.17 A homogeneous linear discrete-time system is described by $\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k)$. Measured outputs are given by $\mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k) + \mathbf{e}(k)$, where $\mathbf{e}(k)$ is an error vector due to imperfect measuring devices. The precise initial conditions for this system, $\mathbf{x}(0)$, are not known, although an estimate, $\hat{\mathbf{x}}(0)$, is available.‡ Use the recursive weighted least-squares technique to develop a scheme for improving on the estimate $\hat{\mathbf{x}}(0)$ each time a new measurement $\mathbf{y}(k)$ becomes available.

In order to place this problem in the framework developed in Sec. 6.8, page 223, all that is required is a few notational changes and the elimination of the difference equation. The solution to the difference equation is $\mathbf{x}(k) = \mathbf{\Phi}(k, 0)\mathbf{x}(0)$, so that the measurement equations can be written as

$$\mathbf{y}(k) = \mathbf{C}(k)\mathbf{\Phi}(k, 0)\mathbf{x}(0) + \mathbf{e}(k)$$

‡ The circumflex ^ is used in this and the next problem to indicate an estimated quantity. It should not be confused with the notation for a unit vector used earlier.

This is in the form used in Sec. 6.8, where $\mathbf{C}(k)\mathbf{\Phi}(k, 0)$ corresponds to $\mathbf{H}_k$ used in that earlier section. In Sec. 6.8 the index k referred to the kth estimate of a constant vector $\mathbf{x}$. To avoid confusion here, $\hat{\mathbf{x}}(0|k)$ will be used to indicate the estimate of $\mathbf{x}(0)$ based on all measurements up to and including $\mathbf{y}(k)$. Assuming that the original estimate $\hat{\mathbf{x}}(0)$ is to be weighted by a nonsingular $n \times n$ matrix $\mathbf{P}_0^{-1}$ and that each succeeding measurement is weighted by the $m \times m$ nonsingular matrix $\mathbf{R}_k^{-1}$ the results of Sec. 6.8 give

$$\hat{\mathbf{x}}(0|1) = \hat{\mathbf{x}}(0) + \mathbf{K}_0[\mathbf{y}(1) - \mathbf{C}(1)\mathbf{\Phi}(1, 0)\hat{\mathbf{x}}(0)]$$

The gain matrix $\mathbf{K}_0$ is given by

$$\mathbf{K}_0 = \mathbf{P}_0\mathbf{\Phi}^T(1, 0)\mathbf{C}^T(1)[\mathbf{C}(1)\mathbf{\Phi}(1, 0)\mathbf{P}_0\mathbf{\Phi}^T(1, 0)\mathbf{C}^T(1) + \mathbf{R}_1]^{-1}$$

The estimate of $\mathbf{x}(0)$ may be further refined each time a new measurement $\mathbf{y}(k)$ is taken by using the recursive relations

$$\hat{\mathbf{x}}(0|k + 1) = \hat{\mathbf{x}}(0|k) + \mathbf{K}_k[\mathbf{y}(k + 1) - \mathbf{C}(k + 1)\mathbf{\Phi}(k + 1, 0)\hat{\mathbf{x}}(0|k)]$$

where

$$K_k = \mathbf{P}_k\mathbf{\Phi}^T(k + 1, 0)\mathbf{C}^T(k + 1)[\mathbf{C}(k + 1)\mathbf{\Phi}(k + 1, 0) \times \mathbf{P}_k\mathbf{\Phi}^T(k + 1, 0)\mathbf{C}^T(k + 1) + \mathbf{R}_{k+1}]^{-1}$$

and where $\mathbf{P}_k$ is computed recursively using Eq. (6.5), page 224, which can be written as

$$\mathbf{P}_{k+1} = \mathbf{P}_k - \mathbf{K}_k\mathbf{C}(k + 1)\mathbf{\Phi}(k + 1, 0)\mathbf{P}_k$$

If the error vector $\mathbf{e}(k)$ and the weighting matrices $\mathbf{P}_k$ and $\mathbf{R}_k$ are given the proper statistical interpretation, the above technique constitutes a simple example of the fixed-point smoothing algorithm [8]. A block diagram of the procedure is given in Figure 6.6.

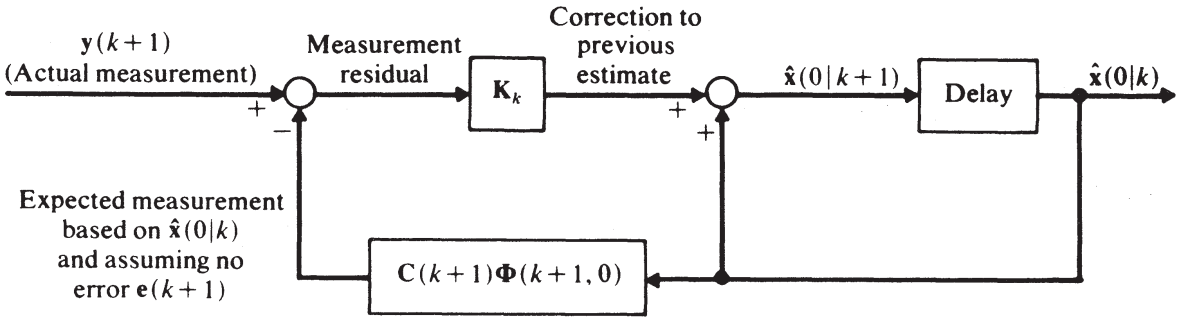


Figure 6.6

6.18 Consider the results of Problem 6.17. Instead of reestimating $\mathbf{x}(0)$ after each measurement, an estimate of the current state is now desired. Let $\hat{\mathbf{x}}(k|k)$ be the estimate of $\mathbf{x}(k)$ based on all measurements up to and including $\mathbf{y}(k)$. For this estimate, use the intuitively reasonable relation

$$\hat{\mathbf{x}}(k + 1|k + 1) = \mathbf{\Phi}(k + 1, 0)\hat{\mathbf{x}}(0|k + 1)$$

Modify the previous block diagram so that $\hat{\mathbf{x}}(k|k)$ is the output.

If the transition matrix $\mathbf{\Phi}(k + 1, 0)$ is inserted into the diagram of Figure 6.6 before the delay, then the output will be $\hat{\mathbf{x}}(k|k)$ as shown in Figure 6.7. To maintain the correct relations in the rest of the diagram, the term $\hat{\mathbf{x}}(0|k) = \mathbf{\Phi}(0, k)\hat{\mathbf{x}}(k|k)$ is needed, so $\mathbf{\Phi}(0, k)$ is inserted in the feedback path as shown in Figure 6.7.

Using standard matrix block diagram manipulations, $\mathbf{\Phi}(k + 1, 0)$ is moved past the summing junction, into both paths. A new gain matrix $\mathbf{K}'_{k+1} = \mathbf{\Phi}(k + 1, 0)\mathbf{K}_k$ is defined. The other $\mathbf{\Phi}(k + 1, 0)$ term is shifted into the feedback path and combined with $\mathbf{\Phi}(0, k)$ to give $\mathbf{\Phi}(k + 1, k) = \mathbf{\Phi}(k + 1, 0)\mathbf{\Phi}(0, k)$. This also removes the $\mathbf{\Phi}(k + 1, 0)$ term multiplying $\mathbf{C}(k + 1)$.

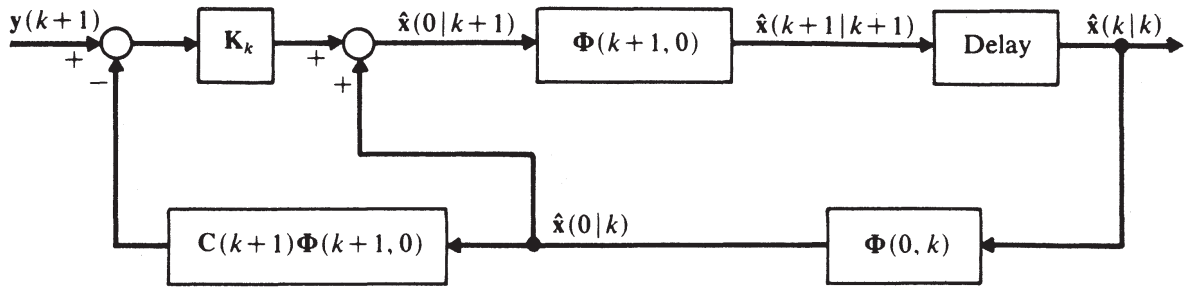


Figure 6.7

This results in the most commonly used form of the discrete Kalman filter [7, 8], shown in Figure 6.8.

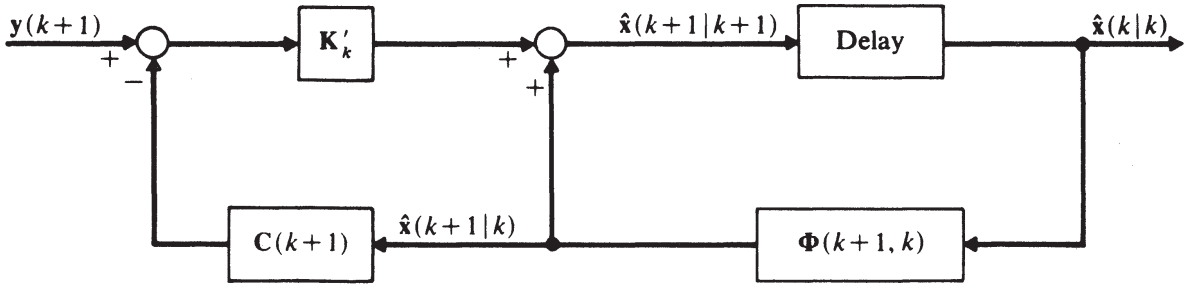


Figure 6.8

For easy reference the five equations which constitute the discrete Kalman filter are summarized below. An additive deweighting matrix $\mathbf{Q}_k$ (see Section 6.8) is included. To appreciate these results fully some knowledge of random processes is required [6]. Lacking this, the procedure can still be interpreted and used successfully as a recursive least-squares algorithm with ad hoc additive deweighting included.

To **find** the gain $\mathbf{K}'_k$, recursively compute

$$\mathbf{M}_k = \mathbf{\Phi}(k, k-1)\mathbf{P}_{k-1}\mathbf{\Phi}^T(k, k-1) + \mathbf{Q}_{k-1} \tag{1}$$

$$K'_k = \mathbf{M}_k\,\mathbf{C}^T(k)[\mathbf{C}(k)\mathbf{M}_k\,\mathbf{C}^T(k) + \mathbf{R}(k)]^{-1} \tag{2}$$

$$\mathbf{P}_k = [\mathbf{I} - \mathbf{K}'_k\,\mathbf{C}(k)]\mathbf{M}_k \tag{3}$$

To **use** the gain to estimate **x**, recursively compute

$$\hat{\mathbf{x}}(k+1|k) = \mathbf{\Phi}(k+1, k)\hat{\mathbf{x}}(k|k) \tag{4}$$

$$\hat{\mathbf{x}}(k+1|k+1) = \hat{\mathbf{x}}(k+1|k) + \mathbf{K}'_{k+1}[\mathbf{y}(k+1) - \mathbf{C}(k+1)\hat{\mathbf{x}}(k+1|k)] \tag{5}$$

To **initialize** the procedure, there are two possibilities:

(i) If $\hat{\mathbf{x}}(k+1|k)$ and $\mathbf{M}_{k+1}$ are given, then start by using equation (*2*) to find $\mathbf{K}'_{k+1}$, then use (*5*), along with the measurement $\mathbf{y}(k+1)$, to find $\hat{\mathbf{x}}(k+1|k+1)$. To get ready for the next cycle, use (*3*), (*4*), and (*1*).

(ii) If $\hat{\mathbf{x}}(k|k)$ and $\mathbf{P}_k$ are given, then start by using (*1*), then (*2*) to find $\mathbf{K}'_{k+1}$. Then use (*4*), followed by (*5*). To complete the first cycle and get ready for the next, (*3*) is then used.

Note that the above algorithm reduces to those at the end of Sec. 6.8 when $\mathbf{\Phi}(k+1, k) = \mathbf{I}$, that is, when $\mathbf{x}(k)$ is just a constant. Also be aware that the above algorithm can be written in several other forms which are *algebraically* equivalent, but which may have different numerical behavior on a finite word-length computer.

Properties of Linear Transformations

6.19 Prove that if $\mathcal{A}$ is any linear transformation there are only two possibilities regarding the null space: either (*1*) $\mathcal{N}(\mathcal{A})$ consists of only the zero vector or (*2*) $\mathcal{N}(\mathcal{A})$ contains an infinite number of vectors.

Consider $\mathcal{A}(\mathbf{x}) = \mathbf{y}$. Since $\mathcal{A}$ is linear, $\mathcal{A}(\mathbf{x} - \mathbf{x}) = \mathcal{A}(\mathbf{x}) - \mathcal{A}(\mathbf{x}) = \mathbf{y} - \mathbf{y}$ so $\mathbf{x} = \mathbf{0}$ is *always* a solution to $\mathcal{A}(\mathbf{x}) = \mathbf{0}$. Thus the null space always contains the zero vector. If $\mathbf{x}_1$ also belongs to the null space, $\mathcal{A}(\mathbf{x}_1) = \mathbf{0}$. Assume $\mathbf{x}_1 \neq \mathbf{0}$. Then the infinite set of vectors defined by $\mathbf{x} = \alpha\mathbf{x}_1$, with α any scalar, satisfies $\mathcal{A}(\mathbf{x}) = \mathcal{A}(\alpha\mathbf{x}_1) = \alpha\mathcal{A}(\mathbf{x}_1) = \mathbf{0}$. Thus if one nonzero vector belongs to the null space, then so does the infinite set of scalar multiples of it. These need not be the only vectors in $\mathcal{N}(\mathcal{A})$.

6.20 Consider the linear transformation $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$. Prove that the null space of $\mathcal{A}$ is a subspace of $\mathcal{X}$.

The solution to this problem consists of combining the results of the previous problem with the definition of a subspace. A set of vectors is a subspace if for each $\mathbf{x}_1, \mathbf{x}_2$ in the set, $\alpha\mathbf{x}_1 + \beta\mathbf{x}_2$ is also in the set, for arbitrary scalars $\alpha, \beta \in \mathcal{F}$. Let $\mathbf{x}_1$ and $\mathbf{x}_2 \in \mathcal{N}(\mathcal{A})$. That is, $\mathcal{A}(\mathbf{x}_1) = \mathbf{0}$ and $\mathcal{A}(\mathbf{x}_2) = \mathbf{0}$. Then $\mathcal{A}(\alpha\mathbf{x}_1 + \beta\mathbf{x}_2) = \alpha\mathcal{A}(\mathbf{x}_1) + \beta\mathcal{A}(\mathbf{x}_2) = \mathbf{0}$. Thus $\alpha\mathbf{x}_1 + \beta\mathbf{x}_2 \in \mathcal{N}(\mathcal{A})$ and so the null space is a subspace. It is a zero dimensional subspace if its only element is the zero vector.

6.21 Let $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ be a linear transformation, with $\mathcal{X}$ and $\mathcal{Y}$ finite dimensional. Prove that the space $\mathcal{X}$ can be written as a direct sum $\mathcal{X} = \mathcal{N}(\mathcal{A}) \oplus \mathcal{R}(\mathcal{A}^*)$.

It was shown in Problem 6.20 that $\mathcal{N}(\mathcal{A})$ is a subspace of $\mathcal{X}$. Let $\mathcal{N}(\mathcal{A})^\perp$ be the orthogonal complement of $\mathcal{N}(\mathcal{A})$. Then from the results of Section 5.9, $\mathcal{X} = \mathcal{N}(\mathcal{A}) \oplus \mathcal{N}(\mathcal{A})^\perp$.

It remains to be shown that $\mathcal{R}(\mathcal{A}^*) = \mathcal{N}(\mathcal{A})^\perp$. Let $\mathbf{x}$ be an arbitrary vector in $\mathcal{N}(\mathcal{A})$ and let $\mathbf{z}$ be an arbitrary vector in $\mathcal{Y}$. Then $\mathcal{A}(\mathbf{x}) = \mathbf{0}$ so that $\langle \mathbf{z}, \mathcal{A}(\mathbf{x}) \rangle = \langle \mathcal{A}^*(\mathbf{z}), \mathbf{x} \rangle = \mathbf{0}$. From this we see that $\mathbf{x}$ is orthogonal to $\mathcal{A}^*(\mathbf{z})$, which shows that $\mathcal{N}(\mathcal{A}) = \mathcal{R}(\mathcal{A}^*)^\perp$. The orthogonal complements are also equal, $\mathcal{N}(\mathcal{A})^\perp = (\mathcal{R}(\mathcal{A}^*)^\perp)^\perp$, but $(\mathcal{R}(\mathcal{A}^*)^\perp)^\perp = \mathcal{R}(\mathcal{A}^*)$. This completes the proof for $\mathcal{X}$ finite dimensional.

For the infinite dimensional case, the decomposition is valid if the closure of $\mathcal{R}(\mathcal{A}^*)$ is used in place of $\mathcal{R}(\mathcal{A}^*)$. Every finite dimensional space is closed.

6.22 Consider the linear transformation $\mathcal{A}(\mathbf{x}) = \mathbf{y}$, where $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$. Show that there is no unique solution for $\mathbf{x}$ if $\mathcal{N}(\mathcal{A})$ contains vectors other than the zero vector.

Suppose $\mathbf{x}_1 \in \mathcal{N}(\mathcal{A})$ and $\mathbf{x}_1 \neq \mathbf{0}$. If $\mathbf{x}$ satisfies $\mathcal{A}(\mathbf{x}) = \mathbf{y}$, then $\mathbf{x} + \mathbf{x}_1$ is also a solution, since $\mathcal{A}(\mathbf{x} + \mathbf{x}_1) = \mathcal{A}(\mathbf{x}) + \mathcal{A}(\mathbf{x}_1) = \mathbf{y} + \mathbf{0} = \mathbf{y}$.

6.23 Prove that the linear transformation $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$, with $\mathcal{A}(\mathbf{x}) = \mathbf{y}$, has a solution for every $\mathbf{y} \in \mathcal{Y}$ if and only if $\mathcal{N}(\mathcal{A}^*) = \{\mathbf{0}\}$. Assume that $\mathcal{Y}$ is finite dimensional.

The linear transformation $\mathcal{A}^*$ and its finite dimensional domain $\mathcal{Y}$ can be used in place of $\mathcal{A}$ and $\mathcal{X}$ in the result of Problem 6.21. That is, $\mathcal{Y} = \mathcal{N}(\mathcal{A}^*) \oplus \mathcal{R}((\mathcal{A}^*)^*) = \mathcal{N}(\mathcal{A}^*) \oplus \mathcal{R}(\mathcal{A})$. If $\mathcal{N}(\mathcal{A}^*) = \{\mathbf{0}\}$, then $\mathcal{Y} = \mathcal{R}(\mathcal{A})$, so that every $\mathbf{y} \in \mathcal{Y}$ is the image of at least one $\mathbf{x} \in \mathcal{X}$. Note that $\mathbf{y} \in \mathcal{R}(\mathcal{A})$ is equivalent to the requirement for existence of solutions given in Section 6.2: rank $[\mathbf{A}]$ = rank $[\mathbf{A} \quad \mathbf{y}]$. If $\mathcal{N}(\mathcal{A}^*) \neq \{\mathbf{0}\}$, then $\mathcal{Y} \neq \mathcal{R}(\mathcal{A})$; so there exists some $\mathbf{y} \in \mathcal{Y}$ but $\mathbf{y} \notin \mathcal{R}(\mathcal{A})$. Such a $\mathbf{y}$ is not the image of any $\mathbf{x} \in \mathcal{X}$.

6.24 Prove: $\mathcal{A}(\mathbf{x}) = \mathbf{y}$ has a *unique* solution for every $\mathbf{y} \in \mathcal{Y}$ if $\mathcal{N}(\mathcal{A}^*) = \{\mathbf{0}\}$ and $\mathcal{N}(\mathcal{A}) = \{\mathbf{0}\}$.

The results of Problem 6.23 guarantee that at least one solution exists. Assume $\mathbf{x}_1$ and $\mathbf{x}_2$ are two solutions, that is, $\mathcal{A}(\mathbf{x}_1) = \mathbf{y}$ and $\mathcal{A}(\mathbf{x}_2) = \mathbf{y}$. Then $\mathcal{A}(\mathbf{x}_1) - \mathcal{A}(\mathbf{x}_2) = \mathbf{0}$ or $\mathcal{A}(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{0}$. But since $\mathcal{N}(\mathcal{A}) = \{\mathbf{0}\}$, this requires that $\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{0}$ or $\mathbf{x}_1 = \mathbf{x}_2$ is the unique solution. Note that if the domain $\mathcal{X}$ is n-dimensional, $\mathcal{N}(\mathcal{A}) = \{\mathbf{0}\}$ implies that $\text{rank}(\mathcal{A}^*) = \dim \mathcal{R}(\mathcal{A}^*) = \dim \mathcal{X} = n$. Also $\mathcal{N}(\mathcal{A}^*) = \{\mathbf{0}\}$ implies that $\text{rank}(\mathcal{A}) = \dim \mathcal{R}(\mathcal{A}) = \dim \mathcal{Y}$. But $\text{rank}(\mathcal{A}) = \text{rank}(\mathcal{A}^*)$, so a unique solution requires $\text{rank}(\mathcal{A}) = n$ as shown in Section 6.2, using a matrix representation $\mathbf{A}$ for $\mathcal{A}$.

6.25 If the domain and codomain for $\mathcal{A}$ are restricted so that $\mathcal{A} : \mathcal{R}(\mathcal{A}^*) \to \mathcal{R}(\mathcal{A})$, show that $\mathcal{A}$ is one-to-one and onto, and thus possesses an inverse.

Since, in general, $\mathcal{X} = \mathcal{N}(\mathcal{A}) \oplus \mathcal{R}(\mathcal{A}^*)$ and $\mathcal{Y} = \mathcal{N}(\mathcal{A}^*) \oplus \mathcal{R}(\mathcal{A})$, restricting $\mathcal{A}$ as stated guarantees that the conditions for a unique solution, as stated in Problem 6.24, are satisfied, so the restriction of $\mathcal{A}$ is one-to-one and onto.

6.26 Let $\mathcal{A}: \mathcal{X} \to \mathcal{Y}$ be a linear transformation for which $\mathcal{N}(\mathcal{A}) \neq \{\mathbf{0}\}$, but $\mathcal{N}(\mathcal{A}^*) = \{\mathbf{0}\}$.

(a) Show that the most general solution to $\mathcal{A}(\mathbf{x}) = \mathbf{y}$ is given by $\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_1$, where $\mathbf{x}_1 \in \mathcal{R}(\mathcal{A}^*)$, $\mathbf{x}_0 \in \mathcal{N}(\mathcal{A})$.

(b) Show that $\mathbf{x}_1$ is the minimum norm solution given by $\mathbf{x}_1 = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathbf{y}$.

(1) For every $\mathbf{x} \in \mathcal{X}$, a unique decomposition is possible, $\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_1$ with $\mathbf{x}_1 \in \mathcal{R}(\mathcal{A}^*)$, $\mathbf{x}_0 \in \mathcal{N}(\mathcal{A})$.

(2) Using this decomposition gives $\mathcal{A}(\mathbf{x}) = \mathcal{A}(\mathbf{x}_0 + \mathbf{x}_1) = \mathcal{A}(\mathbf{x}_1) = \mathbf{y}$. Since $\mathbf{x}_1 \in \mathcal{R}(\mathcal{A}^*)$, there exists a $\mathbf{y}_1 \in \mathcal{Y}$ such that $\mathcal{A}^*(\mathbf{y}_1) = \mathbf{x}_1$, from which $\mathcal{A}(\mathbf{x}_1) = \mathbf{y} = \mathcal{A}\mathcal{A}^*(\mathbf{y}_1)$. Since $\mathcal{A}\mathcal{A}^*$ is a one-to-one linear transformation from $\mathcal{R}(\mathcal{A}) = \mathcal{Y}$ onto itself, and thus possesses an inverse, $\mathbf{y}_1 = (\mathcal{A}\mathcal{A}^*)^{-1}\mathbf{y}$. Using $\mathbf{x}_1 = \mathcal{A}^*(\mathbf{y}_1)$ gives the minimum norm solution $\mathbf{x}_1 = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathbf{y}$. Any vector $\mathbf{x}_0 \in \mathcal{N}(\mathcal{A})$ can be added to $\mathbf{x}_1$ and the result is still a solution, but with a larger norm.

6.27 If $\mathcal{N}(\mathcal{A}^*) \neq \{\mathbf{0}\}$ and $\mathbf{y} \notin \mathcal{R}(\mathcal{A})$, no solution to $\mathcal{A}(\mathbf{x}) = \mathbf{y}$ exists. Give a geometrical interpretation of the least-squares approximate solution, $\mathbf{x} = (\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\mathbf{y}$.

Figure 6.9 illustrates the decomposition of $\mathcal{X}$ and $\mathcal{Y}$. For any $\mathbf{y} \in \mathcal{Y}$, $\mathbf{y} = \mathbf{z} + \mathbf{w}$ with $\mathbf{z} \in \mathcal{R}(\mathcal{A})$ and $\mathbf{w} \in \mathcal{N}(\mathcal{A}^*)$. Therefore

$$\mathcal{A}^*(\mathbf{y}) = \mathcal{A}^*(\mathbf{z} + \mathbf{w}) = \mathcal{A}^*(\mathbf{z})$$

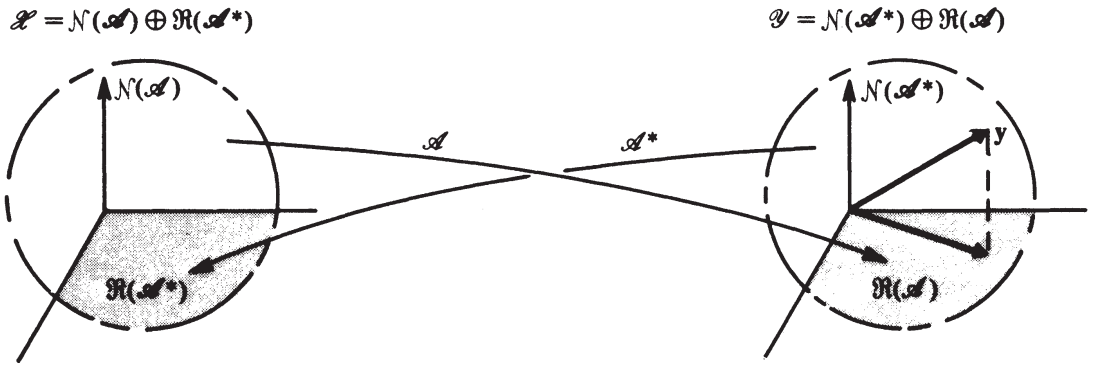


Figure 6.9

Since $\mathbf{z} \in \mathcal{R}(\mathcal{A})$, there exists some $\mathbf{x} \in \mathcal{X}$ such that $\mathcal{A}(\mathbf{x}) = \mathbf{z}$. Combining these results gives

$$\mathcal{A}^*(\mathbf{z}) = \mathcal{A}^*\mathcal{A}(\mathbf{x}) = \mathcal{A}^*(\mathbf{y})$$

If $\mathcal{N}(\mathcal{A}) \neq \{\mathbf{0}\}$ there will be many vectors $\mathbf{x}$ satisfying $\mathcal{A}(\mathbf{x}) = \mathbf{z}$. The one such vector with minimum norm is selected, i.e. the one belonging to $\mathcal{R}(\mathcal{A}^*)$. With this restriction, $\mathcal{A}^*\mathcal{A}$ is a one-to-one transformation from $\mathcal{R}(\mathcal{A}^*)$ onto itself and thus possesses an inverse. The least-squares solution is $\mathbf{x} = (\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(\mathbf{y})$. This solution is the pre-image of the projection of $\mathbf{y}$ on $\mathcal{R}(\mathcal{A})$.

6.28 Prove that every finite dimensional linear transformation is bounded.

Let $\mathcal{A}: \mathcal{X}^n \to \mathcal{X}^m$, with $\mathbf{x} \in \mathcal{X}^n$. Let $\{\mathbf{v}_i, i = 1, 2, \ldots, n\}$ be an orthonormal basis for $\mathcal{X}^n$. Then $\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$ and $\mathcal{A}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \mathcal{A}(\mathbf{v}_i)$. Therefore,

$$\|\mathcal{A}(\mathbf{x})\| = \|\sum_{i=1}^{n} \alpha_i \mathcal{A}(\mathbf{v}_i)\| \leq \sum_{i=1}^{n} |\alpha_i| \|\mathcal{A}(\mathbf{v}_i)\|$$

Since the vectors $\mathbf{y}_i \triangleq \mathcal{A}(\mathbf{v}_i)$ belong to $\mathcal{X}^m$, they have a finite norm. Define $K = \max_{i=1,\ldots,n} \|\mathbf{y}_i\|$. Then

$$\|\mathcal{A}(\mathbf{x})\| \leq \sum_{i=1}^{n} |\alpha_i| \, \|\mathbf{y}_i\| \leq K \sum_{i=1}^{n} |\alpha_i|$$

But $\alpha_i = \langle \mathbf{v}_i, \mathbf{x} \rangle$ so that the Cauchy-Schwarz inequality gives $|\alpha_i| \leq \|\mathbf{v}_i\| \cdot \|\mathbf{x}\| = \|\mathbf{x}\|$ since $\|\mathbf{v}_i\| = 1$. Using this gives $\|\mathcal{A}(\mathbf{x})\| \leq Kn\|\mathbf{x}\|$, so that $\mathcal{A}$ is clearly bounded [Eq. (5.6)]. Notice the dependence on the finite dimension n.

6.29 Let $\mathcal{A}: \mathcal{X} \to \mathcal{Y}$ be a linear transformation such that $(\mathcal{A}^*\mathcal{A})^{-1}$ exists. Prove that $\mathcal{A}(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*$ is the orthogonal projection of $\mathcal{Y}$ onto $\mathcal{R}(\mathcal{A})$.

The indicated transformation is a projection if and only if

$$[\mathcal{A}(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*][\mathcal{A}(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*] = \mathcal{A}(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*$$

This condition is obviously satisfied. Decompose $\mathcal{Y} = \mathcal{N}(\mathcal{A}^*) \oplus \mathcal{R}(\mathcal{A})$ and let $\mathbf{y} = \mathbf{y}_0 + \mathbf{y}_1$, where $\mathbf{y}_0 \in \mathcal{N}(\mathcal{A}^*)$, $\mathbf{y}_1 \in \mathcal{R}(\mathcal{A})$. The orthogonal projection of $\mathbf{y}$ onto $\mathcal{R}(\mathcal{A})$ is $\mathbf{y}_1$ by definition.

We must show that $\mathcal{A}(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\mathbf{y} = \mathbf{y}_1$. First note that

$$\mathcal{A}^*(\mathbf{y}) = \mathcal{A}^*\overbrace{(\mathbf{y}_0)}^{0} + \mathcal{A}^*(\mathbf{y}_1)$$

Both $\mathbf{y}_1$ and $\mathcal{A}(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\mathbf{y}_1 \in \mathcal{R}(\mathcal{A}) = \mathcal{N}(\mathcal{A}^*)^{\perp}$ so their difference, $\mathbf{e}$, does also. Applying the adjoint transformation $\mathcal{A}^*$ to their difference gives

$$\mathcal{A}^*[\mathcal{A}(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\mathbf{y}_1 - \mathbf{y}_1] = \mathbf{0}$$

Since $\mathbf{e} \in \mathcal{N}(\mathcal{A}^*)^{\perp}$ and $\mathcal{A}^*(\mathbf{e}) = \mathbf{0}$, the conclusion is that $\mathbf{e} = \mathbf{0}$. This leads to $\mathbf{y}_1 = \mathcal{A}(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\mathbf{y}$.

PROBLEMS

Miscellaneous

6.30 Solve for $\mathbf{x}$ if $\begin{bmatrix} 8 & 2 & 1 \\ 1 & 1 & 3 \\ 2 & 5 & 4 \end{bmatrix}\mathbf{x} = \begin{bmatrix} 10 \\ 5 \\ 1 \end{bmatrix}$.

6.31 Assume that a dynamic system can be described by a vector of time-varying parameters $\mathbf{x}(t)$, with initial conditions $\mathbf{x}(0)$. The relation between $\mathbf{x}(t)$ and $\mathbf{x}(0)$ is $\mathbf{x}(t) = \mathbf{\Phi}(t)\mathbf{x}(0)$, where $\mathbf{\Phi}(t)$ is an $n \times n$ matrix. Let $\mathbf{x}$ be partitioned into $\left[\begin{array}{c} \mathbf{x}_1(t) \\ \hline \mathbf{x}_2(t) \end{array}\right]$. If $\mathbf{x}_1(0)$ and $\mathbf{x}_2(T)$ are known, find $\mathbf{x}_2(0)$.

Homogeneous Equations

6.32 If $\mathbf{Ax} = \mathbf{0}$ has q linearly independent solutions $\mathbf{x}_i$ and $\mathbf{Ax} = \mathbf{y}$ has $\mathbf{x}_0$ as a solution, show that

(a) $\mathbf{x}_c = \sum_{i=1}^{q} \alpha_i \mathbf{x}_i$ is also a solution of $\mathbf{Ax} = \mathbf{0}$,

(b) $\mathbf{x} = \mathbf{x}_0 + \sum_{i=1}^{q} \alpha_i \mathbf{x}_i$ is a solution of $\mathbf{Ax} = \mathbf{y}$.

6.33 Find the nontrivial solutions for $\begin{bmatrix} 1 & 0 & 4 \\ 2 & 3 & 8 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}$.

6.34 Find all nontrivial solutions for $\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}$.

6.35 Determine whether nontrivial solutions exist for $\begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 0 & 0 \\ -1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}$.

6.36 Find all nontrivial solutions of $\mathbf{Ax} = \mathbf{0}$, i.e, the null space, of

$$\mathbf{A} = \begin{bmatrix} 26 & 17 & 8 & 39 & 35 \\ 17 & 13 & 9 & 29 & 28 \\ 8 & 9 & 10 & 19 & 21 \\ 39 & 29 & 19 & 65 & 62 \\ 35 & 28 & 21 & 62 & 61 \end{bmatrix}$$

Least Squares

6.37 Solve for x_1 and x_2 if
(a) $2x_1 - x_2 = 5, x_1 + 2x_2 = 3$,
(b) in addition to the equations in **a,** a third equation is $-x_1 + x_2 = -1$. Use least squares.

6.38 Given that $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} x + \begin{bmatrix} e_1 \\ e_1 \end{bmatrix}$. Measurements give $[y_1 \quad y_2] = [3 \quad 4]$. Find the least-squares estimate for x. Use a sketch in the y_1, y_2 plane to indicate the geometrical interpretation.

6.39 Verify the result of Problem 5.35, page 201 by determining the least-squares solution for **x**:

$$\begin{bmatrix} 1 & 0 \\ 2 & 1 \\ -3 & 3 \\ 1 & 3 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \\ 4 \\ 2 \\ 8 \end{bmatrix}$$

and then use the fact that the orthogonal projection of **y** on the column space of **A** is $\mathbf{y_p} = \mathbf{Ax}$.

6.40 A physical device is shown in Figure 6.10. It is believed that the output y is linearly related to the input u. That is, $y = au + b$. What are the values of a and b if the following data are taken?

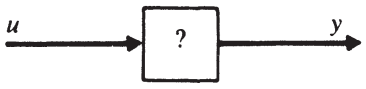


Figure 6.10

u	2	-2
y	5	1

6.41 The same device as in Problem 6.40 is considered. One more set of readings is taken as

$$u = 5, \qquad y = 7$$

Find a least-squares estimate of a and b. Also find the minimum mean-squared error in this straight line fit to the three points.

6.42 Consider the data of Problem 6.41, but assume that the first two equations are much more reliable. Use

$$\mathbf{R}^{-1} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and show that the resultant weighted least-squares estimate is much closer to the values obtained in Problem 6.40.

6.43 Estimate the initial current $i(0)$ in the circuit of Figure 6.11, if $R = 10\ \Omega$, $L = 3.56$ H, and the following voltmeter readings are taken:

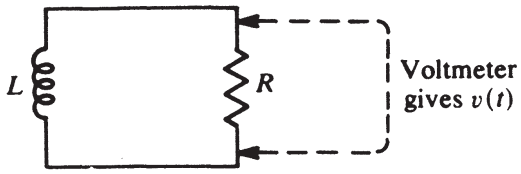


Figure 6.11

t	0	1	2	3
$v(t)$	167.9	95.5	88.8	55.3

6.44 Least-squares fit a quadratic function to the data of Problem 6.15. Determine the coefficients and the norm of the residual error. Then predict the GPA after semester eight.

6.45 An empirical theory used by many distance runners states that the time T_i required to race a distance D_i can be expressed as $T_i = C(D_i)^{\alpha}$, where C and α are constants for a given person, determined by lung capacity, body build, etc. Obtain a least-squares fit to the following data for one middle-aged jogger. (Convert to a linear equation in the unknowns C and α by taking the logarithm of the above expression.) Predict the time for one mile.

Time	185 min	79.6 min	60 min	37.9 min	11.5 min
Distance	26.2 mi	12.4 mi	9.5 mi	6.2 mi	2 mi

6.46 Apply the recursive least squares algorithm to the data of Examples 6.4 and 6.7. Try different starting assumptions to determine their effect on the estimate. Recall that **A** is not full rank and thus a unique least-squares solution does not exist. Add a fifth equation,

$$4x_1 - x_2 + 6x_3 = 2$$

so that the enlarged **A** matrix is of full rank. How do your results compare with Example 6.7?

6.47 How should the results for the minimum norm and least-squares solutions of Problems 6.26 and 6.27 be modified if a weighted norm or weighted least-squares solution is desired?

Lyapunov Equations

6.48 Solve $\mathbf{XA} + \mathbf{BX} = \mathbf{C}$ for $\mathbf{X}$ if

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 0 & 5 \end{bmatrix}, \qquad \mathbf{B} = \mathbf{A}^T, \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

6.49 Solve $\mathbf{XA} + \mathbf{BX} = \mathbf{C}$ for $\mathbf{X}$ if

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 3 & 5 \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

7

Eigenvalues and Eigenvectors

7.1 INTRODUCTION

This chapter defines eigenvalues and eigenvectors (also referred to as proper, or characteristic, values and vectors). Methods of determining eigenvalues and eigenvectors are presented as well as some of their more important properties and uses. For a linear continuous-time system—i.e., Eqs. (*3.11*), (*3.12*)—or for a linear discrete-time system—i.e., Eqs. (*3.13*), (*3.14*)—the eigenvalues of the **A** matrix completely determine system stability. The eigenvectors of **A** form a very convenient choice for basis vectors in state space. When a full set of eigenvectors can be found, it will be shown that the nth-order system can be transformed into an *uncoupled* set of n first-order equations. Each equation describes one natural mode of the system. The uncoupled form allows for easier analysis as well as providing greater insight into the system's structural properties. Unfortunately, not all matrices have a full set of eigenvectors. This can happen only when the matrix has repeated eigenvalues, plus an additional condition to be described in detail later. The most annoying complications that arise in the eigenvalue-eigenvector problem are due to this degenerate case, where less than a full set of eigenvectors exist. It has sometimes been argued (erroneously) that the repeated eigenvalue case is purely academic because small computational differences will always exist between any two eigenvalues. In fact, the degenerate case cannot be avoided so easily. Additional vectors, called generalized eigenvectors, will be defined and used to supplement the eigenvectors when necessary. Doing this will lead to a system which is as close to being uncoupled as possible.

7.2 DEFINITION OF THE EIGENVALUE-EIGENVECTOR PROBLEM

Let $\mathcal{A}$ be any linear transformation with domain $\mathcal{D}(\mathcal{A})$ and range $\mathcal{R}(\mathcal{A})$, both contained within the same linear vector space $\mathcal{X}$. Let elements of $\mathcal{X}$ be denoted as $\mathbf{x}_i$. Those particular elements $\mathbf{x}_i \neq \mathbf{0}$ and the particular scalars $\lambda_i \in \mathcal{F}$ which satisfy

$$\mathcal{A}(\mathbf{x}_i) = \lambda_i \mathbf{x}_i \tag{7.1}$$

are called *eigenvectors* and *eigenvalues,* respectively. Note that the trivial case $\mathbf{x} = \mathbf{0}$ is explicitly excluded. Thus λ_i is an eigenvalue if and only if the transformation $\mathcal{A} - \mathcal{I}\lambda_i$ has no inverse. The set of all scalars λ for which this is true is called the *spectrum of* $\mathcal{A}$.

The eigenvector problem of Eq. (7.1) applies to a more general class of operators than is needed here [1]. With the exception of the material on singular value decomposition, the transformations considered in this chapter map elements in $\mathcal{X}^n$ into other elements in $\mathcal{X}^n$, so $\mathcal{A}$ can be represented by an $n \times n$ matrix **A.** The identity transformation is represented by the unit matrix **I.** The matrix representation of Eq. (7.1) is

$$(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_i = \mathbf{0} \tag{7.2}$$

and the determination of eigenvectors is a matter of finding nontrivial solutions to a set of n homogeneous equations. If scalar eigenvalues λ_i are known, then any of the techniques of the previous chapter for solving simultaneous linear homogeneous equations can be used to find the corresponding eigenvectors $\mathbf{x}_i$. The various row-reduced-echelon- and Gram-Schmidt-based methods are easily adaptable to machine computation for this purpose. However, the eigenvalue-eigenvector problem is actually more difficult than those considered in Sec. 6.6 because the scalar λ_i is also unknown. This leads to a *nonlinear* problem because the product of the unknowns λ_i and $\mathbf{x}_i$ enters into the equations. As a starting point for this discussion, the determination of the eigenvalues is isolated and solved first. Once this is done, the remaining problem of determining the eigenvectors is linear, exactly of the type treated in Sec. 6.6. While splitting the problem this way is a customary method of discussion, it is not necessarily the best computational approach. A direct computational attack on the simultaneous determination of eigenvalues and eigenvectors is more efficient for many problems.

7.3 EIGENVALUES

It was shown in Chapter 6 that a necessary condition for the existence of nontrivial solutions to the set of n homogeneous equations (7.2) is that $\text{rank}(\mathbf{A} - \mathbf{I}\lambda_i) < n$. This is equivalent to requiring $|\mathbf{A} - \mathbf{I}\lambda_i| = 0$. When the determinant is expanded, it yields an nth-degree polynomial in the scalar λ_i—that is,

$$|\mathbf{A} - \mathbf{I}\lambda| = (-\lambda)^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \cdots + c_1\lambda + c_0 = \Delta(\lambda) \tag{7.3}$$

The roots of this algebraic equation are the eigenvalues λ_i. A fundamental result in algebra states that an nth degree polynomial has exactly n roots, so every $n \times n$ matrix **A** has exactly n eigenvalues. The nth-degree polynomial in λ is called the *characteristic polynomial* and the characteristic equation is $\Delta(\lambda) = 0$. In factored form,

$$\Delta(\lambda) = (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n) = 0$$

and the roots are $\lambda_1, \lambda_2, \ldots, \lambda_n$. In general, some of these roots may be equal. If there are $p < n$ distinct roots, $\Delta(\lambda)$ takes the form

$$\Delta(\lambda) = (-1)^n(\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2}\cdots(\lambda - \lambda_p)^{m_p}$$

This indicates that $\lambda = \lambda_1$ is an m_1-order root, $\lambda = \lambda_2$ is an m_2-order root, etc. The integer m_i is called the *algebraic multiplicity* of the eigenvalue λ_i. Of course, $m_1 + m_2 + \cdots + m_p = n$.

The problem of determining eigenvalues for $\mathcal{A}$ amounts to factoring an nth degree polynomial. For large n this is not an easy computational problem, although it is conceptually simple and will not be discussed here. Section 7.6 presents a direct iterative method of determining eigenvalues and eigenvectors, which, when applicable, avoids factoring the characteristic polynomial.

Many relationships exist between $\mathbf{A}$, c_i, and λ_i, three of which are

1. If the scalars c_i in Eq. *(7.3)* are real (the usual case), then if λ_i is a complex eigenvalue, so is $\bar{\lambda}_i$.
2. $\text{Tr}(\mathbf{A}) = \lambda_1 + \lambda_2 + \cdots + \lambda_n = (-1)^{n+1} c_{n-1}$
3. $|\mathbf{A}| = \lambda_1 \lambda_2 \cdots \lambda_n = c_0$.

The characteristic polynomial is often defined by $|\mathbf{I}\lambda_i - \mathbf{A}|$ rather than $|\mathbf{A} - \mathbf{I}\lambda_i|$. This leaves the roots unaltered but changes Eq. *(7.3)* by a factor $(-1)^n$ (see Problem 7.46).

7.4 DETERMINATION OF EIGENVECTORS

The procedure for determining eigenvectors can be divided into two possible cases, depending on the results of the eigenvalue calculations.

Case I: All the eigenvalues are distinct.

Case II: Some eigenvalues are multiple roots of the characteristic equation.

Case I: Distinct Eigenvalues

When each of the eigenvalues has algebraic multiplicity of one (i.e., they are all simple, distinct roots), then $\text{rank}(\mathbf{A} - \mathbf{I}\lambda)$ will be $n - 1$. This means that there is only one independent nontrivial solution to the homogeneous equation

$$(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_i = \mathbf{0}$$

There are any number of methods of solving this equation for the eigenvector $\mathbf{x}_i$. One which is easy to use for hand computations on small matrices is to compute the adjoint matrix $\text{Adj}(\mathbf{A} - \mathbf{I}_n \lambda)$, leaving λ as a parameter. Then, successively substituting the value for each λ_i and selecting any nonzero column will give each $\mathbf{x}_i$ eigenvector in turn. The overhead of computing the adjoint matrix is done only once, and the result gives all simple eigenvectors. Other methods involve reducing $\mathbf{A} - \mathbf{I}_n \lambda_j$ to a purely numeric matrix for each eigenvalue. Then row-reduced-echelon methods, Gram-Schmidt decomposition methods, singular-value decomposition (SVD) methods (see Problems 7.29 through 7.34), or other numerical methods of solution can be applied. The eigenvectors are not unique. If $\mathbf{x}_i$ is an eigenvector, then so is $\alpha\mathbf{x}_i$ for any nonzero scalar α. This fact is often used to normalize the eigenvectors, perhaps so that $\|\mathbf{x}_j\| = 1$. Another

common normalization forces the largest component of $\mathbf{x}_j$ to be unity. Regardless of method, a full set of n linearly independent eigenvectors can always be found for this case. They satisfy $\mathbf{A}\mathbf{x}_1 = \lambda_1 \mathbf{x}_1, \mathbf{A}\mathbf{x}_2 = \lambda_2 \mathbf{x}_2, \ldots, \mathbf{A}\mathbf{x}_n = \lambda_n \mathbf{x}_n$. By defining an $n \times n$ *modal matrix* $\mathbf{M} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \ldots \ \mathbf{x}_n]$ and an $n \times n$ diagonal matrix $\mathbf{\Lambda}$ with the ith eigenvalue in the i, i position, all n eigenvalue-eigenvector equations can be combined into one matrix equation, $\mathbf{AM} = \mathbf{M\Lambda}$. Since the eigenvectors form a linearly independent set, the rank of $\mathbf{M}$ is n, and $\mathbf{M}^{-1}$ exists. Therefore, $\mathbf{\Lambda} = \mathbf{M}^{-1}\mathbf{AM}$. This shows that a matrix $\mathbf{A}$, which has distinct eigenvalues, can always be transformed to a diagonal matrix $\mathbf{\Lambda} = \text{diag}[\lambda_1 \ \lambda_2 \ \lambda_3 \ \ldots \ \lambda_n]$ by a *similarity transformation*. A similarity transformation is a relationship between two square matrices $\mathbf{A}$ and $\mathbf{B}$ of the form $\mathbf{B} = \mathbf{Q}^{-1}\mathbf{AQ}$ for any nonsingular matrix $\mathbf{Q}$. In the particular case before where $\mathbf{B}$ was the diagonal matrix, the modal matrix played the role of $\mathbf{Q}$. In some cases (see Problems 7.25 and 7.27) the eigenvectors are mutually orthogonal and, when normalized, constitute an orthonormal set. In this case $\mathbf{M}$ is an orthogonal matrix, i.e., $\mathbf{M}^{-1} = \mathbf{M}^T$. The similarity transformation simplifies to an *orthogonal* transformation $\mathbf{\Lambda} = \mathbf{M}^T\mathbf{AM}$ in that case.

EXAMPLE 7.1 Consider the unforced portion of the state variable model

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -18 & -27 & -10 \end{bmatrix} \mathbf{x}$$

(These equations are from Problem 3.3.) Find the eigenvalues and eigenvectors of the 3×3 matrix $\mathbf{A}$. Then form the modal matrix.

First find the eigenvalues. The characteristic equation is

$$|\mathbf{A} - \mathbf{I}\lambda| = -\lambda^3 - 10\lambda^2 - 27\lambda - 18 = 0$$

Note the correspondence between the coefficients of the characteristic polynomial and the entries in the last row of $\mathbf{A}$. This is not a coincidence and always occurs when $\mathbf{A}$ is expressed in companion form, as it is here. To eliminate the minus signs, the characteristic equation can obviously be multiplied by -1 without altering its roots. This is equivalent to writing $|\mathbf{I}\lambda - \mathbf{A}| = 0$. The cubic polynomial has three roots, $\lambda_1 = -1$, $\lambda_2 = -3$, and $\lambda_3 = -6$. Note that these three distinct roots have the same values as the poles of the transfer function from which the state equations were derived (Problem 3.3). This is also not a coincidence. Next find the eigenvectors. Four different methods are demonstrated.

1. The adjoint matrix is

$$\text{Adj}(\mathbf{A} - \mathbf{I}\lambda) = \begin{bmatrix} \lambda^2 + 10\lambda + 27 & \lambda + 10 & 1 \\ -18 & \lambda^2 + 10\lambda & \lambda \\ -18\lambda & -27\lambda - 18 & \lambda^2 \end{bmatrix}$$

Any nonzero column can be selected to form the eigenvectors. For example, column 1 with $\lambda = -1$ gives $\mathbf{x}_1 = [18 \quad -18 \quad 18]^T$, column 2 with $\lambda = -3$ gives $\mathbf{x}_2 = [7 \quad -21 \quad 63]^T$, and so on. However, column 3 is clearly the easiest to use for all three eigenvectors. It is seen that $\mathbf{x}_i = [1 \quad \lambda_i \quad \lambda_i^2]^T$, and the modal matrix becomes

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -3 & -6 \\ 1 & 9 & 36 \end{bmatrix}$$

This matrix has a very special form, due to the special companion form of the matrix **A**, and is called a *Vandermonde matrix*. Notice that the numerical values selected before from the other columns are just scalar multiples of the same vectors.

2. The row-reduced echelon form of $\mathbf{A} - \mathbf{I}(-3)$ is easily found to be $\begin{bmatrix} 1 & 0 & -\frac{1}{9} \\ 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 \end{bmatrix}$. From this it is clear that the rank is 2, as it must be for a nonrepeated root, and that the eigenvector for $\lambda = -3$ is $\mathbf{x} = [1 \quad -3 \quad 9]^T$, as before.

3. The Gram-Schmidt-based **QR** decomposition of $\mathbf{A} - \mathbf{I}(-1)$ is found by the computer to be

$$\begin{bmatrix} 0.05547 & -0.44597 & -0.89332 \\ 0 & 0.89470 & -0.44666 \\ -0.99846 & -0.024776 & 0.09629 \end{bmatrix} \begin{bmatrix} 18.0277 & 27.0139 & 8.9861 \\ 0 & 1.11769 & 1.11769 \\ 0 & 0 & 0 \end{bmatrix} = [\mathbf{Q}][\mathbf{R}]$$

Since **Q** is nonsingular, $\mathbf{QRx} = \mathbf{0}$ is equivalent to just $\mathbf{Rx} = \mathbf{0}$. This triangular set of equations has all the same advantages as the RRE form and leads to the eigenvector $\mathbf{x} = [1 \quad -1 \quad 1]^T$ for $\lambda = -1$.

4. The singular-value decomposition of $\mathbf{A} - \mathbf{I}(-6)$ is $\mathbf{U\Sigma V}^T$, where

$$\mathbf{U} = \begin{bmatrix} -0.12408 & 0.79904 & -0.58835 \\ -0.15213 & -0.60122 & -0.78446 \\ 0.98055 & 0.00777 & -0.196116 \end{bmatrix}, \qquad \mathbf{\Sigma} = \begin{bmatrix} 33.3441 & 0 & 0 \\ 0 & 5.5832 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$\mathbf{V} = \begin{bmatrix} -0.55164 & 0.83363 & 0.0273896 \\ -0.82508 & -0.54058 & -0.164337 \\ -0.12219 & -0.11325 & 0.986024 \end{bmatrix}$$

Solving $\mathbf{U\Sigma V}^T\mathbf{x} = \mathbf{0}$ is simplified to solving $\mathbf{\Sigma V}^T\mathbf{x} = \mathbf{0}$ because **U** is orthogonal. Let $\mathbf{V}^T\mathbf{x} = \mathbf{w}$ temporarily. Then, because $\Sigma_{3,3} = 0$, $\mathbf{\Sigma w} = \mathbf{0}$ implies that $\mathbf{w} = [0 \quad 0 \quad 1]^T$ or some scalar multiple. **V** is also orthogonal, so that $\mathbf{x} = \mathbf{Vw}$, that is, the eigenvector **x** is that column (or those columns) of **V** which correspond to the zero elements in $\mathbf{\Sigma}$. In this case, column 3 of **V** is the eigenvector, and it is proportional to $[1 \quad -6 \quad 36]^T$, as found earlier. The extra complications of the **QR** and SVD decompositions would normally rule out these methods for hand calculations. However, they form the basis for reliable machine computations. Section 7.6 shows how the **QR** decomposition can be used to find the eigenvalues as well as the eigenvectors. ■

EXAMPLE 7.2 Use the modal matrix found in the previous example to decouple the modes of the state variable system.

Define a new vector by $\mathbf{x} = \mathbf{Mw}$. Actually this is the same state vector expressed with respect to a new basis set consisting of the eigenvectors. That is, **x** can be written in expanded form as

$$\mathbf{x} = w_1\mathbf{x}_1 + w_2\mathbf{x}_2 + \cdots + w_n\mathbf{x}_n$$

The original state equations become $\mathbf{M}\dot{\mathbf{w}} = \mathbf{AMw}$, and upon premultiplying by $\mathbf{M}^{-1}$, the result is totally uncoupled, $\dot{\mathbf{w}} = \mathbf{\Lambda w}$. The three components of **w** satisfy $\dot{w}_1 = -w_1$, $\dot{w}_2 = -3w_2$, and $\dot{w}_3 = -6w_3$. These uncoupled scalar equations each have an exponential solution of the form $w_i(t) = \exp(\lambda_i t)w_i(0)$. This shows that if any eigenvalue has a positive real part, the correspond-

ing component of **w** will grow without bound, thus forcing the entire vector **w** to infinity. The lesson is that the eigenvalues of the matrix **A** completely determine the stability of a linear, constant coefficient system. This should come as no surprise, since it was shown earlier that transfer function poles are also eigenvalues. Pole locations are at the center of stability discussions in classical control theory, as reviewed in Chapter 2. ■

Case II: Repeated Eigenvalues

When one or more eigenvalues are repeated roots of the characteristic equation, a full set of eigenvectors may or may not exist, and a deeper analysis is required. The question of whether two roots such as 4.000001 and 3.99999 are really numerical approximations of the same root or if they are distinct is postponed temporarily. The clean idealistic case with a binary yes or no answer to the repeated-root question is addressed first. The number of linearly independent eigenvectors associated with an eigenvalue λ_i repeated with an algebraic multiplicity m_i is equal to the dimension of the null space of $\mathbf{A} - \mathbf{I}\lambda_i$. This dimension is given by

$$q_i = n - \text{rank}(\mathbf{A} - \mathbf{I}\lambda_i) \qquad \text{(see Problem 7.18)}$$

and is called the degeneracy of $\mathbf{A} - \mathbf{I}\lambda_i$. The degeneracy is also called the *geometric multiplicity* of λ_i because it is the dimension of the subspace spanned by the eigenvectors. The distinction between the algebraic multiplicity m_i and the geometric multiplicity q_i of a repeated eigenvalue is crucial to finding the associated eigenvectors and, if needed, generalized eigenvectors. First, notice that the range of possible values for the integer q_i is given by $1 \leq q_i \leq m_i$. For example, a given matrix **A** might have λ_i as a triple root of $\Delta(\lambda) = 0$ so that $m_i = 3$. Yet there may only be one eigenvector ($q_i = 1$) or perhaps two eigenvectors ($q_i = 2$) or even a full set of three. An important point is restated for emphasis. *Every $n \times n$ matrix* **A** *always has a full set of n eigenvalues, but it might not have a full set of n independent eigenvectors.* It is convenient to consider three subclassifications for Case II.

***Case II$_1$*: The fully degenerate case, $q_i = m_i$.** The fully degenerate case has a full set of m_i eigenvectors associated with the repeated root λ_i. They can be found by the same types of methods described for Case I, with only minor modifications. In numerical methods such as the row-reduced-echelon, Gram-Schmidt **QR,** or SVD methods, the modification is fairly obvious: There will be q_i independent solutions to $(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_i = \mathbf{0}$ instead of just one, as in Case I and Example 7.1. In the adjoint matrix method the modification is not quite so obvious, but it is based upon Problem 6.4. The m_i independent eigenvectors associated with the repeated root can be selected as independent columns of the differentiated adjoint matrix

$$\frac{1}{(m_i - 1)!}\left\{\frac{d^{m_i - 1}}{d\lambda^{m_i - 1}}[\text{Adj}(\mathbf{A} - \mathbf{I}\lambda)]\right\}_{\lambda = \lambda_i}$$

A further complication here as compared with Case I is that if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ are eigenvectors of **A,** then so is *every* vector **y** in the subspace spanned by the $\mathbf{x}_i$ vectors. That is, any $\mathbf{y} = \Sigma\alpha_i \mathbf{x}_i$ also satisfies $(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{y} = \mathbf{0}$. This makes it more difficult to recognize the

equivalence of the eigenvectors found by different solution methods, as is demonstrated in Example 7.3. Notice that Case I is really a subset of Case II_1 with $m_i = 1$. The category must be determined separately for each eigenvalue. The only time a matrix will have a full set of eigenvectors is when each of its eigenvalues is in either Case I or Case II_1. In this situation the modal matrix **M** is formed as before, and the similarity transformation $\mathbf{M}^{-1}\mathbf{AM}$ will again give a diagonal matrix $\mathbf{\Lambda}$ with λ_i as its diagonal elements. The state equations associated with a matrix **A** with these properties can be fully decoupled, just as in Example 7.2. It is known in advance that real, symmetric matrices and Hermitian matrices will always meet these conditions and thus have a full set of eigenvectors. This property extends to the entire class of *normal* transformations defined in Sec. 5.12 and revisited in Problems 7.26 through 7.28. Many physical matrices, including impedance and admittance matrices of circuit analysis, fall into this group. This might lead to the erroneous conclusion that in practical problems a full set of eigenvectors will always exist. In control theory, the controllable canonical form of the state equations gives a matrix **A** in companion form. A companion form matrix will *always* have just one eigenvector for each eigenvalue, regardless of the multiplicity of the eigenvalues. (See Problem 7.36.)

EXAMPLE 7.3 Find the eigenvalues, eigenvectors, modal matrix, and diagonal form of

$$\mathbf{A} = \begin{bmatrix} \frac{10}{3} & 1 & -1 & -\frac{1}{3} \\ 0 & 4 & 0 & 0 \\ -\frac{2}{3} & 1 & 3 & -\frac{1}{3} \\ -\frac{2}{3} & 1 & -1 & \frac{11}{3} \end{bmatrix}$$

It is impossible to represent this matrix exactly with a finite number of digits. Keeping six-digit input accuracy, a computer routine gave the characteristic equation

$$\Delta(\lambda) = \lambda^4 - 14\lambda^3 + 72\lambda^2 - 160\lambda + 128 = 0$$

and the (approximate) roots were found to be

$$\lambda_1 = 1.999999$$

$$\lambda_2 = 3.999980$$

$$\lambda_3 = 3.999982$$

$$\lambda_4 = 4.000017$$

Is this a case of repeated roots? Here it is known that the *exact* eigenvalues are $\{2, 4, 4, 4\}$, but in a general computer solution how is the question answered? The answer directly relates to the numerical determination of the rank of a matrix. Several solutions to this problem will be presented to demonstrate this point. Actually, $\lambda_1 = 2$ is a simple root, so Case I applies. Also $\lambda_2 = 4$ is a triple root ($m_2 = 3$). In order to determine the degeneracy q_2, rank$(\mathbf{A} - 4\mathbf{I})$ must be found. Three of the four solution procedures will indicate rank automatically as part of the solution process. First the simple root is treated.

$$(2\mathbf{I} - \mathbf{A})\mathbf{x} = \begin{bmatrix} 1.33333 & -1 & 1 & 0.33333 \\ 0 & -2 & 0 & 0 \\ 0.66666 & -1 & -1 & 0.33333 \\ 0.66666 & -1 & 1 & -1.66666 \end{bmatrix} \mathbf{x} = \mathbf{0}$$

gives the solution

$$\mathbf{x}_1 = [-0.99999 \quad 0 \quad -1 \quad -0.999999]^T \approx [-1 \quad 0 \quad -1 \quad -1]^T$$

The repeated root case gives

$$(4\mathbf{I} - \mathbf{A})\mathbf{x} = \begin{bmatrix} \frac{2}{3} & -1 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & 0 \\ \frac{2}{3} & -1 & 1 & \frac{1}{3} \\ \frac{2}{3} & -1 & 1 & \frac{1}{3} \end{bmatrix} \mathbf{x} = \mathbf{0}$$

Solution Method 1: Row-reduced echelon solution shows that the rank is 1, so $q_2 = 3$; this is the fully degenerate case. There is just one independent equation for the four components of **x**, and as a result three independent solutions can be found, all of which satisfy $[\frac{2}{3} \quad -1 \quad 1 \quad \frac{1}{3}]\mathbf{x} = \mathbf{0}$. Among the infinite number of possibilities, the three used here are $\mathbf{x}_2 = [1 \quad 0 \quad 0 \quad -2]^T$, $\mathbf{x}_3 = [0 \quad 1 \quad 1 \quad 0]^T$, and $\mathbf{x}_4 = [0 \quad 1 \quad 0 \quad 3]^T$.

Solution Method 2: The modified Gram-Schmidt process was used to find the **QR** decomposition

$$4\mathbf{I} - \mathbf{A} \approx \begin{bmatrix} 0.5774 & 0.2123 & 0.2634 & 0.7431 \\ 0 & 0.6501 & 0.6384 & -0.4121 \\ 0.5774 & 0.3987 & -0.6258 & -0.3407 \\ 0.5774 & -0.6110 & 0.3624 & -0.4024 \end{bmatrix} \begin{bmatrix} 1.1547 & -1.732 & 1.732 & 0.5773 \\ 0 & 10^{-16} & 10^{-16} & 10^{-6} \\ 0 & 0 & 10^{-16} & 10^{-6} \\ 0 & 0 & 0 & 10^{-6} \end{bmatrix}$$

For the small numbers only the power-of-ten magnitude is shown. The "almost" upper-triangular **R** part shows the type of judgment necessary to determine rank. The input data were accurate only to about six decimal places, so it is reasonable to conclude that the last three rows of **R** are actually zero, giving a rank of 1 to **R**, and hence to $4\mathbf{I} - \mathbf{A}$, since **Q** has full rank. Thus $q_2 = 3$, and there are three independent solutions of

$$[1.1547 \quad -1.732 \quad 1.732 \quad 0.5773]\mathbf{x} = \mathbf{0}$$

This is proportional to the equation found using the row-reduced-echelon form, so the same eigenvectors are again valid. It is very likely that a computer would give three different members of the eigenspace, however. This will be evident in the SVD solution.

Solution Method 3: The SVD decomposition of $4\mathbf{I} - \mathbf{A}$ gave $\Sigma = \text{diag}(2.7688 \quad 10^{-8} \quad 10^{-7} \quad 0)$. The last three singular values are zero to within the accuracy of the input data. The last three columns of **V** are, therefore, eigenvectors for $\lambda = 4$. These columns are

$$\mathbf{x}_2 = [-0.90767 \quad -0.29505 \quad 0.29505 \quad 0.04505]^T$$

$$\mathbf{x}_3 = [0.04715 \quad -0.14711 \quad 0.14711 \quad -0.97698]^T$$

$$\mathbf{x}_4 = [0 \quad -0.7071 \quad -0.7071 \quad 0]^T$$

The last vector is clearly recognizable as being a normalized form of the previously found $\mathbf{x}_3$. The other two are linear combinations of the vectors found using the previous methods, but this is not obvious.

Solution Method 4: The adjoint matrix, as obtained by computer [2], is

$$\text{Adj}(\lambda\mathbf{I} - \mathbf{A}) = \lambda^3\mathbf{I}_4 + \lambda^2\mathbf{F} + \lambda\mathbf{G} + \mathbf{H}$$

where

$$\mathbf{F} = \begin{bmatrix} -10.66667 & 1 & -1 & -0.33333 \\ 0 & -10 & 0 & 0 \\ -0.66667 & 1 & -11 & -0.33333 \\ -0.66667 & 1 & -1 & -10.33333 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 37.3333 & -8 & 8 & 2.6667 \\ 0 & 32 & 0 & 0 \\ 5.3333 & -8 & 40 & 2.6667 \\ 5.3333 & -8 & 8 & 34.6667 \end{bmatrix},$$

and

$$\mathbf{H} = \begin{bmatrix} -42.6667 & 16 & -16 & -5.3333 \\ 0 & -32 & 0 & 0 \\ -10.6667 & 16 & -48 & -5.3333 \\ -10.6667 & 16 & -16 & -5.3333 \end{bmatrix}$$

When $\lambda = 2$ is substituted,

$$\text{Adj}(2\mathbf{I} - \mathbf{A}) = \begin{bmatrix} -2.66667 & 4 & -4 & -1.33333 \\ 0 & 0 & 0 & 0 \\ -2.66667 & 4 & -4 & -1.33333 \\ -2.66667 & 4 & -4 & -1.33333 \end{bmatrix}$$

All columns are proportional to the previously found $\mathbf{x}_1$. Then, with $\lambda = 4$, $\text{Adj}(4\mathbf{I} - \mathbf{A}) = [0]$, and $d/d\lambda[\text{Adj}(\lambda\mathbf{I} - \mathbf{A})] = 3\lambda^2\mathbf{I} + 2\lambda\mathbf{F} + \mathbf{G}$. When $\lambda = 4$ is substituted, this again gives the matrix [0]. Another derivative gives $\frac{1}{2}d^2/d\lambda^2[\text{Adj}(\lambda\mathbf{I} - \mathbf{A})] = 3\lambda\mathbf{I} + \mathbf{F}$. With $\lambda = 4$, this gives the following 4×4 matrix, but only three columns are independent:

$$\begin{bmatrix} 1.3333 & 1 & -1 & -0.3333 \\ 0 & 2 & 0 & 0 \\ -0.6667 & 1 & 1 & -0.3333 \\ -0.6667 & 1 & -1 & 1.6667 \end{bmatrix}$$

Any three of these four columns could be selected as eigenvectors, or any combination of them. For example, the sum of columns 2 and 3 is $[0 \quad 2 \quad 2 \quad 0]^T$, which has appeared as an eigenvector in the other solution methods.

Any valid set of eigenvectors can be used to form $\mathbf{M} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3 \quad \mathbf{x}_4]$, and then $\mathbf{M}^{-1}\mathbf{AM} \approx \text{Diag}[2 \quad 4 \quad 4 \quad 4]$ to within the 10^{-6} accuracy established by the input. ■

***Case II₂*: Simple Degeneracy, $q_i = 1$.** For this case there is just one eigenvector for each eigenvalue, regardless of the algebraic multiplicity. It can be found by using any of the methods mentioned in Case I. The more interesting question here is how to fill in for the missing eigenvectors. If the purpose is to construct a basis set, then the eigenvectors must be augmented with additional linearly independent vectors. This can be done in various ways. The additional vectors could be constructed to be orthogonal to all of the eigenvectors by using a Gram-Schmidt process. Assume this is done and the resulting set of vectors is used to form columns of the $n \times n$ matrix **T**. The result of the similarity transformation $\mathbf{T}^{-1}\mathbf{AT}$ will not be diagonal, although it will be upper triangular and perhaps close to diagonal, depending upon how many non-eigenvectors are included in **T**. This means that state equations with this **A** matrix cannot be fully decoupled. In fact, no similarity transformation exists which will diagonalize **A** in Case II_2 or Case II_3 to follow. If a special class of augmenting vectors, called *generalized eigenvectors,* is used instead of constructing some arbitrary orthogonal set, the diagonal matrix (i.e., the possibility of decoupling) is more nearly achieved. From here forward it is assumed that generalized eigenvectors will be used to fill in where needed. The matrix formed from the set of n eigenvectors and generalized eigenvectors will again be referred to as the modal matrix **M** rather than the matrix **T** just used. The claim is that $\mathbf{M}^{-1}\mathbf{AM} = \mathbf{J}$ will be as nearly diagonal as possible. The matrix **J** is called the *Jordan form.*

The determination of generalized eigenvectors and the Jordan form is now discussed in detail. Suppose that λ_i has an algebraic multiplicity m_i. Since $q_i = 1$ by assumption in Case II_2, there is one eigenvector $\mathbf{x}_1$ and $m_i - 1$ generalized eigenvectors are required. They are defined by the string or chain of equations

$$\mathbf{A}\mathbf{x}_1 = \lambda_i \mathbf{x}_1 \qquad \text{(the usual eigenvalue equation)}$$

$$\mathbf{A}\mathbf{x}_2 = \lambda_i \mathbf{x}_2 + \mathbf{x}_1, \qquad \mathbf{A}\mathbf{x}_3 = \lambda_i \mathbf{x}_3 + \mathbf{x}_2, \ldots, \mathbf{A}\mathbf{x}_{mi} = \lambda_i \mathbf{x}_{mi} + \mathbf{x}_{mi-1}$$

Each equation in the chain except the first is coupled to the preceding equation. Assume for the moment that there are no other eigenvalues, that is $m_i = n$. The preceding chain of equations can be written as one matrix equation:

$$\mathbf{A}[\mathbf{x}_1 \quad \mathbf{x}_2 \ldots \mathbf{x}_{n-1} \quad \mathbf{x}_n] = [\mathbf{x}_1 \quad \mathbf{x}_2 \ldots \mathbf{x}_{n-1} \quad \mathbf{x}_n] \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & & 0 \\ 0 & \lambda_i & 1 & \cdots & & 0 \\ 0 & 0 & \lambda_i & \cdots & & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & \lambda_i & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_i \end{bmatrix}$$

This explicitly shows one example of the Jordan form matrix $\mathbf{J}$. It has the same repeated eigenvalue in every diagonal position and a 1 in every position above the main diagonal. In the more general case where there are other eigenvalues in addition to the m_i multiple root, the Jordan form will be a block diagonal matrix

$$\mathbf{J} = \text{Diag}[\mathbf{J}_1, \mathbf{J}_2, \ldots, \mathbf{J}_p]$$

with each of the $\mathbf{J}_i$ submatrices, called *Jordan blocks,* having the structure just shown explicitly. There will be one $m_i \times m_i$ Jordan block associated with each eigenvalue of multiplicity m_i that satisfies the conditions of Case II_2. Repeated eigenvalues satisfying Case II_1 will have m_i separate 1×1 Jordan blocks $\mathbf{J}_i = [\lambda_i]$, and the nonrepeated Case I will also have separate 1×1 blocks along the diagonal. That is, the diagonal matrix $\mathbf{\Lambda}$ is included in the definition of the Jordan form $\mathbf{J}$ as a special case. From what has been presented so far, it may be surmised that the Jordan form of a matrix will have as many separate Jordan blocks as there are eigenvectors and as many ones just above the main diagonal as there are generalized eigenvectors. This observation is true in general, even for Case II_3, which is discussed shortly.

EXAMPLE 7.4 Find the eigenvalues, eigenvectors, generalized eigenvectors, if needed, and the Jordan form for the companion form matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -8 & -20 & -18 & -7 \end{bmatrix}$$

The characteristic equation $\lambda^4 + 7\lambda^3 + 18\lambda^2 + 20\lambda + 8 = 0$ has roots $\lambda_i = \{-1, -2, -2, -2\}$. The simple root $\lambda_1 = -1$ belongs to Case I, and the corresponding eigenvector is easily found to be

$$\mathbf{x}_1 = [-1 \quad 1 \quad -1 \quad 1]^T$$

For $\lambda_2 = -2$, $m_2 = 3$, and $q_2 = 1$, since the row-reduced-echelon form of

$$\mathbf{A}-\mathbf{I}\lambda_2=\begin{bmatrix}2&1&0&0\\0&2&1&0\\0&0&2&1\\-8&-20&-18&-5\end{bmatrix}\quad\text{is}\quad\begin{bmatrix}1&0&0&0.125\\0&1&0&-0.25\\0&0&1&0.5\\0&0&0&0\end{bmatrix}$$

From this, the only eigenvector for λ_2 is $\mathbf{x}_2=[0.125\quad -0.25\quad 0.5\quad -1]^T$. The generalized eigenvector $\mathbf{x}_3$ must satisfy $(\mathbf{A}-\mathbf{I}\lambda_2)\mathbf{x}_3=\mathbf{x}_2$, which gives $\mathbf{x}_3=[0.1875\quad -0.25\quad 0.25\quad 0]^T$. Then $(\mathbf{A}-\mathbf{I}\lambda_2)\mathbf{x}_4=\mathbf{x}_3$, giving $\mathbf{x}_4=[0.1875\quad -0.1875\quad 0.125\quad 0]^T$. There are many other valid answers. Using the four $\mathbf{x}_i$ vectors as columns in $\mathbf{M}$ gives

$$\mathbf{M}^{-1}\mathbf{AM}=\left[\begin{array}{c|ccc}-1&0&0&0\\\hline 0&-2&1&0\\0&0&-2&1\\0&0&0&-2\end{array}\right]=\mathbf{J}=\text{Diag}[-1,\mathbf{J}_2]$$

where $\mathbf{J}_2$ is a 3×3 Jordan block for $\lambda_2=-2$. ■

Case II₃. The general case for an eigenvalue of algebraic multiplicity m_i and degeneracy q_i satisfying $1\le q_i\le m_i$ still has q_i eigenvectors associated with λ_i. There will be one Jordan block for each eigenvector; that is, λ_i will have q_i blocks associated with it. Case II_3 is really just a combination of the previous two cases, but knowledge of m_i and q_i still leaves some ambiguity. Assume λ_1 is a fourth-order root of the characteristic equation and assume $q_1=2$. Then it is known that there are two eigenvectors and two generalized eigenvectors. The eigenvectors satisfy $\mathbf{Ax}_a=\lambda_1\mathbf{x}_a$ and $\mathbf{Ax}_b=\lambda_1\mathbf{x}_b$, but it is still uncertain whether the generalized eigenvectors are both associated with $\mathbf{x}_a$ or both with $\mathbf{x}_b$ or one with each. That is, the two Jordan blocks could take one of the following forms:

$$\mathbf{J}_1=\begin{bmatrix}\lambda_1&1&0\\0&\lambda_1&1\\0&0&\lambda_1\end{bmatrix},\qquad \mathbf{J}_2=[\lambda_1]\quad\text{or}\quad \mathbf{J}_1=\begin{bmatrix}\lambda_1&1\\0&\lambda_1\end{bmatrix},\qquad \mathbf{J}_2=\begin{bmatrix}\lambda_1&1\\0&\lambda_1\end{bmatrix}$$

The first pair corresponds to the equations

$$\mathbf{Ax}_1=\lambda_1\mathbf{x}_1,\qquad \mathbf{Ax}_2=\lambda_1\mathbf{x}_2+\mathbf{x}_1,\qquad \mathbf{Ax}_3=\lambda_1\mathbf{x}_3+\mathbf{x}_2,\qquad \mathbf{Ax}_4=\lambda_1\mathbf{x}_4$$

The second pair corresponds to

$$\mathbf{Ax}_1=\lambda_1\mathbf{x}_1,\qquad \mathbf{Ax}_2=\lambda_1\mathbf{x}_2+\mathbf{x}_1,\qquad \mathbf{Ax}_3=\lambda_1\mathbf{x}_3,\qquad \mathbf{Ax}_4=\lambda_1\mathbf{x}_4+\mathbf{x}_3$$

Ambiguities such as this can be resolved by a trial-and-error process [3] or they can be avoided by using a systematic method from Section 7.5. Combinations of the preceding cases may be required for a given $n\times n$ matrix **A.** The applicable case can be different for each eigenvalue. For example, if the eigenvalues for some 9×9 matrix were $\{2,3,3,5,5,5,6,6,6\}$, $\lambda=2$ is of necessity an example of Case I. $\lambda=3$ might have two eigenvectors, i.e., Case II_1; $\lambda=5$ might have only one eigenvector, i.e., Case II_2; and $\lambda=6$ might have two eigenvectors and one generalized eigenvector, i.e., Case II_3. Finding a total of n vectors, m_i for each eigenvalue with multiplicity m_i, allows the nonsingular modal matrix **M** to be formed. The similarity transformation $\mathbf{M}^{-1}\mathbf{AM}$ again gives the Jordan form **J.** The diagonal matrix $\mathbf{\Lambda}$ of Case I is considered a special case of the Jordan form with all of its Jordan blocks being 1×1.

Summary. Every $n \times n$ matrix has n eigenvalues and n linearly independent vectors, either eigenvectors or generalized eigenvectors. The eigenvalues are roots of an nth degree polynomial. For each repeated eigenvalue the degeneracy q_i should be found. There will be q_i eigenvectors and Jordan blocks associated with λ_i. If $q_i < m_i$, then generalized eigenvectors will be required. This type of analysis makes it clear how many eigenvectors, generalized eigenvectors, and Jordan blocks there are, as demonstrated in Example 7.5. The actual determination of the generalized eigenvectors and the removal of any remaining ambiguities are discussed in more detail in Section 7.5.

EXAMPLE 7.5 Let $\mathbf{A}$ be an 8×8 matrix and assume that the eigenvalues have been found as $\lambda_1 = \lambda_2 = 2, \lambda_3 = \lambda_4 = \lambda_5 = \lambda_6 = -3, \lambda_7 = \lambda_8 = 4$. If $\text{rank}[\mathbf{A} - 2\mathbf{I}] = 7$, $\text{rank}[\mathbf{A} + 3\mathbf{I}] = 6$, and $\text{rank}[\mathbf{A} - 4\mathbf{I}] = 6$, find the degeneracies and determine how many eigenvectors and generalized eigenvectors there are. Also, write down the Jordan form.

For $\lambda_1 = 2, q_1 = 8 - 7 = 1$. This is the simple degeneracy Case II_2, so $\mathbf{x}_1$ is one eigenvector and $\mathbf{x}_2$ must be a generalized eigenvector. For $\lambda_3 = -3, q_3 = 8 - 6 = 2$. This falls into Case II_3 and there are two eigenvectors (and Jordan blocks) and two generalized eigenvectors must be associated with this root. For $\lambda_7 = 4, q_7 = 8 - 6 = 2$. This is Case II_1, since $q_7 = m_7 = 2$. There are two eigenvectors and no generalized eigenvectors associated with this eigenvalue. There are a total of five eigenvectors (and Jordan blocks), and three generalized eigenvectors. The Jordan form is either

$$\mathbf{J} = \begin{bmatrix} 2 & 1 & & & & & & \\ 0 & 2 & & & & & & \\ & & -3 & 1 & & \mathbf{0} & & \\ & & 0 & -3 & & & & \\ & & & & -3 & 1 & & \\ & & \mathbf{0} & & 0 & -3 & & \\ & & & & & & 4 & \\ & & & & & & & 4 \end{bmatrix} \quad \text{or} \quad \mathbf{J} = \begin{bmatrix} 2 & 1 & & & & & & \\ 0 & 2 & & & & \mathbf{0} & & \\ & & -3 & 1 & 0 & & & \\ & & 0 & -3 & 1 & & & \\ & & 0 & 0 & -3 & & & \\ & & & & & -3 & & \\ & \mathbf{0} & & & & & 4 & \\ & & & & & & & 4 \end{bmatrix} \blacksquare$$

7.5 DETERMINATION OF GENERALIZED EIGENVECTORS

It is assumed throughout this section that one or more multiple eigenvalues exist for a matrix $\mathbf{A}$ and that a need for generalized eigenvectors has already been established. Three alternative methods are presented for finding generalized eigenvectors.

1. *The first method is a bottom-up method in that the eigenvectors are found first and then a chain of one or more generalized eigenvectors is built up from these.* That is, first find all solutions of the homogeneous equation

$$(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_i = \mathbf{0}$$

for the repeated eigenvalue λ_i. For each $\mathbf{x}_i$ thus determined, try to construct a generalized eigenvector using

$$(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_{i+1} = \mathbf{x}_i$$

If the resultant vector $\mathbf{x}_{i+1}$ is linearly independent of all vectors already found, it is a valid generalized eigenvector. If still more generalized eigenvectors are needed for λ_i, then solve

$$(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_{i+2} = \mathbf{x}_{i+1}$$

and so on until all needed vectors are found. This method can be efficiently used in the simply degenerate cases such as Example 7.4 because there is only a single eigenvector and a single chain of generalized eigenvector equations. Because of possible ambiguities about how the chains of equations are connected in the general case, a more systematic method is desirable.

2. A second method is to use the adjoint matrix $\text{Adj}(\mathbf{I}\lambda - \mathbf{A})$, which is also called the *resolvent* matrix of **A,** and its various derivatives. Effective algorithms for computing the resolvent matrix are available [2]. This is a bottom-up method also, since eigenvectors are found first. This is done by selecting linearly independent columns of $\text{Adj}(\mathbf{I}\lambda - \mathbf{A})$ with a particular eigenvalue λ_i. Judgment is withheld about the final set of vectors to be retained, since repetitions often will occur in the process to follow. If λ_i is an m_i-repeated eigenvalue, it is possible that fewer than the required m_i vectors will be found on the first step. Some or even all columns of the resolvent matrix may be zero, and others may be linearly dependent. The derivative of the resolvent matrix is then evaluated at the repeated eigenvalue. Some columns may still be zero on this second step. If a given column j is not zero on step 2, then (1) it is an eigenvector if column j was zero on the previous step and (2) it is a generalized eigenvector if column j was not zero on the previous step. Step 2 may still not yield the required m_i independent vectors, so the second derivative of the resolvent matrix is taken. Again, any column j which is nonzero is either an eigenvector or a generalized eigenvector, depending on whether column j was zero or not on the previous step. These relationships require the retention of the factorial divisor, which appeared to be an irrelevant scale factor in Case II_1. The vectors obtained by this process cannot be arbitrarily rescaled on a given step, since they are all tied together in an interdependent chain. Also note that the same eigenvector can appear more than once. That is, column j on one step may yield the same eigenvector as column k on some other step. For this reason the final selection of eigenvectors should be made only after seeing all the columns from all of the steps.

EXAMPLE 7.6 Consider the matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$. Clearly $\lambda_i = 1$ has algebraic multiplicity $m = 3$, and the rank of $\mathbf{A} - \mathbf{I}\lambda$ is 2, so $q = 1$. There is just one eigenvector, so two generalized eigenvectors are required (Case II_2). The adjoint matrix is

$$\text{Adj}(\mathbf{A} - \mathbf{I}\lambda) = \begin{bmatrix} (1-\lambda)^2 & -2(1-\lambda) & 8 - 3(1-\lambda) \\ 0 & (1-\lambda)^2 & -4(1-\lambda) \\ 0 & 0 & (1-\lambda)^2 \end{bmatrix} = \lambda^2\mathbf{I} + \lambda\mathbf{F} + \mathbf{G}$$

where

$$\mathbf{F} = \begin{bmatrix} -2 & 2 & 3 \\ 0 & -2 & 4 \\ 0 & 0 & -2 \end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} 1 & -2 & 5 \\ 0 & 1 & -4 \\ 0 & 0 & 1 \end{bmatrix}$$

With $\lambda = 1$, this gives $\begin{bmatrix} 0 & 0 & 8 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, and one eigenvector is evident in column 3. The derivative

gives $d[\text{Adj}(\mathbf{A} - \mathbf{I}\lambda)/d\lambda = 2\lambda\mathbf{I} + \mathbf{F}$. When evaluated at $\lambda = 1$, this gives $\begin{bmatrix} 0 & 2 & 3 \\ 0 & 0 & 4 \\ 0 & 0 & 0 \end{bmatrix}$. Column 2 is nonzero for the first time and hence is an eigenvector. However, it is just a rescaled copy of the one found on the first step in column 3. On this second step column 3 is a generalized eigenvector. The final derivative (the m_ith for this Case II_2) is $\frac{1}{2}d^2[\text{Adj}(\mathbf{A} - \mathbf{I}\lambda)]/d\lambda^2 = \mathbf{I}$. Column 1 is nonzero for the first time and hence is an eigenvector. But, except for scaling, it is the same one found twice before. Column 2 is a generalized eigenvector if column 2 of step 2 is used as the eigenvector. This would still leave us one short of the needed three vectors, and further derivatives will only give zero columns. Therefore, column 2 must be rejected. From column 3 the final set is $\mathbf{x}_1 = [8 \quad 0 \quad 0]^T$, $\mathbf{x}_2 = [3 \quad 4 \quad 0]^T$, and $\mathbf{x}_3 = [0 \quad 0 \quad 1]^T$. ■

The adjoint (or resolvent matrix) method is workable for small hand calculations and can be adapted to machine calculation as well. The next method may be better suited to machine implementation for larger problems and can also be used for hand calculation with small problems.

3. The third method of finding eigenvectors and generalized eigenvectors is a *top-down* method. Rather than finding all the eigenvectors first and then building the necessary chains of generalized eigenvectors on them, we find the maximum number m_i of linearly independent vector solutions to a modified problem $(\mathbf{A} - \mathbf{I}\lambda_i)^k \mathbf{x} = \mathbf{0}$. All the eigenvectors and generalized eigenvectors associated with λ_i belong to the m_i-dimensional space spanned by the m_i solution vectors. The eigenvectors belong, because for any integer $j > 1$, $(\mathbf{A} - \mathbf{I}\lambda_i)^j \mathbf{x} = \mathbf{0}$ if it is true for $j = 1$. A jth-order generalized eigenvector must satisfy $(\mathbf{A} - \mathbf{I}\lambda_i)^j \mathbf{x} = \mathbf{0}$ and $(\mathbf{A} - \mathbf{I}\lambda_i)^{j-1} \mathbf{x} = \mathbf{x}_c \neq \mathbf{0}$, for $j = k, k-1, \ldots, 2$. This is consistent with the bottom-up construction equations for generalized eigenvectors:

$$
\begin{aligned}
&(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_1 = \mathbf{0} \\
&(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_2 = \mathbf{x}_1 \Rightarrow (\mathbf{A} - \mathbf{I}\lambda_i)^2 \mathbf{x}_2 = (\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_1 = \mathbf{0} \\
&(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_3 = \mathbf{x}_2 \Rightarrow (\mathbf{A} - \mathbf{I}\lambda_i)^2 \mathbf{x}_3 = (\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_2 = \mathbf{x}_1 \neq \mathbf{0} \\
&\qquad\qquad (\mathbf{A} - \mathbf{I}\lambda_i)^3 \mathbf{x}_3 = (\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_1 = \mathbf{0} \qquad (7.4) \\
&(\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_4 = \mathbf{x}_3 \Rightarrow (\mathbf{A} - \mathbf{I}\lambda_i)^2 \mathbf{x}_4 = (\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_3 = \mathbf{x}_2 \neq \mathbf{0} \\
&\qquad\qquad (\mathbf{A} - \mathbf{I}\lambda_i)^3 \mathbf{x}_4 = (\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_2 = \mathbf{x}_1 \neq \mathbf{0} \\
&\qquad\qquad (\mathbf{A} - \mathbf{I}\lambda_i)^4 \mathbf{x}_4 = (\mathbf{A} - \mathbf{I}\lambda_i)\mathbf{x}_1 = \mathbf{0}
\end{aligned}
$$

This pattern continues up to some maximum integer k_i, the index of λ_i. The key to the top-down method is finding the correct integer k. It is the *index* of the eigenvalue and is the smallest integer such that

$$\text{rank}(\mathbf{A} - \mathbf{I}\lambda_i)^k = n - m_i$$

The index k_i indicates the length of the longest chain of eigenvectors-generalized eigenvectors for λ_i. It also is the size of the largest Jordan block for λ_i in the Jordan form.

After finding the index and the m_i independent solution vectors, a simple testing procedure, consisting of matrix multiplications from the left side of Eq. (7.4), indicates whether each vector is a generalized eigenvector or an eigenvector. The same matrix multiplications also give each successive vector in the chain until the final member—i.e., the eigenvector—is found. The procedure is complete if it is explicitly ensured that the eigenvectors are included in the set of m_i vectors found at the top.

Finding the index is the key to avoiding the ambiguities about how the various chains should be formed. If in Example 7.5 the index for $\lambda_i = -3$ were found to be 2, then the first form for $\mathbf{J}$ would be correct. If the index were 3, then the second Jordan form would be correct.

EXAMPLE 7.7 Find the eigenvalues, eigenvectors, generalized eigenvectors, and Jordan form for

$$\mathbf{A} = \begin{bmatrix} -5 & \frac{1}{6} & -\frac{1}{6} & 0 \\ -\frac{1}{2} & -\frac{16}{3} & \frac{1}{3} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{3} & -\frac{14}{3} & \frac{1}{2} \\ 0 & -\frac{1}{6} & \frac{1}{6} & -5 \end{bmatrix}$$

The characteristic equation is $\Delta(\lambda) = (\lambda + 5)^4 = 0$, so $\lambda = -5$ has algebraic multiplicity $m = 4$. Since $n = 4$, the index k must be found for which $\text{rank}(\mathbf{A} - \mathbf{I}\lambda)^k = 0$. Singular-value decomposition, Gram-Schmidt **QR** decomposition, or RRE methods can be applied to find that $\text{rank}(\mathbf{A} - \mathbf{I}\lambda) = 2$. To be specific, the SVD decomposition gives $\mathbf{\Sigma} = \text{diag}(1.2176, 0.27395, 0, 0)$, and the last two columns of $\mathbf{V}$ give eigenvectors $\mathbf{x}_a = [0 \quad 1 \quad 1 \quad 0]^T$ and $\mathbf{x}_b = [1 \quad 0 \quad 0 \quad 1]^T$. These are the only two independent eigenvectors, but so far it is not clear if the two generalized eigenvectors are connected in a single chain to one eigenvector (giving a 3×3 Jordan block and a 1×1 Jordan block) or if they form two separate chains (giving two 2×2 Jordan blocks). Forming

$$(\mathbf{A} - \mathbf{I}\lambda)^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{6} & \frac{1}{6} & 0 \\ 0 & -\frac{1}{6} & \frac{1}{6} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

reveals that its rank is 1. So far it has been found that the index k is neither 1 nor 2. Forming $(\mathbf{A} - \mathbf{I}\lambda)^3 = [\mathbf{0}]$ (on the computer it was zero to within order 10^{-7}) shows that $k = 3$. It also indicates that one chain is of length 3. The other will be just an isolated eigenvector, and $\mathbf{J}$ will have 1×1 and 3×3 Jordan blocks. All ambiguity has been removed, except for the unimportant order of the Jordan blocks within the Jordan form. This depends only on the order in which the isolated eigenvector and the chain of three are placed in the modal matrix $\mathbf{M}$. The actual finding of these vectors is now demonstrated using the top-down method. Any nonzero vector is a nontrivial solution of $(\mathbf{A} - \mathbf{I}\lambda)^3 \mathbf{x} = \mathbf{0}$ in this case. Choices consisting of the two known eigenvectors plus any two additional independent vectors will suffice. (Actually, if the eigenvectors are not explicitly included, one will be found automatically at the end of the chain of length 3. The eigenvector equation can then be used to find the other one.) Four convenient linearly independent vectors for this problem are

$$\mathbf{x}_a = [1 \quad 0 \quad 0 \quad 1]^T, \qquad \mathbf{x}_b = [0 \quad 1 \quad 1 \quad 0]^T$$

$$\mathbf{x}_c = [0 \quad 1 \quad 0 \quad 0]^T, \qquad \mathbf{x}_d = [0 \quad 0 \quad 0 \quad 1]^T$$

It is not clear yet how these vectors chain together, so the testing procedure is invoked. Let $\mathbf{C} = (\mathbf{A} - \mathbf{I}\lambda_i)$, with $\lambda_i = -5$. Then $\mathbf{C}^2\mathbf{x}_a$ and $\mathbf{C}^2\mathbf{x}_b$ are zero, confirming that $\mathbf{x}_a$ and $\mathbf{x}_b$ are not generalized eigenvectors. It is also found that $\mathbf{C}^2\mathbf{x}_d = \mathbf{0}$, so $\mathbf{x}_d$ is also not a generalized eigenvector. Only $\mathbf{C}^2\mathbf{x}_c$ is nonzero, so $\mathbf{x}_c$ is the generalized eigenvector which starts the chain of three. Call it $\mathbf{x}_4$, thus indicating its ultimate column position in the modal matrix. Then $(\mathbf{A} - \mathbf{I}\lambda)\mathbf{x}_4 = \mathbf{x}_3 = [0.16667 \quad -0.33333 \quad -0.33333 \quad -0.166667]^T$. Next in the chain is $(\mathbf{A} - \mathbf{I}\lambda)\mathbf{x}_3 = \mathbf{x}_2 = [0 \quad -0.16667 \quad -0.16667 \quad 0]^T$. This is a multiple of $\mathbf{x}_b$ found earlier and is an eigenvector (not a generalized eigenvector), as seen at the next step, $(\mathbf{A} - \mathbf{I}\lambda)\mathbf{x}_2 = \mathbf{0}$. The zero vector always signals the end of the chain. The fourth vector is a stand-alone eigenvector $\mathbf{x}_a = [1 \quad 0 \quad 0 \quad 1]^T$, renamed $\mathbf{x}_1$. Using these four vectors as columns in $\mathbf{M}$ gives

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \left[\begin{array}{c|ccc} -5 & 0 & 0 & 0 \\ \hline 0 & -5 & 1 & 0 \\ 0 & 0 & -5 & 1 \\ 0 & 0 & 0 & -5 \end{array}\right]$$

where the zero elements all had magnitude on the order of 10^{-7} or less. ■

EXAMPLE 7.8 Let

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The characteristic equation is $\lambda^4 = 0$, so $\lambda_1 = 0$ with $m_1 = 4$. Since $\text{rank}(\mathbf{A} - \mathbf{I}\lambda_1) = 2$, there are $q = 2$ eigenvectors and also 2 generalized eigenvectors. Since $n - m = 0$, and since $\text{rank}[\mathbf{A} - \mathbf{I}\lambda_1]^2 = 0$, $k_1 = 2$. The largest Jordan block is 2×2. Since there are just two blocks, they both must be 2×2. To find the eigenvectors and generalized eigenvectors, consider $(\mathbf{A} - \mathbf{I}\lambda_1)^2\mathbf{x} = \mathbf{0}$. Any vector satisfies this equation, but there are at most four linearly independent solutions. Select $\mathbf{x}_a = [1 \quad 0 \quad 0 \quad 0]^T$. This is *not* a generalized eigenvector since $(\mathbf{A} - \mathbf{I}\lambda_1)\mathbf{x}_a = \mathbf{0}$. Similarly $\mathbf{x}_b = [0 \quad 1 \quad 0 \quad 0]^T$ is not a generalized eigenvector. Select $\mathbf{x}_c = [0 \quad 0 \quad 1 \quad 0]^T$. Then since $[\mathbf{A} - \mathbf{I}\lambda_1]\mathbf{x}_c = [1 \quad 0 \quad 0 \quad 0]^T \neq \mathbf{0}$, $\mathbf{x}_c$ is a generalized eigenvector associated with the eigenvector $[1 \quad 0 \quad 0 \quad 0]^T$. Finally $\mathbf{x}_d = [0 \quad 0 \quad 0 \quad 1]^T$ is a generalized eigenvector and $[\mathbf{A} - \mathbf{I}\lambda_1]\mathbf{x}_d = [0 \quad 1 \quad 0 \quad 0]^T$ is the associated eigenvector. Thus the modal matrix and the Jordan form are

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{J} = \left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

See page 259 of Reference 3 for a trial-and-error solution to the same problem. ■

7.6 ITERATIVE COMPUTER METHODS FOR DETERMINING EIGENVALUES AND EIGENVECTORS

In all discussions to this point the eigenvalue-eigenvector problem has been split into two parts. First the roots of the characteristic equation—i.e., the eigenvalues—were found by some sort of polynomial root-finding routine such as Newton-Raphson. Only

then was the eigenvector problem considered. Because of the difficulty in accurately factoring high-degree polynomials, other iterative computer algorithms are often used to determine the eigenvalues more directly. In some cases, such as real, symmetric matrices, the eigenvectors are found simultaneously with the eigenvectors. In other procedures the determination of the eigenvectors is still a separate calculation. Two general methods are presented in this section. The first method is restricted to real, symmetric matrices. Thus it is known at the outset that (1) all λ_i will be real, (2) even if some eigenvalue is repeated, a full set of eigenvectors will always exist, and (3) the eigenvectors form an orthogonal set. The simple version of the first algorithm to be presented here assumes that no two eigenvalues have the same magnitude, that is, $|\lambda_i| \neq |\lambda_j|$ if $i \neq j$. Any vector $\mathbf{z}_0$ can be written as

$$\mathbf{z}_0 = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_n \mathbf{x}_n$$

Therefore,

$$\mathbf{A}\mathbf{z}_0 = \alpha_1 \lambda_1 \mathbf{x}_1 + \alpha_2 \lambda_2 \mathbf{x}_2 + \cdots + \alpha_n \lambda_n \mathbf{x}_n \triangleq \mathbf{z}_1$$

$$\mathbf{A}\mathbf{z}_1 = \alpha_1 \lambda_1^2 \mathbf{x}_1 + \alpha_2 \lambda_2^2 \mathbf{x}_2 + \cdots + \alpha_n \lambda_n^2 \mathbf{x}_n \triangleq \mathbf{z}_2$$

$$\vdots$$

$$\mathbf{A}\mathbf{z}_k = \alpha_1 \lambda_1^{k+1} \mathbf{x}_1 + \alpha_2 \lambda_2^{k+1} \mathbf{x}_2 + \cdots + \alpha_n \lambda_n^{k+1} \mathbf{x}_n \triangleq \mathbf{z}_{k+1}$$

If λ_1 is the eigenvalue with the largest absolute value, then for k sufficiently large,

$$\mathbf{z}_{k+1} \cong \alpha_1 \lambda_1^{k+1} \mathbf{x}_1 = \beta \mathbf{x}_1 \quad \text{and} \quad \mathbf{A}\mathbf{z}_{k+1} \cong \lambda_1 \mathbf{z}_{k+1} = \lambda_1 \beta \mathbf{x}_1$$

Hence starting with an arbitrary vector $\mathbf{z}_0$ and repeatedly calculating $\mathbf{z}_{\text{new}} = \mathbf{A}\mathbf{z}_{\text{old}}$ until $\mathbf{z}_{\text{new}}$ is proportional to $\mathbf{z}_{\text{old}}$ leads to the maximum magnitude eigenvalue (the constant of proportionality) and the corresponding eigenvector. At each step of the iterative calculations, the vectors $\mathbf{z}$ can be normalized in any number of ways. In the subsequent discussion it is assumed that the final vector is normalized to a unit vector.

The next largest eigenvalue and its eigenvector can be found by constraining all vectors in the iteration process to be orthogonal to the first eigenvector. From Chapter 5, the projection operator $\mathbf{P}_1 = \mathbf{I} - \mathbf{x}_1 \mathbf{x}_1^T$ takes any arbitrary vector $\mathbf{z}$ into the subspace orthogonal to $\mathbf{x}_1$. Thus an arbitrary $\mathbf{z}_{\text{free}}$ gets mapped into $\mathbf{P}_1 \mathbf{z}_{\text{free}} = \mathbf{z}_{\text{constrained}}$ and $\mathbf{A}\mathbf{z}_{\text{constrained}}$ is equivalent to defining $\mathbf{A}_1 = \mathbf{A}\mathbf{P}_1$. Then the same iteration process is performed with $\mathbf{A}_1$ and freely selected $\mathbf{z}$ vectors, $\mathbf{A}_1 \mathbf{z}_k = \mathbf{z}_{k+1}$. The end results will be the next largest $|\lambda_2|$ and its eigenvector $\mathbf{x}_2$. For the eigenvalue with the third largest magnitude, iteration vectors $\mathbf{z}$ are restricted to be orthogonal to both $\mathbf{x}_1$ and $\mathbf{x}_2$. This can be done by defining a new projection matrix $\mathbf{P}_2 = \mathbf{I} - \mathbf{x}_1 \mathbf{x}_1^T - \mathbf{x}_2 \mathbf{x}_2^T$ and then using $\mathbf{A}_2 = \mathbf{A}\mathbf{P}_2$ in place of $\mathbf{A}$. The sequence of $\mathbf{P}_i$ matrices are often called *sweep matrices*. The similarity with the vector version of the Gram-Schmidt construction process of Problem 5.17 is noted. The process continues in an obvious way until all eigenvalues and eigenvectors are found. This is a very rapid and effective calculation method for the limited class of matrices to which it applies. Ordinarily the eigenvalues are not known at the outset, so it is not clear whether this method applies or not. Actually, the method sometimes gives the correct answers even when two eigenvalues have the same magnitude. For a trivial example, consider $\mathbf{A} = \text{Diag}(2, 2)$. If the starting vector is

$\mathbf{z}_0 = [1 \quad 0]^T$, then one iteration gives $\mathbf{z}_1 = \mathbf{z}_0$ and $\lambda_1 = 2$. Then $\mathbf{P}_1$ is found to be Diag(0, 1) and $\mathbf{A}_1 = \text{Diag}(0, 2)$. If the same $\mathbf{z}_0 = [1 \quad 0]^T$ is used to start the search for $\mathbf{x}_2$ and λ_2, one iteration gives the *wrong* final answer $\lambda_2 = 0$. If $\mathbf{z}_0 = [1 \quad 1]^T$ is used to start the second stage, two iterations give the *correct* final answer $\mathbf{x}_2 = [0 \quad 1]^T$ and $\lambda_2 = 2$. This type of dangerous behavior can be minimized but not avoided by starting with randomly selected $\mathbf{z}_0$ vectors. Another type of failure, which is less dangerous because it is recognized as a failure, is illustrated by

$$\mathbf{A} = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}$$

This matrix has $\lambda = \pm 2$. For *any* initial vector, the two components flip-flop back and forth each iteration, and convergence never occurs.

Another commonly used method of finding the eigenvalues by iteration is the so-called **QR** method. Only the rudiments of the method are given here. It, like many other methods, depends on transforming the original matrix to a form in which the eigenvalues are obvious. For example, if a similarity transformation could be found such that $\mathbf{T}^{-1}\mathbf{AT} = \mathbf{D}$ is diagonal, then those diagonal elements are the eigenvalues of **A. D** and **A** have the same eigenvalues, since

$$|\mathbf{D} - \mathbf{I}\lambda| = |\mathbf{T}^{-1}\mathbf{AT} - \mathbf{T}^{-1}\mathbf{T}\lambda| = |\mathbf{T}^{-1}||\mathbf{A} - \mathbf{I}\lambda||\mathbf{T}| = |\mathbf{A} - \mathbf{I}\lambda|$$

The eigenvectors of **D** are not the same as the eigenvectors of **A,** however. If the decomposition $\mathbf{A} = \mathbf{QR}$ is found and then a new matrix $\mathbf{A}_1 = \mathbf{RQ}$ is formed, $\mathbf{A}_1$ and the original **A** are related by a similarity transformation. Since **Q** is orthogonal, it can be inverted to give $\mathbf{R} = \mathbf{Q}^{-1}\mathbf{A}$. Using this in the reversed-order product gives $\mathbf{A}_1 = \mathbf{Q}^{-1}\mathbf{AQ}$. If $\mathbf{A}_1$ is now decomposed into $\mathbf{A}_1 = \mathbf{Q}_1\mathbf{R}_1$ and then $\mathbf{A}_2 = \mathbf{R}_1\mathbf{Q}_1$ is formed, it can be seen that $\mathbf{A}_2 = \mathbf{Q}_1^{-1}\mathbf{Q}^{-1}\mathbf{AQQ}_1 = (\mathbf{QQ}_1)^{-1}\mathbf{A}(\mathbf{QQ}_1)$. Thus $\mathbf{A}_2$ is also related to **A** by a similarity transformation, and they have the same eigenvalues. This remains true for any number of steps. It is an interesting fact that because of the upper triangular nature of $\mathbf{R}_i$ at each step, this procedure will (usually) converge to a matrix with either a 1×1 or a 2×2 block **E** in the lower-right corner, $\mathbf{A}_k = \begin{bmatrix} \mathbf{F} & \mathbf{G} \\ \mathbf{0} & \mathbf{E} \end{bmatrix}$. The eigenvalues of this block-triangular structure are the eigenvalues of **F** and of **E.** The eigenvalues of **E** are just **E** itself if it is 1×1. If **E** is 2×2, a simple quadratic equation can be solved to find its eigenvalues. Complex conjugate pairs of eigenvalues can be found using only real arithmetic by this method. In either case, the **E** portion can be stripped out, and the **QR**$\rightarrow$**RQ** procedure can be continued on just the **F** portion. As just described, the convergence would be very slow. Good **QR** eigenvalue procedures have various refinements, including initial conditioning on **A** to speed convergence [4, 5]. Another modification which speeds convergence is to do the $\mathbf{Q}_{k+1}\mathbf{R}_{k+1}$ decomposition on $\mathbf{R}_k\mathbf{Q}_k - \mathbf{I}\alpha$ rather than on $\mathbf{R}_k\mathbf{Q}_k$, for some judicious choice of α. A common choice for α is the current lower-right corner element in $\mathbf{R}_k\mathbf{Q}_k$. Once all the eigenvalues are found, one more **QR** decomposition of $\mathbf{A} - \mathbf{I}\lambda$ can be carried out to find each eigenvector. Since **Q** is nonsingular, $(\mathbf{A} - \mathbf{I}\lambda)\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{Rx} = \mathbf{0}$, and this is easily solved due to the triangular nature of **R.**

7.7 SPECTRAL DECOMPOSITION AND INVARIANCE PROPERTIES

Definition 7.1. Let $\mathscr{X}$ be a linear vector space defined over the complex number field. Let $\mathscr{A}: \mathscr{X} \rightarrow \mathscr{X}$ be a linear transformation and let $\mathscr{X}_1$ be a subspace of $\mathscr{X}$. Then $\mathscr{X}_1$ is said to be *invariant* under the transformation $\mathscr{A}$ if for every $\mathbf{x} \in \mathscr{X}_1$, $\mathscr{A}(\mathbf{x})$ also belongs to $\mathscr{X}_1$.

Definition 7.2. The set of all vectors $\mathbf{x}_i$ satisfying

$$\mathscr{A}(\mathbf{x}_i) = \lambda_i \mathbf{x}_i$$

for a particular λ_i is called the *eigenspace* of λ_i.

It consists of all the eigenvectors associated with that particular eigenvalue λ_i, plus the zero vector. The eigenspace of λ_i is a subspace of $\mathscr{X}$, and may alternatively be characterized as the null space of the transformation $(\mathscr{A} - \mathscr{I}\lambda_i)$, denoted as $\mathscr{N}_i$ for brevity.

Theorem 7.1. $\mathscr{N}_i$ is a q_i dimensional subspace of $\mathscr{X}$ which is invariant under $\mathscr{A}$, where q_i is the degeneracy, $q_i = n - \text{rank}(\mathscr{A} - \mathscr{I}\lambda_i)$.

Definition 7.3. If the linear transformation $\mathscr{A}$ has a complete set of n linearly independent eigenvectors (i.e., Case I or Case II_1), then $\mathscr{A}$ is said to be a *simple* linear transformation.

Theorem 7.2. If $\mathscr{A}$ is simple, then

$$\mathscr{X} = \mathscr{N}_1 \oplus \mathscr{N}_2 \oplus \cdots \oplus \mathscr{N}_p$$

where the direct sum is taken over the p distinct eigenvalues. (Note that $p < n$ for Case II_1.)

Theorem 7.3. If $\mathscr{A}$ is normal, $\mathscr{N}_i$ and $\mathscr{N}_j$ are orthogonal to each other, for all $i \neq j$. Note that normal transformations are a subset of simple transformations.

Let $\mathscr{A}: \mathscr{X} \rightarrow \mathscr{X}$ be a simple linear transformation, with the matrix representation $\mathbf{A}$. Then the eigenvectors of $\mathbf{A}$, $\{\mathbf{x}_i\}$, form a basis for $\mathscr{X}$. Let $\{\mathbf{r}_i\}$ be the reciprocal basis vectors. Then, for every $\mathbf{z} \in \mathscr{X}$,

$$\mathbf{z} = \sum_{i=1}^{n} \langle \mathbf{r}_i, \mathbf{z} \rangle \mathbf{x}_i \quad \text{and} \quad \mathbf{A}\mathbf{z} = \sum_{i=1}^{n} \langle \mathbf{r}_i, \mathbf{z} \rangle \mathbf{A}\mathbf{x}_i = \sum_{i=1}^{n} \lambda_i \langle \mathbf{r}_i, \mathbf{z} \rangle \mathbf{x}_i$$

This allows $\mathbf{A}$ to be written as

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \langle \mathbf{r}_i \tag{7.5}$$

This is called the *spectral representation* of $\mathbf{A}$. If $\mathscr{A}$ is normal, then its eigenvectors are mutually orthogonal (see Problem 7.27) so that the reciprocal basis vector $\mathbf{r}_i$ can be made equal to $\mathbf{x}_i$ by normalizing the eigenvectors. Then

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \rangle\langle \mathbf{x}_i \tag{7.6}$$

When a linear transformation is not simple, its eigenvectors do not form a basis for $\mathscr{X}$ (Cases II_2 and II_3). It is still possible to construct a basis by adding generalized eigenvectors, as discussed in Section 7.5. These vectors, along with the eigenvectors, belong to $\mathscr{N}_i^{k_i}$, the null space of $(\mathscr{A} - \mathscr{I}\lambda_i)^{k_i}$ (see Problem 7.35). The power k_i is called the *index* of the eigenvalue λ_i and is one for simple transformations. Using this generalization, it is again possible to write $\mathscr{X}$ as a direct sum of invariant subspaces of $\mathscr{A}$.

$$\mathscr{X} = \mathscr{N}_1^{k_1} \oplus \mathscr{N}_2^{k_2} \oplus \cdots \oplus \mathscr{N}_p^{k_p} \tag{7.7}$$

Alternatively, the space $\mathscr{X}$ can be decomposed into

$$\mathscr{X} = \mathscr{N}_i^{k_i} \oplus \mathscr{R}_i^{k_i \perp}$$

Since all m_i eigenvectors and generalized eigenvectors associated with λ_i belong to $\mathscr{N}_i^{k_i}$, $\dim(\mathscr{N}_i^{k}) = m_i$. Thus, $\text{rank}(\mathbf{A} - \mathbf{I}\lambda_i)^{k_i} = n - m_i$ and the index k_i is the smallest integer for which this is true.

Theorem 7.4. If $\mathscr{A} : \mathscr{X} \to \mathscr{X}$, with $\dim(\mathscr{X}) = n$, and if $\mathscr{X}$ can be expressed as the direct sum of p invariant subspaces, as in Eq. (7.7), then $\mathscr{A}$ can be represented by a block diagonal matrix, with p blocks, each of dimension k_i, provided a suitable basis is selected.

The block diagonal representation for $\mathscr{A}$ is the Jordan form, and "the suitable basis" consists of eigenvectors, and if necessary, generalized eigenvectors. This provides the simplest possible representation for a linear transformation, and will be most useful in analyzing systems in later chapters.

7.8 BILINEAR AND QUADRATIC FORMS

The expression $\langle \mathbf{y}, \mathbf{Ax} \rangle$ is called a *bilinear form,* since if $\mathbf{y}$ is held fixed, it is linear in $\mathbf{x}$: and if $\mathbf{x}$ is held fixed, it is linear in $\mathbf{y}$. When $\mathbf{x} = \mathbf{y}$, the result is the *quadratic form,* $Q(\mathbf{x}) = \langle \mathbf{x}, \mathbf{Ax} \rangle$. Every matrix $\mathbf{A}$ can be written as the sum of a Hermitian matrix and a skew-Hermitian matrix. It will be assumed that all quadratic forms are defined in terms of a Hermitian matrix $\mathbf{A}$. For real quadratic forms, there is no loss of generality since $\langle \mathbf{x}, \mathbf{Ax} \rangle = 0$ for all $\mathbf{x}$ if $\mathbf{A}$ is skew-symmetric.

Quadratic forms arise in connection with performance criteria in optimal control problems, in consideration of system stability, and in other applications. Here the several types of quadratic forms are defined and means for establishing the type of a given quadratic form are summarized.

Definitions.

1. Q (or the defining matrix $\mathbf{A}$) is said to be *positive definite* if and only if $\langle \mathbf{x}, \mathbf{Ax} \rangle > 0$ for all $\mathbf{x} \neq 0$.

2. Q is *positive semidefinite* if $\langle \mathbf{x}, \mathbf{Ax} \rangle \geq 0$ for all $\mathbf{x}$. That is, $Q = 0$ is possible for some $\mathbf{x} \neq \mathbf{0}$.
3. Q is *negative definite* if and only if $\langle \mathbf{x}, \mathbf{Ax} \rangle < 0$ for all $\mathbf{x} \neq \mathbf{0}$.
4. Q is *negative semidefinite* if $\langle \mathbf{x}, \mathbf{Ax} \rangle \leq 0$ for all $\mathbf{x}$.
5. Q is said to be *indefinite* if $\langle \mathbf{x}, \mathbf{Ax} \rangle > 0$ for some $\mathbf{x}$ and $\langle \mathbf{x}, \mathbf{Ax} \rangle < 0$ for other $\mathbf{x}$.

Tests for Definiteness. Let $\mathbf{A}$ be an $n \times n$ real symmetric matrix with eigenvalues λ_i. Define

$$\Delta_1 = a_{11}, \qquad \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \qquad \Delta_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \ldots, \Delta_n = |\mathbf{A}|$$

The Δ_i are called the *principal minors* of $\mathbf{A}$. Two possible methods of determining the definiteness of a Hermitian matrix $\mathbf{A}$ are given in Table 7.1.

TABLE 7.1

Class	Tests Using:	
	Eigenvalues of $\mathbf{A}$	Principal Minors of $\mathbf{A}$ (for real symmetric $\mathbf{A}$)
1. Positive definite	All $\lambda_i > 0$	$\Delta_1 > 0, \Delta_2 > 0, \ldots, \Delta_n > 0$
2. Positive semidefinite	All $\lambda_i \geq 0$	$\Delta_1 \geq 0, \Delta_2 \geq 0, \ldots, \Delta_n \geq 0$
3. Negative definite	All $\lambda_i < 0$	$\Delta_1 < 0, \Delta_2 > 0, \Delta_3 < 0, \ldots$ (note alternating signs)
4. Negative semidefinite	All $\lambda_i \leq 0$	$\Delta_1 \leq 0, \Delta_2 \geq 0, \Delta_3 \leq 0, \ldots$
5. Indefinite	Some $\lambda_i > 0$, some $\lambda_j < 0$	None of the above

7.9 MISCELLANEOUS USES OF EIGENVALUES AND EIGENVECTORS

Eigenvalues and eigenvectors are useful in many contexts. Four of the more important uses in modern control theory are mentioned.

1. *Existence of solutions for sets of linear equations:* In Chapter 6, the existence of nontrivial solutions for homogeneous equations and of a unique solution for nonhomogeneous equations was seen to depend upon whether or not the coefficient matrix $\mathbf{A}$ had a zero determinant. Stated differently, the existence of unique solutions depended upon whether or not the null space of a linear transformation contained nonzero vectors. Both of these conditions are related to the question of whether or not zero is an eigenvalue. Even when the transformation maps vectors from a space of one dimension to a space of another dimension (and thus cannot define an eigenvalue problem), conditions can be expressed in terms of the eigenvalues of transformations $\mathcal{A}\mathcal{A}^*$ and/or $\mathcal{A}^*\mathcal{A}$. This will be done in Chapter 11 when discussing controllability and observability.

2. *Stability of linear differential and difference equations:* It is not difficult to show that the characteristic equation of the companion matrix (Problem 7.36) is the

same as the denominator of the input-output transfer function for the nth order differential equation. Thus the system poles are the same as the eigenvalues of the matrix. The influence of pole locations on system stability has been discussed in Chapter 2. It will be seen in Chapter 10 that the eigenvalues of a system matrix, not necessarily the companion matrix, determine the stability of linear systems described either by differential or difference equations.

3. *Eigenvectors are convenient basis vectors:* When eigenvectors and, if needed, generalized eigenvectors are used as basis vectors, a linear transformation assumes its simplest possible form. In this simple form independent modes of system behavior become apparent. As a simple example, consider the equation

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

where $\mathbf{A}$ is an $n \times n$ matrix, and $\mathbf{y}$ might be a set of time derivatives or any other $n \times 1$ vector. If a change of basis is used, $\mathbf{x} = \mathbf{M}\mathbf{z}$, $\mathbf{y} = \mathbf{M}\mathbf{w}$, where $\mathbf{M}$ is the modal matrix, then

$$\mathbf{M}\mathbf{w} = \mathbf{A}\mathbf{M}\mathbf{z} \quad \text{or} \quad \mathbf{w} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}\mathbf{z} = \mathbf{J}\mathbf{z}$$

$\mathbf{J}$ is the Jordan form in general, but will simply be the diagonal matrix $\mathbf{\Lambda}$ in many cases. The simultaneous equations are now as nearly uncoupled as is possible.

When considering the real quadratic form $Q = \langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle$, with $\mathbf{A}$ symmetric, all second-order products of the components of $\mathbf{x}$ are usually present. If a change of basis $\mathbf{x} = \mathbf{M}\mathbf{z}$ is used, then $Q = \langle \mathbf{M}\mathbf{z}, \mathbf{A}\mathbf{M}\mathbf{z} \rangle = \langle \mathbf{z}, \mathbf{M}^T\mathbf{A}\mathbf{M}\mathbf{z} \rangle$. Since $\mathbf{A}$ is real and symmetric, it can always be diagonalized by an orthogonal transformation, so $Q = \langle \mathbf{z}, \mathbf{\Lambda}\mathbf{z} \rangle$ reduces to the sum of the squares of the z_i components, weighted by the eigenvalues λ_i. This makes the relationships between eigenvalues and the various kinds of definiteness rather transparent.

4. *Sufficient conditions for relative maximum or minimum:* When considering a smooth function of a single variable on an open interval, the necessary condition for a relative maximum or minimum is that the first derivative vanish. To determine whether a maximum, minimum, or saddle point exists, the sign of the second derivative must be determined. In multidimensional cases it is again necessary that the first derivatives (all of them) vanish at a point of relative maximum or minimum. The test of the sign of the second derivative in the scalar case is replaced by a test for positive or negative definiteness of a matrix of second derivative terms. Eigenvalue-eigenvector theory plays an important part in the investigation of these and many other questions.

Additional material related to this chapter may be found in References 3 through 7.

REFERENCES

1. Brogan, W. L.: "Optimal Control Theory Applied to Systems Described by Partial Differential Equations," *Advances in Control Systems,* Vol. 6, C. T. Leondes, Ed., Academic Press, New York, 1968.
2. Melsa, J. L. and S. K. Jones: *Computer Programs for Computational Assistance in the Study of Linear Control Theory,* 2nd ed., McGraw-Hill, New York, 1973.

3. De Russo, P. M., R. J. Roy, and C. M. Close: *State Variables for Engineers,* John Wiley, New York, 1965.
4. Strang, G.: *Linear Algebra and Its Applications,* Academic Press, New York, 1980.
5. Golub, G. H. and C. F. Van Loan: *Matrix Computations,* Johns Hopkins University Press, Baltimore, 1983.
6. Forsythe, G. F., M. A. Malcolm, and C. Moler: *Computer Methods for Mathematical Computations,* Prentice Hall, Englewood Cliffs, N.J., 1977.
7. Courant, R. and D. Hilbert: *Methods of Mathematical Physics,* Vol. 1, Interscience: John Wiley, New York, 1953.

ILLUSTRATIVE PROBLEMS

Determination of Eigenvalues, Eigenvectors, and the Jordan Form

7.1 Find the eigenvalues and eigenvectors and then use a similarity transformation to diagonalize $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -3 & -4 \end{bmatrix}$.

The characteristic equation is $|\mathbf{A} - \mathbf{I}\lambda| = \lambda^2 + 4\lambda + 3 = 0$. Therefore $\lambda_1 = -1, \lambda_2 = -3$. For simple roots (Case I) compute

$$\text{Adj}[\mathbf{A} - \mathbf{I}\lambda] = \begin{bmatrix} -4 - \lambda & -1 \\ 3 & -\lambda \end{bmatrix}$$

Substituting in $\lambda = -1$ gives $\mathbf{x}_1 = [-1 \quad 1]^T$ or any vector proportional to this. Using $\lambda = -3$ gives $\mathbf{x}_2 = [-1 \quad 3]^T$:

$$\Lambda = \begin{bmatrix} -1 & -1 \\ 1 & 3 \end{bmatrix}^{-1} \mathbf{A} \begin{bmatrix} -1 & -1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix}$$

7.2 Consider the eigenvalue-eigenvector problem for $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ -2 & -3 \end{bmatrix}$.

The characteristic equation is $|\mathbf{A} - \mathbf{I}\lambda| = \lambda^2 + 2\lambda + 1 = (\lambda + 1)^2$. Therefore, $\lambda = -1$ with algebraic multiplicity 2. $\text{Rank}[\mathbf{A} - \mathbf{I}\lambda]|_{\lambda = -1} = 1$ and degeneracy $q = 2 - 1 = 1$. This is an example of simple degeneracy, so there is one eigenvector $\mathbf{x}_1$ and one generalized eigenvector $\mathbf{x}_2$:

$$\text{Adj}[\mathbf{A} - \mathbf{I}\lambda]\Big|_{\lambda = -1} = \begin{bmatrix} -2 & -2 \\ 2 & 2 \end{bmatrix}$$

Select $\mathbf{x}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Generalized eigenvector, method 1:

Set $\mathbf{A}\mathbf{x}_2 = -\mathbf{x}_2 + \mathbf{x}_1$ and let $\mathbf{x}_2 = [a \quad b]^T$. Then $a + 2b = -a - 1$ or $2a + 2b = -1$. We could set $a = 1$, then $b = -\frac{3}{2}$ and $\mathbf{x}_2 = [1 \quad -\frac{3}{2}]^T$, or if $a = -1, b = \frac{1}{2}$. If $\mathbf{x}_1$ was selected as $\mathbf{x}_1 = [2 \quad -2]^T$, then $\mathbf{x}_2 = [1 \quad 0]^T$. Either of these choices is valid and each gives $\mathbf{J} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix}$.

Generalized eigenvector, method 2:

If $\mathbf{x}_1$ is selected as column 1 of $\text{Adj}[\mathbf{A} + \mathbf{I}]$, i.e., $\mathbf{x}_1 = [-2 \quad 2]^T$, then $\mathbf{x}_2$ is column 1 of the differentiated adjoint matrix

$$\mathbf{x}_2 = \frac{d}{d\lambda}\begin{bmatrix} -3 - \lambda \\ 2 \end{bmatrix}\Big|_{\lambda = -1} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

If $\mathbf{x}_1 = [-2 \quad 2]^T$ (column 2 of Adj$[\mathbf{A} - \mathbf{I}]$), then $\mathbf{x}_2$ is column 2 of the differentiated adjoint matrix

$$\mathbf{x}_2 = \frac{d}{d\lambda}\begin{bmatrix} -2 \\ 1-\lambda \end{bmatrix}_{\lambda=-1} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

Eigenvector and generalized eigenvector, method 3:

Since $n = m = 2$, we seek the smallest k such that $\text{rank}[\mathbf{A} - \mathbf{I}\lambda]^k|_{\lambda_1=-1} = 0$. The index k is 2. Then $(\mathbf{A} - \mathbf{I}\lambda_1)^2\mathbf{x} = \mathbf{0}$ has two independent solutions, one of which is $\mathbf{x}_2 = [1 \quad 0]^T$. Since $[\mathbf{A} - \mathbf{I}\lambda_1]\mathbf{x}_2 = [2 \quad -2]^T \neq \mathbf{0}$, $\mathbf{x}_2$ is a generalized eigenvector and $\mathbf{x}_1$ is the eigenvector. Alternatively, one could select $\mathbf{x}_2 = [0 \quad 1]^T$ and this also leads to $\mathbf{x}_1 = [2 \quad -2]^T$. All of these methods give the same Jordan form.

7.3 Find the eigenvalues-eigenvectors and the Jordan form for $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

The characteristic equation is $|\mathbf{A} - \mathbf{I}\lambda| = \lambda^2 - 2\lambda = \lambda(\lambda - 2)$. Therefore, $\lambda_1 = 0, \lambda_2 = 2$. Since $\lambda_1 \neq \lambda_2$, Case I applies:

$$\text{Adj}[\mathbf{A} - \mathbf{I}\lambda] = \begin{bmatrix} 1-\lambda & -1 \\ -1 & 1-\lambda \end{bmatrix}$$

Using the first column with $\lambda = 0$ gives $\mathbf{x}_1 = [1 \quad -1]^T$. Using the first column with $\lambda = 2$ gives $\mathbf{x}_2 = [-1 \quad -1]^T$ or $[1 \quad 1]^T$:

$$\mathbf{M} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \qquad \mathbf{M}^{-1} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{J} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$$

7.4 Find the eigenvalues, eigenvectors, and Jordan form for

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & -3 \\ 0 & 1 & -3 & 0 \\ -0.5 & -3 & 1 & 0.5 \\ -3 & 0 & 0 & 1 \end{bmatrix}$$

The characteristic equation is $\lambda^4 - 4\lambda^3 - 12\lambda^2 + 32\lambda + 64 = 0$. The roots are found to be $\lambda_i = -2, -2, 4, 4$. With $\lambda = -2$

$$\mathbf{A} - \lambda\mathbf{I} = \mathbf{A} + 2\mathbf{I} = \begin{bmatrix} 3 & 0 & 0 & -3 \\ 0 & 3 & -3 & 0 \\ -0.5 & -3 & 3 & 0.5 \\ -3 & 0 & 0 & 3 \end{bmatrix}$$

This matrix has rank 2, so $q = n - r = 2$. This shows that there are two eigenvectors associated with $\lambda = -2$, and since the multiplicity of that root is also 2, no generalized eigenvectors are needed for this eigenvalue. Two linearly independent solutions of $[\mathbf{A} + 2\mathbf{I}]\mathbf{x}_i = \mathbf{0}$ are $\mathbf{x}_1 = [0 \quad 1 \quad 1 \quad 0]^T$ and $\mathbf{x}_2 = [1 \quad 0 \quad 0 \quad 1]^T$. With $\lambda = 4$, $\mathbf{A} - \lambda\mathbf{I} = \begin{bmatrix} -3 & 0 & 0 & -3 \\ 0 & -3 & -3 & 0 \\ -0.5 & -3 & -3 & 0.5 \\ -3 & 0 & 0 & -3 \end{bmatrix}$. The rank is 3 and $q = 1$. There is only one eigenvector, and since the algebraic multiplicity $m = 2$, a generalized eigenvector is needed. Solving $[\mathbf{A} - 4\mathbf{I}]\mathbf{x}_i = 0$ gives only one independent solution, $\mathbf{x}_3 = [0 \quad 1 \quad -1 \quad 0]^T$. Therefore, a generalized eigenvector is needed and it can be found by solving

$$[\mathbf{A} - 4\mathbf{I}]\mathbf{x}_4 = \mathbf{x}_3$$

The result is $\mathbf{x}_4 = [2 \quad -\frac{1}{3} \quad 0 \quad -2]^T$. The modal matrix is thus

$$\mathbf{M} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3 \quad \mathbf{x}_4]$$

and the Jordan form is

$$\mathbf{J} = \begin{bmatrix} -2 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

7.5 **A** is a 5×5 matrix for which the following information has been found:

$$\lambda_1 = \lambda_2 = 2, \qquad \text{Rank}[\mathbf{A} - 2\mathbf{I}] = 4$$

$$\lambda_3 = \lambda_4 = \lambda_5 = -2, \qquad \text{Rank}[\mathbf{A} + 2\mathbf{I}] = 3$$

Determine the Jordan form for **A.**

For $\lambda = 2$, the degeneracy is $q_1 = 1$, so there is a single Jordan block $\mathbf{J}_1 = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$. For $\lambda = -2$, the degeneracy is $q_3 = 2$. Thus there are two Jordan blocks $\mathbf{J}_2 = \begin{bmatrix} -2 & 1 \\ 0 & -2 \end{bmatrix}$ and $\mathbf{J}_3 = [-2]$. The arrangement of these blocks within the Jordan form depends upon the ordering of the eigenvectors and generalized eigenvectors within **M.** Assuming that $\mathbf{x}_1$ and $\mathbf{x}_2$ are an eigenvector and generalized eigenvector for $\lambda = 2$, $\mathbf{x}_3$ and $\mathbf{x}_4$ are an eigenvector and generalized eigenvector for $\lambda = -2$, and $\mathbf{x}_5$ is the second eigenvector associated with $\lambda = -2$, then

$$\text{diag}\,[\mathbf{J}_1, \mathbf{J}_2, \mathbf{J}_3] = \left[\begin{array}{cc:cc:c} 2 & 1 & & & \\ 0 & 2 & & \mathbf{0} & \\ \hdashline & & -2 & 1 & \\ & \mathbf{0} & 0 & -2 & \\ \hdashline & & & & -2 \end{array}\right] = \mathbf{J}$$

7.6 Let $\mathbf{A} = \begin{bmatrix} 3 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$.

(a) What are the eigenvalues?
(b) How many linearly independent eigenvectors does **A** have?
(c) How many generalized eigenvectors?

a. **A** is upper triangular and so is $\mathbf{A} - \mathbf{I}\lambda$. Thus the eigenvalues are the diagonal elements of **A,** $\lambda = 3, 3, 3, 4, 4, 4, 4$. b. The matrix **A** is already in Jordan form with four Jordan blocks. There are four eigenvectors. c. There are three generalized eigenvectors. The number of "ones" above the main diagonal is always equal to the number of generalized eigenvectors.

7.7 Find the eigenvalues, eigenvectors, and, if needed, the generalized eigenvectors for $\mathbf{A} = \begin{bmatrix} 4 & 2 & 1 \\ 0 & 6 & 1 \\ 0 & -4 & 2 \end{bmatrix}$. Also find the Jordan form.

The characteristic equation is $|\mathbf{A} - \mathbf{I}\lambda| = (4 - \lambda)^3$. Then $\lambda_1 = \lambda_2 = \lambda_3 = 4$ with algebraic multiplicity 3.

$[\mathbf{A} - 4\mathbf{I}] = \begin{bmatrix} 0 & 2 & 1 \\ 0 & 2 & 1 \\ 0 & -4 & -2 \end{bmatrix}$ has rank $r = 1$. The degeneracy is $q = 2$. This is an example of the general Case II_3 with two eigenvectors and one generalized eigenvector.

Method 1:

$$\text{Adj}[\mathbf{A}-\mathbf{I}\lambda]=\begin{bmatrix}(6-\lambda)(2-\lambda)+4 & 2\lambda-8 & \lambda-4\\ 0 & (4-\lambda)(2-\lambda) & \lambda-4\\ 0 & 16-4\lambda & (4-\lambda)(6-\lambda)\end{bmatrix}$$

With $\lambda = 4$, this reduces to the null matrix.

$$\frac{d}{d\lambda}\{\text{Adj}[\mathbf{A}-\mathbf{I}\lambda]\}\bigg|_{\lambda=4}=\begin{bmatrix}2\lambda-8 & 2 & 1\\ 0 & 2\lambda-6 & 1\\ 0 & -4 & 2\lambda-10\end{bmatrix}\bigg|_{\lambda=4}=\begin{bmatrix}0 & 2 & 1\\ 0 & 2 & 1\\ 0 & -4 & -2\end{bmatrix}$$

One eigenvector can be selected as $\mathbf{x}_2 = [1 \quad 1 \quad -2]^T$.

$$\frac{1}{2}\frac{d^2}{d\lambda^2}\{\text{Adj}[\mathbf{A}-\mathbf{I}\lambda]\}\bigg|_{\lambda=4}=\frac{1}{2}\begin{bmatrix}2 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 2\end{bmatrix}$$

The first column can be selected as another eigenvector, call it $\mathbf{x}_1 = [1 \quad 0 \quad 0]^T$. A generalized eigenvector is given by the third column $\mathbf{x}_3 = [0 \quad 0 \quad 1]^T$. Note that $\mathbf{x}_1$ and $\mathbf{x}_2$ are selected as columns that are nonzero for the first time as the adjoint matrix is repeatedly differentiated. Note also that $\mathbf{x}_3$ is selected from the same column that gave $\mathbf{x}_2$, but with one more differentiation.

Method 2:

We require two independent solutions of $[\mathbf{A}-4\mathbf{I}]\mathbf{x}=\mathbf{0}$. Let $\mathbf{x}=[a \quad b \quad c]^T$. Then $2b + c = 0$ is the only restriction placed on a, b and c. Since a is arbitrary, set $a = 1$ and $b = c = 0$, or $\mathbf{x}_1 = [1 \quad \mathbf{0} \quad 0]^T$. Another solution is $a = 1$, $b = 1$, $c = -2$, or $\mathbf{x}_2 = [1 \quad 1 \quad -2]^T$.

A generalized eigenvector is needed, and it must satisfy $(\mathbf{A}-4\mathbf{I})\mathbf{x}_3 = \mathbf{x}_2$ or $(\mathbf{A}-4\mathbf{I})^2\mathbf{x}_3 = (\mathbf{A}-4\mathbf{I})\mathbf{x}_2 = \mathbf{0}$. This reduces to $[\mathbf{0}]\mathbf{x}_3 = \mathbf{0}$, so $\mathbf{x}_3$ is arbitrary, except it must be nonzero and linearly independent of $\mathbf{x}_1$ and $\mathbf{x}_2$. $\mathbf{x}_3 = [0 \quad 0 \quad 1]^T$ is one such vector.

$$\text{Setting } \mathbf{M}=\begin{bmatrix}1 & 1 & 0\\ 0 & 1 & 0\\ 0 & -2 & 1\end{bmatrix} \text{ gives } \mathbf{M}^{-1}=\begin{bmatrix}1 & -1 & 0\\ 0 & 1 & 0\\ 0 & 2 & 1\end{bmatrix}, \text{ so that } \mathbf{J}=\mathbf{M}^{-1}\mathbf{A}\mathbf{M}=\left[\begin{array}{c|cc}4 & 0 & 0\\ \hline 0 & 4 & 1\\ 0 & 0 & 4\end{array}\right].$$

Method 3:

Alternatively, the index of $\lambda = 4$ is $k = 2$, since $\text{rank}[\mathbf{A}-4\mathbf{I}]^2 = n - m = \mathbf{0}$. A generalized eigenvector satisfying $[\mathbf{A}-4\mathbf{I}]^2\mathbf{x}_3 = \mathbf{0}$, $[\mathbf{A}-4\mathbf{I}]\mathbf{x}_3 \neq \mathbf{0}$ is $\mathbf{x}_3 = [0 \quad 0 \quad 1]^T$. Then $\mathbf{x}_2 = [\mathbf{A}-4\mathbf{I}]\mathbf{x}_3 = [1 \quad 1 \quad -2]^T$. This is one eigenvector, with the other one obviously being $\mathbf{x}_1 = [1 \quad 0 \quad 0]^T$.

7.8 Rework Example 7.6 using the top-down method of finding the eigenvectors or generalized eigenvectors.

Since $\lambda = 1$ has algebraic multiplicity $m = 3$ and since the matrix $\mathbf{A}$ has $n = 3$, the index k must give $\text{rank}\{[\mathbf{A}-\mathbf{I}]^k\} = 0$. It is easily verified that $k = 3$ and that $[\mathbf{A}-\mathbf{I}]^3 = [0]$. Therefore, any vector will satisfy the kth-order generalized eigenvector equation. Select $\mathbf{x}_a$, $\mathbf{x}_b$, and $\mathbf{x}_c$ as columns of the unit matrix. Define $\mathbf{C} = [\mathbf{A}-\mathbf{I}]$. The testing procedure shows that $\mathbf{C}\mathbf{x}_a = \mathbf{0}$, indicating that $\mathbf{x}_a$ is not the sought-for generalized eigenvector. It is an eigenvector, but we do not select it at this point because of scaling considerations. Then $\mathbf{C}\mathbf{x}_b = [2 \quad 0 \quad 0]^T \neq \mathbf{0}$ is found, indicating that $\mathbf{x}_b$ is a generalized eigenvector. It is not selected yet either. Continuing with the testing shows $\mathbf{C}\mathbf{x}_c = [3 \quad 4 \quad 0]^T = \mathbf{x}_d \neq \mathbf{0}$, so $\mathbf{x}_c$ is also a generalized eigenvector. Since $\mathbf{C}\mathbf{x}_b$ gives an eigenvector (and not a generalized eigenvector), it cannot be used to start our chain of three vectors. Testing at the next level shows that $\mathbf{C}\mathbf{x}_d = [8 \quad 0 \quad 0]^T$, another copy of the eigenvector. The final selection can now be made as $\mathbf{x}_3 = \mathbf{x}_c$, $\mathbf{x}_2 = \mathbf{x}_d$, and $\mathbf{x}_1 = 8\mathbf{x}_a$. This is the same set found in Example 7.6.

7.9 Find the eigenvalues, eigenvectors, and if needed, generalized eigenvectors, and the Jordan form for

$$\mathbf{A} = \begin{bmatrix} -4.5 & -1 & 0.5 & 0.5 \\ 0.333333 & -5.33333 & 0.666667 & 0 \\ -0.083333 & 0.33333 & -4.916667 & -0.25 \\ 0.25 & 0 & 0.75 & -5.25 \end{bmatrix}$$

Computer solution gives

$$\lambda_i = \{-4.99959, -5.00063, -4.99958, -5.00035\}$$

It seems likely but not certain that this is a case of repeated roots, but with rounding and truncation errors. The matrix $\mathbf{C} = 5\mathbf{I} - \mathbf{A}$ was formed and subjected to both SVD and **QR** decomposition. The rank was determined to be two, since the singular values Σ_{ii} were 1.5759, 0.8989, 10^{-5}, and 10^{-5}. This indicates that we have at least two repeated roots. Next $\mathbf{C}^2$ was computed and found to be [0] to within 10^{-6}. This indicates that the second-order generalized eigenvector problem has four independent solutions for $\lambda \approx -5$. The conclusion is that $m = 4$, $n = 4$, $q = 2$, and $k = 2$. There are two eigenvectors and two generalized eigenvectors and the Jordan form will have two 2×2 Jordan blocks $\mathbf{J} = \text{Diag}[\mathbf{J}_1, \mathbf{J}_2]$ with $\mathbf{J}_1 = \mathbf{J}_2 = \begin{bmatrix} -5 & 1 \\ 0 & -5 \end{bmatrix}$.

There are many ways to find the eigenvectors. The last two columns of the SVD **V** matrix could be used. Here we note that $\mathbf{C}^2 = [\mathbf{0}]$, so any column of **C** could be selected as an eigenvector. Rank(**C**) = 2, so only two independent columns exist. We select $\mathbf{x}_a = [0.5 \quad 0.33333 \quad -0.08333 \quad 0.25]^T$ and $\mathbf{x}_b = [0.5 \quad 0.66667 \quad 0.08333 \quad 0.75]^T$. Two other independent vectors $\mathbf{x}_c = [1 \quad 0 \quad 0 \quad 0]^T$ and $\mathbf{x}_d = [0 \quad 0 \quad 1 \quad 0]^T$ are selected. They all satisfy the 2nd-order generalized eigenvector problem. The top-down testing procedure shows that $\mathbf{C}\mathbf{x}_c = \mathbf{x}_a$ and that $\mathbf{C}\mathbf{x}_d = \mathbf{x}_b$. The columns of the modal matrix are thus selected as $\mathbf{x}_1 = \mathbf{x}_a$, $\mathbf{x}_2 = \mathbf{x}_c$, $\mathbf{x}_3 = \mathbf{x}_b$, and $\mathbf{x}_4 = \mathbf{x}_d$.

7.10 Noting that $\mathbf{AM} - \mathbf{MJ} = [0]$ is in the class of problems considered in Section 6.10, the solution for the vectorized version of **M** can be found by solving the $n^2 \times n^2$ problem $[\mathbf{I}_n \otimes \mathbf{A} - \mathbf{J}^T \otimes \mathbf{I}_n](\mathbf{M}) = (0)$.

(a) Apply this approach to Problems 7.1, 7.2, and 7.13 assuming **J** is known.

(b) Use the same approach on Problem 7.2, but this time try using $\mathbf{J} = \text{Diag}[-1, -1]$.

(a) Computer solution of the four simultaneous equations obtained for Problem 7.1 gives two independent nontrivial solutions, $(\mathbf{M})_1 = [1 \quad -1 \quad 0 \quad 0]^T$ and $(\mathbf{M})_2 = [0 \quad 0 \quad \frac{1}{3} \quad -1]^T$. Neither of these provides a valid nonsingular matrix **M,** but any linear combination of $(\mathbf{M})_1$ and $(\mathbf{M})_2$ is also a solution. Using $(\mathbf{M}) = -(\mathbf{M})_1 - 3(\mathbf{M})_2$ gives the previously obtained modal matrix. For Problem 7.2 there are also two solutions to the 4×4 problem, $(\mathbf{M})_1 = [1 \quad -1 \quad 0.5 \quad 0]^T$ and $(\mathbf{M})_2 = [0 \quad 0 \quad 1 \quad -1]^T$. The first solution gives an acceptable modal matrix $\mathbf{M} = \begin{bmatrix} 1 & 0.5 \\ -1 & 0 \end{bmatrix}$. For Problem 7.3 a similar result is obtained. Two vectorized solutions $(\mathbf{M})_1 = [1 \quad -1 \quad 0 \quad 0]^T$ and $(\mathbf{M})_2 = [0 \quad 0 \quad -1 \quad -1]^T$ are obtained. The former answer is obtained from $(\mathbf{M}) = (\mathbf{M})_1 - (\mathbf{M})_2$.

(b) When the wrong **J** matrix is used, two solutions are obtained; $(\mathbf{M})_1$ is the same as when the correct **J** was used in (a) and $(\mathbf{M})_2$ is $[0 \quad 0 \quad 1 \quad -1]^T$. No linear combination of these will give a nonsingular matrix **M** because the two nonzero columns are linearly dependent. This problem suggests another approach to determining eigenvectors and generalized eigenvectors, based on assuming **J** and testing the resulting **M.** Although it works, the dimension of the vectorized problems quickly get out of hand. Using the method on Problem 7.4 requires solution of a 16×16 set of equations. When this was done, six linearly independent $(\mathbf{M})_i$ vectors were found, and the previous answer was a linear combination of these six.

7.11 If $\lambda_i = \sigma + j\omega$ is a complex eigenvalue for **A,** with the associated complex eigenvector $\mathbf{x}_i = \mathbf{x}\mathbf{r}_i + \mathbf{x}\mathbf{c}_i j$, then the complex conjugate of λ_i is also an eigenvalue associated with an eigenvector $\mathbf{x}_{i+1} = \mathbf{x}\mathbf{r}_i - \mathbf{x}\mathbf{c}_i j$. The methods already presented for solving the eigenvalue-eigenvector problem apply on the complex number field. Most examples have been restricted to real eigenvalues to maintain simplicity. In the complex case, purely real arithmetic can again be used

on a double-sized problem. By combining $\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i$ and $\mathbf{A}\bar{\mathbf{x}}_i = \bar{\lambda}_i \bar{\mathbf{x}}_i$ and equating real parts and imaginary parts, the following a set of equations are obtained:

$$\mathbf{A}[\mathbf{xr}_i \quad \mathbf{xc}_i] = [\mathbf{xr}_i \quad \mathbf{xc}_i]\mathbf{E}$$

where

$$\mathbf{E} = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}$$

This is the type of equation dealt with in Sec. 6.10; it can be written in vectorized form as a set of six simultaneous equations:

$$[(\mathbf{I}_2 \otimes \mathbf{A}) - (\mathbf{E}^T \otimes \mathbf{I}_2)]\begin{bmatrix} \mathbf{xr}_i \\ \mathbf{xc}_i \end{bmatrix} = (0)$$

Find the eigenvalues, eigenvectors, and Jordan form for

$$\mathbf{A} = \begin{bmatrix} -4 & -2 & 1 \\ -1 & -2 & 1 \\ -1 & -4 & -6 \end{bmatrix}$$

The iterative **QR** method and direct solution for the characteristic equation roots both give eigenvalues as $\lambda = \{-1.56516, -5.21742 \pm 1.85843j\}$. The eigenvector associated with the real root is found by solving $(\mathbf{A} - \mathbf{I}\lambda_1)\mathbf{x}_1 = \mathbf{0}$ and is $\mathbf{x}_1 = [-1 \quad 0.916739 \quad -0.60136]^T$. With $\mathbf{E} = \begin{bmatrix} -5.21742 & -1.85643 \\ 1.85643 & -5.21742 \end{bmatrix}$ the 6×6 coefficient matrix $(\mathbf{I}_2 \otimes \mathbf{A}) - (\mathbf{E}^T \otimes \mathbf{I}_3)$ is

Row 1	1.217420E + 00 0.000000E + 00	−2.000000E + 00 0.000000E + 00	1.000000E + 00	−1.856430E + 00
Row 2	−1.000000E + 00 −1.856430E + 00	3.217420E + 00 0.000000E + 00	1.000000E + 00	0.000000E + 00
Row 3	−1.000000E + 00 0.000000E + 00	−4.000000E + 00 −1.856430E + 00	−7.825799E − 01	0.000000E + 00
Row 4	1.856430E + 00 −2.000000E + 00	0.000000E + 00 1.000000E + 00	0.000000E + 00	1.217420E + 00
Row 5	0.000000E + 00 3.217420E + 00	1.856430E + 00 1.000000E + 00	0.000000E + 00	−1.000000E + 00
Row 6	0.000000E + 00 −4.000000E + 00	0.000000E + 00 −7.825799E − 01	1.856430E + 00	−1.000000E + 00

Its rank is 4, yielding two independent solutions for the stacked eigenvector,

Row 1	1.041709E − 01	6.530732E − 01
Row 2	1.696022E − 01	3.008392E − 01
Row 3	−1.000000E + 00	0.000000E + 00
Row 4	−6.530732E − 01	1.041709E − 01
Row 5	−3.008392E − 01	1.696022E − 01
Row 6	0.000000E + 00	−1.000000E + 00

From this, $\mathbf{x}_2 = [0.10417 - 0.65307j \quad 0.169602 - 0.30084j \quad -1]^T$, and $\mathbf{x}_3 = \bar{\mathbf{x}}_2$. The complex modal matrix is formed from these columns, and it yields

$$\mathbf{J} = \mathbf{M}^{-1}\mathbf{AM} = \begin{bmatrix} -1.56516 & 0 & 0 \\ 0 & -5.21742 - 1.85643j & 0 \\ 0 & 0 & -5.21742 + 1.85643j \end{bmatrix}$$

Note that if columns of $\mathbf{T} = [\mathbf{x}_1 \quad \mathbf{xr}_2 \quad \mathbf{xc}_2]$ are used as basis vectors instead of **M**, a block diagonal real matrix is obtained in place of **J**, namely, $\mathbf{T}^{-1}\mathbf{AT} = \text{Diag}[\lambda_1, \mathbf{E}]$.

Similar Matrices

7.12 Prove that two similar matrices have the same eigenvalues.

A and **B** are similar matrices if they are related by $\mathbf{A} = \mathbf{Q}^{-1}\mathbf{BQ}$ for some nonsingular matrix **Q**. The characteristic equation for **A** is

$$|\mathbf{A} - \mathbf{I}\lambda| = |\mathbf{Q}^{-1}\mathbf{BQ} - \mathbf{Q}^{-1}\mathbf{Q}\lambda| = 0$$

or

$$|\mathbf{Q}^{-1}[\mathbf{B} - \mathbf{I}\lambda]\mathbf{Q}| = |\mathbf{Q}^{-1}|\cdot|\mathbf{Q}||\mathbf{B} - \mathbf{I}\lambda| = |\mathbf{B} - \mathbf{I}\lambda| = 0$$

The characteristic equations for **A** and **B** are the same so they have the same eigenvalues.

7.13 Are the following matrices similar?

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

All four matrices have the same characteristic equation $(2 - \lambda)^4 = 0$, so $\lambda = 2$ is the eigenvalue with algebraic multiplicity $m = 4$. The given matrices are expressed in Jordan form. Since similar matrices must have the same Jordan form, the answer is no. Note that the index of the eigenvalue is 1, 2, 3, and 4, respectively, for these matrices.

Miscellaneous Properties

7.14 Prove that $|\mathbf{A}| = \lambda_1\lambda_2\cdots\lambda_n$.

Any $n \times n$ matrix **A** can be reduced to the Jordan form $\mathbf{J} = \mathbf{M}^{-1}\mathbf{AM}$; so $\mathbf{A} = \mathbf{MJM}^{-1}$. From this $|\mathbf{A}| = |\mathbf{MJM}^{-1}| = |\mathbf{M}||\mathbf{J}||\mathbf{M}^{-1}| = |\mathbf{J}|$. Since **J** is upper triangular, with the eigenvalues on the main diagonal, $|\mathbf{A}| = |\mathbf{J}| = \lambda_1\lambda_2\cdots\lambda_n$. Thus $|\mathbf{A}| = 0 \Leftrightarrow$ at least one $\lambda_i = 0$.

7.15 Prove $\text{Tr}(\mathbf{A}) = \lambda_i + \lambda_2 + \cdots + \lambda_n$.

Since $\mathbf{A} = \mathbf{MJM}^{-1}$, $\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{MJM}^{-1})$.

But $\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA})$, so $\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{JM}^{-1}\mathbf{M})$ or $\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{J}) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$.

7.16 Prove that if **A** is nonsingular with eigenvalues λ_i, then $1/\lambda_i$ are the eigenvalues of $\mathbf{A}^{-1}$.

The n roots λ_i are defined by $|\mathbf{A} - \mathbf{I}\lambda| = 0$. But $|\mathbf{A} - \mathbf{I}\lambda| = |\mathbf{A}[\mathbf{I} - \mathbf{A}^{-1}\lambda]| = |\mathbf{A}|\cdot|\mathbf{I} - \mathbf{A}^{-1}\lambda| = 0$. Since **A** is nonsingular, $|\mathbf{A}|$ can be divided out leaving $|\mathbf{I} - \mathbf{A}^{-1}\lambda| = |\mathbf{I}(1/\lambda) - \mathbf{A}^{-1}|\lambda^n = 0$. Since $|\mathbf{A}|$, and hence λ, are not zero, the characteristic equation for **A** leads to $|\mathbf{A}^{-1} - \mathbf{I}(1/\lambda)| = 0$. Thus if λ_i is an eigenvalue of **A**, then $1/\lambda_i$ is an eigenvalue of $\mathbf{A}^{-1}$.

7.17 Let **A** be an $n \times n$ matrix with n distinct eigenvalues. Prove that the set of n eigenvectors $\mathbf{x}_i$ are linearly independent.

Let

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n = \mathbf{0} \qquad (1)$$

If it can be shown that this implies that $a_1 = a_2 = \cdots = a_n = 0$, then the set $\{\mathbf{x}_i\}$ is linearly independent. Define $\mathbf{T}_i = \mathbf{A} - \mathbf{I}\lambda_i$ and note that $\mathbf{T}_i\mathbf{x}_i = \mathbf{0}$. $\mathbf{T}_i\mathbf{x}_j = (\lambda_j - \lambda_i)\mathbf{x}_j$ if $i \neq j$. Multiplying equation (*1*) by $\mathbf{T}_1$ gives

$$a_2(\lambda_2 - \lambda_1)\mathbf{x}_2 + a_3(\lambda_3 - \lambda_1)\mathbf{x}_3 + \cdots + a_n(\lambda_n - \lambda_1)\mathbf{x}_n = \mathbf{0}$$

Multiplying this in turn by $\mathbf{T}_2$, then $\mathbf{T}_3, \ldots, \mathbf{T}_{n-1}$ gives

$$a_3(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)\mathbf{x}_3 + \cdots + a_n(\lambda_n - \lambda_1)(\lambda_n - \lambda_2)\mathbf{x}_n = \mathbf{0}$$

$$\vdots$$

$$a_{n-1}(\lambda_{n-1} - \lambda_1)(\lambda_{n-1} - \lambda_2)\cdots(\lambda_{n-1} - \lambda_{n-2})\mathbf{x}_{n-1} + a_n(\lambda_n - \lambda_1)(\lambda_n - \lambda_2)\cdots(\lambda_n - \lambda_{n-2})\mathbf{x}_n = \mathbf{0} \tag{2}$$

$$a_n(\lambda_n - \lambda_1)(\lambda_n - \lambda_2)\cdots(\lambda_n - \lambda_{n-2})(\lambda_n - \lambda_{n-1})\mathbf{x}_n = \mathbf{0} \tag{3}$$

Since $\mathbf{x}_n \neq \mathbf{0}$, and $\lambda_n \neq \lambda_i$ for $i \neq n$, equation (3) requires that $a_n = 0$. This plus Eq. (2) requires that $a_{n-1} = 0$. Continuing this reasoning shows that Eq. (1) requires $a_i = 0$ for $i = 1, 2, \ldots, n$, so the eigenvectors are linearly independent.

7.18 Let $\mathbf{T}_i : \mathscr{X} \to \mathscr{X}$ be defined by $\mathbf{T}_i = \mathbf{A} - \mathbf{I}\lambda_i$, where $\mathscr{X}$ is an n-dimensional space. Prove that there are always $q_i = n - \text{rank}(\mathbf{T}_i)$ linearly independent eigenvectors associated with eigenvalue λ_i.

The space $\mathscr{X}$ can be written as the direct sum $\mathscr{X} = \mathscr{R}(\mathbf{T}_i^*) \oplus \mathscr{N}(\mathbf{T}_i)$ and $\dim(\mathscr{X}) = \dim(\mathscr{R}(\mathbf{T}_i^*)) + \dim(\mathscr{N}(\mathbf{T}_i))$ or $n = \text{rank}(\mathbf{T}_i^*) + \dim(\mathscr{N}(\mathbf{T}_i))$. But since $\text{rank}(\mathbf{T}_i) = \text{rank}(\mathbf{T}_i^*)$, $n - \text{rank}(\mathbf{T}_i) = q_i = \dim(\mathscr{N}(\mathbf{T}_i))$. The null space of $\mathbf{T}_i$ is of dimension q_i and, therefore, it contains q_i linearly independent vectors, all of which are eigenvectors.

7.19 Let **A** be an arbitrary $n \times r$ matrix and let **B** be an arbitrary $r \times n$ matrix, so that **AB** and **BA** are $n \times n$ and $r \times r$ matrices respectively. Assume that $n \geq r$ and prove:

(a) The scalar λ is a nonzero eigenvalue of **AB** if and only if it is a nonzero eigenvalue of **BA.**

(b) If $\mathbf{x}_i$ is an eigenvector (or generalized eigenvector) of **AB** associated with a nonzero eigenvalue, then $\zeta_i \triangleq \mathbf{B}\mathbf{x}_i$ is an eigenvector (or generalized eigenvector) or **BA.**

(c) **AB** has at least $n - r$ zero eigenvalues.

Assume $\mathbf{A}\mathbf{B}\mathbf{x}_i = \lambda\mathbf{x}_i$ with $\lambda \neq 0, \mathbf{x}_i \neq \mathbf{0}$. Then multiplying by **B** gives $\mathbf{B}\mathbf{A}(\mathbf{B}\mathbf{x}_i) = \lambda\mathbf{B}\mathbf{x}_i$ or $\mathbf{B}\mathbf{A}\zeta_i = \lambda\zeta_i$. Thus λ and ζ_i are an eigenvalue and eigenvector of **BA** provided $\zeta_i \neq \mathbf{0}$. But since $\lambda\mathbf{x}_i \neq \mathbf{0}$, $\mathbf{B}\mathbf{x}_i \neq \mathbf{0}$; otherwise $\mathbf{A}\mathbf{B}\mathbf{x}_i = \mathbf{0}$. Therefore, λ and ζ_i are an eigenvalue and eigenvector of **BA,** provided $\lambda \neq 0$ and $\mathbf{x}_i$ are an eigenvalue and eigenvector of **AB.** Now assume $\mathbf{B}\mathbf{A}\zeta_i = \lambda\zeta_i$, $\lambda \neq 0$, and $\zeta_i \neq \mathbf{0}$. Using the same kind of arguments as above show that λ and $\mathbf{x}_i = \mathbf{A}\zeta_i$ are an eigenvalue and eigenvector of **AB.** These results can be generalized for the case of generalized eigenvectors. This proves a and b. To prove c, it is only necessary to note that **AB** has n eigenvalues and each nonzero eigenvalue is simultaneously an eigenvalue of **BA.** Since **BA** has r eigenvalues, **AB** has at most r nonzero eigenvalues and, therefore, at least $n - r$ zero eigenvalues.

7.20 Let **A** and **B** be defined as in Problem 7.19. Define $\mathbf{N} = \mathbf{I}_n + \mathbf{A}\mathbf{B}$ and $\mathbf{R} = \mathbf{I}_r + \mathbf{B}\mathbf{A}$. Prove:

(a) $\mathbf{x}_i$ is an eigenvector of **AB** if and only if it is an eigenvector of **N** and λ is an eigenvalue of **AB** if and only if $1 + \lambda$ is an eigenvalue of **N.**

(b) ζ_i is an eigenvector of **BA** if and only if it is an eigenvector of **R,** and λ is an eigenvalue of **BA** if and only if $1 + \lambda$ is an eigenvalue of **R.**

(a) The proof requires showing that

$$(\mathbf{A}\mathbf{B} - \lambda_i \mathbf{I}_n)\mathbf{x}_i = \mathbf{0} \Leftrightarrow (\mathbf{N} - (\lambda_i + 1)\mathbf{I}_n)\mathbf{x}_i = \mathbf{0}$$

But $\mathbf{N} - (\lambda_i + 1)\mathbf{I}_n = \mathbf{A}\mathbf{B} + \mathbf{I}_n - \lambda_i \mathbf{I}_n - \mathbf{I}_n = \mathbf{A}\mathbf{B} - \lambda_i \mathbf{I}_n$.

(b) The proof is a simple matter of applying the definitions of the two eigenvalue problems in question, just as in part a.

7.21 Show that the r eigenvalues of **R** defined in Problem 7.20 are also eigenvalues of **N.** The remaining $n - r$ eigenvalues of **N** are all equal to one.

If λ is an eigenvalue of **BA,** then it is also an eigenvalue of **AB.** Since the eigenvalues of **N** and **R** are shifted by one from these eigenvalues, the result is proven for the r eigenvalues of **BA.** It was shown in Problem 7.19 that the remaining $n - r$ eigenvalues of **AB** must be zero, so the corresponding eigenvalues of **N** are one.

7.22 Let **A** and **B** be as defined in Problem 7.19. Prove the determinant identity of Problem 4.5, page 142, i.e., prove $|\mathbf{I}_n \pm \mathbf{AB}| = |\mathbf{I}_r \pm \mathbf{BA}|$.

Since the determinant is equal to the product of the eigenvalues (see Problem 7.14) and since $\mathbf{N} = \mathbf{I}_n + \mathbf{AB}$ and $\mathbf{R} = \mathbf{I}_r + \mathbf{BA}$ have been shown to have r eigenvalues in common with the rest of the $n - r$ eigenvalues equal to one, the identity is proven for the plus sign. The identity also holds for the minus sign, since we can always define $\mathbf{B}_1 = -\mathbf{B}$ or $\mathbf{A}_1 = -\mathbf{A}$. The previous results placed no restrictions on **A** and **B** other than their dimensions. Another form of the same identity is

$$|\mathbf{A}_1\mathbf{B} \pm \lambda\mathbf{I}_n| = (\pm\lambda)^{n-r}|\mathbf{B}\mathbf{A}_1 \pm \lambda\mathbf{I}_r|$$

which is established by defining $\lambda\mathbf{A} = \mathbf{A}_1$ and using the rule for multiplying a determinant by a scalar.

7.23 Let **P** be a nonsingular $n \times n$ matrix whose determinant and inverse are known. Let **C** and **D** be arbitrary $n \times r$ and $r \times n$ matrices, respectively. Show that $|\mathbf{P} + \mathbf{CD}| = |\mathbf{P}|\cdot|\mathbf{I}_r + \mathbf{D}\mathbf{P}^{-1}\mathbf{C}|$.

Simple manipulations show $|\mathbf{P} + \mathbf{CD}| = |\mathbf{P}[\mathbf{I}_n + \mathbf{P}^{-1}\mathbf{CD}]| = |\mathbf{P}|\cdot|\mathbf{I}_n + \mathbf{P}^{-1}\mathbf{CD}|$. Using the result of Problem 7.22 allows the interchange of factors $(\mathbf{P}^{-1}\mathbf{C})$ and $(\mathbf{D})$ and the corresponding change from an $n \times n$ determinant to an $r \times r$ determinant.

7.24 Let **N** be an $n \times n$ matrix given by $\mathbf{N} = \mathbf{I}_n + \mathbf{AB}$, where **A** and **B** are $n \times 1$ and $1 \times n$ matrices. Show that the $n \times n$ determinant can be expressed in terms of the easily evaluated trace: $|\mathbf{N}| = \mathrm{Tr}(\mathbf{N}) + 1 - n$.

Since $|\mathbf{N}| = |\mathbf{I}_n + \mathbf{AB}| = |\mathbf{I}_r + \mathbf{BA}|$ and since in this case the dimension r of **BA** is one, $|\mathbf{N}| = 1 + \mathbf{BA} = 1 + \mathrm{Tr}(\mathbf{BA})$. But $\mathrm{Tr}(\mathbf{BA}) = \mathrm{Tr}(\mathbf{AB})$ and $\mathrm{Tr}(\mathbf{N}) = \mathrm{Tr}(\mathbf{AB}) + \mathrm{Tr}(\mathbf{I}_n)$ or $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{N}) - n$, so $|\mathbf{N}| = 1 + \mathrm{Tr}(\mathbf{N}) - n$.

Self-Adjoint Transformation

7.25 If $\mathscr{A}$ is a self-adjoint transformation (see Section 5.12), show that all of its eigenvalues are real and that the eigenvectors associated with two different eigenvalues are orthogonal.

Consider $\mathscr{A}(\mathbf{x}_i) = \lambda_i\,\mathbf{x}_i$ and form the inner product $\langle \mathbf{x}_i, \mathscr{A}(\mathbf{x}_i)\rangle = \langle \mathbf{x}_i, \lambda_i\,\mathbf{x}_i\rangle = \lambda_i\langle \mathbf{x}_i, \mathbf{x}_i\rangle$. The definition of $\mathscr{A}^*$ ensures that $\langle \mathbf{x}_i, \mathscr{A}(\mathbf{x}_i)\rangle = \langle \mathscr{A}^*(\mathbf{x}_i), \mathbf{x}_i\rangle$ and if $\mathscr{A} = \mathscr{A}^*$, this gives $\langle \bar{\lambda}_i\,\mathbf{x}_i, \mathbf{x}_i\rangle = \bar{\lambda}_i\langle \mathbf{x}_i, \mathbf{x}_i\rangle$. Subtracting gives $0 = (\bar{\lambda}_i - \lambda_i)\langle \mathbf{x}_i, \mathbf{x}_i\rangle$. Since $\mathbf{x}_i$ is an eigenvector $\|\mathbf{x}_i\|^2 \neq 0$, so $\bar{\lambda}_i = \lambda_i$ and all eigenvalues are real.

Now consider $\mathscr{A}(\mathbf{x}_i) = \lambda_i\,\mathbf{x}_i$, $\mathscr{A}(\mathbf{x}_j) = \lambda_j\,\mathbf{x}_j$ with $\lambda_i \neq \lambda_j$. Then $\langle \mathbf{x}_j, \mathscr{A}(\mathbf{x}_i)\rangle = \lambda_i\langle \mathbf{x}_j, \mathbf{x}_i\rangle$. Also $\langle \mathbf{x}_j, \mathscr{A}(\mathbf{x}_i)\rangle = \langle \mathscr{A}^*(\mathbf{x}_j), \mathbf{x}_i\rangle = \langle \mathscr{A}(\mathbf{x}_j), \mathbf{x}_i\rangle = \bar{\lambda}_j\langle \mathbf{x}_j, \mathbf{x}_i\rangle$. But $\bar{\lambda}_j = \lambda_j$, so substracting gives $0 = (\lambda_i - \lambda_j)\langle \mathbf{x}_j, \mathbf{x}_i\rangle$. Since $\lambda_i \neq \lambda_j$, we have $\langle \mathbf{x}_j, \mathbf{x}_i\rangle = 0$ and $\mathbf{x}_j$ is orthogonal to $\mathbf{x}_i$.

Normal Transformation

7.26 Let $\mathscr{A}$ be a normal transformation (see Section 5.12). Prove that $\mathscr{A}(\mathbf{x}_i) = \lambda_i\,\mathbf{x}_i$ if and only if $\mathscr{A}^*(\mathbf{x}_i) = \bar{\lambda}_i\,\mathbf{x}_i$.

This is equivalent to showing $(\mathscr{A} - \mathscr{I}\lambda_i)\mathbf{x}_i = 0 \Leftrightarrow (\mathscr{A}^* - \mathscr{I}\bar{\lambda}_i)\mathbf{x}_i = \mathbf{0}$.

$$\begin{aligned}\langle(\mathscr{A} - \mathscr{I}\lambda_i)\mathbf{x}_i,\ (\mathscr{A} - \mathscr{I}\lambda_i)\mathbf{x}_i\rangle &= \langle \mathscr{A}(\mathbf{x}_i), \mathscr{A}(\mathbf{x}_i)\rangle - \langle \lambda_i\,\mathbf{x}_i, \mathscr{A}(\mathbf{x}_i)\rangle \\ &\quad - \langle \mathscr{A}(\mathbf{x}_i), \lambda_i\,\mathbf{x}_i\rangle + \langle \lambda_i\,\mathbf{x}_i, \lambda_i\,\mathbf{x}_i\rangle \\ &= \langle \mathscr{A}^*\mathscr{A}(\mathbf{x}_i), \mathbf{x}_i\rangle - \bar{\lambda}_i\langle \mathscr{A}^*(\mathbf{x}_i), \mathbf{x}_i\rangle \\ &\quad - \lambda_i\langle \mathbf{x}_i, \mathscr{A}^*(\mathbf{x}_i)\rangle - \bar{\lambda}_i\,\lambda_i\langle \mathbf{x}_i, \mathbf{x}_i\rangle\end{aligned}$$

Using $\mathscr{A}^*\mathscr{A} = \mathscr{A}\mathscr{A}^*$ allows this to be rewritten as

$$\begin{aligned}&\langle \mathscr{A}^*(\mathbf{x}_i), \mathscr{A}^*(\mathbf{x}_i)\rangle - \langle \mathscr{A}^*(\mathbf{x}_i), \bar{\lambda}_i\,\mathbf{x}_i\rangle - \langle \bar{\lambda}_i\,\mathbf{x}_i, \mathscr{A}^*(\mathbf{x}_i)\rangle + \langle \bar{\lambda}_i\,\mathbf{x}_i, \bar{\lambda}_i\,\mathbf{x}_i\rangle \\ &\qquad = \langle(\mathscr{A}^* - \mathscr{I}\bar{\lambda}_i)\mathbf{x}_i,\ (\mathscr{A}^* - \mathscr{I}\bar{\lambda}_i)\mathbf{x}_i\rangle\end{aligned}$$

or

$$\|(\mathcal{A} - \mathcal{I}\lambda_i)\mathbf{x}_i\|^2 = \|(\mathcal{A}^* - \mathcal{I}\bar{\lambda}_i)\mathbf{x}_i\|^2$$

The desired result follows.

7.27 Prove that the eigenvectors $\mathbf{x}_i$ and $\mathbf{x}_j$ associated with eigenvalues λ_i and λ_j are orthogonal for any normal transformation, provided $\lambda_i \neq \lambda_j$.

A normal transformation satisfies $\mathcal{A}^*\mathcal{A} = \mathcal{A}\mathcal{A}^*$, and therefore the class of normal transformations includes self-adjoint transformations as a subclass. Consider $\mathcal{A}(\mathbf{x}_i) = \lambda_i \mathbf{x}_i$ and $\mathcal{A}(\mathbf{x}_j) = \lambda_j \mathbf{x}_j$ with $\lambda_i \neq \lambda_j$. Then

$$\langle \mathbf{x}_j, \mathcal{A}(\mathbf{x}_i)\rangle = \lambda_i \langle \mathbf{x}_j, \mathbf{x}_i\rangle = \langle \mathcal{A}^*(\mathbf{x}_j), \mathbf{x}_i\rangle \tag{1}$$

From the previous problem $\mathcal{A}^*(\mathbf{x}_j) = \bar{\lambda}_j \mathbf{x}_j$, so

$$\langle \mathcal{A}^*(\mathbf{x}_j), \mathbf{x}_i\rangle = \langle \bar{\lambda}_j \mathbf{x}_j, \mathbf{x}_i\rangle = \lambda_j \langle \mathbf{x}_j, \mathbf{x}_i\rangle \tag{2}$$

Subtracting Eq. (2) from Eq. (1) gives $0 = (\lambda_i - \lambda_j)\langle \mathbf{x}_j, \mathbf{x}_i\rangle$. Since $\lambda_i \neq \lambda_j$, it follows that $\langle \mathbf{x}_j, \mathbf{x}_i\rangle = 0$ and, therefore, $\mathbf{x}_i$ and $\mathbf{x}_j$ are orthogonal.

7.28 Prove that if $\mathbf{A}$ is the $n \times n$ matrix representation of a normal transformation, then $\mathbf{A}$ has a full set of n linearly independent eigenvectors, regardless of the multiplicity of the eigenvalues.

Let $\mathbf{T}_i \triangleq \mathbf{A} - \mathbf{I}\lambda_i$. The eigenvectors satisfy $\mathbf{T}_i \mathbf{x}_i = \mathbf{0}$, and a generalized eigenvector $\mathbf{x}_{i+1}$ must satisfy $\mathbf{T}_i \mathbf{x}_{i+1} = \mathbf{x}_i$. The required proof consists of showing that if $\mathbf{A}$ is normal, the condition on $\mathbf{x}_{i+1}$ leads to a contradiction and thus cannot be satisfied. For $\mathbf{A}$ normal, $\mathbf{A}^*\mathbf{x}_i = \bar{\lambda}_i \mathbf{x}_i$ for each eigenvector $\mathbf{x}_i$; that is, $\mathbf{T}_i^* \mathbf{x}_i = \mathbf{0}$. Then $\mathbf{T}_i^* \mathbf{T}_i \mathbf{x}_{i+1} = \mathbf{T}_i^* \mathbf{x}_i = \mathbf{0}$. This means that

$$\langle \mathbf{x}_{i+1}, \mathbf{T}_i^* \mathbf{T}_i \mathbf{x}_{i+1}\rangle = \langle \mathbf{T}_i \mathbf{x}_{i+1}, \mathbf{T}_i \mathbf{x}_{i+1}\rangle = 0 \quad \text{or} \quad \|\mathbf{T}_i \mathbf{x}_{i+1}\|^2 = 0$$

This requires that $\mathbf{T}_i \mathbf{x}_{i+1} = \mathbf{0}$, but this contradicts the original assumption, since $\mathbf{x}_i \neq \mathbf{0}$. When $\mathbf{A}$ is normal, it cannot have generalized eigenvectors and, therefore, must have a full set of n linearly independent eigenvectors.

Singular Value Decomposition SVD [6]

7.29 Consider an $m \times n$ matrix $\mathbf{A}$ with $m \geq n$, and with $\text{rank}(\mathbf{A}) = r$. Show that $\mathbf{A}$ can be written as $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are $m \times m$ and $n \times n$ orthogonal matrices respectively and where $\boldsymbol{\Sigma}$ is $m \times n$ and "diagonal." (A nonsquare matrix is diagonal if all i, j entries are zero for $i \neq j$.)

Since $\mathbf{A}$ is not square, it cannot be used directly in an eigenvalue problem. However, two related problems are pertinent. Consider $\mathbf{A}\mathbf{A}^T\xi_1 = \sigma_i^2 \xi_i$ and $\mathbf{A}^T\mathbf{A}\eta_i = \lambda_i^2 \eta_i$. $\mathbf{A}\mathbf{A}^T$ is $m \times m$, symmetric, and hence normal. It is also positive semidefinite, and thus the σ_i^2 notation for the eigenvalue is justified. From Problem 7.28 there is a full set of m eigenvectors. From Problem 7.25 or 7.26 these eigenvectors are mutually orthogonal, at least for two different eigenvalues. They still can be selected as orthogonal even if there are repeated eigenvalues. For a multiplicity k we are assured there are k independent eigenvectors, and Gram-Schmidt can be used to construct k orthogonal vectors from them. The new vectors are still eigenvectors. By proper normalization, all ξ_i are also unit vectors, i.e., they are orthonormal. These vectors form the columns of an $m \times m$ orthogonal matrix $\mathbf{U}$.

$\mathbf{A}^T\mathbf{A}$ is $n \times n$ symmetric and at least positive semidefinite. Thus it also has a full set of n orthonormal eigenvectors η_i and nonnegative eigenvalues λ_i^2. Use the η_i vectors to form columns of an $n \times n$ orthogonal matrix $\mathbf{V}$. From Problem 7.19 it is known that $\mathbf{A}^T\xi_i \triangleq \zeta_i$ will be an eigenvector of $\mathbf{A}^T\mathbf{A}$, at least in the case where $\sigma_i \neq 0$. This is still true even for $\sigma_i = 0$, as will be seen when the length of ζ_i is computed below. From that same problem the nonzero values of σ_i and λ_i are the same. Thus $\mathbf{A}\mathbf{A}^T\xi_i = \sigma_i^2 \xi_i$ becomes $\mathbf{A}\zeta_i = \sigma_i^2 \xi_i$. But ζ_i is generally not a unit vector. In fact its length is found from $\|\zeta_i\|^2 = \langle \zeta_i, \zeta_i\rangle = \langle \mathbf{A}^T\xi_i, \mathbf{A}^T\xi_i\rangle = \langle \mathbf{A}\mathbf{A}^T\xi_i, \xi_i\rangle = \sigma_i^2\langle \xi_i, \xi_i\rangle = \sigma_i^2$. Therefore, dividing the earlier equation by σ_i gives $\mathbf{A}\eta_i = \sigma_i \xi_i$. The entire set of such equations is

$$\mathbf{A}[\eta_1 \eta_2 \ldots \eta_n] = [\sigma_1 \xi_1 \ldots \sigma_m \xi_m] = [\xi_1 \xi_2 \ldots \xi_m] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & \sigma_n \\ & & \mathbf{0} & \end{bmatrix} \begin{matrix} \} n \times n \text{ diagonal, rank } r \\ \} (m-n) \times n \text{ zero} \end{matrix}$$

or $\mathbf{AV} = \mathbf{U\Sigma}$. Using the orthogonality of $\mathbf{V}$ gives $\mathbf{V}^{-1} = \mathbf{V}^T$. The final result is $\mathbf{A} = \mathbf{U\Sigma V}^T$. For exposition purposes it was assumed that $m > n$. This was not at all essential. For example, if $m < n$, define $\mathbf{B} = \mathbf{A}^T$ and then all the above applies to $\mathbf{B}$.

The positive square roots of the eigenvalues of $\mathbf{A}^T\mathbf{A}$, namely the $\sigma_i = \lambda_i$, are called the singular values of $\mathbf{A}$. The eigenvectors of $\mathbf{AA}^T$, namely the ξ_i, are called the left singular vectors of $\mathbf{A}$ and the η_i are called the right singular vectors of $\mathbf{A}$.

7.30 Show that any $m \times n$ matrix $\mathbf{A}$ of rank r can be written as

$$\mathbf{A} = \mathbf{U}'\mathbf{\Sigma}'\mathbf{V}'^T$$

where $\mathbf{U}'$ and $\mathbf{V}'$ are $m \times r$ and $r \times n$ matrices, respectively, with orthonormal columns, and where $\mathbf{\Sigma}'$ is an $r \times r$ full rank diagonal matrix.

Starting with the previous problem results, all the zero columns of $\mathbf{\Sigma}$ can be deleted as long as the corresponding rows of $\mathbf{V}^T$ are also deleted to maintain conformability. The values in the resulting matrix product are unchanged. Likewise, all the zero rows of $\mathbf{\Sigma}$ can be deleted without changing the answer, so long as the corresponding columns of $\mathbf{U}$ are deleted to maintain a conformable product. Actually this last set of deletions can be done in every case, whether $\mathbf{A}$ is full rank n or not. The first set of deletions only applies when $\mathbf{A}$ is of less than full rank, say r, because in the full rank case there are no zero columns in $\mathbf{\Sigma}$. The row-deleted version of $\mathbf{V}^T$ is $\mathbf{V}'^T$ and the column-deleted version of $\mathbf{U}$ is $\mathbf{U}'$. Likewise for $\mathbf{\Sigma}$ and $\mathbf{\Sigma}'$. The primed matrices form what has been called the economy-sized version of singular value decomposition. It can save a lot of computer storage. Even though $\mathbf{U}'$ and $\mathbf{V}'$ are no longer square, it is still true that $\mathbf{U}'^T\mathbf{U}' = \mathbf{I}$ and $\mathbf{V}'^T\mathbf{V}' = \mathbf{I}$. In both this form of the singular value decomposition and the previous full-sized form, the rank of $\mathbf{A}$ is the number of nonzero singular values in $\mathbf{\Sigma}$ or $\mathbf{\Sigma}'$.

7.31 Show how SVD can be used to solve simultaneous linear equations $\mathbf{Ax} = \mathbf{y}$.

Using the SVD form for $\mathbf{A}$ gives $\mathbf{U\Sigma V}^T\mathbf{x} = \mathbf{y}$. Using the orthogonality of $\mathbf{U}$ and defining $\mathbf{U}^T\mathbf{y} \triangleq \mathbf{w}$ and $\mathbf{V}^T\mathbf{x} \triangleq \mathbf{v}$ gives $\mathbf{\Sigma v} = \mathbf{w}$. Because of the diagonal nature of $\mathbf{\Sigma}$, these are easily solved for $\mathbf{v}$ in most cases. Then a simple matrix product gives $\mathbf{x} = \mathbf{Vv}$. An expanded form of the crucial equation is

$$\mathbf{\Sigma v} = \mathbf{w} \Rightarrow \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_r & \\ \hline \multicolumn{2}{c|}{\mathbf{0}} & & \mathbf{0} \end{array}\right] \left[\begin{array}{c} v_1 \\ \vdots \\ v_r \\ \hline v_{r+1} \\ \vdots \\ v_n \end{array}\right] = \left[\begin{array}{c} w_1 \\ \vdots \\ w_r \\ \hline w_{r+1} \\ \vdots \\ w_m \end{array}\right]$$

From this it can be seen that the original equations are inconsistent and have no solution if $\mathbf{A}$ and $\mathbf{\Sigma}$ have rank $r < n$ unless $\mathbf{w}$ also has these last $m - r$ rows zero. A least-squares solution is still possible. The solution (or least-squares solution) for $\mathbf{v}$ will have some arbitrary components whenever there are zero columns in $\mathbf{\Sigma}$. Setting these components of $\mathbf{v}$ to zero will give the minimum norm solution (or least-squares solution if required) for $\mathbf{v}$. Since $\mathbf{x}$ and $\mathbf{v}$ are related by an orthogonal matrix, they have the same norm, so $\mathbf{x}$ is also minimum norm in that case.

7.32 Show that the SVD provides the means for extending the spectral representation of Eq. (7.5) to nonsquare matrices.

Starting with $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ as defined in Problem 7.29,

$$\mathbf{A} = [\xi_1\, \xi_2 \ldots \xi_m][\boldsymbol{\Sigma}]\begin{bmatrix}\eta_1^T \\ \vdots \\ \eta_n^T\end{bmatrix} = [\sigma_1\xi_1 \quad \sigma_2\xi_2 \quad \cdots \quad \sigma_n\xi_n]\begin{bmatrix}\eta_1^T \\ \vdots \\ \eta_n^T\end{bmatrix}$$

$$= \sum_{i=1}^{n} \sigma_i\, \xi_i\, \eta_i^T = \sum_{i=1}^{n} \sigma_i\, \xi_i \rangle\langle \eta_i$$

This is of the form of Eq. (7.5). It has similar uses. For example, in approximation theory this series can be truncated prior to including all n terms if some values of σ_i are considered to be sufficiently small to be neglected.

7.33 Find the singular value decomposition for the matrix $\mathbf{A}$ of Problem 6.14.

$$\mathbf{A} = \begin{bmatrix}1 & 2 \\ 3 & 4 \\ 5 & 6\end{bmatrix} \quad \text{so} \quad \mathbf{A}^T\mathbf{A} = \begin{bmatrix}35 & 44 \\ 44 & 56\end{bmatrix} \quad \text{and} \quad \mathbf{A}\mathbf{A}^T = \begin{bmatrix}5 & 11 & 17 \\ 11 & 25 & 39 \\ 17 & 39 & 61\end{bmatrix}$$

The nonzero eigenvalues are approximately $\sigma_1^2 = 90.7355$ and $\sigma_2^2 = 0.2645$. The square roots of these form the diagonal terms in $\boldsymbol{\Sigma}$ below, and the two sets of normalized eigenvectors are shown as columns of $\mathbf{U}$ and rows of $\mathbf{V}^T$ next.

$$\mathbf{A} = \overbrace{\begin{bmatrix}0.2298 & -0.8835 & 0.4082 \\ 0.5247 & -0.2408 & -0.8165 \\ 0.8196 & 0.4019 & 0.4082\end{bmatrix}}^{\mathbf{U}} \overbrace{\begin{bmatrix}9.5255 & 0 \\ 0 & 0.5143 \\ 0 & 0\end{bmatrix}}^{\boldsymbol{\Sigma}} \overbrace{\begin{bmatrix}0.6196 & 0.7849 \\ 0.7849 & -0.6196\end{bmatrix}}^{\mathbf{V}^T}$$

The efficient computational determination of the SVD form is crucial if it is to be useful. The indicated eigenvectors can be determined by the means presented in this chapter. This easily leads to the SVD form in simple cases like this one. However, Reference 6 should be consulted for a superior algorithm for use in more realistic cases. The real value of the discussion of this and the four previous problems is in understanding the concepts of the method, not in developing a general-purpose algorithm.

7.34 Resolve the equations of Problem 6.14 using the SVD results of Problem 7.33 and the method of Problem 7.31.

The simultaneous equations are

$$\begin{bmatrix}1 & 2 \\ 3 & 4 \\ 5 & 6\end{bmatrix}\mathbf{x} = \begin{bmatrix}2 \\ 3 \\ 14\end{bmatrix}$$

Using $\mathbf{U}$ from Problem 7.33 gives $\mathbf{w} \triangleq \mathbf{U}^T\mathbf{y} = \begin{bmatrix}13.5095 \\ 3.1372\end{bmatrix}$.

Then $\mathbf{v}_1 = 13.5095/\sigma_1 = 1.4182$ an $\mathbf{v}_2 = 3.1372/\sigma_2 = 6.0999$. A matrix product then gives $\mathbf{x} = \mathbf{V}\mathbf{v} = \begin{bmatrix}5.666 \\ -2.666\end{bmatrix}$.

Independence of Generalized Eigenvectors

7.35 If $\mathbf{x}_1$ is an eigenvector and $\mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_k$ are generalized eigenvectors, all associated with the same eigenvalue λ_1, show that:

(a) All of these vectors belong to the null space of $(\mathbf{A} - \mathbf{I}\lambda_1)^k$.
(b) This set of vectors is linearly independent.
(a) The defining equations for the set of vectors are

$$\mathbf{A}\mathbf{x}_1 = \lambda_1 \mathbf{x}_1, \quad \mathbf{x}_1 \neq \mathbf{0}, \quad \text{or} \quad (\mathbf{A} - \mathbf{I}\lambda_1)\mathbf{x}_1 = \mathbf{0} \tag{1}$$

$$\mathbf{A}\mathbf{x}_2 = \lambda_1 \mathbf{x}_2 + \mathbf{x}_1 \qquad (\mathbf{A} - \mathbf{I}\lambda_1)\mathbf{x}_2 = \mathbf{x}_1 \tag{2}$$

$$\mathbf{A}\mathbf{x}_3 = \lambda_1 \mathbf{x}_3 + \mathbf{x}_2 \qquad (\mathbf{A} - \mathbf{I}\lambda_1)\mathbf{x}_3 = \mathbf{x}_2 \tag{3}$$

$$\vdots \qquad\qquad \vdots$$

$$\mathbf{A}\mathbf{x}_k = \lambda_1 \mathbf{x}_k + \mathbf{x}_{k-1} \qquad (\mathbf{A} - \mathbf{I}\lambda_1)\mathbf{x}_k = \mathbf{x}_{k-1}$$

From Eqs. (*1*) and (*2*), $(\mathbf{A} - \mathbf{I}\lambda_1)^2 \mathbf{x}_2 = (\mathbf{A} - \mathbf{I}\lambda_1)\mathbf{x}_1 = \mathbf{0}$. Multiplying Eq. (*3*) by $(\mathbf{A} - \mathbf{I}\lambda_1)^2$ gives $(\mathbf{A} - \mathbf{I}\lambda_1)^3 \mathbf{x}_3 = (\mathbf{A} - \mathbf{I}\lambda_1)^2 \mathbf{x}_2 = \mathbf{0}$. In general, it can be seen that $(\mathbf{A} - \mathbf{I}\lambda_1)^p \mathbf{x}_p = \mathbf{0}$, and $(\mathbf{A} - \mathbf{I}\lambda_1)^{p-1} \mathbf{x}_p = \mathbf{x}_1$. Since $(\mathbf{A} - \mathbf{I}\lambda_1)^k = (\mathbf{A} - \mathbf{I}\lambda_1)^{k-p}(\mathbf{A} - \mathbf{I}\lambda_1)^p$, $(\mathbf{A} - \mathbf{I}\lambda_1)^k \mathbf{x}_p = \mathbf{0}$ for $p = 1, 2, \ldots, k$ and part a is proven.

(b) Let

$$a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + \cdots + a_k \mathbf{x}_k = \mathbf{0} \tag{4}$$

and show that this implies that each $a_i = 0$. Multiplying Eq. (*4*) by $(\mathbf{A} - \mathbf{I}\lambda_1)^{k-1}$ gives $a_k(\mathbf{A} - \mathbf{I}\lambda_1)^{k-1} \mathbf{x}_k = \mathbf{0}$. Since $(\mathbf{A} - \mathbf{I}\lambda_1)^{k-1} \mathbf{x}_k = \mathbf{x}_1 \neq \mathbf{0}$, a_k must be zero. Using this fact and then multiplying equation (*4*) by $(\mathbf{A} - \mathbf{I}\lambda_1)^{k-2}$ shows that $a_{k-1} = 0$. Continuing this process shows that if $\sum_{i=1}^{k} a_i \mathbf{x}_i = \mathbf{0}$, then $a_i = 0$ for $i = 1, 2, \ldots, k$. This means the set $\{\mathbf{x}_i\}$ is linearly independent.

Companion Matrix

7.36 When considering nth-order linear differential equations of the type

$$\frac{d^n x}{dt^n} + a_{n-1}\frac{d^{n-1} x}{dt^{n-1}} + a_{n-2}\frac{d^{n-2} x}{dt^{n-2}} + \cdots + a_1 \frac{dx}{dt} + a_0 x = u(t)$$

the $n \times n$ matrix $\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & & 0 \\ 0 & 0 & 0 & 1 & & 0 \\ \vdots & & & & & \vdots \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} \end{bmatrix}$

will often arise. $\mathbf{A}$ is called the companion matrix. Show that the companion matrix always has just one eigenvector for each eigenvalue, regardless of its algebraic multiplicity.

The matrix $\mathbf{A} - \mathbf{I}\lambda$ always has rank r, which satisfies $r \geq n - 1$. To see this, delete the first column and the last row, leaving a lower triangular $n - 1 \times n - 1$ matrix with ones on the diagonal. If λ is an eigenvalue, $\text{rank}(\mathbf{A} - \mathbf{I}\lambda) < n$. Together, these results imply $\text{rank}(\mathbf{A} - \mathbf{I}\lambda_i) = n - 1$, so the degeneracy is $q = n - r = 1$. The case of simple degeneracy always applies, so there is exactly one eigenvector for each eigenvalue.

Quadratic Form

7.37 Show that the eigenvalue problem for a real, symmetric matrix can be characterized as one of maximizing or minimizing a quadratic form subject to the constraint that $\mathbf{x}$ be a unit vector.

Consider $Q = \langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle$. If the change of basis $\mathbf{x} = \mathbf{M}\mathbf{z}$ is used, $Q = \sum_{i=1}^{n} \lambda_i z_i^2$. Because of the orthogonal transformation, $\mathbf{z}$ is also a unit vector. Since all $z_i^2 \leq 1$, this suggests that the

maximum value of Q will be attained if all $z_i = 0$ except $\mathbf{z}_k^2 = 1$, where $\lambda_k = \lambda_{\max}$. Then $Q = \lambda_{\max}$. Alternatively, adjoining the constraint $\|\mathbf{x}\|^2 = 1$ to Q by means of the Lagrange multiplier λ shows this directly. That is, maximizing $\mathbf{x}^T\mathbf{A}\mathbf{x} - \lambda(\mathbf{x}^T\mathbf{x} - 1)$ requires that the derivative with respect to each component x_i must vanish. This gives the set $\mathbf{A}\mathbf{x} - \lambda\mathbf{x} = \mathbf{0}$, which is the eigenvalue-eigenvector equation. If $\mathbf{x}$ satisfies this condition, then $Q = \mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{x}\lambda = \lambda$, so $Q_{\max} = \lambda_{\max}$. Also $Q_{\min} = \lambda_{\min}$.

If $\mathbf{x}_1$ is the eigenvector associated with $\lambda_{\max}$, then selecting $\mathbf{x}$ to maximize Q subject to $\langle \mathbf{x}_1, \mathbf{x} \rangle = 0$ and $\|\mathbf{x}\| = 1$ will lead to the second largest eigenvalue and its associated eigenvector. The remaining eigenvalues-eigenvectors are found in a similar way by requiring orthogonality with all previously found eigenvectors.

7.38 Reduce the quadratic form $Q = \frac{1}{3}[16y_1^2 + 10y_2^2 + 16y_3^2 - 4y_1y_2 + 16y_1y_3 + 4y_2y_3]$ to a sum of squared terms only by selecting a suitable change of coordinates.

This quadratic form can be expressed in matrix form as $Q = \mathbf{y}^T\mathbf{A}\mathbf{y}$, where $\mathbf{y} = [y_1 \quad y_2 \quad y_3]^T$ and $\mathbf{A} = \frac{1}{3}\begin{bmatrix} 16 & -2 & 8 \\ -2 & 10 & 2 \\ 8 & 2 & 16 \end{bmatrix}$. The eigenvalues of $\mathbf{A}$ are $\lambda_1 = 8, \lambda_2 = 4, \lambda_3 = 2$, and since $\mathbf{A}$ is real and symmetric, a set of orthonormal eigenvectors can be found. They are used as columns of the modal matrix

$$\mathbf{M} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{6} & -1/\sqrt{3} \\ 0 & 2/\sqrt{6} & -1/\sqrt{3} \\ 1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{3} \end{bmatrix}$$

Since $\mathbf{M}$ is orthogonal, $\mathbf{M}^{-1} = \mathbf{M}^T$ and $\mathbf{A}$ is diagonalized by the orthogonal transformation $\mathbf{M}^T\mathbf{A}\mathbf{M} = \text{diag}[8, 4, 2]$. If the change of variables $\mathbf{y} = \mathbf{M}\mathbf{z}$ is used, then

$$Q = \mathbf{z}^T\mathbf{M}^T\mathbf{A}\mathbf{M}\mathbf{z} = 8z_1^2 + 4z_2^2 + 2z_3^2$$

PROBLEMS

7.39 Find the eigenvalues, eigenvectors, and Jordan form for $\mathbf{A} = \begin{bmatrix} 2 & -2 & 3 \\ 1 & 1 & 1 \\ 1 & 3 & -1 \end{bmatrix}$.

7.40 Find the eigenvectors of $\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$.

7.41 Analyze the eigenvalue-eigenvector problem for $\mathbf{A} = \begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$.

7.42 Compute the eigenvalues, eigenvectors, and Jordan form for $\mathbf{A} = \begin{bmatrix} 4 & -2 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 6 \end{bmatrix}$.

7.43 Are $\mathbf{A} = \frac{1}{2}\begin{bmatrix} 3 & 1 \\ -1 & 5 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ similar matrices?

7.44 Use the iterative technique of Sec. 7.6 to find approximate eigenvalues and eigenvectors for $\mathbf{A} = \begin{bmatrix} 8 & 2 & -5 \\ 2 & 11 & -2 \\ -5 & -2 & 8 \end{bmatrix}$.

7.45 Find an approximate set of eigenvalues and eigenvectors for $\mathbf{A} = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ using iteration.

7.46 If $\Delta(\lambda)$ is defined as in equation (*7.3*) and if

$$\Delta'(\lambda) \triangleq |\mathbf{I}\lambda - \mathbf{A}| = \lambda^n + c'_{n-1}\lambda^{n-1} + c'_{n-2}\lambda^{n-2} + \cdots + c'_1\lambda + c'_0$$

show that

(a) $c_0 = |\mathbf{A}|$; $c'_0 = |-\mathbf{A}| = (-1)^n|\mathbf{A}|$.

(b) $(-1)^{n-1}c_{n-1} = \text{Tr}(\mathbf{A})$; $c'_{n-1} = -\text{Tr}(\mathbf{A})$.

7.47 Find the eigenvalues, eigenvectors, and Jordan form for

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ -2 & 3 & -4 \\ 1 & 1 & -4 \end{bmatrix}$$

7.48 Find the eigenvalues, eigenvectors, and Jordan form for

$$\mathbf{A} = \begin{bmatrix} 4 & 2 & 1 \\ 1 & 2 & 1 \\ -1 & -4 & 8 \end{bmatrix}$$

7.49 Draw conclusions about the sign definiteness of

(a) $\mathbf{A} = \begin{bmatrix} -6 & 2 \\ 2 & -1 \end{bmatrix}$,

(b) $\mathbf{A} = \begin{bmatrix} 13 & 4 & -13 \\ 4 & 22 & -4 \\ -13 & -4 & 13 \end{bmatrix}$,

(c) $\mathbf{A} = \begin{bmatrix} -1 & 3 & 0 & 0 \\ 3 & -9 & 0 & 0 \\ 0 & 0 & -6 & 2 \\ 0 & 0 & 2 & -1 \end{bmatrix}$,

(d) $\mathbf{A} = \begin{bmatrix} 8 & 2 & -5 \\ 2 & 11 & -2 \\ -5 & -2 & 8 \end{bmatrix}$,

(e) $\mathbf{A} = \begin{bmatrix} -3 & 2 & 0 & 1 & 7 \\ 2 & 1 & -2 & 1 & 0 \\ 0 & -2 & 6 & 3 & 8 \\ 1 & 1 & 3 & 2 & 4 \\ 7 & 0 & 8 & 4 & 5 \end{bmatrix}$.

7.50 Let $\mathbf{A} = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$ and let $\mathbf{x} = [x_1 \quad x_2]^T$ be any *unit* vector. Consider $Q = \langle \mathbf{x}, \mathbf{Ax} \rangle$ as a scalar function of $\mathbf{x}$. From among all unit vectors find the one which gives Q its maximum value. Also determine $Q_{\max}$.

7.51 Show that any simple linear transformation can be represented as $\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{E}_i$, where $\mathbf{E}_i$ is a projection onto $\mathcal{N}(\mathbf{A} - \mathbf{I}\lambda_i)$.

7.52 Show that any normal linear transformation can be written as a sum of orthogonal projection transformations.

7.53 Analyze the simultaneous equations of Example 6.4 using the SVD method.